DISSERTAÇÃO DE MESTRADO

Estudo da Empregabilidade Feminina Utilizando Técnicas de Mineração de Dados e Algoritmos de Machine Learning

Ludimila de Oliveira Félix

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Estudo da Empregabilidade Feminina Utilizando Técnicas de Mineração de Dados e Algoritmos de Machine Learning

Ludimila de Oliveira Félix

ORIENTADOR: PROF. DR. GEORGES DANIEL AMVAME NZE, ENE/UNB

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO: PPGENE.DM - 765/2021
BRASÍLIA-DF, MARÇO DE 2021.

FICHA CATALOGRÁFICA

FÉLIX, LUDIMILA DE OLIVEIRA
Estudo da Empregabilidade Feminina Utilizando Técnicas de Mineração de Dados e Algoritmos de Machine Learning [Distrito Federal] 2020.
xvi, 52 p., 210 x 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2020).
Dissertação de Mestrado - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica

1. Empregabilidade	2. Árvore de Decisão
3. Machine Learning	4. Classificação Supervisionada
I. ENE/FT/UnB	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

FÉLIX, L.O. (2020). *Estudo da Empregabilidade Feminina Utilizando Técnicas de Mineração de Dados e Algoritmos de Machine Learning*. Dissertação de Mestrado, **PPGENE.DM-765/21**, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 52 p.

CESSÃO DE DIREITOS

AUTOR: Ludimila de Oliveira Félix
TÍTULO: Estudo da Empregabilidade Feminina Utilizando Técnicas de Mineração de Dados e Algoritmos de Machine Learning.
GRAU: Mestre em Engenharia Elétrica ANO: 2020

Ludimila de Oliveira Félix
Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

DEDICATÓRIA

Às minhas amadas avós serigueiras(in memoriam), as mulheres mais fortes que já conheci.

Für meine geliebten Omas (in memoriam), Die stärksten Frauen, die ich kennengelernt habe

AGRADECIMENTOS

O trabalho de dissertação é essencialmente solitário, porém, a realização é rodeada de uma grande rede de apoio, mas também de muita pressão e cobranças. Os últimos anos foram anos muito difíceis, no entanto, fui agraciada com muitas bençãs divinas e incentivo, os quais levarei sempre comigo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, o qual, agradeço pelo subsídio da bolsa de estudos concedida. Quero agradecer também:

Ao meu professor orientador Georges, pela confiança, orientação e paciência durante o processo desse trabalho;

As pessoas da secretária do PPGEE por toda ajuda ao logo desses dois anos de adversidades;

À minha mãe e ao meu pai deixo um agradecimento especial, sinto-me orgulhoso e privilegiado por ter pais tão especiais e companheiros. Obrigada por serem a certeza do alicerce e do amparo nessa vida;

À minha irmã, querida, meu maior presente e orgulho, obrigada por ser minha maior parceria e motivadora;

Às minhas tias Mana e Graça, por todos os bolinhos, risadas e orações para esse trabalho;

Às minhas grandes amigas, Vívian Varela e Marina Herejk, que me acompanham desde a graduação, mulheres que me inspiram, companheiras de conquistas e agonias;

Às minhas amigas de infância, Renata Café, Maria Valente, Larisa Santos, Gabriela Evora, Amanda Meneguzzo e Jéssica Querino por serem um porto seguro apesar das distâncias físicas, obrigada por me encorajarem quando nem eu mesma acreditava ser possível;

Aos meus amigos do Handebol-UnB, em especial Eliza Gabriela, Roberta Ramos e Cecília Cabral, por serem minha válvula de escape, companheiros dentro e fora de quadra;

À tapioca e Valdemar, companheiros de quatro patas que sempre estão ao meu lado, inclusive nas madrugadas de estudo;

À minha psicologa por tentar me ajudar a ressignificar o viver e a *Gelassenheit*.

Por fim, a todos aqueles que contribuíram, direta ou indiretamente, para a realização desta dissertação, o meu sincero agradecimento.

RESUMO

O atual cenário da empregabilidade feminina no Brasil é preocupante, apesar de serem a maioria da população, mulheres representam 41,2% da população economicamente ativa, enquanto homens são 58,8%. Além disso, mulheres lideram as taxas de desemprego. No contexto mundial, a ONU orienta os países membros em promover a igualdade de gênero e empoderamento. Nesse sentido, é necessário ações governamentais que viabilizem a inclusão feminina no mercado de trabalho e a equidade salarial. A utilização de tecnologias para análise do grande volume de dados é realidade e pode gerar informações essenciais para criação de políticas públicas e sociais. A proposta da pesquisa é utilizar as informações bases de dados do ano base 2018 compostas pela Relação Anual de Informações Sociais (RAIS) e pelo Cadastro Geral de Empregados e Desempregados (CAGED) para estudar o perfil da empregabilidade feminina no contexto do mercado de trabalho brasileiro utilizando técnicas de aprendizado de máquina. E assim, gerar informações para futuras políticas governamentais e empresariais. Como resultado, conseguimos observar de forma automatizada e rápida fatores que afetam a empregabilidade feminina no país.

ABSTRACT

Currently Brazil's female employability scenario concerns, despite being the majority of population, women lead the unemployment rates and men still represent the major portion of the economically active people, 55.08 %. UN (United Nations) advises the member countries to promote gender equality and women empowerment. Therefore, government actions need to encourage women's labour market inclusion and wage equity. Technology usage for data mining provides essential information which allows public and social policies conceiving. This study aims to use the data collected on 2018's Annual List of Social Information (RAIS) and General Register of Employed and Unemployed (CAGED) to study the profile of female employability applied to Brazilian labour market using machine learning techniques. Thus, generating material for future government and business policies conception. Results of this paper includes exposure of variables presents impacts on women employability.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

Siglas

AM	Aprendizado de Máquina
AUC	*Area Under the ROC curve*
BD	Banco de Dados
BI	*Business Intelligence*
CAGED	Cadastro Geral de Empregados e Desempregados
CBO	Classificação Brasileira de Ocupações
CLT	Consolidação das Leis do Trabalho
CMIG	Conjunto Mínimo de Indicadores de Gênero
CNAE	Classificação Nacional de Atividades Econômicas
EIACBD	Extração Inteligente e Automática de Conhecimento em Bancos de Dados
FN	Falsos Negativos
FP	Falsos Positivos
IBGE	Instituto Brasileiro de Geografia e Estatística
KDD	*Knowledge Discovery in Database*
MD	Mineração de Dados
ODM	Objetivos de Desenvolvimento do Milênio
ONU	Organização das Nações Unidas
PCM	Propensão Marginal a Consumir
PDET	Programa de Disseminação das Estatísticas do trabalho
PEA	População Economicamente Ativa
PMC	Propensão Marginal de Consumo
PNAD	Pesquisa Nacional por Amostra de Domicílios Contínua
RAIS	Relação Anual de Informações Socias
RF	Randon Florest
ROC	*Receiver Operating Characteristics*
RFE	*Recursive Feature Elimination*
SGDB	Sistema de Gerenciamento de Banco de Dados
SIG	Sistema de Informação Geográfico
SIG	Sistema de Informação Geográfico
UE	União Europeia
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos
XGBoost	Extreme Gradient Boosting

1 INTRODUÇÃO

Existem diversos fatores que afetam a empregabilidade de um país ou região. A empregabilidade está relacionada com o perfil da oferta de trabalho e da mão de obra disponível. Portanto, o conceito de empregabilidade é um importante demonstrativo da estabilidade econômica e social em um país. Um país em crise econômica possui oferta reduzida de vagas de emprego para a população economicamente ativa.

Entre os oito Objetivos de Desenvolvimento do Milênio (ODM) apresentados pela Organização da Nações Unidas(ONU) está promover a igualdade de gênero e empoderamento feminino. A ONU defende a participação equitativa das mulheres em todos os aspectos da vida e elenca cinco áreas prioritárias: Aumentar a liderança e a participação das mulheres; eliminar a violência contra as mulheres e meninas; engajar as mulheres em todos os aspectos dos processos de paz e segurança; aprimorar o empoderamento econômico das mulheres e colocar a igualdade de gênero no centro do planejamento e dos orçamentos de desenvolvimento nacional (Franco 2018).

Nesse contexto, a ONU estabeleceu padrões globais para alcançar a igualdade de gênero. os Estados-membro devem formular leis, políticas públicas, programas e serviços necessários à implementação desses padrões e consequentemente alcançar o objetivo (Medeiros e Pinheiro 2018, Franco 2018).

Um importante ponto na desigualdade de gênero é a estrutura econômica e a disparidade no mercado de trabalho. Ademais, a empregabilidade é um fator considerável na redução das desigualdade de gênero na estrutura econômica. O estudo do perfil da empregabilidade feminina pode auxiliar na redução dos índices de desempregos feminino e mitigação das disparidades salarias (Oliveira, Scorzafave e Pazello 2009).

Vale ressaltar que as pesquisas aplicadas à empregabilidade têm, ainda hoje, uma abordagem exploratória. Isso ocorre porque esta área de pesquisa apresenta dificuldades em ter dados adequados, confiáveis e atualizados. Além disso, não há um consenso dos resultados das pesquisas, mas também o aparecimento de questionamentos da metodologias e forma de abordagem mais adequada para estas questões. Por esses motivos, a área ainda está crescendo e precisa levar os resultados a novos níveis (García-Peñalvo et al. 2018).

Em 2013, a Comissão de Estatística das Nações Unidas organizou o Conjunto Mínimo de Indicadores de Gênero - CMIG, constituído por 63 indicadores (52 quantitativos e 11 qualitativos) que refletem o esforço de sistematização de informações destinadas à produção nacional e à harmonização internacional de estatísticas de países e regiões relativamente à igualdade de gênero e ao empoderamento feminino (Boletim Geográfico 2018).

No Brasil, o IBGE possui um Sistema Nacional de Informações de Gênero (SNIG) e com base nessas informações divulgou os resultados de grande partes dos indicadores CMIG (Boletim Geográfico 2018). As informações datam de 2014 e foram originadas da Pesquisa Nacional por Amostra de Domicílios (PNAD) e estão organizadas segundo os cinco domínios estabelecidos no CMIG.

Os domínios se dividem em: Estruturas econômicas; participação em atividades produtivas e acesso a recursos; Educação; Saúde e serviços relacionados; Vida pública e tomada de decisão e Direitos humanos das mulheres e meninas. A partir deles é possível fornecer um panorama, ainda que sucinto, das desigual-

dades de gênero no País, com valiosos elementos para reflexão de estudiosos e formuladores de políticas públicas (Boletim Geográfico 2018).

Ademais, a gestão governamental do setor do trabalho conta com o importante instrumento de coleta de dados denominado de Relação Anual de Informações Sociais (RAIS). Instituída pelo Decreto nº 76.900, a RAIS tem como principais objetivos: O suprimento às necessidades de controle da atividade trabalhista no País, o provimento de dados para a elaboração de estatísticas do trabalho e a disponibilização de informações do mercado de trabalho às entidades governamentais.

1.1 CONTRIBUIÇÕES

Esta pesquisa concentra-se no desenvolvimento de estratégias e resultados rápidos e automatizados que possam auxiliar as equipes e instituições responsáveis pela criação de políticas públicas a aprimorarem suas soluções para promoverem a geração de emprego e distribuição justa do mercado de trabalho. E desse modo reduzir as desigualdades de gênero.

1.2 MOTIVAÇÕES E JUSTIFICATIVA

Segundo dados da PNAD Contínua (Pesquisa Nacional por Amostra de Domicílios Contínua) de 2019, o número de mulheres no Brasil é superior ao de homens, representam 51,8% da população. Na faixa etária até 24 anos, os homens tiveram estimativa superior a das mulheres. Contudo, a partir dos 25 anos de idade, a proporção de mulheres é maior que a dos homens em todos os grupos de idade (IBGE 2019).

De acordo com a ONU Mulher, apesar de representarmos maioria da população no Brasil apenas 50% das mulheres em idade economicamente ativa (idade entre 10 e 65 anos) participam do mercado de trabalho. Todavia, entre os homens, o índice sobe para 76%, demonstrando uma concentração de renda no sexo masculino. Nesse contexto, outro aspecto da desigualdade de gênero é a disparidade salarial: Em média, homens ganham 23% a mais do que as mulheres (IBGE 2019, ONU 2017).

A motivação do trabalho veio de uma inquietação pessoal. No ano de 2016, ano da minha conclusão em Engenharia de Redes, 1,1 milhão de estudantes concluíram a educação superior e a maioria eram mulheres: 62,2% do total de formandos, segundo dados do Censo da Educação Superior do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Dados daquele mesmo ano apontavam que a conquista da ocupação das salas de aulas pelas mulheres ainda não havia gerado reflexos significativos em sua participação no mercado de trabalho: do total de brasileiros ocupantes do trabalho formal, 58,8% eram homens e apenas 41,2% eram mulheres, de acordo com dados do PNAD.

Essa inquietação se intensificou ao adentrar o mercado de trabalho. Na área de TI mulheres são minoria, homens costumam ganhar mais e ocuparem cargos de chefia. Infelizmente essa realidade não é exclusiva da área de formação da autora.

Partindo do pressuposto de que empregabilidade está diretamente ligada à educação, presumi-se que,

com mais mulheres se formando e se dedicando a obter qualificação profissional de alto nível, logo também teríamos mais mulheres ocupando postos de trabalho. No entanto, os dados demonstram resultados contrários.

Para reverter esse cenário, a ONU Mulheres recomenda o acompanhamento estatístico do mercado de trabalho feminino, a criação e o cumprimento de leis que garantam a igualdade salarial para trabalhos de igual valor. Além disso, políticas públicas também devem estabelecer licenças parentais remuneradas e voltadas tanto para pais quanto para mães. Outra medida é investir na oferta de serviços de cuidado infantil (ONU 2017).

Nesse contexto, este trabalho concentrar-se no estudo da aplicação das técnicas de aprendizado de máquina, utilizando dados oriundos do observatório do trabalho do governo para auxiliar o governo a criar políticas publicas de igualdade de gênero e desmistificar o contexto de oportunidades do mercado de trabalho feminino e assim, reduzir esse cenário díspar.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Este trabalho tem como objetivo analisar fatores de empregabilidade feminina e analisar o perfil da mão de obra feminina brasileira. Para isso, a dissertação propõe uma abordagem baseada em mineração de dados e aprendizado de máquina que produz modelos preditivos de empregabilidade, fornecendo os principais fatores que afetam o modelo preditivo e os mais relevantes utilizando as informações da RAIS e Cadastro Geral de Empregados e Desempregados (CAGED).

1.3.2 Objetivos Específicos

Os objetivos específicos a serem atingidos são os seguintes:

- Analisar o problema de empregabilidade da distribuição de empregadas e desempregadas por região;
- Analisar e propor uma metodologia para extração das informações dos fatores de empregabilidade e analisar como as mulheres são contratadas;
- Utilizando python e aprendizado de máquina para a modelagem preditiva dos fatores relevantes da empregabilidade feminina;
- Analisar e comparar os resultados a partir das técnicas de classificação supervisionada;

A partir dos objetivos expostos, foram realizados experimentos apresentados que serão apresentados no Capítulo 4 com o intuito de responder as seguintes questões:

1. Qual e a melhor forma de composição dos *datasets* ?
2. Qual a melhor forma de visualização dos dados?

3. O modelo proposto alcançou um bom desempenho para os padrões da literatura?

4. E possível propor políticas públicas com base nos indicadores obtidos ?

1.4 ESTRUTURA DO TRABALHO

Neste primeiro Capítulo foi realizado uma breve contextualização e foram determinados objetivos a serem alcançados. O restante deste trabalho é estruturado de forma que:

No Segundo Capítulo será apresentado o estado da arte das pesquisas de questão de gênero e empregabilidade. Além disso, algumas pesquisas relacionadas que auxiliaram na resolução do problema proposto.

Em seguida, no Capítulo 3 é apresentado a fundamentação teórica para o problema de empregabilidade, além da definição do problema, abordando-se conhecimentos relacionados à estratégia de solução: aprendizado de máquina e mineração de dados.

O Capítulo 4 define a forma como o trabalho foi desenvolvido. Apresenta-se detalhes sobre as bases de dados utilizadas no trabalho; as ferramentas usadas na implementação do trabalho: o Sistema de Gerenciamento de Banco de Dados, linguagens de programação e bibliotecas utilizadas, sistema de visualização de dados georreferenciados. Por fim, a forma como os dados foram modelados.

No Capítulo 5 são apresentados e discutidos os resultados obtidos.

Por fim, no Capítulo 6, é feito algumas considerações finais e apresenta-se possíveis propostas de trabalhos futuros.

2 ESTADO DA ARTE E TRABALHOS CORRELATOS

Este capítulo apresenta o levantamento dos principais artigos científicos, trabalhos de mestrado e doutorado que contribuíram no desenvolvimento desse trabalho.

2.1 EMPREGABILIDADE FEMININA

No Capítulo anterior, iniciou-se a apresentação da importância da produção de dados de empregabilidade para um país. Devido à sua importância, o estudo do tema atrai a atenção de pesquisadores de diferentes áreas com diferentes abordagens e metodologias.

A maioria dos estudos concentram-se em identificar as competências mais exigidas pelos empregadores, ou elementos que influenciam a empregabilidade. Outro tema comum é a relação entre a formação educacional oferecida pelas instituições acadêmicas e como os ex-alunos atuam no mercado de trabalho após o egresso.

Assim, essas pesquisas têm o objetivo de avaliar a compatibilidade entre as competências adquiridas e as exigidas no meio profissional. Além disso, existem estudos que tentam relacionar o contexto sociodemográfico, a estrutura institucional de cada empresa e a estrutura do mercado de trabalho.

As pesquisas (García-Aracil e Velden 2008), (Biesma et al. 2007), (Kelly, O'Connell e Smyth 2010) realizam um monitoramento da inserção do jovem no mercado de trabalho em diferentes países europeus, em especial os recém graduados. Ademais, esses trabalhos permitiram aprofundar o estudo da relação educação-mercado de trabalho, bem como as compatibilidades do ensino universitário e seus efeitos na empregabilidade.

Na maioria das pesquisas são aplicadas metodologias analíticas, embora não exista uma prática comum. Nas citadas acima são utilizados técnicas e ferramentas mais sofisticadas para a aplicação das metodologias analíticas vinculadas à econometria, psicometria e outros métodos quantitativos e qualitativos da pesquisa em ciências sociais (McQuaid e Lindsa 2005).

Vale ressaltar que esses trabalhos utilizam principalmente estatística descritiva, aplicando medidas básicas de tendência central e distribuição de frequências. A apresentação dos dados é composta por gráficos de clara visualização e interpretação: Histogramas, gráficos de barras e gráficos de pizza. E mesmo assim, apresentam ótimos resultados(Schomburg e Teichler 2007).

Os dados das pesquisas fazem parte do acompanhamento do observatório do trabalho da União Europeia. Contudo, os atributos das bases de dados que compõem o estudo diferenciam-se muito entre si. As informações analisadas mais recorrentes são a taxa de sucesso na forma utilizada para procurar emprego, o tempo (em meses) necessário para encontrar o primeiro emprego, o salário médio, o percentual de satisfação com seus empregos ou estudos, a distribuição dos graduados de acordo com a escolaridade e a correspondência, o nível médio de competências exigidas pelos empregadores e adquiridas nas universida-

des.

Essas pesquisas concentram-se no desenvolvimento de estratégias e formulação de políticas para colocação dos jovens europeus recém-formados no mercado de trabalho. Desafio para todas as economias. Mas também, buscam auxiliar as instituições a aprimorarem as disciplinas ofertas e promover a formação plena dos futuros empregados, intensificando a capacitação nos elementos detectados como primordiais para obtenção de emprego, melhores salários e satisfação.

Como jovens recém formados, a colocação feminina no mercado formal de trabalho é um desafio, mulheres lideram o mercado informal e recebem menores salários. A dificuldade ainda se agrava durante crises econômicas, cenário das pesquisas citadas (McQuaid e Lindsa 2005).

Além disso, em um contexto de suscetíveis crises econômicas, os economistas estão sempre buscando prever os comportamentos da economia, baseando-se, por exemplo, em dados de de desemprego e assim, antecipar possíveis ações e evitar catástrofes. O artigo (McQuaid e Lindsa 2005) tenta prever o desemprego através de um modelo de regressão linear usando erro de correlação linear. Esse modelo é muito utilizado para previsão de desemprego e é possível encontrar alguns estudos com abordagem semelhante. Nesse viés, o artigo (Koyanagi 2010) tenta prevê a eficiência do seguro desemprego através de estatística descritiva e gráficos de barras. Similar a essas propostas, encontram-se alguns trabalho que analisam a rotatividade do empregado em um empresa, como (Punnoose e Pankaj 2016).

No campo de estudos de ciências social e psicologia existem diversas pesquisas relativas às questões de equidade de gênero e empoderamento feminino. Um destaque é a tese de doutorado (Franco 2018) que estuda o fenômeno *Opt-out* no brasil. A pesquisa tenta buscar fatores psicológicos que levam uma mulher extremamente qualificada a abandonam o mercado de trabalho para se dedicar aos cuidado de pessoas por prazo indeterminado.

Surpreendentemente, (Osborne 2015) e (Pearson, Frehill e McNeely 2015) mostram que no Brasil a produção científica é igualitária, ou seja, possuímos um número similar de homens e mulheres com produção acadêmicas relevantes. Além disso, a proporção de artigos por área de estudo não é tão discrepante. Conforme Figuras 2.1 e 2.2. Infere-se, portanto, que a comunidade acadêmica é mais igualitária que o mercado de trabalho formal. Além disso, mostra que mulheres estão tão qualificadas quanto os homens, porém, não conseguem efetivamente adentrar o mercado formal de trabalho ou posteriormente abandonam-o.

Figura 2.1: Proporção e número de pesquisadores por gênero no Brasil. Os valores destacados em verde são para mulheres e roxo para homens (Osborne 2015)

Em geral, a produção acadêmica das mulheres conta muito mais com elementos interdisciplinar do que a masculina. No japão, mulheres publicam mais artigos que homens, porém homens publicam mais livros acadêmicos. Outro ponto relevante, é que mulheres cientista tender a não ter filhos, ou reduz sua produção durante os dois primeiros anos da criança. Em contra partida, homens cientistas com filhos até dois anos possuem um crescimento em sua produção acadêmica durante esse período(Osborne 2015).

A pesquisa rastreou várias revistas cientificas por 20 anos (1995-2015), de mais de 12 países, e 27 áreas de pesquisa. Portanto, contou com um volume maciço de dados. É utilizado estatística descritiva e gráficos

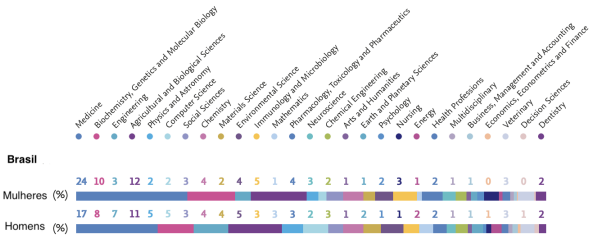


Figura 2.2: Proporção de pesquisadores por área de pesquisa para cada gênero no Brasil (Osborne 2015)

de barras para apresentar os dados, como os apresentados nas Figuras 2.1 e 2.2 para cada país e área.

As pesquisas (Dolado e Felgueroso 2004) e (Dolado e Felgueroso 2001) examinam o papel da educação e de outros fatores socioeconômicos na segregação ocupacional por gênero na União Europeia(UE) e os EUA. Posteriormente, tentam verificar se existem relação entre as características do trabalho, remuneração e oportunidades de promoção e gênero na empregabilidade. O trabalho (Dolado e Felgueroso 2001) faz a comparação entre UE e EUA utilizando os resultados obtidos em (Dolado e Felgueroso 2004).

Os dados comparativos não são de qualidade, segundo o próprio autor. No entanto, é apresentado inicialmente informações amostrais comparativas entre entre a taxa de empregabilidade e o nível educacional para diferentes faixas de idade para EUA e alguns países da União Europeia. Em seguida, informações da participação feminina no mercado de trabalho. Esse resultado é extraído da participação feminina nos órgãos legislativos, cargos de chefia corporativos e cargos de chefia em geral dos diferentes países (Dolado e Felgueroso 2001).

Posteriormente, algumas análises de regressão são realizadas a fim de descobrir possíveis correlações entre o nível de segregação ocupacional (por país, nível educacional e faixa de idade) e outras variáveis os quais foram julgadas potencialmente relacionadas a segregação, como participação no mercado de trabalho de um país e se trabalha em tempo integral ou parcial.

Um resultado importante da pesquisa foi os coeficientes estimados das regressões para a escolaridade e idade: Apontam que as mulheres mais jovens e mais educadas se saem muito melhor do que suas mais velhas de mesmo nível educacional. Apesar de mulheres mais novas participarem mais do mercado de trabalho, a participação masculina mostra-se muito superiores em todas as faixas etárias.

Outro resultado importante é que a segregação ocupacional por gênero parece estar positivamente correlacionada com a participação de empregos de meio período na economia, pois esses são predominantemente femininos. Existem fatores sociológicos para acreditar que as mulheres preferem racionalmente esses arranjos de trabalho de horário flexível. E são confirmados, pois a pesquisa aponta que o grau de satisfação profissional das mulheres é alta nesses empregos (Dolado e Felgueroso 2001).

A disparidade salarial de gênero diminuiu tanto na UE quanto nos EUA. A correlação entre o nível de ocupação feminina e a disparidade de gênero na remuneração é fraco para as características pessoais e do trabalho. Porém, os postos de trabalho onde as mulheres estão pouco representadas tendem a ser aquelas em que a proporção de mulheres com funções de chefia é menor. (Dolado e Felgueroso 2004).

O artigo (Haynes, Western e Spallek 2005) utiliza dados do órgão público de monitoramento do tra-

balho da Austrália (em inglês, the Household, Income and Labour Dynamics in Australia (HILDA)). A proposta é estudar o perfil da empregabilidade feminina do país das mulheres de 20 a 55 anos entre 2001-2003. O estudo conta com informações de 3755 mulheres, empregadas e desempregadas.

A pesquisa compara 3 trabalho realizados no departamento da Universidade de Melbourne, cada trabalho realizado em um ano. É utilizado regressão linear para avaliar fatores que influenciam na empregabilidade feminina, como idade, número de filhos, idade dos filhos, educação, entre outros.

O estudo acaba avaliando o fenômeno *Opt-out* nas mulheres australianas com alto grau de escolaridade com crianças em idade de pré-escola. Como nas pesquisas realizada no USA e UE, mulheres mais jovens se saem muito melhor do que suas mais velhas de mesmo nível educacional, mesmo nas mais velhas sem filhos. Mostrando que o fator idade é importantíssimo nesse países desenvolvidos, porém filhos não influenciam tanto.

Os resultados obtidos nesses estudos, mostram a relevância do acompanhamento constante da empregabilidade feminina. Para isso, é necessário dados consistentes para resultados cofiáveis e comparativos. Portanto, um monitoramento governamental que produza uma base de dados de volume considerável e confiável é essencial. No entanto, volumes expressivos de dados necessitam de uma capacidade computacional alta para seu processamento.

Nesse contexto, é necessário a integração dos estudos das ciências humanas e o avanço na ciência de dados. Apesar de muitas destas pesquisas se limitarem a análises descritivas dos dados, os resultados apresentados são valiosos e importantes que auxiliam muito no entendimento das disparidades de gênero e que podem auxiliar na criação de politicas públicas para reduzi-las.

2.2 UTILIZAÇÃO DE MINERAÇÃO DE DADOS E APRENDIZADO DE MÁQUINA NO ESTUDO DA EMPREGABILIDADE

Os trabalhos (Shabana, Gracious e Subramonian 2016) e (Pawha e Kamthania 2019) têm o objetivo de estudar o mercado de trabalho indiano para engenheiros recém formados. No primeiro artigo, o autor identifica fatores que influem no salários oferecidos para os engenheiros na Índia. A proposta foi utilizar técnicas de aprendizado de máquina para identificar lacunas e oportunidades para os futuros engenheiros naquele país. O autor tentou ainda relacionar fatores que garantem maiores salários, como, por exemplo, a universidade de formação e cidade de residência. A base de dados é composta por atributos quantitativos e inteiros. Foram comparados o desempenhos de quatro algorítimos de aprendizagem de máquina: Regressão linear, árvore de decisão, *feature analysis*, análise de correlação e teste t. (Pawha e Kamthania 2019).

O Segundo trabalho realiza um estudo semelhante ao primeiro, utilizando inclusive de vários atributos semelhantes. No entanto, a pesquisa é expandida para todos os recém formados e tem foco na extração dos elementos mais relevantes utilizando árvore de decisão.

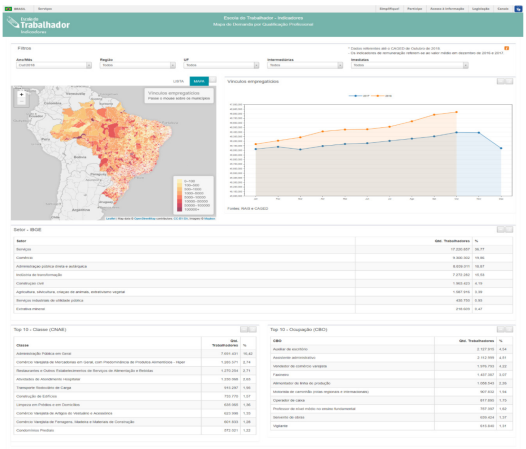
No artigo (Punnoose e Pankaj 2016) foram analisados quais os fatores de evasão dos funcionários de uma empresa, um dos coeficientes que impactam na produtividade das instituições. O interessante desse artigo é a comparação do XGBoost com outros algorítimos. O XGBoost é um algoritimo de apredizado de

máquina baseado em árvore de decisão que utiliza uma estrutura de *Gradient Boosting*, o qual se destaca nas competições de cientistas de dados para resolver desafios por meio de *Machine Learning*.

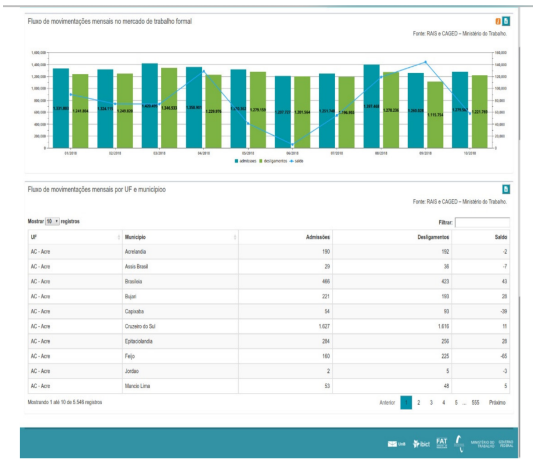
No contexto brasileiro, temos o artigo (Silva et al. 2019) realizado em parceria com o Ministério do Trabalho e o Programa de Disseminação das Estatísticas do Trabalho (PDET). O trabalho tem o proposito de produzir indicadores de empregabilidade brasileira utilizando t*Business Intelligence*-BI.

Os indicadores produzidos são oriundos de dois registros administrativos nacionais e obrigatórios: RAIS - Relação Anual de Informações Sociais - e CAGED - Cadastro Geral de Empregados e Desempregados, à sociedade civil. As duas Bases são as mesmas utilizadas nessa dissertação.

O estudo utiliza-se da ferramenta de BI *Pentaho6* em sua distribuição comercial. A versão comercial mostrou-se mais adequada do que a *freware* para a proposta, pois conta com suporte técnico do fabricante e algumas funcionalidades adicionais. O processo ETL utiliza o *Pentaho Data Integration (PDI)*, com interface de apresentação dos dados, como nas Figuras 2.3 . As informação são dispostas em painéis via web de forma discriminada e intuitiva, em mapas georreferenciáveis e gráficas de piza e barra, facilitando a visualização e avaliação do panorama do mercado de trabalho.



(a) Mapa de Calor com Distribuição CNAE



(b) Admitidos e Desligados Mensalmente

Figura 2.3: Fonte (Silva et al. 2019): http://indicadores.escoladotrabalhador.gov.br/ - Acesso em: Jan. 2019

3 FUNDAMENTAÇÃO TEÓRICA

Nessa seção será apresentado o embasamento teórico necessário no desenvolvimento dessa dissertação. Inicialmente, será exposto um panorama da empregabilidade feminina do Brasil. Posteriormente, discute-se o estudo da empregabilidade e apresenta-se as técnicas de mineração de dados e algoritmos de aprendizado de máquina.

3.1 EMPREGABILIDADE FEMININA NO BRASIL

No ano de 2014, o Instituto Brasileiro de Geografia e Estatística - IBGE realizou um estudo sobre estatística de gênero baseado nos resultados do Censo Demográfico de 2010. O estudo produziu indicadores sociais das mulheres no Brasil, enriquecendo o debate da desigualdade de gênero. Nesse viés, corrobora com dados importantes, demonstrando a necessidade de manter uma agenda pública permanente, que coloque a igualdade de gênero como um dos eixos estruturantes da formulação de políticas públicas no Brasil. Os resultados são dispostos na Cartilha Estatísticas de Gênero Indicadores Sociais das Mulheres no Brasil (Boletim Geográfico 2018)

As estatísticas de gênero devem refletir, segundo informações do Manual de Gênero da Divisão de Estatísticas das Nações Unidas (*United Nations Statistics Division* - UNSD), as questões relacionadas aos aspectos da vida de mulheres e homens, incluindo as suas necessidades, oportunidades ou contribuições para a sociedade (ONU 2017).

As diferenças e desigualdades entre mulheres e homens são moldadas ao longo da história das relações sociais humanas. É importante ressaltar que as diferenças e desigualdades estão relacionadas as atribuições, acesso à recursos e oportunidade de decisão em cada sociedade. Dessarte, o estudo dessas diferenças auxilia a compreensão da estrutura social de uma sociedade. (Franco 2018).

Os parâmetros utilizados para a construção dos indicadores brasileiros estão baseados no Conjunto Mínimo de Indicadores de Gênero - CMIG (*Minimum Set of Gender Indicators - MSGI*), organizado pela Comissão de Estatística das Nações Unidas (*United Nations Statistical Commission*) . Os indicadores foram organizados em cinco domínios: estruturas econômicas, participação em atividade produtivas e acesso a recursos; educação; saúde e serviços relacionados; vida pública e tomada de decisão; e direitos humanos das mulheres e meninas(Boletim Geográfico 2018).

O presente trabalho relaciona-se ao domínio dos indicadores de estrutura econômicas, participação em atividades produtivas e acesso a recursos, ou seja, o monitoramento do mercado de trabalho feminino. Nesse contexto, o estudo da empregabilidade feminina no Brasil avalia a participação da mulher em atividades produtivas e consequentemente sua capacidade de acesso a recursos e sua colocação na estrutura econômica brasileira.

Segundo dados da PNAD Contínua 2019, o número de mulheres no Brasil é superior ao de homens, representam 51,8% da população. Na faixa etária até 24 anos, os homens tiveram estimativa superior a das

mulheres. Contudo, a partir dos 25 anos de idade, a proporção de mulheres é maior que a dos homens em todos os grupos de idade (IBGE 2019). Portanto, o número de mulheres em idade economicamente ativa (15 e 65 anos) é superior ao de homens. Porém, apenas 41,2% dos ocupantes de postos de trabalhos no mercado formal são mulheres, enquanto, 58,8% do total são homens.

No Brasil, as mulheres dedicam-se aos cuidados de pessoas e/ou afazeres domésticos cerca de 73% a mais de horas do que os homens (18,1 horas contra 10,5 horas), conforme 3.1. Observando a realidade regional, verifica-se uma maior desigualdade na distribuição de horas dedicadas a estas atividades. A maior discrepância é na região nordeste, onde as mulheres dedicam cerca de 80% a mais de horas do que os homens, 19 horas semanais (Boletim Geográfico 2018).

Além disso, analisando a dedicação aos cuidados de pessoas e/ou afazeres domésticos por cor ou raça, a questão racial não é indicador de grande discrepância. Observa-se que o indicador pouco varia para os homens quando se considera a cor ou raça. E entre as mulheres,pretas ou pardas são as que mais se dedicam aos cuidados de pessoas e/ou aos afazeres domésticos, com o registro de 18,6 horas semanais em 2016 (Boletim Geográfico 2018).

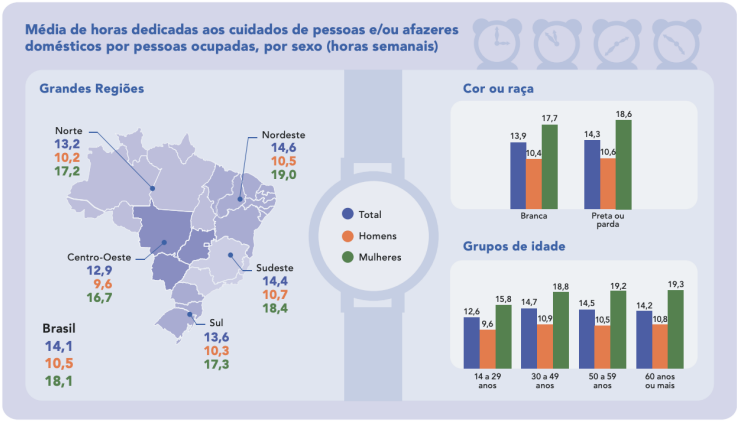


Figura 3.1: Indicador média de horas dedicadas aos cuidados de pessoas e/ou afazeres domésticos por pessoas ocupadas , por sexo (horas semanais) (Boletim Geográfico 2018)

Mesmo em meio a tantas transformações sociais ocorridas no último século sob a perspectiva de gênero (maior participação das mulheres no mercado de trabalho, aumenta a escolaridade, redução da fecundidade), as mulheres seguem dedicando relativamente mais tempo aos afazeres domésticos e/ou cuidados de pessoas (ONU 2017). Nesse viés, a participação feminina no mercado de trabalho é menos expressiva que a masculina no mercado de trabalho.

O indicador acima mostra uma realidade das mulheres brasileiras, mesmo conquistando o mercado de trabalho, a mulher continua conciliando o trabalho remunerado com os afazeres domésticos e/ou cuidados de pessoas, e para isso, acabam trabalhando em ocupações com carga horária reduzida. O indicador proporção de ocupados em trabalho por tempo parcial, Figura 3.2, mostra um percentual mais elevado de mulheres que trabalha em período parcial, de até 30 horas semanais, quando comparado com os homens.

Na observação por regiões mais uma vez as desigualdades regionais são marcantes, as Regiões Norte e Nordeste concentram um maior número de mulheres que trabalham em período parcial. Ademais, na desagregação do indicador por cor ou raça, as desigualdades de gênero é ainda mais expressiva, as mulheres

pretas ou pardas as que mais exercem ocupação por tempo parcial, alcançando 31,3% do total, enquanto 25,0% das mulheres brancas se ocuparam desta forma, em 2016. Para os homens, somente 11,9% dos brancos ocupam-se por tempo parcial, ao passo que a proporção de pretos ou pardos alcançou 16,0% (Boletim Geográfico 2018).

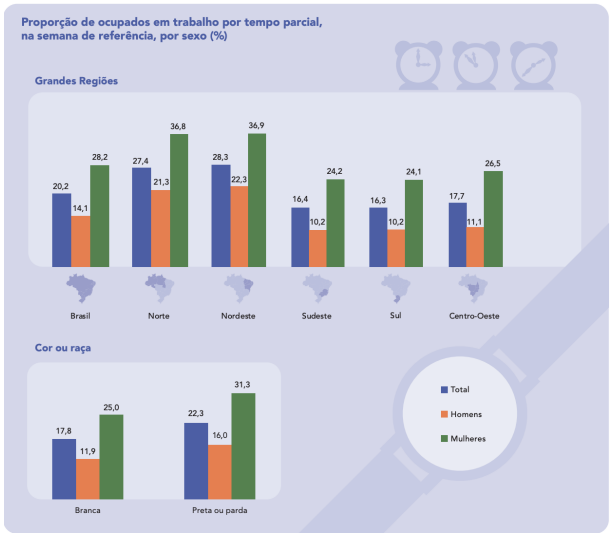


Figura 3.2: Proporção de ocupados em trabalho por tempo parcial, na semana de referência, por sexo (%) (Boletim Geográfico 2018)

Com relação ao indicador dos rendimentos médios do trabalho, as mulheres seguem recebendo cerca de $3/4$ do que os homens recebem. O que pode contribuir para esse resultado é a própria natureza dos postos de trabalho ocupados pelas mulheres, em que se destaca a maior proporção dedicada ao trabalho em tempo parcial.

A Figura 3.3 mostra a razão do rendimento habitual por horas trabalhadas de homens e mulheres. Apesar da disparidade salarial de gênero vir diminuído, ela ainda é expressiva. Essa desigualdade pode estar relacionada com a segregação ocupacional e discriminação salarial das mulheres no mercado de trabalho, conforme vasta literatura e indicadores divulgados acerca das desigualdades de inserção ocupacional das mulheres (ONU 2017).

Nesta comparação, os resultados desagregados por nível de instrução apontam que o diferencial de rendimentos é mais elevado na categoria ensino superior completo ou mais, em que as mulheres recebem 63,4% do que os homens. Essa desigualdade é ainda maior nas mulheres com pós graduação, recebem 78%. Ou seja, mulheres mais instruídas possuem uma disparidade salarial maior do que as menos instruídas.

3.2 ECONOMETRIA E O ESTUDO DA EMPREGABILIDADE

O estudo da empregabilidade é realizada pela econometria, área de estudo das ciências econômicas. Em uma interpretação literal, econometria significa "medição econômica". Ela é a aplicação de modelos estatísticos matemáticos a dados econômicos para dar suporte empírico aos modelos formulados pela economia matemática e , portanto, obter resultados numéricos (Gujarati e Porter 2008).

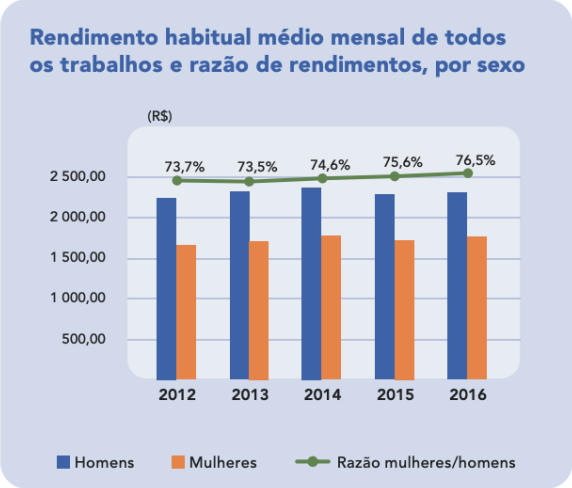


Figura 3.3: Rendimento habitual médio mensal de todos os trabalhos e razão de rendimentos, por sexo (Boletim Geográfico 2018)

A econometria pode ser definida como a ciência social em que as ferramentas da teoria econômica, da matemática e da inferência estatística são aplicadas à análise dos fenômenos econômicos. Utiliza-se de premissas estatísticas para testar hipóteses e prever tendências de possíveis fenômenos econômicos (Wooldridge 2003).

A arte do econometrista está em encontrar o conjunto de soluções ótimas suficientemente específicas e realistas que lhe permitam tirar o melhor proveito dos dados de que dispõe (Theil 1971).

As teorias econômicas fazem declarações ou hipóteses principalmente de natureza qualitativa, ou seja, baseia-se em variáveis qualitativas, aquelas as quais não podem ser mesuradas numericamente, fundamenta-se em qualificar as variáveis(Samuelson 1955).

Por exemplo, a teoria microeconômica afirma que, se todos os demais fatores permanecerem inalterados, uma redução no preço de uma mercadoria deve resultar no aumento da quantidade demandada por esta mercadoria. Então, essa teoria postula uma relação negativa ou inversa entre o preço e a quantidade demandada de uma mercadoria. Mas a teoria em si não oferece nenhuma medida quantitativa da relação entre as duas variáveis, ou seja, não informa quanto a quantidade aumentará ou diminuirá em consequência de determinada variação no preço da mercadoria. Cabe ao econometrista oferecer essas estimativas numéricas. Em outras palavras, o econometrista proporciona conteúdo prático à maior parte das teorias econômicas (Samuelson 1955).

A principal preocupação da economia matemática é expressar de forma matemática um fenômeno, através de equações, sem levar em conta se a teoria pode ser medida ou verificada empiricamente. A estatística econômica busca principalmente a coleta, processamento e apresentação dos dados econômicos na forma de gráficos e tabelas. Essa é a tarefa do estatístico econômico. É ele o principal responsável por coleta e análise dos dados brutos dos fenômeno sobre o produto nacional bruto (PNB), o emprego, o desemprego, os preços e etc (Gujarati e Porter 2008).

Para tanto, a metodologia econométrica segue os seguintes passos (Gujarati e Porter 2008):

1. Exposição da teoria ou hipótese;
2. Especificação do modelo matemático da teoria;
3. Especificação do modelo estatístico ou econométrico;
4. Obtenção dos dados;
5. Estimação dos parâmetros do modelo econométrico;
6. Teste de hipóteses;
7. Projeção ou previsão;
8. Uso do modelo para fins de controle ou de política pública.

A qualidade dos dados coletados é fundamental. Muitas vezes eles são qualitativos (idade e gênero, por exemplo) e precisam ser relacionados com dados numéricos, ou monetários, dados chamados quantitativos (Samuelson 1955).

Na econometria, nem sempre os resultados gerados encontram respaldo nas teorias econômicas clássicas. Diante desse fato, é necessário uma auditória na teoria ou hipótese proposta ou nos próprios dados coletados. E posteriormente, pode surgir novas teorias que justifique os resultados obtidos, levando em conta a qualidade e integridade dos dados (Wooldridge 2003).

A ciência estatística possui muitas técnicas de análise de dados. Uma delas é conhecida como análise de regressão. O termo regressão foi primeiramente utilizado por Sir Francis Galton (1822 – 1911), que estudou a relação entre as estaturas de crianças e as estaturas de seus pais (Gujarati e Porter 2008).

A análise de regressão é a principal ferramenta utilizada na econometria. Os resultados desta análise estão mais precisos devido aos avanços computacionais e dos softwares estatísticos (Wooldridge 2003).

Um outro problema é as próprias limitações da regressão linear: Como é um modelo matemático não faz juízo de valor, ele tende a indicar uma relação entre as variáveis mensuradas, independentemente da hipótese testada ser totalmente absurda. Por isso, é importante uma limpeza prévia da base de dados de dados esdrúxulos, do inglês *outliers* (Wooldridge 2003).

3.2.1 Análise de Regressão

A análise de regressão diz respeito ao estudo da dependência de uma variável em relação a uma ou mais variáveis. A proposta é estimar e/ou prever o valor médio (da população) da primeira em termos dos valores conhecidos ou fixados (em amostragens repetidas) das segundas (Gujarati e Porter 2008).

Um exemplo de aplicação da regressão na economia é o estudo da relação de dependência das despesas de consumo pessoal e a renda pessoal disponível, após o pagamento de impostos. Essa análise é útil para estimar a propensão marginal a consumir (PMC), isto é, a variação média nas despesas de consumo de uma população, com relação à variação do dólar na renda real(Theil 1971).

Outro exemplo é a possibilidade de fixar um preço ou a produção (mas não ambos) e descobrir a resposta da demanda por um produto perante variações nos preços. Isso permite estimar a elasticidade-preço (isto é, a resposta dos preços) da demanda pelo produto e a contribuição na determinação do preço mais lucrativo (Theil 1971).

Um economista do trabalho pode querer estudar a relação entre a variação dos salários nominais e a taxa de desemprego. A Figura 3.4 mostra um exemplo de diagrama de dispersão dos dados históricos da relação entre taxa de desemprego e taxa de variação dos salários nominais. A curva traçada é um exemplo hipotético da famosa curva de Phillips. Esse diagrama de dispersão permitiria ao economista prever a variação média dos salários para uma dada taxa de desemprego. Tal conhecimento poderia contribuir para esclarecer o processo inflacionário de uma economia, visto que o aumento dos salários nominais tende a refletir-se em aumento de preços (Theil 1971).

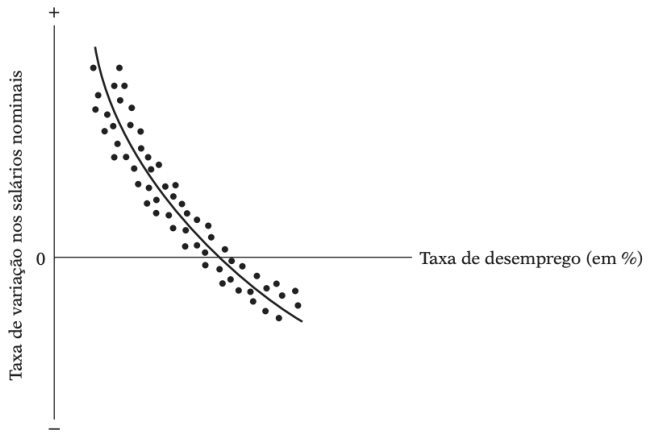


Figura 3.4: Curva de Phillips hipotética (Theil 1971)

De forma simplificada, a análise de regressão ou modelo de regressão corresponde a uma equação matemática, que descreve a relação entre duas ou mais variáveis. Nesse sentido, o modelo estuda um fator o qual é causador de outro fenômeno. Nesse caso, chama-se variável independente a variável considerada causa de um efeito e variável dependente a variável considerada efeito de uma causa. E assim, prever comportamentos com base na associação entre duas variáveis que geralmente possuem uma boa correlação (Morettin e Bussab 2007).

Os efeitos de duas ou mais variáveis independentes sobre uma variável dependente é um problema de análise de regressão múltipla. O estudo de uma única variável independente (geralmente a mais importante) sobre uma variável dependente é chamado de regressão linear (Morettin e Bussab 2007).

Quando a relação entre as duas variáveis, x independente e y dependente é uma regressão linear pode ser descrita pela equação, $y = \beta_0 + \beta_1 x$, ou melhor:

$$y_i = \beta_0 + \beta_1 x_i \qquad i = 1, 2, ..., n \tag{3.1}$$

Os β_0 e β_1 são os parâmetros da população. Como os dados da população são difíceis de obter, utiliza-se valores estimados, que são calculados utilizando-se de dados gerados pelas amostras.Cada um dos conjuntos de valores do estimador de β_0 e β_1 fornece uma linha reta diferente. O intercepto é forne-

cido com base no termo constante na equação e corresponde ao valor do estimador de y quando x é zero (Morettin e Bussab 2007).

Entretanto, em muitos casos, a relação entre duas variáveis não é exata, não inclui todas as variáveis que influência no comportamento da variável dependente. Assim, é preciso adicionar um termo ϵ, um erro de estimação que levar em consideração a ausência de outras variáveis no proposto. Considerando os dois fenômenos a representação do modelo mais realista é:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon \qquad i = 1, 2, ..., n \tag{3.2}$$

Ademais, o grau de relação entre as variáveis é calculado pelo coeficiente de Pearson. Assim, para selecionar a variável mais importante, basta calcular o coeficiente de Pearson para cada uma das variáveis disponíveis e verificar qual a mais correlação(Morettin e Bussab 2007).

Por fim, para testar o erro aleatório é necessário validar as premissas do modelo, por exemplo, erro aleatório com média zero, erro com distribuição normal, ser independente, ter variância constante.

3.3 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

Modernas tecnologias computacionais e estatísticas têm sido desenvolvidas para suprir as necessidades de descobrir relações significativas, mas desconhecidas, em Bancos de Dados (BD). Esta área de pesquisa é denominada de Extração Inteligente e Automática de Conhecimento em Bancos de Dados (EIACBD). A EIACBD focaliza o desenvolvimento de métodos de extração eficientes e que possam ser utilizados em escala de bancos de dados de diversas dimensões, ou seja, proporcionam o escalonamento. A EIACBD é uma forma de expressar o processo de descoberta de conhecimento em BD (*Knowledge Discovery in Database* - KDD) que também é conhecido como mineração de dados (Data Mining) (Han, Kamber e Pei 2000).

Os dados estruturados, armazenados na maioria dos Sistemas de Gerência de Bancos de Dados(SGBD), são mais fáceis de serem tratados por meios computacionais, porque existem linguagens formais, como SQL, que permitem sua manipulação e consulta de forma mais concisa e precisa (Tan, Steinbach e al. 2009). Os dados não estruturados, por outro lado, necessitam de mecanismos computacionais diferentes dos tradicionalmente usados, para que possam ser coletados, armazenados, manipulados e consultados.

O desafio principal da extração de conhecimento é processar automaticamente e inteligentemente grandes quantidades de dados brutos(dados operacionais), identificar os padrões mais significantes e representativos, e apresentar estes modelos ou padrões como conhecimento apropriado para alcançar os objetivos do usuário(Phridvi e Guru 2013).

Para isso, os dados devem passar por uma transformação em sua forma e conteúdo. Os dados processados são chamados de transacionais e devem ser convertidos em informação e disponibilizados em um ambiente adequado de coleta, armazenamento e publicação. Essas informações transacionadas possibilitam a utilização de técnicas para descoberta de conhecimento, gerando insumos informacionais de acordo com o domínio necessário na aplicação(Tan, Steinbach e al. 2009).

A figura 3.5 apresenta a sintetização do processo de extração do conhecimento em um banco de dados até a experiência:

Figura 3.5: Processo de Extração de Dados até a Experiência

3.3.1 *Knowledge Discovery in Database* - KDD

O KDD foi proposto em 1989 para referir-se às etapas que produzem conhecimentos a partir de dados relacionados, sendo a mineração de dados a etapa que transforma dados em informações. O KDD refere-se ao processo de extração da informação relevante ou de padrões nos dados contidos em grandes BD e que sejam: não-triviais, implícitos, previamente desconhecidos e potencialmente úteis (Fayyad et al. 1996).

A expressão Mineração de Dados (DM) surge inicialmente, como um sinônimo de KDD, mas é apenas uma das etapas da descoberta de conhecimento em bases de dados no processo global do KDD (Goldschmidt e al. 2015). O processo de KDD é constituído pelas seguintes fases: Obtenção, Pré - Processamento, Transformação, Mineração de Dados e Pós-Processamento. Conforme figura 3.6.

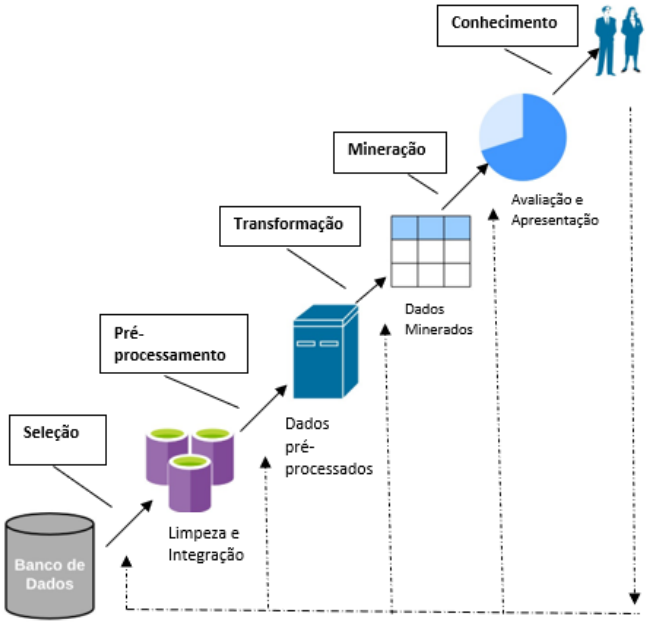


Figura 3.6: Processo KDD - Adaptado de (Fayyad et al. 1996)

Na primeira fase, os dados são obtidos e é efeito um entendimento da situação do problema. Em seguida, a fase de pré-processamento trata da limpeza da base e integração dos dados. Realiza-se a remoção de valores esdrúxulos (do inglês,*outliers*) e dados irrelevantes, definição de estratégias de tratamento de informações ausentes (do inglês, *missing values*) e seleção dos dados relevantes (Goldschmidt e al. 2015).

Na fase de transformação, os dados são transformados ou consolidados no formato apropriado para mineração, para isso, é realizado redução da dimensão dos dados e da quantidade efetiva de variáveis. Agora

com os dados tratados e transformados é realizado a Mineração dos Dados, é nessa fase que é selecionado o melhor algoritmos de aprendizado de máquina para atingir o objetivo definido (Wu et al. 2008).

Por fim, na fase de pós-processamento, técnicas de visualização e representação do resultados obtidos são aplicados. E a partir da avaliação desses resultados, o conhecimento é extraído pelos usuários (Phridvi e Guru 2013).

3.4 MINERAÇÃO DE DADOS

Segundo (Fayyad et al. 1996) a Mineração de Dados(MD) (do inglês, *Data Mining*) consiste na aplicação de algoritmos para análise e descoberta de conhecimento e na produção de padrões ocultos e previamente desconhecidos ou modelos a partir de grandes Base de Dados, afim de transformar em informações úteis(Han, Kamber e Pei 2000). A preparação dos dados é parte essencial de um projeto de MD: Inspeção e limpeza, as quais são responsáveis pela maior parte do tempo de um projeto de MD (Tan, Steinbach e al. 2009).

A expressão MD surge inicialmente, como um sinônimo de KDD, mas é apenas uma das etapas da descoberta de conhecimento em bases de dados no processo global do KDD, de acordo com a processo apresentado na Figura 3.6 (Goldschmidt e al. 2015). É um campo multidisciplinar que teve suas origens nas tecnologias de banco de dados, aprendizado de máquina, inteligência artificial e estatística entre outras áreas.

Apesar das definições sobre a Mineração de Dados levar a crer que o processo de extração de conhecimento se dá de uma forma totalmente automática, sabe-se hoje que de fato isso não é verdade. Apesar de encontrarmos diversas ferramentas que nos auxiliam na execução dos algoritmos de mineração, os resultados ainda precisam de uma análise humana. Porém, ainda assim, a mineração contribui de forma significativa no processo de descoberta de conhecimento, permitindo aos especialistas concentrarem esforços apenas em partes mais significativa dos dados (Lorose 2005).

É importante distinguir o que é tarefa e uma técnica de mineração. A tarefa consiste na especificação do que estamos querendo buscar nos dados, que tipo de regularidades ou categoria de padrões temos interesse em encontrar, ou que tipo de padrões poderiam nos surpreender (por exemplo, um gasto exagerado de um cliente de cartão de crédito, fora dos padrões usuais de seus gastos). A técnica de mineração consiste na especificação de métodos que nos garantam como descobrir os padrões que nos interessam. Dentre as principais técnicas utilizadas em mineração de dados, temos técnicas estatísticas, aprendizado de máquina e técnicas baseadas em crescimento poda-validação (Hand, Mannil e Smyth 2001).

A Mineração de Dados é comumente classificada pela sua capacidade em realizar determinadas tarefas (Lorose 2005). As tarefas mais comuns são: Descrição, Classificação, Estimação ou Regressão, predição, agrupamento e associação.

Descrição e a tarefa utilizada para descrever os padrões e tendências revelados pelos dados. A descrição geralmente oferece uma possível interpretação para os resultados obtidos. A tarefa de descrição é muito utilizada em conjunto com as técnicas de análise exploratória de dados, para comprovar a influência de

certas variáveis no resultado obtido (Hand, Mannil e Smyth 2001).

A classificação é uma das tarefas mais comuns, a classificação, visa identificar a classe a qual um determinado registro pertence. Nesta tarefa, o modelo analisa um conjunto de registros fornecidos, como treinamento, o qual os registros já contendo a indicação à qual classe pertence, a fim de "aprender"como classificar um novo registro (aprendizado supervisionado)(Hand, Mannil e Smyth 2001).

A estimação ou regressão é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais (Hand, Mannil e Smyth 2001). Na seção 3.2.1 foi apresentado a aplicação de regressão linear para resolver problemas econométricos.

A tarefa de predição é similar às tarefas de classificação e estimação,porém ela visa descobrir o valor futuro de um determinado atributo. A tarefa de agrupamento visa identificar e aproximar os registros similares. Um agrupamento (ou cluster) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado). Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares (Hand, Mannil e Smyth 2001).

Por fim, a tarefa de associação consiste em identificar quais atributos estão relacionados. Identificam se um dado atributo X está relacionado com um dado atributo Y. É uma das tarefas mais conhecidas devido aos bons resultados obtidos, principalmente nas análises da "Cestas de Compras"(*Market Basket*), onde identificamos quais produtos são levados juntos pelos consumidores (Hand, Mannil e Smyth 2001).

Segundo (Alpaydin 2010), a aplicação de métodos de aprendizado de máquina para grandes bancos de dados é chamado de Mineração de dados. Portanto, MD está diretamente relacionada com o Aprendizado de Máquina(AM), o objetivo é aplicação de algoritmos específicos para extração de padrões em base de dados. Assim, a ênfase do MD está na aplicação de algoritmos de AM como ferramenta para descobrir padrões que sejam potencialmente valiosos para o processo KDD (Fayyad et al. 1996).

3.5 APRENDIZADO DE MÁQUINA

Definimos o Aprendizado de Máquina (AM) como um processo computacional que busca realizar uma tarefa, aprendendo a partir de uma experiência(um grande volume de dados), procurando melhorar a performance de um processo, através de dois objetivos principais: O desempenho preditivo de modelos e o processo de extração de conhecimento de uma base de dados. Existem diversos algorítimos de aprendizado de máquina do inglês, *Machine Learning - (ML)* (Alpaydin 2010).

O aprendizado de máquina é multidisciplinar. Ele se baseia em resultados de inteligência artificial, probabilidade e estatística, teoria da complexidade computacional, teoria do controle, teoria da informação, filosofia, psicologia, neurobiologia e outros campos (Mitchell 1997).

Segundo (Mitchell 1997, pag 3) define:

> Diz-se que um programa de computador aprende com a experiência E em relação a alguma classe de tarefas T e medida de desempenho P, se seu desempenho nas tarefas em T, medido por P, melhora com a experiência E.

Assim, o papel do aprendizado de máquina pode ser dividido em duas etapas: primeiramente, a etapa de treinamento e posteriormente, necessita-se de um modelo de inferência da eficiência do algoritmo. A etapa de treinamento consiste em escolher um algoritmo eficiente e treiná-lo para resolver o problema de otimização, além de armazenar e processar a enorme quantidade de dados. Na etapa de inferência, algumas métricas de cálculo do erro de validação e previsão (do inglês, *out of sample*) são realizados. Em certas aplicações, a eficiência do algoritmo de aprendizado ou inferência, pode está relacionada a sua complexidade de espaço e tempo (Alpaydin 2010).

A literatura apresenta diversas classificações dos algoritmos de AM, as principais são Aprendizado Supervisionado (do inglês *Supervised Learning*), Aprendizado Não-Supervisionado (do inglês *Unsupervised Learning*) e Aprendizado por Reforço (do inglês *Reinforcement Learning*).

3.6 APRENDIZADO SUPERVISIONADO

No Aprendizado de Máquina Supervisionado há uma entrada X e uma saída Y e a tarefa é aprender o mapeamento da entrada para a saída (Alpaydin 2010). Ou seja, prever ou classificar uma variável dependente a partir de uma lista de variáveis independentes. Regressões e classificações são exemplos de aprendizado de máquina supervisionado.

A abordagem supervisionada no aprendizado assume um modelo definido em um conjunto de parâmetros:

$$y = g(x|\theta) \tag{3.3}$$

Sendo $g(.)$ o modelo e θ os parâmetros do modelo. Y é um número em regressão e a saída da classe (por exemplo, 0/1) no caso de classificações. O modelo $g(.)$ é a função de regressão ou na classificação, a função discriminante que separa as instâncias de diferentes classes. Os algoritmos de AM otimizam os parâmetros, θ, de modo que o erro de previsão/classificação seja minimizado, ou seja, as estimativas são o mais próximas possível dos valores corretos fornecidos no conjunto de treinamento (Alpaydin 2010).

Dentre as técnicas mais conhecidas para resolver problemas de aprendizado supervisionado estão Regressão Linear, Regressão Logística, Redes Neurais Artificiais, Máquina de Suporte Vetorial (ou Máquinas kernel), Árvores de Decisão, entre outros.

Na seção 3.2.1 foi apresentado uma importante aplicação das regressões na economia. Nas próximas seções será feito uma revisão desse algorítimo e apresentado dois outros: Árvore de Decisão e Máquina de Suporte Vetorial.

3.6.1 Regressão

Na seção 3.2.1, o econometrista conhece uma função da economia clássica o qual representa o fenômeno, seu trabalho é avaliar a hipótese a partir de um conjunto de dados presentados pela função. No aprendizado de máquina, a função não é conhecida, mas é apresentado um conjunto de exemplos do tipo:

$$X = \{x^t, r^t\}_{t=1}{}^N \tag{3.4}$$

Sendo o rótulo dado por $r^t \in \mathbb{R}$.

Caso, não exista ruídos (Elementos ocultos, por exemplo) é realizado uma interpolação para extrair a função $f(x)$ que passa entre os pontos do conjunto de dados (Alpaydin 2010):

$$r^t = f(x^t) \tag{3.5}$$

Uma interpolação polinomial consiste em encontrar um polinômio de grau $(N-1)$ dados N pontos que pode ser usado para achar as saída para qualquer x.

Caso, o x esteja fora do intervalo de x^t (Dados de entrada) é realizado uma extrapolação, ou sejam é possível prever um valor futuro usando uma regressão (Alpaydin 2010).

Como dito na seção 3.2.1, normalmente existem ruídos na regressão e é necessário adicionado à saída da função um ruído aleatório ϵ:

$$r^t = f(x^t) + \epsilon \tag{3.6}$$

Sendo z^t o conjunto de variáveis ocultas do problema $g(x)$ A função extraída dos dados apresentados é:

$$r^t = f(x^t, z^t) \tag{3.7}$$

O objetivo da regressão é encontrar o melhor modelo $g(.)$ que minimizar o erro empírico, dado por:

$$E(g|X) = [r^t - g(x^t)]^2 \tag{3.8}$$

Como r e $g(x)$ são valores numéricas, o problema de minimização do erro pode ser definido como a distância entre os valores,por exemplo quadrado da diferenças (Alpaydin 2010).

3.6.2 Árvore de Decisão

Uma árvore de decisão é uma estrutura hierárquica de dados que implementa a estratégia de dividir e conquistar, isto é, um problema complexo é decomposto em subproblemas mais simples e recursivamente a mesma estrategia é aplicada a cada subproblema. É um método não paramétrico eficiente, que pode ser

utilizado como classificação e regressão (Alpaydin 2010), ou seja, um teste de hipótese que testa outras situações de relação dos dados que não parâmetros populacionais (média, variância, por exemplo).

O objetivo de qualquer árvore de decisão é criar um modelo viável que preveja o valor de uma variável de destino com base no conjunto de variáveis de entrada (Gollapudi 2016). os dados de entrada descrevem um conjunto de propriedades para produzir, e podem ser discretos ou contínuos, por exemplo, uma decisão booliana - sim ou não.

O processo de indução de árvores de decisão particiona recursivamente um conjunto de treinamento até que cada subconjunto obtido deste particionamento contenha casos de uma única classe ou atributo alvo. Uma árvore de decisão toma como entrada um objeto ou situação descrito por um conjunto de atributos e retorna uma decisão (Russell 2004).

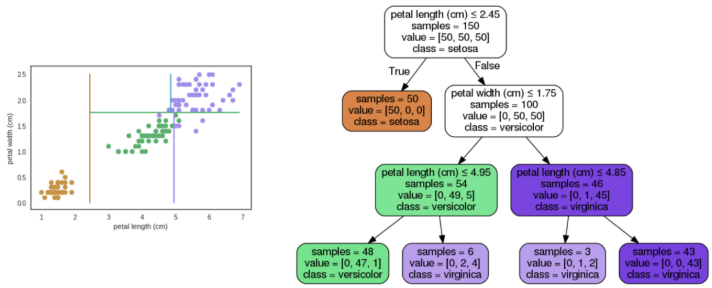


Figura 3.7: Exemplo de Árvore de Decisão do Conjunto de Dados Flores Iris

Uma das principais questões das árvores de decisão é que elas podem criar modelos excessivamente complexos, dependendo dos dados apresentados no conjunto de treinamento (Alpaydin 2010). Para evitar que ocorra o sobre ajuste (do inglês, *overfitting*) do algoritmo de aprendizado da máquina, pode-se rever os dados de treinamento e podar os valores para as categorias, o que produzirá um modelo mais refinado e melhor ajustado.

Existem diversos algoritmos de indução de árvore para o aprendizado, a maioria deles são variações de um algoritmo ID3 e seu sucessor C4.5, principal algoritmo de aprendizado da árvore de decisão (Wu et al. 2008).

A escolha central no algoritmos de aprendizado de árvore de decisão é selecionar qual atributo testar em cada nó na árvore. Gostaríamos de selecionar o atributo mais útil para classificar. O questionamento é qual uma boa medida quantitativa de um atributo. A resposta é diferente para cada algorítimo de indução. Por exemplo, O ID3 utilizada o chamada Ganho de Informação (do inglês, *info gain*) e C4.5 usa Razão de Ganho (do inglês, Gain Ration). A proposta é semelhante em ambas os algorítimos, avaliar quão bom um determinado atributo separa os exemplos de treinamento de acordo com sua classificação de alvo (Mitchell 1997).

3.6.2.1 Algorítimo C4.5

O algorítimo C4.5 é resultado do trabalho publicando em 1993 por Quinlan, o C4.5 é uma evolução do ID3(Quinlan 1993).

Não é necessário fazer uma rotulação dos atributos binariamente. Ele consegue lidar tanto com atributos categóricos (ordinais ou não-ordinais) como com atributos contínuos. Nos atributos contínuos, o algoritmo define um limiar e então recursivamente divide, agora, de forma binária: aqueles cujo valor do atributo é maior que o limiar e aqueles cujo valor do atributo e menor ou igual ao limiar (Quinlan 1993).

Além disso, o algorítimo ignora valores desconhecidos de atributos, os conhecidos no inglês *Missing Values*, nos cálculos de decisão em cada nó. Como mencionado anteriormente, utiliza a medida de Razão de ganho para melhor dividir os exemplos de treinamento. Essa medida demonstraram melhores resultados que seu antecessor ID3, o qual utilizava o ganho de informação (Quinlan 1993).

Para calcular o índice da Razão de Ganho e necessário encontrar o valor da Entropia. A entropia caracteriza a impureza dos dados, num conjunto de dados, determina a falta de homogeneidade dos dados de entrada em relação a sua classificação. Por exemplo, a entropia é máxima (igual a 1) quando o conjunto de dados é heterogêneo (Mitchell 1997).

Dado um conjunto de entrada S que pode ter c classes distintas, a entropia de S será dada por:

$$Entropia(S) = \sum_{i=1}^{c} -p_i log p_i \tag{3.9}$$

Dado p_i como a proporção de dados em S que pertence a classe i.

Assim, é calculado o ganho de informação para um atributo A de um conjunto de dados S, o qual é a medida de diminuição da entropia esperada quando o atributo A for utilizado na partição do conjunto de dados.

Se $P(A)$ é o conjunto de valores que o atributo A pode ter, x um elemento desse conjunto e S_x um subconjunto de S formado pelos dados em que $A = x$, a entropia obtida ao particionar S em função do atributo A é dada por:

$$Entropia(A) = \sum_{x \in P(A)} \frac{S_x}{S} Entropia S_x \tag{3.10}$$

Portanto, a razão do ganho será:

$$Ganho(S, A) = Entropia(S) - Entropia(A) \tag{3.11}$$

Existe um viés natural na medida de ganho de informação que favorece atributos com muitos valores sobre aqueles com poucos valores. Por exemplo, um atributo Data possui um número muito grande de valores possíveis (por exemplo, 4 de março de 1979). Ele teria um alto ganho de informação em relação ao treinamento exemplos e apesar de ser um preditor muito pobre da função alvo em relação às demais instâncias (Mitchell 1997).

Para evitar esse problema (Quinlan 1993) utilizou a taxa de razão do ganho, a qual penaliza atributos como Data, incorporando um termo relacionado à informação divida (do inglês, *Split Information*), que é sensível à amplitude e uniformidade do atributo que divide os dados:

$$SplitInformation(S, A) = \sum_{i=1}^{c} \frac{|S_i|}{S} log_2 \frac{|S_i|}{S} \tag{3.12}$$

Sendo S_1 e S subconjuntos de c exemplos resultantes do particionamento S.

O $SplitInformation(S, A)$ é realmente a entropia de S_i em relação a informação gerada das c partições . Anteriormente considerava-se apenas a $Entropia(S)$ com relação a atributos de destino. Agora, a $Entropia(S)$ é calculada em relação ao valor previsto pela árvore aprendida.

Assim, a medida de Razão de Ganho é definida em termos $SplitInformation(S, A)$ e o $Ganho(S, A)$, medido anteriormente (Quinlan 1993), para selecionar o atributo de acordo com a razão entre seu ganho e seu efetiva quantidade de informação e portanto tentar medir com eficiência se um atributo fornece informações relevantes sobre a classificação correta de um exemplo, conforme a equação:

$$RazaoGanho(S, A) = \frac{Ganho(S, A)}{SplitInformation(S, A)} \tag{3.13}$$

Assim, os atributo que representam os melhores ganhos de informação são selecionado como raiz da árvore. E transforma em folhas os atributos que não representam ganhos significativos. Logo, a proposta inicial de "dividir para conquistar"é cumprida, o problema é retornado em formato de árvore que facilita a interpretação. No entanto, a estratégia da indução do algoritmo C4.5 é gananciosa, pois executa sempre o melhor passo avaliado localmente, sem se preocupar na obtenção do melhor resultado no final, podendo geral árvores gigantescas e com folhas irrelevantes (Quinlan 1993).

3.6.2.2 A Poda de Árvores de Decisão

Todo algoritmo de construção de árvores de decisão tenta encontrar associações entre os atributos dos objetos e a classificação dos mesmos. A força dessas associações varia de acordo com o número de exemplos que suportam ou negam a relação observada. Algumas dessas associações refletirão características genuínas do domínio. Outras serão encontradas ao acaso, devido a efeitos aleatórios ou à escolha particular dos exemplos presentes no sistema (Mitchell 1997).

Quando árvores de decisão são construídas, muitas das arestas ou sub-árvores podem refletir ruídos ou erros. Isso acarreta em um problema conhecido como sobre ajuste (do inglês, overfitting),que significa um aprendizado muito específico do conjunto de treinamento, não permitindo ao modelo generalizar (Breiman 2001).

Nesse sentido, é necessário encontrar o tamanho ideal da árvore. Para detectar e excluir essas arestas e sub-árvores são utilizados métodos de poda da árvore (do inglês, pruning). Ademais, existem duas abordagens de poda: pré-poda ou pós-poda (Quinlan 1993). O emprego de podas pode melhorar o desempenho de generalização de uma árvore de decisão, melhorando a taxa de acerto do modelo para novos exemplos, os quais não foram utilizados no conjunto de treinamento. Consequentemente, a árvore podada se torna mais simples, facilitando a interpretação do usuário (Han, Kamber e Pei 2000).

A pré-poda é realizada durante o processo de construção da árvore. São os processos de divisão do

conjunto de elementos e transformação em nó raiz e folhas de uma árvore, os métodos utilizados na sessão anterior (ganho de informação e razão do ganho) são critérios de poda (Breiman 2001).

A pós-poda realiza-se após a construção da árvore. Nessa abordagem avalia-se os nó da árvore construída e poda-se ramos completos, transformando o nó em folhas, representando a classe mais frequente.

Para cada nó interno da árvore, o algoritmo calcula a taxa de erro caso a sub-árvore abaixo desse nó seja podada, assim como a taxa de erro caso a sub-árvore não seja podada. Se a diferença entre as duas taxas de erro for menor que um valor pré-estabelecida na construção da árvore, então a árvore é podada. Caso contrário, não ocorre a poda (Breiman 2001).

Esse processo se repete progressivamente, gerando um conjunto de árvores podadas. Por fim, para cada uma delas é calculado a acurácia na classificação de um conjunto de dados independente dos dados de treinamento. Por exemplo, um conjunto de validação. E assim, a árvore que obtiver a melhor acurácia será escolhida (Breiman 2001).

Os métodos de poda são amplamente utilizados e eficazes na solução do problema de sobre ajuste. No entanto, deve-se se atentar para não podar demais a árvore. Quando isso ocorre, tem-se o problema de sub-ajuste, em que o modelo de classificação não aprendeu o suficiente sobre o conjunto de dados de treinamento (Breiman 2001).

Alguns metódos de poda são *Cost Complexity Pruning, Reduced Error Pruning, Minimum Error Pruning (MEP), Pessimistic Pruning, ErrorBased Pruning (EBP), Minimum Description Length (MDL) Pruning*.

3.6.3 Máquina de Suporte Vetorial

Uma Máquina de Suporte Vetorial ou Máquina de Kernnel (do inglês, *Kernnel Machine*) é um algoritmo de aprendizado supervisionado que implementa os princípios da teoria estatística da aprendizagem e pode resolver problemas de classificações binárias lineares e não lineares. É um método discriminatório e usa o princípio de Vapnik para nunca resolver um problema mais complexo como o primeiro passo antes do problema real (Alpaydin 2010).

Máquinas de kernel são métodos de margem máxima que permitem que o modelo seja escrito como uma soma das influências de um subconjunto das instâncias de treinamento. Ela desenvolve esse modelo tomando as entradas de treinamento, mapeando elas no espaço multidimensional e utilizando regressão para encontrar um hiperplano (um hiperplano é uma superfície em espaço de n dimensões que o separa em duas metades de espaço) que melhor separa duas classes de entradas. Uma vez que a máquina de vetores de suporte tenha sido treinada, ela é capaz de avaliar novas entradas em relação ao hiperplano divisor e classificá-las em uma entre duas categorias (Alpaydin 2010).

3.7 APRENDIZADO NÃO SUPERVISIONADO

Nesse tipo de aprendizagem, o conjunto de dados utilizado não possui nenhum tipo de rótulo, categorizados previamente. Os algorítimos de aprendizados não supervisionados têm o objetivo de descobrir similaridades entre os objetos analisados a fim de detectar similaridades e anomalias (Lorose 2005).

Os algoritmos avaliam as hipóteses usando critérios como simplicidade, generalidade e performance, para testar hipóteses por meio de experimentos que os próprios algoritmos adquire em seu uso. É possível utilizar essa estrutura, agrupando os dados com base em relações entre as variáveis nos dados. Também pode ser usada para reduzir o número de dimensões em um conjunto de dados para concentrar somente nos atributos mais úteis, ou para detectar tendências (Mitchell 1997).

O grande objetivo desta técnica é agrupar objetos com alto grau de semelhança, que a similaridade é alguma função de distância,por exemplo, distância euclidiana. O algoritmo k-means se destaca no aprendizado não supervisionado (Lorose 2005).

3.7.1 *K-Means* (KNN)

O algoritmo de Analise de Agrupamento *k-means* foi apresentado por JB. MacQueen em 1967 e é um dos mais famosos algoritmos de agrupamento de dados, este algoritmo tenta fornecer uma classificação de acordo com os próprios dados sendo a classificação feita por similaridade de grupos, os quais o objetivo é atribuído ao grupo (*cluster*) ao qual é mais semelhante (MacQueen 1967).

O k-means escolhe k objetos (aleatoriamente ou relacionado à uma escolha heurística) que serão à base de cada grupo (denominados centroides), os demais objetos são associados o centroide mais próximo. A cada passo os centróides são recalculados dentro os objetos de seu próprio grupo e os objetos são realocados para o centroide mais próximo, este procedimento é repetido até que o nível de convergência seja satisfatório de acordo com alguma heurística estabelecida (Lorose 2005).

Algoritmo 1: Pseudo Código k-mens

Result: Agrupa Dados Semelhantes

Entrada: Conjunto de Dados de Treinamento D

while *not at end of this document* **do**

    Selecione k pontos como centroides iniciais.

    repeat;

    Atribua cada objeto ao cluster mais próximo;

end

Recalcula cada centroide de cada cluster.

Nesse tipo de agrupamento exigi-se que as variáveis sejam numéricas ou binárias. No entanto, grande parte das aplicações envolvem dados categorizados. Uma alternativa é converter os dados categorizados em valores numéricos. Além disso, o algoritmo *K-mens* possui alta sensibilidade à valores esdrúxulos (*outliers*). Um objeto discrepante pode modificar substancialmente o agrupamento, prejudicando a clusterização.

3.8 APRENDIZADO POR REFORÇO

No aprendizado por reforço, a máquina tenta aprender qual é a melhor ação a ser tomada, dependendo das circunstâncias na qual essa ação será executada. E assim, diante de cada tomada de decisão uma recompensa ou punição é dada conforme a decisão tomada (Alpaydin 2010).

Nessa abordagem de aprendizado o ajuste dos parâmetros é feito pela interação contínua com o ambiente para minimizar (ou maximizar) um determinado índice de desempenho. Assim, não há um supervisor indicando a saída esperada a cada estímulo fornecido como entrada, mas sim um "crítico" que atribui uma nota para a resposta da máquina de aprendizado dado um circunstância de estimulo, com o objetivo de alcançar o nível máximo de sucesso no seu funcionamento com base em um índice estabelecido (Smola e Vishwanathan 2008).

3.9 MÉTODOS DE REDUÇÃO DE DIMENSÃO DOS DADOS

A complexidade de um problema de classificação ou regressão relaciona-se ao número de entradas. Isso é, determinará do tempo, complexidade computacional e o número necessário de exemplos de treinamento para treinar o algoritmo (Alpaydin 2010).

Nos algoritmos de aprendizado são fornecidos dados de entrada que contém informação para tomada de decisão. Idealmente, não seria necessário uma seleção ou extração de dos dados previamente, afinal o próprio algoritmo deveria receber todos os dados de entrada e descartando dados irrelevante para o processo de aprendizado. No entanto, existem várias razões pelas quais é importante a dimensão dos dados como uma etapa separada de pré-processamento (Alpaydin 2010).

A complexidade dos algoritmos depende da dimensões do número de dados de entrada, d, como também do tamanho da amostra de dados, N. A complexidade é limitante da capacidade computacional, por essa razão é interessados reduzir a dimensão do problema. Diminuir d também diminui a complexidade do algoritmo de inferência durante o teste.

Quando uma entrada é considerada desnecessária, economizamos o custo computacional ao extraí-la. Modelos mais simples são mais robustos em conjuntos de dados pequenos. Modelos mais simples têm menos variação, ou seja, variam menos dependendo dos detalhes de uma amostra, incluindo ruído, *outliers*. Se um problema pode ser representados em uma dimensões reduzida sem perda de informações, sua visualização também é facilitada (Alpaydin 2010).

Para reduzir a dimensão de um conjunto de dados é possível ser realizado por duas maneiras: Seleção de atributos, do inglês *features* e Extração de *features*.

A seleção de *features* é interessada em encontrar k das dimensões d que são as informações mais relevantes e as outras $(d - k)$ dimensões são descartadas. Já extração de *features*, é construído um novo conjunto com dimensão k que é uma combinação das d dimensões do conjunto (Alpaydin 2010).

Os métodos de extração de *features* mais conhecidos são: Análise de Componentes Principais (PCA) e Análise Discriminante Linear (LDA), que são métodos de projeção linear, não supervisionados e supervisi-

onados, respectivamente. O PCA possui muita semelhança com outros dois métodos de projeção linear não supervisionados, a análise fatorial (FA) e a escala multidimensional (MDS). Como exemplos de redução de dimensionalidade não linear: o mapeamento de recursos isométricos (Isomap) e incorporação localmente linear (LLE).

3.10 AVALIAÇÃO DE DESEMPENHO DOS ALGORITMOS DE AM

Dado a aplicação de um modelo de AM, é necessário mostrar sua eficiência por meio de testes de performance em relação aos resultados obtidos. Existem diversas abordagens para validação e testes dos resultados obtidos. Dentre essas metodologias, a validação cruzado (do inglês *cross validation*) é a que mais se destaca. Ademais, outros exemplos são *hold-out*, amostragem aleatória (MacQueen 1967).

A Validação Cruzada consiste em particionar a base de dados de tamanho n em K conjuntos mutualmente exclusivos de tamanho aproximadamente igual a $\frac{n}{K}$. Assim, um subconjunto será utilizado para validação do modelo e os $(K-1)$ subconjuntos restantes treinaram o modelo. O processo é repetido K vezes. Nota-se, que cada um dos K subconjuntos sejam utilizados exatamente uma vez como teste para validação do modelo. A medida de eficiência é a média do desempenho do modelo nos K testes (Duchesne e Rémilland 2005).

No método Hold-out, a base de dados é dividida em duas partes: uma para treino e a outra para teste. A divisão dos dados é em porcentagem fixa p para treinamento e $1-p$ para teste, considerando normalmente $p > \frac{1}{2}$. Valores típicos são $p = 30\%$ dos dados e $1-p = 70\%$ para teste. Uma vantagem do modelo hold-out é que o tempo necessário para aprender o modelo é relativamente menor do que o tempo necessário para a aprendizagem do modelo usando a validação cruzada, apesar de não chegar a resultados tão precisos quanto a validação (Duchesne e Rémilland 2005).

Já na amostragem aleatória, são criados K partições do conjunto de dados. Cada partição é criada de forma aleatória e sem reposição dos exemplos para treinamento. Assim, são realizados K experimentos e a medida de eficiência é a média das medidas de eficiência obtidas em cada experimento. A amostragem aleatória ontem melhores resultados que o método *hold-out* (Duchesne e Rémilland 2005).

Existem diversas métricas para avaliação de desempenho de um algoritmo de AM: Acurácia, Precisão, *Recall*,Matriz de Confusão, Área sob a curva ROC (do inglês *Receiver Operating Characteristics*), Índice de Correção, Incorreção de instâncias Mineradas, Estatística Kappa, Erro Médio Absoluto, Erro Relativo Médio, entre outras (MacQueen 1967).

A Matriz de Confusão, também conhecida como matriz de erro ou tabela de contingência, é uma tabela que permite a avaliação do desempenho do problema sob vários aspectos. A Figura 3.8 apresenta um exemplo de Matriz de Confusão, nela são dispostas as previsões e os valores reais em linhas e colunas, tais como (MacQueen 1967):

- **Verdadeiros Positivos (VP)**: apresenta a quantidade de classificações corretas e que a classe classificada corretamente é positiva.
- **Verdadeiros Negativos (VN)**: aqui encontra-se a quantidade de classificações corretas e que a classe

de classificação correta é a negativa.

- **Falsos Positivos (FP)**: número de classificações erronias, as quais foram classificadas como Positivas, porém deveriam ser classificadas como falsas.
- **Falsos Negativos (FN)**: Quantidade de classificações erradas para negativo, as quais deveriam ser classificadas como positivas.

Figura 3.8: Exemplo de Matriz de Confusão para duas Classe

A acurácia é a métrica mais simples. É basicamente o número de acertos (VP e VN) divido pelo número total de exemplos (Positivos e Negativos). Ela deve ser usada em conjuntos de dados com a mesma proporção de exemplos para cada classe, e quando as penalidades de acerto e erro para cada classe forem as mesmas. Em problemas com classes desproporcionais, ela causa uma falsa impressão de bom desempenho. Por exemplo, em um conjuntos de dados com 80% dos exemplos pertençam a uma classe, só de classificar todos os exemplos naquela classe já se atinge uma precisão de 80%, mesmo que todos os exemplos da outra classe estejam classificados incorretamente (MacQueen 1967).

$$Acuracia = \frac{VP + VN}{Positivos + Negativos} \tag{3.14}$$

A precisão é o número de exemplos classificados como pertencentes a uma classe, que realmente são daquela classe (Verdadeiros Positivos - (VP)), dividido pela soma entre este número, e o número de exemplos classificados nesta classe, mas que pertencem a outras (Falsos Positivos - FP) (MacQueen 1967).

$$Precisao = \frac{TP}{TP + FP} \tag{3.15}$$

O *Recall* consiste em identificar a razão entre os valores previstos como "Positivos"e toda população positiva(Verdadeiros Positivos(TP) + Falsos Negativos (FN)) (MacQueen 1967).

$$Recall = \frac{Positivos}{TP + FN} \tag{3.16}$$

Uma das formas mais importantes de avaliação de modelos de classificação é baseada na matriz de contingência (ou de suas derivações, exemplo, curva ROC). Uma alternativa à avaliação utilizando medidas é o uso de gráficos e/ou diagramas. Gráficos permitem uma melhor visualização da multi-dimensionalidade do problema de avaliação, por exemplo, a curva ROC do inglês *Receiver Operating Characteristics*. A curva ROC, mostra o quão bom o modelo de classificação pode distinguir as classes, traçando a relação

entre a sensibilidade (Taxa de Verdadeiros Positivos) e a especificidade (Taxa de Falsos Positivos) de uma classificação. Consequentemente, a AUC é área sob a curva ROC(do inglês, *Area Under the ROC curve*), importante métrica para avaliação de classificadores.

Diante disso, A ROC, usa como parâmetro de validação, dados referentes a tabela de confusão, vistos anteriormente nos parágrafos acima. A visualização gráfica da ROC deriva das seguintes fórmulas:

$$TaxaFalsosPositivos = \frac{FP}{FP + FN} \tag{3.17}$$

$$Sensibilidade = \frac{VP}{VP + VN} \tag{3.18}$$

$$Especificidade = \frac{VN}{FP + VN} = 1 - TaxaFalsosPositivos \tag{3.19}$$

Além disso, é definido um parâmetro para a curva AUC, chamado *Threshold* ou parâmetro T, geralmente escolhido com valor de 0,5. Nesse Caso, representado pela reta diagonal da Figura 3.9

Quanto mais distante da linha diagonal a sua curva estiver, melhor, ou seja, AUC mais próxima de 1. Na Figura 3.9 foi traçado 3 curvas ROC para diferentes classificadores. Nota-se que a discriminação mais baixa (representada por quadrados) tem AUC menor que as demais, ou seja, tem resultados piores, porém acima da *Threshold*, que é o limite aceitável. Além disso, a curva mais elevada (representada por x) possui AUC aproximadamente 1, um resultado excelente.

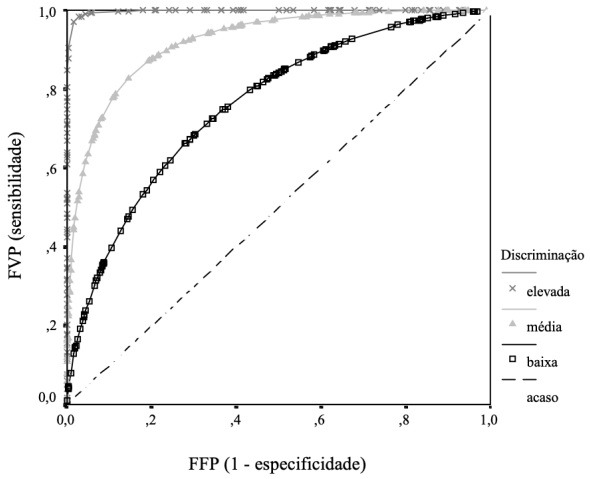


Figura 3.9: Curvas ROC comparativa de três modelos, adaptada (Fawcett 2006)

4 METODOLOGIA

Este capítulo apresenta-se a metodologia da pesquisa. Será mostrado a aplicação do processo de extração do conhecimento apresentado no Capítulo 3.3. Além disso, as ferramentas utilizadas na pesquisa serão informados.

4.1 FERRAMENTAS UTILIZADAS NO TRABALHO

Para armazenar os dados do Banco de Dados foi utilizado o Sistema Gerenciador de Banco de Dados (SGBD) PostgreSQL. O SGDB é responsável pelo armazenamento e gerenciamento das tabelas. Aqui temos todas as tabelas das duas bases de dados e as tabelas dimensões dos dados.

Os dados foram manipulados através da linguagem de programação Python e suas biblioteca de manipulação e análise de dados.

4.2 BASE DE DADOS

O trabalho é um direito social assegurado pela constituição brasileira, além disso, é um objetivo fundamental a redução das desigualdades sociais e regionais. Portanto, é um dever governamental acompanhar o mercado de trabalho brasileiro e criar políticas públicas de redução de possíveis desigualdades. Para isso, é necessário elementos os quais proporcionem a criação de políticas públicas sociais baseadas em estudos concretos e relevantes.

Para suprir a necessidade da coleta de dados empregatício o governo instituiu em 1975 que as empresas preenchessem anualmente a Relação Anual de Informações Sociais – (RAIS), informando as alguns elementos destinados as necessidades de controle, controle estatístico das entidades governamentais da área social no mercado de trabalho formal brasileiro. (DOU 1975).

Além disso, todas as admissões e dispensas de empregados sob regime da Consolidação das Leis do Trabalho (CLT) são registrados no registro permanente do Cadastro Geral de Empregados e Desempregados (CAGED). Esse cadastro é utilizado pelo Programa de Seguro-Desemprego, para conferir os dados referentes aos vínculos trabalhistas, além de outros programas sociais (Ministério da Economia 2019).

As duas bases são públicas e disponíveis pelo Ministério da Fazenda pelo Programa de Disseminação das Estatísticas do trabalho (PDET). O consolidado da informação das duas bases foi utilizada nessa pesquisa. Nas próximas seções será apresentado a forma de coleta das informação de alimentação das bases de dados.

4.2.1 Relação Anual de Informações Sociais - RAIS

O Decreto n.o 76.900, de 23 de dezembro de 1975 obriga que todos os estabelecimentos forneçam ao Ministério da Economia as informações de seus funcionários, por meio da Relação Anual de Informações Sociais (RAIS). Todos os anos é realizado um manual com orientações aos empregadores no correto preenchimento das informações da RAIS (DOU 1975).

A RAIS tem por objetivo suprir às necessidades de controle da atividade trabalhista no país, para identificação dos trabalhadores com direito ao recebimento do Abono Salarial. Além disso, promove dados para a elaboração da estatísticas do trabalho ao disponibilizar informações do mercado de trabalho brasileiro às entidades governamentais (Ministério da Economia 2019).

Segundo o Decreto, os dados coletados pela RAIS constituem expressivos insumos para atendimento das necessidades: da legislação da nacionalização do trabalho; controle dos registros do FGTS ; dos Sistemas de Arrecadação e de Concessão e Benefícios Previdenciários; de estudos técnicos de natureza estatística e atuarial; de identificação do trabalhador com direito ao abono salarial PIS/PASEP; criação de políticas públicas (Ministério da Economia 2019).

O estabelecimento/entidade sem vínculo empregatício preenchem a RAIS negativa e as empresas com empregados a RAIS normal. As RAIS negativas não serão objeto de estudo nessa pesquisa.

Na RAIS são relacionados empregados com vinculo CLT; servidores da administração pública direta ou indireta, federal, estadual ou municipal, bem como das fundações supervisionadas; trabalhadores avulsos (aqueles que prestam serviços de natureza urbana ou rural a diversas empresas, sem vínculo empregatício, com a intermediação obrigatória do órgão gestor de mão-de-obra, nos termos da Lei nº 8.630, de 25 de fevereiro de 1993, ou do sindicato da categoria); empregados de cartório extrajudiciais; trabalhadores temporários para diferentes regimes de contratação; diretores sem vínculo empregatício, para os quais o estabelecimento/entidade tenha optado pelo recolhimento do FGTS ; servidores públicos não-efetivos (demissíveis ad nutum ou admitidos por meio de legislação especial, não regidos pela CLT); trabalhadores regidos pelo Estatuto do Trabalhador Rural ; jovens aprendizes (maior de 14 anos e menor de 24 anos); servidores e trabalhadores licenciados (Ministério da Economia 2019).

Cada empregador deve informar seu tipo de inscrição: CNPJ, CEI/CNO ou CAEPF. Além disso, sua razão social e informações cadastrais do estabelecimento: endereço, município de localização, telefone e e-mail (Ministério da Economia 2019).

Ademais, deve ser informado obrigatoriamente sua principal atividade econômica, de acordo com a Classificação Nacional de Atividades Econômicas (CNAE), tal como o porte da empresa (microempresa, empresa de pequeno porte, empresa/órgão que não se classifica nos itens anteriores).

Depois de cadastrado a empresa, os empregadores devem cadastrar todos empregados efetivos ou desligados no ano-base da RAIS. Inicialmente as informações gerais: nome, sexo, nacionalidade, raça/cor, escolaridade e se é portador de deficiência habilitada ou beneficiário reabilitado, além do tipo de deficiência conforme as seguintes categorias: física, auditiva, visual, intelectual, múltipla ou reabilitado.

As informações de cada empregado/servidor devem constar na RAIS de todos os estabelecimentos da empresa/entidade aos quais ele esteve vinculado durante o ano-base, cabendo a cada estabelecimento

fornecer as informações referentes ao período em que o empregado esteve a ele vinculado, seja como "transferido", "cedido" ou na categoria de "contratado". O vinculo estabelecido deve seguir o código e descrição da Classificação Brasileira de Ocupações(CBO). Outro ponto de informação é relativa a remuneração e controle da jornada dos trabalhadores: horas semanais trabalhadas, a periodicidade do pagamento dos vencimento e o valor do salário contratual básico (Ministério da Economia 2019).

Hoje, a CLT (Consolidação das Leis do Trabalho) define jornada máxima semanal de 44 horas regulares. A jornada mensal, por sua vez, fica limitada a 220 horas. A lei também determina que o trabalhador não pode fazer mais de duas horas extras por dia.

Quando o empregado/servidor possuir mais de um contrato ou ocupação com o mesmo estabelecimento/órgão, as informações de cada vínculo devem ser declaradas separadamente e as horas semanais devem ser informadas de acordo com o contrato. No caso de empregado desligado e readmitido no decorrer do ano-base, as informações referentes a cada um dos períodos deverão ser fornecidas separadamente. Ou seja, é possível que um mesmo emprego tenha múltiplos registros da RAIs em um único ano-base.

4.2.2 Cadastro Geral de Empregados e Desempregados - CAGED

O Decreto n.º 4923, de 23 de dezembro de 1965 instituiu o registro permanente de admissões e desligamentos de empregados sob regime da Consolidação das Leis do Trabalho (CLT) (DOU 1965).

Os estabelecimentos informam mensalmente ao Ministério do Trabalho e Emprego os novos admitidos, desligados ou transferidos. Ao fim do ano é gerado o Cadastro Geral. As informações do CAGED são utilizadas pelo Programa de Seguro-Desemprego para conferir os dados referentes aos vínculos trabalhistas e liberar os benefícios. É também com base nestas informações que o Governo Federal e a sociedade como um todo contam com estatísticas para elaboração de Políticas de Emprego e Salário, bem como pesquisas e estudos sobre mercado de trabalho (Ministério da Economia 2019).

Inicialmente o decreto previa apenas o registro dos empregados empregados celetistas, no entanto, atualmente, além dos celetistas são declarados na CAGED os trabalhadores de contrato temporário, trabalhadores regidos pelo Estatuto do Trabalhador Rural e aprendizes contratados nos termos do art. 428 da CLT (Ministério da Economia 2019).

Admitindo, desligando ou transferindo um funcionário, o empregador preenche informações gerais do contratado: PIS/PASEP, nome, informações da carteira de trabalho, CPF, data de nascimento, raça, deficiência e o tipo de deficiência, sexo, grau de instrução. O preenchimento é muito semelhante à RAIS, descrito na seção anterior.

Além disso, as informações do trabalho desempenhado, conforme CBO e informações contratuais(Data da Admissão, Horas Contratuais e Remuneração ,e quando caber, Data de Desligamento).

O ponto mais importante de registro na CAGED é o tipo de movimentação: Admissões ou Desligamentos. Em Admissões são informados informações cruciais das características do trabalho no pais, são informados se a admissão é do primeiro emprego, reemprego, contrato por prazo determinado, reintegração ou transferência de entrada.

No caso de desligamento, é informado o motivo do desligamento. O desligamento pode ser por dis-

pensa sem justa causa por iniciativa do empregador, a pedido por iniciativa do empregado (espontâneo), término de contrato por prazo determinado, término de contrato, aposentado, morte ou transferência de saída.

4.2.3 Definição dos Dados

O presente trabalho usa as informações do ano-base da declaração de 2018 das RAIS, sem as informações da RAIS negativa e CAGED. Na seção anterior foi apresentado a composição das bases de dados RAIS e CAGED. Além disso, a base RAIS é preenchida anualmente, enquanto CAGED mensalmente.

Graças as similaridades entre as várias colunas de RAIS e CAGED foi criado uma base que é a combinação das informações do ano de 2018 dos dados da RAIS com os dados do CAGED. A fim de obter informações mais atualizadas, contanto com 319.985.655 registros de empregados respeitando as condições de registro apresentados nas seções 4.2.2 e 4.2.1.

É importante destacar que a base não totaliza a informação de 319.985.655 empregados, valor que é quase o dobro da população brasileira. O valor apresentado é o número de registros totais da junção das duas bases de dados. E um mesmo empregado pode apresentar múltiplos registros. Afinal, durante um ano um mesmo empregado pode ser empregado, demitido e transferido em diferentes empresas.

A Figura 4.1 apresenta um Modelo Dimensional da base de dados utilizada no trabalho. O Modelo Dimensional simplifica as informações da base de dado, a tabela fato (tabela principal) é RAIS_CAGED e recebe como entrada as informação das tabelas dimensão.

Vale observar que algumas dessas tabelas dimensão são referente a outras bases de dados importantes: o Cadastro de Municípios do IBGE, Cadastro Brasileiro de Ocupações(CBO) e Classificação Nacional de Atividades Econômicas (CNAE) do próprio Ministério do Trabalho.

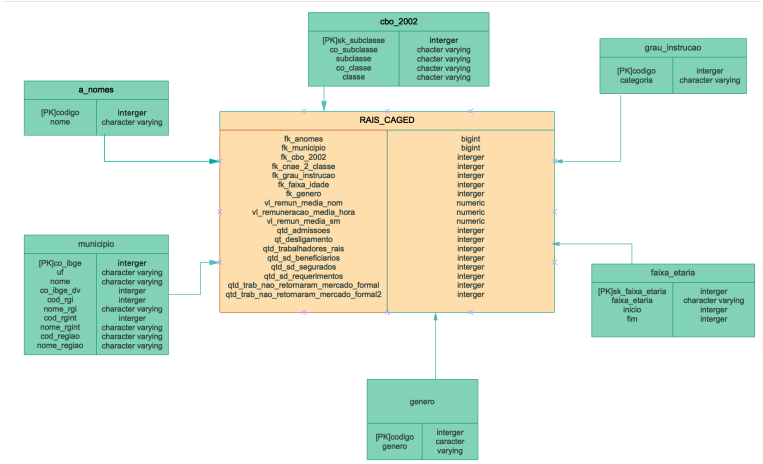


Figura 4.1: Modelo Dimensional da Base de Dados do Trabalho Normalizada

4.3 PREPARAÇÃO DOS DADOS

O objetivo nesse momento é conhecer a base de dados para melhor entendimento do problema e realizar o melhor processo de aprendizado. Nessa fase é feito a seleção dos dados dentro da base de dados que efetivamente trarão informação no processo de aprendizagem (Goldschmidt e al. 2015).

As bases de dados completas contam com todos os registros do sexo feminino e masculino durante o ano de 2018. Para essa pesquisa foram desconsideradas as entradas sem as informações de gênero. Ademais, para melhores comparações, foi desmembrado a tabela fato em 3 tabelas no postgres para facilitar o processo de modelagem: Primeiro os registros femininos, os registros masculinos e os registros conjuntos.

4.3.1 Estatística Descritiva dos Dados

A Figura 4.2 mostra um gráfico de barras das quantidades de registros para homens e mulheres presentes na base de dados Rais_Caged. No caso, são 40,25% de registros de mulheres e 59,74% de homens. Diante disso, é importante destacar que a composição dos registros é muito similar ao encontrado na participação no mercado de trabalho dos dados do PNAD, 58,8% eram homens e apenas 41,2% eram mulheres, de acordo com dados do PNAD. Ademais para o ano de 2018, temos 95,95% registros de mulheres empregadas e 4,04% demitidas. Similarmente, 95,57% de registros de mulheres empregadas e 4,42% de homens.

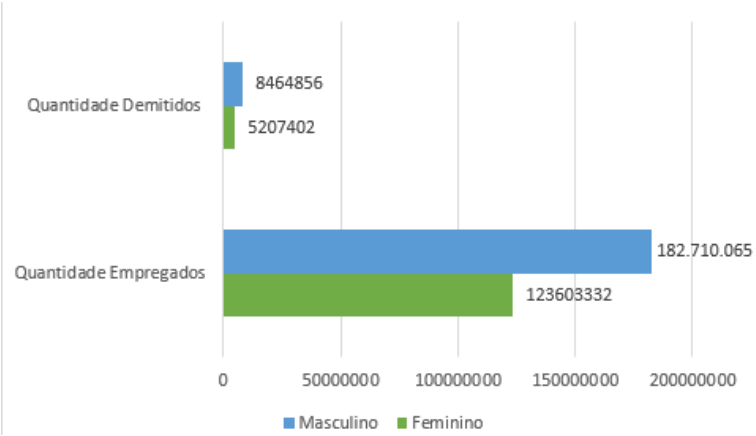


Figura 4.2: Número de Homens em Mulheres na Base de Dados Rais_Caged

O Brasil é um país de dimensão continental e é importante conhecer a realidade regional, afinal, cada região possui suas particularidades de composição populacional e características da mão de obra requisitada. Conforme os dados do PNAD, a maior parte da população brasileira concentra-se na região sudeste, cerca de 42,04%, enquanto a região centro-oeste possui a menor parcela da população, 7,79%.

Sendo assim, a Figura 4.3 apresenta o número de registros do sexo feminino e masculino na base por região. Nota-se um desbalanceamento entre as regiões, afinal a população brasileira é heterogeneamente distribuída.

Nesse contexto, 47% dos empregados registrados pertencem à região. Além disso, os 3 maiores salários registrados na base de dados também integram essa região. Essa concentração era esperada, pois maior parte da população localiza-se nessa região e também os maiores centros urbanos, responsáveis pela maior

parte dos empregos formais, segundo PNAD.

Por outro lado, apenas 5% dos registros dos trabalhos formais estão na região norte. Apesar do centro-oeste contar com a menor concentração populacional, possui mais registros que a região norte. Está falta de registro está relacionado ao perfil das característica da empregabilidade da região. Conforme o PNAD, a maioria dos trabalhadores da região norte estão fora do mercado formal, tipo de mão de obra registrada pelas RAIS e CAGED.

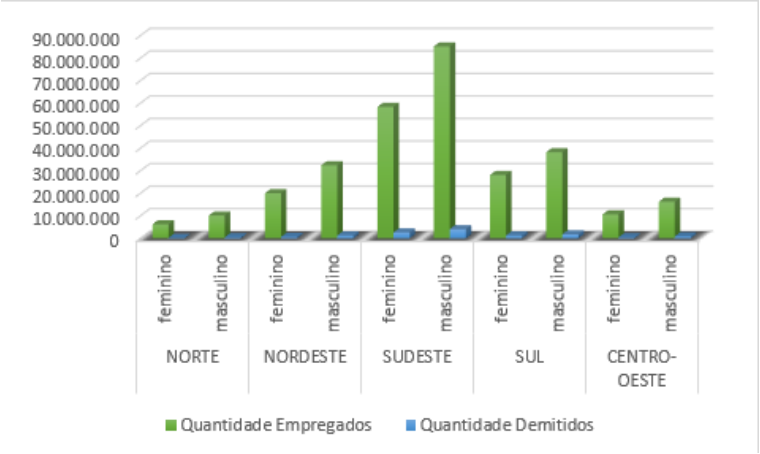


Figura 4.3: Distribuição de homens e mulheres por região

As Figuras 4.4 e 4.5 mostram a características dos registros de mulheres e homem por escolaridade. Tanto para mulheres e para homens, os registros de escolaridade Ensino Médio Completo e Ensino Superior Completo são as que mais aparecem. Um outro ponto a evidenciar é em relação as escolaridades referentes à pós-graduação.

Embora homens sejam maioria nos registros nas bases de dados, o número de registros para Doutorado, Especialização e Mestrado é muito similar ao número de registros femininos. Essa informação é muito importante e concorda com os resultados de (Osborne 2015) e (Pearson, Frehill e McNeely 2015) , os quais mostram que no Brasil a produção científica é igualitária em relação a gênero.

Porém, quando avaliado o percentual de registros de desemprego em relação ao topo das escolaridades (pós-graduação), os resultados são desanimadores para mulheres. 62,8% dos registros do tipo para mulheres aparecem como desempregadas, contra 38,4% para homens. Conquanto, esse fenômeno, *Opt-out*, ocorre no mundo todo, inclusive no Brasil. E deve ser melhor estudado. Na seção de trabalhos correlatos foi destacado a tese de doutorado (Franco 2018) que estuda o evento *Opt-out* no Brasil.

Observando as informações dos registros por Faixa Etária, pode-se notar que homens entram mais cedo no mercado de trabalho. Proporcionalmente a maior parte dos registros masculinos estão entre as faixas de 18-39 anos, enquanto mulheres 25-39 anos. Para os dois casos a maior parte dos registros está na faixa de 30-39 anos. É válido ressaltar que segundo o IBGE, a faixa de 30-39 anos concentra a maior porção da população que está com idade e no ápice da suas condições de trabalho. Assim, é esperado que a maior parte dos registros estivessem nessa faixa etária.

Já em relação aos registros de remuneração média mensal nacional para homens e mulheres observa-se uma disparidade salária, como esperado: 2104,17 reais para mulheres e 2518,53 reais para homens. Outro

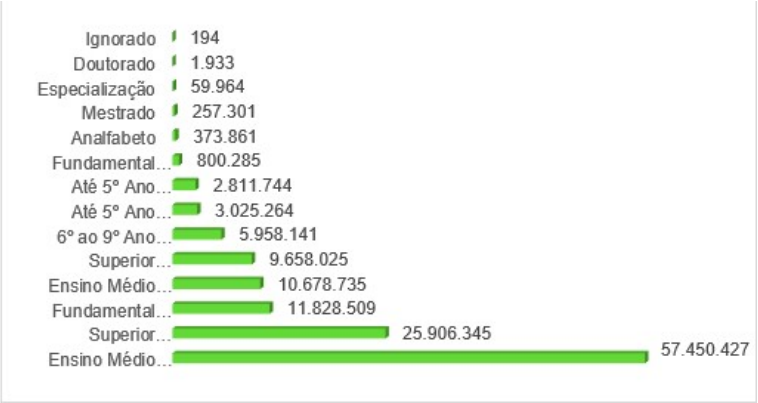


Figura 4.4: Distribuição de Registro de Mulheres por Escolaridade

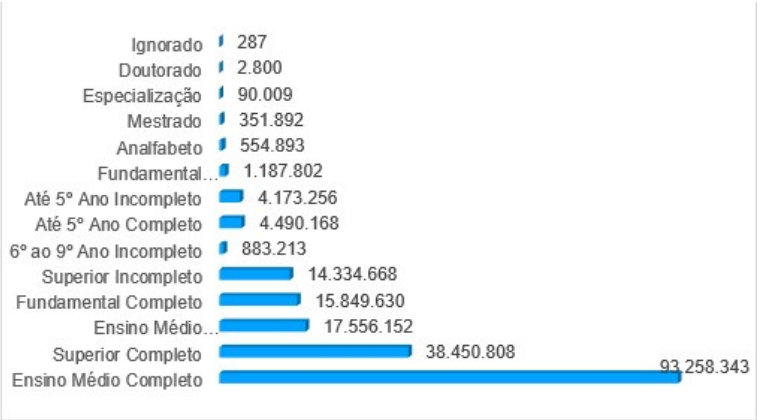


Figura 4.5: Distribuição de Registros de Homem por Escolaridade

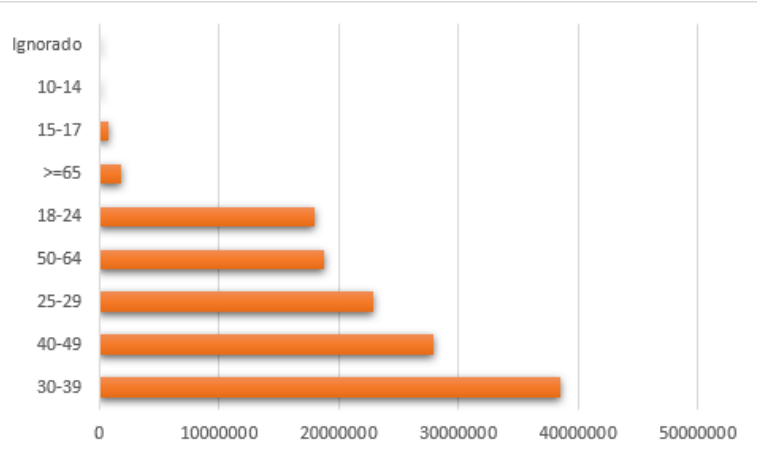


Figura 4.6: Distribuição de Registros de Mulheres por Faixa Etária

aspecto não menos relevante nessa disparidade é que os 3 maiores salários são masculinos e os 3 menores são femininos.

Realizando a avaliação por região das remuneração dos registros, homens sempre ganham mais. A região com a maior média salarial é a sudeste 2597,80 reais. No entanto, apresenta a maior disparidade salarial com relação a gênero. Homens ganham 57% a mais, homens ganham próximo da média salarial conjunta para a região, 2590,50 reais, enquanto mulheres 1512,30 reais. A região nordeste apresenta a

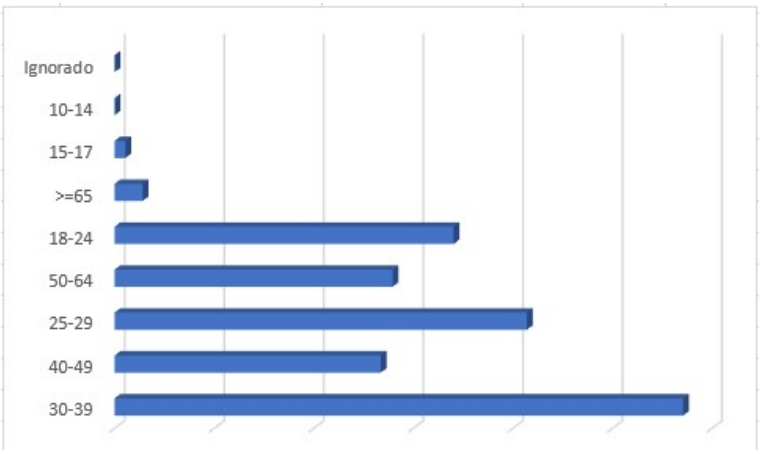


Figura 4.7: Distribuição de Registros de Homem por Faixa Etária

menor média salaria, 1862,07 reais. Entretanto, possui a menor desigualdade por gênero. A Figura 4.8 apresenta a média salarial por região e gênero.

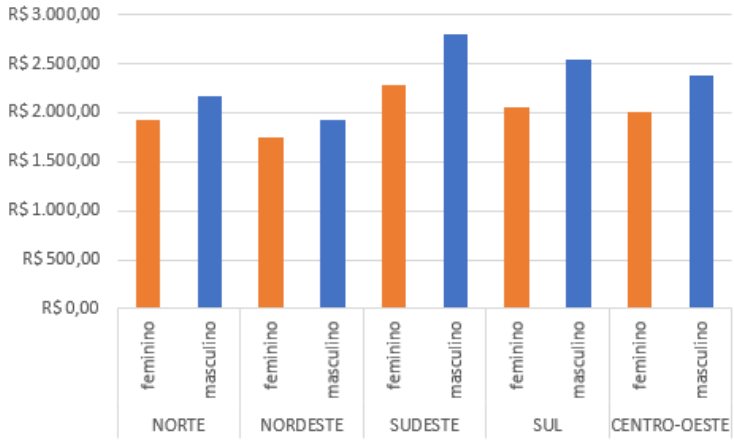


Figura 4.8: Média Salarial por Região e Gênero

Ademais, os postos de trabalho formais mais frequentes na base de dados ocupados por mulheres e homens é muito diferente. Os postos mais recorrentes para mulheres são, respectivamente, auxiliar de escritório, educação infantil e auxiliar de limpeza. Por outro lado, os mais recorrentes respectivamente são vendedores de comércio varejista, auxiliar de serviços gerais na construção civil e auxiliar de serviços gerais área industrial.

Os resultados para os postos de trabalho foram os que mais discreparam do PNAD, porém ainda muito semelhantes. O PNAD mulheres ocupam mais frequentemente postos de trabalho na área de saúde e educação, enquanto homens construção e indústria extrativista.

Por fim, foi avaliado a média de horas trabalharam para os registros de mulheres e homens. As mulheres trabalham em média 22horas semanais, enquanto homem 44,06 horas semanais, essa inferência é muito similar aos dados coletados no PNAD, onde mulheres tendem a preferir trabalhar em postos de trabalho com jornadas de meio-período, pois acabam reservando mais tempo com cuidado de pessoas e coisas do que homens.

4.4 PRÉ-PROCESSAMENTO

Concluído a analise prévia da base de dados da seção anterior, partiu-se para a construção do *dataset* utilizado no processo de mineração e construção do conhecimento. Essa fase é conhecida com o pré-processamento dos dados, o qual é feita a limpeza da base e integralização dos dados, conforme (Goldschmidt e al. 2015).

Destaca-se o exorbitante volume de dados. Esse fator foi muito desafiador para o trabalho. Buscar a melhor forma de reduzir esse volume de dados foi excecional para sua viabilidade.

Inicialmente, foi retirado os dados esdrúxulos e o tratamento dos dados ausentes. Todas as entradas com informação de gênero, faixa de idade, grau de instrução sem informações ou marcadas como ignoradas foram desconsideradas. Os registros com essas informações, porém com as demais entradas ausentes dessas informações foi utilizado biblioteca do Python para lidar com essas ausências. Foi usado PCA para extração a informação principal e posteriormente foi inserido como rótulo das entradas ausentes.

Como informado na seção anterior, um mesmo empregado pode ter múltiplos registros nas bases de dados. Primeiramente, foi lidado com essas entradas buscando identificar as condições finais dos empregados e deixando apenas a informação consolidada. Para isso, foi necessário um grande custo computacional e um trabalho manual da autora. Saindo de 319.985.655 registros para aproximadamente 90 mil registros.

Além das técnicas de selecionar *features*, foi considerado as informações utilizadas dos trabalhos correlatos e as informações estudadas pela cartilha de monitoramento de gênero do IBGE. A maioria dos trabalhos conta com informações de escolaridade, remuneração, faixa de idade, horas trabalhadas e remuneração média.

A base de dados RAIS_CAGED já possui todas essas informações, exceto horas trabalhadas. Entretanto foi possível extrair da própria base essa informações, pois ela possui as informações da remuneração média mensal(vl_remuneracao_media_nom), remuneração média semanal(vl_remuneracao_media_sm) e remuneração média por hora (vl_remuneracao_media_hora).

Para melhor atingir o objetivo desse trabalho, foi realizada três composições de *datasets*: Primeiramente um com todos os *data-points* da base de dados; posteriormente uma com os dados do sexo feminino ; por fim, uma contato apenas com os dados do sexo masculino.

4.5 MODELAGEM

Nessa etapa foi efetivamente aplicado os algorítimos de aprendizado de máquina para a realização da mineração dos dados e geração de resultados que poderão influenciar na avaliação do estudo da empregabilidade feminina no Brasil. Além disso, nessa etapa é avaliado o desempenho do algoritmo na proposta na mineração.

O presente trabalho usa técnicas de aprendizado de máquina supervisionadas, as árvores de decisão. Foi utilizado a linguagem de programação python e suas diversas bibliotecas. Destacando-se o sklearn, pandase e Numpy.

Os modelos baseados em árvores de decisão receberam diversas modificações. Nesse trabalho foi utilizado o método XGBoost (*Extreme Gradient Boosting*), método baseado em florestas aleatórias.

Na literatura de ciência de dados existem alguns recursos para selecionar as melhores *features* e reduzir a dimensão da base, por exemplo, RFE (*Recursive Feature Elimination*) e o *(Random Forest)*. O RFE treina o modelo utilizando o conjunto inicial com todas as *features* e *data points* que vierem nele. Na segunda rodada de treinamento, o RFE verifica a importância das features (utilizando atributos como coef_ ou feature_importances) e, recursivamente, remove *features* menos importantes do *dataset* e treinar o modelo novamente. Até chegar a um número ideal de *features* (Alpaydin 2010).

Como o volume de *data points* é enorme, o custo computacional também é muito alto, sem outros recursos, fica inviável a aplicação dos modelos para selecionar as melhores textitfeatures e também reduzir a dimensão da base. Para contornar esses problemas, o trabalho foi implementado utilizando o método XGBoost, que utiliza uma estrutura de *Gradient boosting* (XGBOOST 2019).

Com o XGBoost, o problema do custo computacional, consequente do grande volume de dados é mitigado, pois utiliza a aceleração de GPU. Para o treinamento de GPU, quanto maior o conjunto de dados, maior a aceleração. Não faz muito sentido usar o treinamento de GPU para objetos de mil ou menos, mas a partir de 10.000 você terá uma boa aceleração, o que não é o caso do trabalho (XGBOOST 2019).

Então, XGBoost foi uma boa opção para lidar com o problema do grande volume de dados, pois pode lidar com milhares de variáveis de entrada e identificar as variáveis mais significativas. Além disso o modelo produz uma tabela com a listagem do grau de importância das variáveis, baseada na estimativa de correlação das variáveis. Por essa razão, é um dos métodos de aplicação de algoritmos de *machine learning* atualmente.

Com base na identificação das variáveis mais significativas gerado pelo XGBoost, que é produzida de forma semelhante ao RFE, as **features** foram escolhidas. A Tabela 4.1 mostra as *feautures* mais recorrente entres as três composições de *dataset*.

Tabela 4.1: *Dataset* selecionado para o processo de mineração

features	**Descrição**
fk_genero	Gênero do trabalhador
fk_cbo_2002	Informações da ocupação da trabalhadores
fk_grau_instrucao	Informação da Escolaridade da trabalhadores
fk_faixa_idade	Faixas de Idade da trabalhadores
vl_remun_media	Valor da Remuneração Média mensal dos Trabalhadores no ano Base
horas_trab	Horas Trabalhadas
vl_remun_hora	Valor da Remuneração por hora
cd_regiao	Código da Região Trabalhada

4.5.1 Comparativo Desempenho Algoritmos

Na próxima seção será apresentado os resultados. Nessa etapa, queremos avaliar o desempenho da classificação que será melhor explicada na próximo capítulo. Vale ressaltar, que os dados foram rotulados como empregado e desempregado.

Foi realizado dois procedimentos de classificação. Primeiramente, com os dados referentes ao sexo feminino e posteriormente com os do sexo masculino.

A biblioteca sklearn possui uma função *classification_report*, a qual retorna as informação das métricas de avaliação de desempenho comentadas na fundamentação teórica (3.10). Nesse sentido, informações como precisão, *recall* e f1-score são retornadas automaticamente, facilitando a construção de uma matriz de confusão e assim, estimar se o algoritmo de aprendizado de máquina foi uma boa escolha para o nosso problema.

Conforme especificado na seção de pré processamento(seção 4.4) foram criados três composições de *dataset*. As Tabelas 4.2, 4.3 e 4.4 apresentam o desempenho do algoritmo de árvore de decisão usando o método XGBoost para todos os *data-point*, para os dados do sexo feminino e para o sexo masculino respectivamente.

Tabela 4.2: Desempenho da Classificação do *Dataset* composto por todos os *data-points*

Classes	**Precisão**	**Recall**	**f1-score**	**Suporte**
Empregados	0,83	0,85	0,84	90.225
Desempregados	0,67	0,63	0,65	10.215

Tabela 4.3: Desempenho da Classificação para o *Dataset* Composto pelos Dados do Sexo Feminino

Classes	**Precisão**	**Recall**	**f1-score**	**Suporte**
Empregados	0,93	0,91	0,87	20.091
Desempregados	0,83	0,81	0.85	16.005

Tabela 4.4: Desempenho da Classificação para o *Dataset* Composto pelos Dados do Sexo Masculino

Classes	**Precisão**	**Recall**	**f1-score**	**Suporte**
Empregados	0,73	0,75	0,74	50.001
Desempregados	0,67	0,63	0,65	4.140

Avaliando as três tabelas é possível observar um bom desempenhos do algoritmo para todas as composições de *dataset*. E assim, é razoável a utilizar desse algoritmos para a proposta do trabalho. No entanto, o *dataset* que inclui os empregados de ambos os gêneros teve uma alta taxa de classificações errada (FP e FN), já quando separados os gêneros a taxa de erros de classificação teve uma diminuição expressiva. Mostrando que a observação macro do problema não responderá os problemas particulares de cada gênero.

Afinal, a quantidade de entradas do gênero feminino e masculino é bem desbalanceada e mesmo o XGBoost possuindo métodos para equilibrar erros em conjuntos de dados onde as classes são desequilibradas, o número um número alto comparado as outras composições de *dataset*.

É importante ressaltar que foi utilizado a função nativa *xgb.cv* do XGBoost, que calcula o melhor número de rodadas (*nrounds*) de validação cruzada para a melhor estimativa das classes e assim atingir o melhor desempenho de predição do modelo.

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Nesse capítulo discute-se os resultados da aplicação do algorítimos de aprendizado de máquina discutido na seção de modelagem (seção 4.5). Após o processo de treinamento e validação do algorítimo foram obtidos alguns resultados que serão discutidos a seguir. A seção anterior (seção 3.10) foi disponibilizado a avaliação do desempenho desses resultados.

Na fundamentação teórica, a seção referente a árvores de decisão (3.6.2) comentou-se que somente os atributos com capacidade de generalização são utilizado na construção das árvores de decisão. No caso do algoritmo C4.5, os atributos são escolhidos a partir do valor da informação referente ao valor do seu ganho de informação.

Posterior o processo de treinamento e validação do algoritmo, XGboost pode retorno um arquivo *"fmap"* com as informações relevantes de cada atributo na árvore. Assim, utilizando o método *get_score()* é possível retornar a importância de cada atributo, a qual é definida de acordo com alguns parâmetros de tipo (XGBOOST 2019). São eles:

- *"weight"*: número de vezes que um atributo é usado para separar os dados em todas as árvores da floresta;
- *"weight"*: o ganho médio em todas as divisões em que o atributo é usado para separação dos dados;
- *"cover"*: A cobertura média em que o atributo é usado;
- *"Total_gain"*: o ganho total em todas as divisões em que o recurso é usado.

As Figuras 5.1, 5.2 e 5.3 representam o resultado da importância dos atributos baseado no seu ganho de informação, ou seja, o resultado do comando: *get_score(fmap=", importance_type='gain')*.Ademais, os atributos que não foram listados possuem ganho irrelevante. Nota-se que a importância dos atributos difere muito entre cada composição de *dataset*. Demonstrando a necessidade de análise desmembrada para atingir a melhor resultado e consequentemente criar políticas públicas mais efetivas.

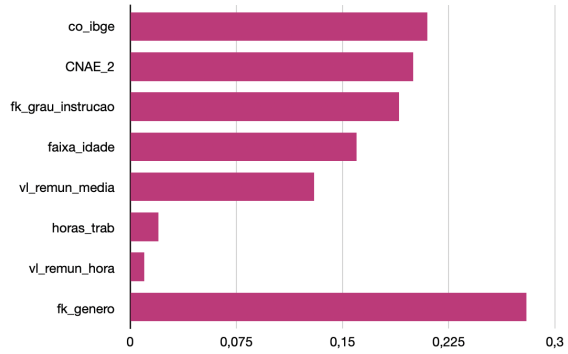


Figura 5.1: Listagem do Ganho para cada atributo para todos os *Data-Point*

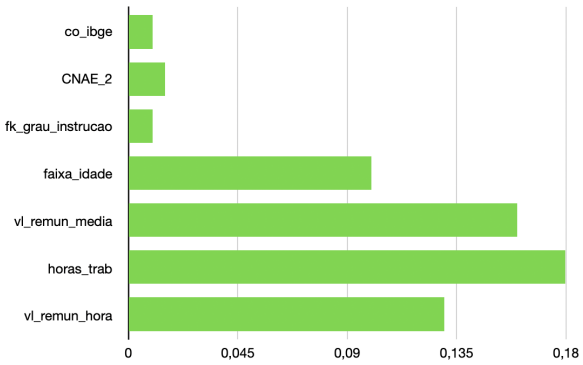


Figura 5.2: Listagem do Ganho para cada atributo para Mulheres

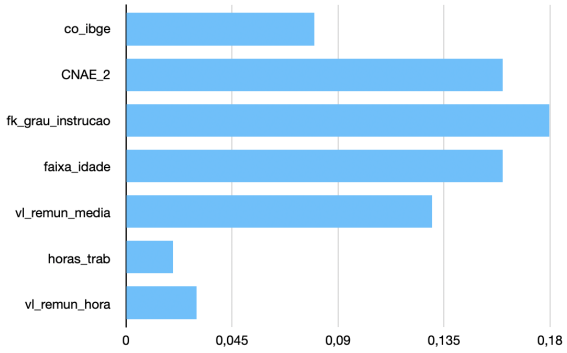


Figura 5.3: Listagem do Ganho para cada atributo para Homens

As Figuras 5.4, 5.5 e 5.6 são as árvores de decisão das referentes composições do *datasets* comentadas posteriormente.

Inicialmente, vale destacar que as três composições de *datasets* são extremamente balanceadas, o número de entradas da classe empregada é maior do que a classe desempregada. Além disso, o *datasets* geral, contendo os dois gêneros, possui muito mais homens que mulheres.

Nota-se que quando analisando a árvore de decisão geral, o gênero é a raiz da árvore, sendo o fator com o maior ganho de informação. Essa informação é condizente, afinal, o percentual de mulheres desempregadas é muito maior do que o de homens. Outro atributo com um ganho alto é *"co_ibge"*. Atributo referente à região trabalhada pelo empregado. Como comentado anteriormente, o Brasil é extremamente desproporcional regionalmente, fazendo sentindo esse ser um dos atributos principais quando avaliado a situação de empregados e desempregado no Brasil.

Outro que vale destacar é o atributo relacionada a idade dos empregados. A dificuldade de incluir os extremos da carreira (jovens em início de carreira e adultos no fim de carreira) é um problema enfrentado por vários países. Observando os resultados do *dataset* geral o fator idade é muito mais relevante que no *dataset* feminino, no entanto, no masculino já é mais relevante.

A presente pesquisa considerou a possibilidade de estudo regionalizado das informações, porém os dados ofertados para a pesquisa os resultado entre as regiões ficaram muito semelhantes. Seria necessário

uma quantidade maior de informações para a avaliação regionalizada.

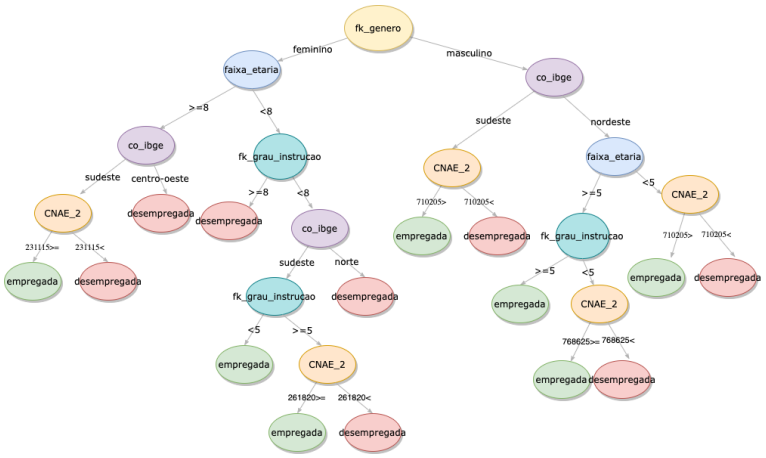


Figura 5.4: Árvore de Decisão para os empregados brasileiros

Agora, avaliando efetivamente os resultados do sexo feminino, pode-se inferir que o fator horas trabalhadas é muito importante na empregabilidade feminina. Essa informação pode parece óbvio, afinal, quanto mais se trabalha, mais se ganha. No entanto, mulheres, segundo a fundamentação teórica, preferem postos de trabalho com cargo horária reduzida, um dos motivos, segundo o PNAD. Além disso, a média de horas trabalhadas dos registros femininos é proporcional a jornada de trabalho de meio período da legislação vigente.

Conforme a fundamentação teórica, a preferência por postos de carga de horária reduzida relaciona-se com o tempo dedicado aos cuidados de pessoas e/ou afazeres domésticos, o qual é quase o dobro para mulheres com relação aos homens. Porém, alguns postos de trabalham não permitem que a mulher exerça jornadas inferiores. Portanto, mesmo que a mulher possua os requisitos para o trabalho, ela poderá preterir esse posto, por necessitar/querer dedicar um tempo maior aos cuidados de pessoas e/ou afazeres domésticos.

Outro ponto que merece destaque é a pouca influência da escolaridade na empregabilidade feminina, para mulheres o ganho de informação para o atributo de escolaridade é quase irrelevante. E nem aparece como fator de decisão na árvore em nenhum discriminante. Essa informação pode está relacionada a dois fatores: O fenômeno *opt-out* e os efetivos postos de trabalhos ofertados para as mulheres.

Avaliando a quantidade de registros de pessoas com grau de escolaridade maior de 8 (Pós graduação - Mestrado, Especialização e Doutorado) desempregadas do sexo feminino é infinitamente maior do que o masculino. Na literatura sabe-se que o fenômeno *opt-out* é uma realidade muito expressiva em países desenvolvidos. Além disso, a baixa influência da escolaridade na empregabilidade pode está relacionado aos cargos ocupados. Avaliando os postos de trabalho mais recorrente nos registros da base dados percebe-se que os postos ocupados não exigem alto grau de escolaridade.

É necessário compreender melhor o fenômeno *opt-out* no Brasil, verificar se a saída dessa mão-de-obra extremamente qualificada é espontânea ou os contexto de oferta e realidade socioeconômico permitem a mulher competir de maneira igual à mão de obra de masculina. E caso, essa evasão não seja espontânea deve criar maneiras de isentiva a contratação feminina e também criar uma realidade socioeconômica que integre a realidade da mulher ao mercado de trabalho. Por exemplo, criar possibilidade para que mães

possam cuidar de seus filhos sem medo de ser excluída do mercado de trabalho. Para isso, é necessário a criação de creches, escolas de período integral, isentivo a empresas que criam programas de inclusão de mães à realidade da empresa. O estado tem o dever de buscar uma solução de mitigar essa desigualdade.

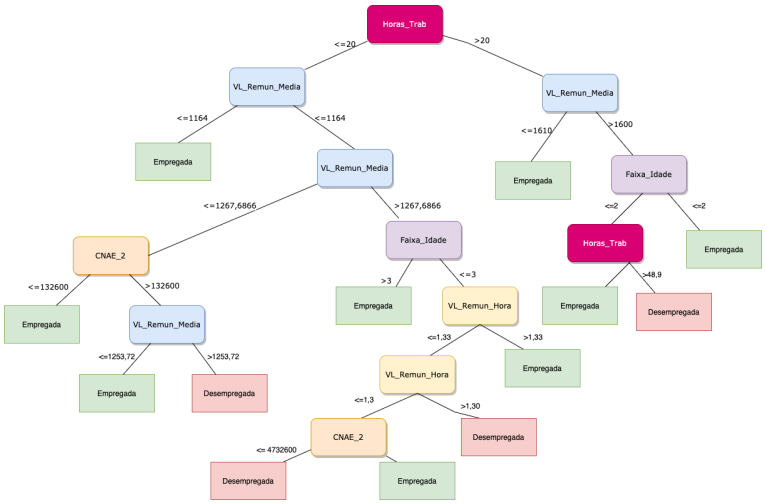


Figura 5.5: Árvore de Decisão para Empregadas do sexo Feminino

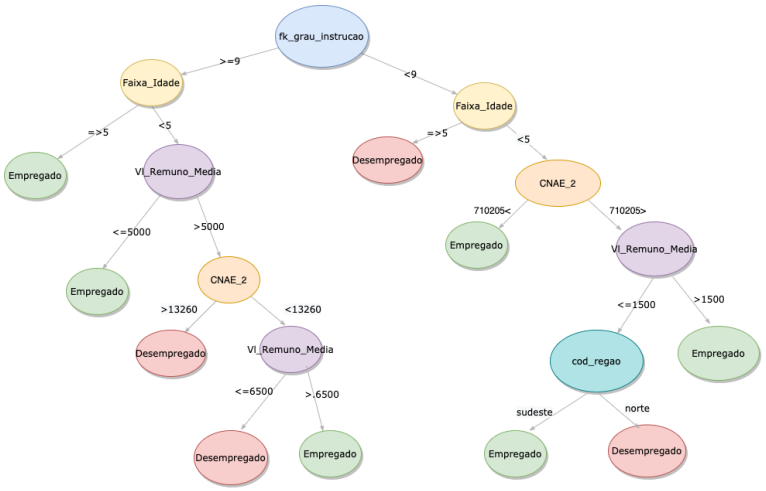


Figura 5.6: Árvore de Decisão para Empregados do Sexo Masculino

Contundo, quando avalia-se os resultados do sexo masculino percebe-se que o grau de instrução é o fator mais relevante da empregabilidade masculina, diferente do que o sexo feminino, sendo o nó raiz da árvore de decisão. No entanto, homens com baixa instrução e jovens possuem chances altíssimas de estarem empregados. O que é um realidade da desigualdade de gênero brasileira. Além disso, homens com alta escolaridade tem mais chances de estarem empregados do que mulheres de alta escolaridade.

Analisando minuciosamente os resultados, nota-se que a faixa de idade e escolaridade não influenciam efetivamente em sua empregabilidade. Na verdade, irão influenciar no valor de sua remuneração. Contudo, o tipo de atividade desempenhada é fator decisivo para sua empregabilidade. Ou seja, o que realmente influenciará a empregabilidade masculina é a realidade econômica brasileira. O que é notório nos resultados: Os resultados mostram um número de desempregados alto na construção civil e setor industrial, o que faz muito sentido, afinal vive-se uma crise econômica no período dos dados analisados.

Outro ponto a se destacar, que a idade é um ponto relevante da empregabilidade masculina. Olhando para pirâmide etária masculina, verifica-se que o topo etário masculina é muito menor do que o topo feminino. Grande parte da população masculina não chega a terceira idade. No ponto de vista da empregabilidade temos um grande número de homens em idade economicamente ativa. Além disso, verificou-se que homens iniciam a carreira antes das mulheres.

Essa realidade é mais um demostrativo da desigualdade de gênero no país. Além dos fatos econômicos do país, a mulher tem que lidar com outros elementos que irão influenciar diretamente em sua empregabilidade. É necessário o estudo constante dessa realidade e acrescentar mais fatores para entender melhor as causas do desemprego feminino.

6 CONCLUSÕES

Mesmo em meio a tantas transformações sociais ocorridas no último século sob a perspectiva de gênero (maior participação das mulheres no mercado de trabalho, aumenta a escolaridade, redução da fecundidade), as mulheres seguem dedicando relativamente mais tempo aos afazeres domésticos e/ou cuidados de pessoas (ONU 2017). Nesse viés, a participação feminina no mercado de trabalho é menos expressiva que masculina no mercado de trabalho.

Este trabalho de mestrado estudou as técnicas de Mineração de Dados, especificamente a classificação supervisionada com árvores de decisão, para estudo da empregabilidade feminina no Brasil. O estudo se mostrou eficiente na geração de conhecimento de forma automatizada e rápida para formação de conhecimento para a construção de políticas públicas pelo governo.

No trabalho tentou-se elencar fatores que influenciam a empregabilidade feminina. É notório que mulheres ganham menos do que homem, porém trabalham menos horas semanais no trabalho formal. Além disso, trabalham em atividades que não exigem um alto grau de instrução. Um dado interessante é que mulheres analfabetas possuem mais posto de trabalhos formais do que mulheres com alto grau de escolaridade, porém, isso mora no fato na baixa incidência de mulheres com esse alto grau de escolaridade. Porém, avaliando esse comparativo por idades, notamos que idade é um fator muito importante na empregabilidade feminina, mulheres entre 18-35 anos são mais empregadas que as demais.

Apesar de extrairmos muitas informações das bases ofertadas, os resultados obtidos demonstram que as informações coletadas pelo governo ainda tem muito o que melhorar para compreender a empregabilidade feminina. É necessário trazer informações adicionais dos empregados e empregadores para essas bases de dados. O estudo da empregabilidade é complexo e está atrelado a algumas relações sociais históricas difíceis de mensurar em dados. Por isso, é necessário adicionais algumas informações sociais desses empregados na RAIS, como número de filhos, renda familiar para entender melhor a realidade desses trabalhadores.

Além disso, a utilização do XGBoost possibilitou a formação de *datasets* eficientes que retornaram resultados com bom desempenho do algoritmo. A ideia de utilização de árvore de decisão garante uma visualização clara e intuitiva dos resultados, possibilitando que qualquer um possa interpretar os resultados.

6.1 TRABALHOS FUTUROS

Como perspectivas de trabalhos futuros, foram listados as seguintes propostas:

- Validação dos resultados obtidos com base em dados de outras fontes e enriquecer o *Dataware* com mais informações;
- Realização um estudo temporal para avaliar a evolução da empregabilidade feminina no Brasil;

- Avaliar a informação econômicos importantes, como inflação e recessão, na empregabilidade;
- Integralizar as funcionalidades de ETL e Aprendizado de Máquina ao site do observatório do trabalho;
- Realizar o estudo regionalizado da empregabilidade feminina.

REFERÊNCIAS BIBLIOGRÁFICAS

Alpaydin 2010 ALPAYDIN, E. *Introduction to machine learning*. 2. ed. [S.l.]: Massachusetts Institute of Technology, 2010.

Biesma et al. 2007 BIESMA, R. G.; PAVLOVA, M.; MERODE, G. V.; GROOT, W. Using conjoint analysis to estimate employers preferences for key competencies of master level dutch graduates entering the public health field. Economics of Education Review, v. 26, n. 3, p. 375–386, 2007.

Boletim Geográfico 2018 BOLETIM GEOGRÁFICO. Estatísticas de gênero indicadores sociais das mulheres no brasil. IBGE, 2018. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv101551_notas_tecnicas.pdf>. Acesso em: 29 dez. 2019.

Breiman 2001 BREIMAN, L. *Random forests*. [S.l.: s.n.], 2001.

Dolado e Felgueroso 2001 DOLADO, J.; FELGUEROSO, F. Female employment and occupational changes in the 1990s: How is the eu performing relative to the us? European Economic Review, v. 45, p. 875–889, 2001.

Dolado e Felgueroso 2004 DOLADO, J. J.; FELGUEROSO, F. Where do women work? analysing patterns in occupational segregation by gender. Annales d'Économie et de Statistique, v. 71/72, p. 293–315, 2004.

DOU 1965 DOU. Decreto nº 76.900, de 23 de dezembro de 1965.- institui o cadastro permanente das admissões e dispensas de empregados, estabelece medidas contra o desemprego e de assistência aos desempregados, e dá outras providências. 1965. Disponível em: <http://www.planalto.gov.br/ccivil_03/LEIS/L4923.htm>. Acesso em: 20 dez. 2019.

DOU 1975 DOU. Decreto nº 76.900, de 23 de dezembro de 1975.-institui a relação anual de informações sociais – rais e dá outras providência. 1975. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/Antigos/D76900.htm>. Acesso em: 20 dez. 2019.

Duchesne e Rémilland 2005 DUCHESNE, P.; RéMILLAND, B. Statistical modeling and analysis for complex data problems. Springer Science and Business Media, 2005.

Fawcett 2006 FAWCETT, T. An introduction to roc analysis. Pattern recognition letters, v. 27, n. 8, p. 861–874, 2006.

Fayyad et al. 1996 FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. *Advances in knowledge discovery and data mining*. [S.l.]: AAAI press, 1996. v. 21.

Franco 2018 FRANCO, S. A. Mulheres que abandonam a carreira profissional: Uma análise da ocorrência do fenômeno opt-out entre brasileiras. *Universidade de Brasília - UnB*, 2018.

García-Peñalvo et al. 2018 GARCÍA-PEñALVO, F. J.; CRUZ-BENITO, J.; MARTÍN-GONZÁLEZ, M.; VÁZQUEZ-INGELMO, A.; SÁNCHEZ-PRIETO, J. C.; THERÓN, R. Proposing a machine learning approach to analyze and predict employment and its factors. International Journal of Interactive Multimedia and Artificial Intelligence, v. 5, n. 7, p. 39–45, 2018.

García-Aracil e Velden 2008 GARCíA-ARACIL, A.; VELDEN, R. V. der. Competencies for young european higher education graduates: labor market mismatches and their payoffs. Higher Education, v. 55, n. 2, p. 219–239, 2008.

Goldschmidt e al. 2015 GOLDSCHMIDT, R.; AL. et. *Data Mining - Conceitos, técnicas, algoritmos, orientações e aplicações*. 2. ed. [S.l.]: Elsevier, 2015.

Gollapudi 2016 GOLLAPUDI, S. *Practical Machine Learning*. [S.l.]: Packt Publishing Ltd, 2016.

Gujarati e Porter 2008 GUJARATI, D. N.; PORTER, D. C. *Econometria Básica*. 5. ed. [S.l.]: McGraw-Hill Kogakush,LTD, 2008.

Han, Kamber e Pei 2000 HAN, J.; KAMBER, M.; PEI, J. *Data Mining - Concepts and Techniques*. 3. ed. [S.l.]: Morgan Kaufmann, 2000.

Hand, Mannil e Smyth 2001 HAND, D.; MANNIL, H.; SMYTH, P. *Principles of data mining*. [S.l.]: MIT press, 2001.

Haynes, Western e Spallek 2005 HAYNES, M.; WESTERN, M.; SPALLEK, M. Methods for categorical longitudinal survey data: Understanding employment status of australian women. HILDA Survey Research Conference, University of Melbourne, Austr'ali, 2005.

IBGE 2019 IBGE. Pesquisa nacional por amostra de domicílios contínua. 2019. Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv101707_informativo.pdf>. Acesso em: 29 dez. 2019.

Kelly, O'Connell e Smyth 2010 KELLY, E.; O'CONNELL, P. J.; SMYTH, E. The economic returns to field of study and competencies among higher education graduates in irelandl. Economics of Education Review, v. 29, n. 4, p. 650–657, 2010.

Koyanagi 2010 KOYANAGI, R. *Programa seguro-desemprego: combinação de eficiência econômica e proteção social*. 102 f. Dissertação (Mestrado em Ciências Sociais) — Universidade de Brasília, Brasília, 2010.

Lorose 2005 LOROSE, D. *Discovering Knowledge in Data: An Introduction to Data Mining*. [S.l.]: John Wiley and Son, 2005.

MacQueen 1967 MACQUEEN, J. Some methods for classification and analysis of multivariate observations. symposium on mathematical statistics and probability, v. 1, p. 281–297, 1967.

MacQueen 1967 MACQUEEN, J. Some methods for classification and analysis of multivariate observations. symposium on mathematical statistics and probability, v. 1, p. 281–297, 1967.

McQuaid e Lindsa 2005 MCQUAID, R. W.; LINDSA, C. The concept of employability. International Journal of Interactive Multimedia and Artificial Intelligence, v. 42, n. 2, p. 197–2195, 2005.

Medeiros e Pinheiro 2018 MEDEIROS, M.; PINHEIRO, L. S. Desigualdades de gênero em tempo de trabalho pago e não pago no brasil. *Sociedade e Estado*, v. 33, p. 159 – 185, 04 2018.

Ministério da Economia 2019 MINISTÉRIO DA ECONOMIA. Manual de orientação da relação anual de informações sociais (rais). 2019. Disponível em: <http://www.rais.gov.br>. Acesso em: 20 dez. 2019.

Ministério da Economia 2019 MINISTÉRIO DA ECONOMIA. Manual de orientação do cadastro geral de empregados e desempregados (caged). 2019. Disponível em: <http://portalfat.mte.gov.br/programas-e-acoes-2/caged-3/>. Acesso em: 20 dez. 2019.

Mitchell 1997 MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill Science, 1997.

Morettin e Bussab 2007 MORETTIN, P. A.; BUSSAB, W. de O. *Estatística Básica*. 9. ed. [S.l.]: Saraiva, 2007.

Oliveira, Scorzafave e Pazello 2009 OLIVEIRA, P. R. de; SCORZAFAVE, L. G.; PAZELLO, E. T. Desemprego e inatividade nas metrópoles brasileiras: as diferenças entre homens e mulheres. *Nova Economia*, v. 19, p. 291 – 324, 09 2009.

ONU 2017 ONU. Glossário de termos do objetivo de desenvolvimento sustentável 5: Alcançar a igualdade de gênero e empoderar todas as mulheres e meninas. Organização da Unidas Unidas para as Mulheres- ONU Mulher, 2017. Disponível em: <http://www.onumulheres.org.br/onu-mulheres/documentos-de-referencia/#>.

Osborne 2015 OSBORNE, J. W. *Gender in the Global Research Landscape*. [S.l.]: Elsevier, 2015.

Pawha e Kamthania 2019 PAWHA, A.; KAMTHANIA, D. Quantitative analysis of historical data for prediction of job salary in india - a case study. *Journal of Statistics and Management Systems*, v. 22, n. 2, p. 187 – 198, 09 2019.

Pearson, Frehill e McNeely 2015 PEARSON, W. J.; FREHILL, L. M.; MCNEELY, C. L. *An International Perspective on Advancing Women in Science*. [S.l.]: Springer, 2015.

Phridvi e Guru 2013 PHRIDVI, R. M.; GURU, R. C. Data mining – past, present and future – a typical survey on data streams. The 7th International Conference Interdisciplinarity in Engineering, 2013.

Punnoose e Pankaj 2016 PUNNOOSE, R.; PANKAJ, A. Prediction of employee turnover in organizations using machine learning algorithms. *International Journal of Advanced Research in Artificial Intelligence*, v. 5, 10 2016.

Quinlan 1993 QUINLAN, J. *C4.5: Programs for Machine Learning*. [S.l.]: Morgan Kaufman, 1993.

Russell 2004 RUSSELL, S. *Inteligência Artificial*. [S.l.]: ed. Elsevier, 2004.

Samuelson 1955 SAMUELSON, P. A. *Economics*. 9. ed. [S.l.]: McGraw-Hill Kogakush,LTD, 1955.

Schomburg e Teichler 2007 SCHOMBURG, H.; TEICHLER, U. Higher education and graduate employment in europe: results from graduates surveys from twelve countries. Springer Science and Business Media, 2007.

Shabana, Gracious e Subramonian 2016 Shabana, K. M.; Gracious, T.; Subramonian, H. Understanding the indian labour market: A data-centric approach. *2016 International Conference on Data Science and Engineering (ICDSE)*, p. 1–6, 2016.

Silva et al. 2019 SILVA, D. A. da; MACHADO, P. L.; COELHO, V. C. G.; MENDONçA, R. V. B. F. L. L. de; SANTOS, D. P. dos; JúNIOR, R. T. de S. Produção de indicadores de empregabilidade com base em técnicas de mineração de big data e business intelligence. *Inclusão Social*, v. 12, n. 2, p. 141 – 155, 2019.

Smola e Vishwanathan 2008 SMOLA, A.; VISHWANATHAN, S. *Introduction to machine learning*. 1. ed. [S.l.]: Cambridge University Press, 2008.

Tan, Steinbach e al. 2009 TAN, P.; STEINBACH, M.; AL. et. *Introdução ao data mining: Mineração de dados*. [S.l.]: Ciência Moderna, 2009.

Theil 1971 THEIL, H. *Principles of econometrics*. 1. ed. [S.l.]: Nova York: John Wiley and Sons, 1971.

Wooldridge 2003 WOOLDRIDGE, J. M. *Introdução à econometria - Uma Abordagem Moderna*. 1. ed. [S.l.: s.n.], 2003.

Wu et al. 2008 WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; MOTODA, Q. Y. nad H.; MCLACHLAN, G. J.; NG, A.; LIU, B. Top 10 algorithms in data mining knowledge and information systemsg. v. 14, n. 1, p. 1–37, 2008.

XGBOOST 2019 XGBOOST. Xgboost documentation. 2019. Disponível em: <https://xgboost.readthedocs.io/en/latest/index.htm>. Acesso em: 20 dez. 2019.