UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Alexandre Seidy Ioshisaqui

Investigação de perfis ocupacionais no mercado de trabalho brasileiro através da análise de dados

Campinas

2021

Alexandre Seidy Ioshisaqui

Investigação de perfis ocupacionais no mercado de trabalho brasileiro através da análise de dados

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na Área de Engenharia de Computação.

Orientador: Prof. Dr. Romis Ribeiro de Faissol Attux

Coorientadora: Profa. Dra. Ivette Raymunda Luna Huamaní

Este trabalho corresponde à versão final da dissertação defendida pelo aluno Alexandre Seidy Ioshisaqui, e orientada pelo Prof. Dr. Romis Ribeiro de Faissol Attux.

Campinas

2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Ioshisaqui, Alexandre Seidy, 1992-
Io7i Investigação de perfis ocupacionais no mercado de trabalho brasileiro através da análise de dados / Alexandre Seidy Ioshisaqui. – Campinas, SP : [s.n.], 2021.

Orientador: Romis Ribeiro de Faissol Attux.
Coorientador: Ivette Raymunda Luna Huamaní.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Mercado de trabalho. 2. Indústria 4.0. 3. Aprendizado de máquina. I. Attux, Romis Ribeiro de Faissol, 1978-. II. Luna Huamaní, Ivette Raymunda, 1978-. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Investigation of occupational profiles in the Brazilian labor market based on data analysis
Palavras-chave em inglês:
Job market
Industry 4.0
Machine learning
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Romis Ribeiro de Faissol Attux [Orientador]
Leonardo Tomazeli Duarte
Diego Jair Vicentin
Data de defesa: 21-12-2021
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: 0000-0003-2887-3295
- Currículo Lattes do autor: http://lattes.cnpq.br/1399668695309204

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidato: Alexandre Seidy Ioshisaqui RA: 137943
Data de defesa: 21 de Dezembro de 2021
Título da Tese: "Investigação de perfis ocupacionais no mercado de trabalho brasileiro através da análise de dados"

Prof. Dr. Romis Ribeiro de Faissol Attux (Presidente)
Prof. Dr. Leonardo Tomazeli Duarte (FCA/UNICAMP)
Prof. Dr. Diego Jair Vicentin (FCA/UNICAMP)

A Ata de Defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

Agradecimentos

Sou muito grato por todos os apoios que recebi, de todas as formas e de todas as pessoas com quem convivi, que me trouxeram à finalização deste mestrado. Primeiramente, pelo apoio incondicional à minha trajetória e pelo suporte inabalável em todos os momentos, agradeço à minha mãe Mitiko, ao meu pai, Joel e ao meu irmão, Victor.

Agradeço ao meu orientador, professor Romis Attux, quem me guiou e acompanhou nesta trajetória de formação e descoberta acadêmica, por sua generosidade e determinação, em orientar um aluno buscando por temas tão pouco estabelecidos nas engenharias, como são os que buscam a integração destas às faculdades humanas e sociais.

Agradeço à minha co-orientadora, professora Ivette Luna, quem me sugeriu o tema da pesquisa dos perfis ocupacionais dos trabalhadores brasileiros, pela confiança que conferiu a mim para realizarmos este trabalho, por todas as conversas e trocas de ideias, pelas informações e recomendações e pelas cuidadosas revisões e correções ao longo do mestrado.

Agradeço também aos professores Leonardo Tomazeli e Diego Vicentin, que trouxeram diversas reflexões e sugestões que me ajudaram não só a esclarecer ideias e melhorar a qualidade deste trabalho, mas também me ajudaram a pensar sobre a realidade social dos trabalhadores em unidade com o contexto de transformações tecnológicas atual.

Aos meus amigos do DSPCom, tenho profundo apreço por todos os momentos, de conversas, de trocas de ideias, de ajudas, dos cafés, das brincadeiras e de tantas outras vivências que compartilhamos.

Finalmente, agradeço aos funcionários da Unicamp e da FEEC, especialmente à equipe da Coordenadoria de Pós-Graduação da FEEC, pelo trabalho tenaz e contínuo que realizam de manter em funcionamento as nossas atividades dentro do espaço universitário. Também agradeço a CAPES, que tornou possível que eu me dedicasse integralmente à minha formação acadêmica.

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

Resumo

No contexto da expansão e aprofundamento das transformações tecnológicas em diversos aspectos da sociedade contemporânea, levanta-se a discussão acerca dos possíveis impactos destas transformações na realidade dos trabalhadores, em suas relações de trabalho e sua empregabilidade. Particularmente, o processo de digitalização e a indústria 4.0 trazem questões sobre a capacidade de automação de tarefas cada vez mais complexas.

Neste trabalho, realizamos uma investigação dos perfis ocupacionais no mercado de trabalho brasileiro através da análise de dados. Para tal, implementamos a clusterização via Fatoração de Matrizes Não-Negativas sobre os dados do Quadro Brasileiro de Qualificações (QBQ), uma base de dados que contém representações numéricas das atividades, conhecimentos e atitudes de cada uma das ocupações brasileiras, desenvolvida pelo Ministério da Economia. Para considerar os possíveis efeitos de automação de tarefas devidos ao processo de digitalização, utilizamos os dados de probabilidade de automação de ocupações publicados no trabalho de Frey e Osborne, originalmente para o contexto estadunidense, para extrapolar um modelo para as ocupações brasileiras.

Combinando os resultados da clusterização e do modelo de probabilidade de automação, sintetizamos as tendências e padrões encontradas nos dados, discutimos sobre os possíveis significados dos resultados e consideramos algumas possíveis limitações das procedimentos realizados no trabalho. A análise dos resultados indica que os *clusters* encontrados, que foram construídos a partir das representações numéricas das ocupações, de fato refletem em determinados perfis ocupacionais, com variadas comunalidades e diferenças em relação às categorias utilizadas oficialmente pela Classificação Brasileira de Ocupações (CBO).

Palavras-chaves: Mercado de trabalho, digitalização, indústria 4.0, aprendizado de máquina.

Abstract

In the context of the expansion and deepening of technological transformations within various aspects of contemporary society, a discussion arises about the possible impacts of these transformations on the reality of workers, on their work relationships and their employability. Particularly, the process of digitalization and the Industry 4.0 brings questions about the ability to automate increasingly complex tasks.

In this work, we investigate occupational profiles in the Brazilian labor market through data analysis. To this end, we implement clustering via Non-Negative Matrix Factorization on the data from the Brazilian Qualifications Framework (QBQ), a database that contains numerical representations for the activities, knowledges and behaviors of each occupation, developed by the Ministry of Economy. To consider the possible effects of task automation due to the digitalization process, we used the data on the probability of automation of occupations published in the work by Frey and Osborne, originally for the context of the United States, to extrapolate a model for the Brazilian occupations.

Combining the results of clustering and the automation probability model, we synthesize trends and patterns found in the data, discuss the possible meanings of the results and consider some possible limitations of the procedures performed in the work. The analysis of the results indicates that the clusters we found, which were constructed from the numerical representations of the occupations, do in fact reflect on certain occupational profiles, with varied commonalities and differences in relation to the categories officially used by the Brazilian Classification of Occupations (CBO).

Keywords: Job market, digitalization, industry 4.0, machine learning.

Lista de ilustrações

Lista de abreviaturas e siglas

AM Aprendizado de Máquina.

CAGED Cadastro Geral de Empregados e Desempregados.

CBO Classificação Brasileira de Ocupações.

CIUO Classificação Internacional Uniforme de Ocupações.

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico.

COP Classificação Ocupacional Padrão.

GB Grupo de base ou família.

GG Grande grupo.

IBGE Instituto Brasileiro de Geografia e Estatística.

ICA Independent Component Analysis.

Ipea Instituto de Pesquisa Econômica Aplicada.

MTST Movimento dos Trabalhadores Sem Teto.

NMF *Non-negative Matrix Factorization*.

PCA Principal Component Analysis.

QBQ Quadro Brasileiro de Qualificações.

QEQ Quadro Europeu de Qualificações.

QNQ Quadro Nacional de Qualificações.

RAIS Relação Anual de Informações Sociais.

ReLU *Rectified Linear Unit*.

SG Subgrupo.

SGP Subgrupo principal.

SINE Sistema Nacional de Empregos.

Sumário

1 Introdução

Existem frequentes discussões sobre como tecnologias modernas de informação e comunicação se tornaram elementos-chave em várias esferas da vida, desde atividades simples do cotidiano até grandes cadeias produtivas das nossas economias. O desenvolvimento de robôs cada vez mais versáteis em fábricas, a ubiquidade de plataformas online, a aplicação de aprendizado de máquina na automação de uma quantidade crescente de tarefas — esses são temas recorrentes em assuntos como o processo da digitalização, a chamada indústria 4.0 e o advento da uberização. Isso estabelece a necessidade de reflexão sobre novas dinâmicas nas estruturas ocupacionais (BAKHSHI et al., 2017) e de produção (CEPAL, 2014), e levanta preocupações sobre o futuro das relações e condições de trabalho.

Trabalhos na literatura recente focaram em estimar os impactos destas tendências tecnológicas no deslocamento de trabalhadores. O trabalho de (FREY; OSBORNE, 2017) apresenta estimativas da probabilidade de automação dos trabalhos nos Estados Unidos como consequência da computadorização. Isto é feito através da anotação manual de uma quantidade de ocupações como "automatizáveis" ou não e desenvolvendo um modelo treinado via aprendizado supervisionado baseado nas descrições numéricas de cada ocupação, fornecidas pela base de dados da O*NET (O*NET, 2019). Outro exemplo é o trabalho publicado pelo Instituto de Pesquisa Econômica Aplicada (Ipea) (ALBUQUERQUE, 2019), que, similarmente, buscou estimar o risco de automação dos trabalhos no Brasil através da anotação manual de uma porção das ocupações brasileiras e generalizando estas avaliações preliminares, treinando um modelo supervisionado baseado nas descrições textuais fornecidas pela Classificação Brasileira de Ocupações (CBO) (CBO, 2002a).

Neste trabalho, utilizamos a nova base de dados do Quadro Brasileiro de Qualificações (QBQ) (QBQ, 2020), que fornece descrições para as ocupações no mercado de trabalho brasileiro em termos de habilidades, conhecimentos e atitudes requeridas para executar um trabalho. Nosso objetivo principal é realizar uma análise de dados que ajude a visualizar padrões gerais e caracterizar perfis específicos. Dado um condicionamento inicial dos dados do QBQ, utilizamos a Fatoração de Matrizes Não-negativas (*Non-negative Matrix Factorization* (NMF), do inglês *Non-negative Matrix Factorization*) como meio de realizar clusterização de dadaos. Os clusters resultantes serão a base para a análise de padrões nos dados. Em seguida, estendemos os dados da QBQ com as estimativas de risco de automação para cada ocupação, que será baseada nos resultados publicados por Frey e Osborne (FREY; OSBORNE, 2017). Esta camada adicional de informação nos permitirá realizar novas análises e nos dará novas perspectivas sobre os padrões e perfis no contexto da automação de trabalhos devida à computadorização.

Estas análises serão realizadas com o objetivo de encontrar padrões expressos nos dados que nos permitam refletir sobre possíveis relações entre o trabalho e as tecnologias de computadorização e automação, além de também nos ajudar a entender em mais detalhes a composição destas bases de dados em si. Realizamos estas análises sem pretensão de estabelecer predições sobre a realidade futura das ocupações e do mercado de trabalho brasileiro, nem de caracterizar as relações e padrões encontrados nos dados como estruturas que descrevam, rigorosamente, processos ou fenômenos socio-econômicos.

O restante deste trabalho é dividido em 5 capítulos. O Capítulo 2 apresenta uma breve contextualização sobre o cenário dos empregos em relação às tecnologias recentes de computadorização e automação. O Capítulo 3 introduz as diferentes técnicas de análise de dados e de extração de informação utilizadas ao longo deste trabalho. O Capítulo 4 é dedicado à descrição e discussão das bases de dados utilizadas e das estratégias utilizadas preparar os dados para análise. Obtidos os resultados, no Capítulo 5, são feitas discussões e análises sobre as diferentes tendências e padrões observadas nestes dados. Por fim, o Capítulo 6 traz as principais conclusões e considerações finais do trabalho.

2 Indústria 4.0 e Mercado de Trabalho

2.1 Transformações Produtivas e Indústria 4.0

As três primeiras revoluções industriais na história da humanidade foram importantes momentos de mudanças de paradigma nas formas de produção, e se caracterizaram por diferentes saltos tecnológicos que acabaram afetando significativamente as nossas atividades. A primeira revolução foi marcada pelas tecnologias de mecanização, a segunda pelo uso intensivo de energia elétrica e a terceira pela ampla digitalização (LASI et al., 2014). Todos estes processos promoveram não só importantes mudanças nas dinâmicas e estruturas produtivas (CEPAL, 2014) — levando a maiores níveis de produtividade e eficiência nas indústrias — como também geraram amplas mudanças nas estruturas ocupacionais (LUNA, 2019) dos países que passaram por estes processos de transformação.

O momento mais recente dos processos de mudanças tecnológicas, não obstante, traz características distintas em relação aos impactos na estrutura ocupacional. Em períodos anteriores, foram observadas substituições por máquinas (mecânicas, elétricas ou computacionais, dependendo da época) em tarefas predominantemente rotineiras e simples. Mas, com os desenvolvimentos tecnológicos mais recentes, tarefas mais estreitamente associadas a capacidades cognitivas, até então consideradas exclusivamente humanas, começam a ser consideradas passíveis de automação, completa ou parcial, por vias computacionais ou robóticas. De fato, as mais recentes mudanças de paradigma tecnológico são tais que alguns autores chegam a considerá-las base de uma nova revolução tecnológica (CEPAL, 2014). É no contexto da expectativa desta revolução que o termo "Indústria 4.0" foi estabelecido, como um planejamento para a "quarta revolução industrial" (LASI et al., 2014).

Encontramo-nos num momento de rápida expansão e aprofundamento da adesão às tecnologias de informação e comunicação, tanto nas esferas de produção quanto nas esferas de atividades cotidianas. Tecnologias como robótica, aprendizagem de máquina e Internet das Coisas seguem recebendo investimentos expressivos para pesquisa e desenvolvimento. Um dos fatores que articulam a interação destas tecnologias é o processo de coleta e produção de grandes e complexas bases de dados, que configura o que conhecemos por *big data*. É a geração e o uso de grandes bases de dados que torna possível o desenvolvimento de algoritmos pertencentes aos modernos paradigmas de aprendizagem de máquina, algoritmos que buscam otimizar determinadas tarefas — programadas em termos matemáticos — em função de um grande número de registros de algum determinado fenômeno, processo, grupo de objetos ou população.

2.2 Potenciais Impactos nos Empregos

Neste contexto, podemos vislumbrar um cenário em que novas esferas de atividades produtivas, previamente consideradas exclusivamente de capacidade humana devido a suas complexidades ou dimensões de cognição, tornam-se passíveis de automação através de algoritmos e máquinas. São diversos os exemplos: carros autônomos (SELF-DRIVING..., 2016), chatbots para atendimento de clientes (ex.: serviços de e-banking (WALCH, 2019; Chatbot..., 2018)), algoritmos de auxílio em tomadas de decisão (ex.: recomendações para a área de saúde (AI..., 2021; DeepMind..., 2019) e jurídica (MARR, 2020)) e recomendação de produtos e serviços (ex.: serviços imobiliários (OLICK, 2021) e lojas (MORGAN, 2018)).

As potencialidades dos efeitos das tecnologias vindas dos novos desenvolvimentos nas áreas de aprendizagem de máquina e *big data* é diversa e possui diferentes nuances. Como nos exemplos citados anteriormente, podem substituir desde ocupações inteiras quanto apenas tarefas específicas. Também existem possibilidades mais sinergéticas, como nas próprias áreas relacionadas à análise de dados ou em áreas de pesquisa científica (ex.: descoberta de proteínas pelo AlphaFold pelo método desenvolvido pela DeepMind (SENIOR et al., 2020; JUMPER et al., 2021)).

Também existem diferentes nuances nas maneiras pelas quais estas tecnologias podem afetar as relações e condições do trabalho. Em um debate promovido pelo Instituto de Economia da Unicamp (Instituto de Economia da Unicamp, 2020) acerca da precarização do trabalho, Gabriel Simeone, organizador do Movimento dos Trabalhadores Sem Teto (MTST) e ex-trabalhador de entrega por motos, relata que não é exatamente a relação de trabalho por aplicativo em si (como Rappi, iFood ou Uber Eats) o fator que causa revolta entre os trabalhadores da categoria dos motoboys, já que este tipo de relação já existiam antes com empresas como a Loggi. Segundo Gabriel, houve um primeiro elemento de tirar parte do trabalho intelectual do motorista. Os motoristas, ao receberem pedidos de entrega, planejavam o roteiro sequenciando as entregas de maneira que formassem uma rota linear, ou seja, que levasse em consideração a minimização de seus gastos enquanto buscavam maximizar a quantidade de entregas. Os aplicativos, abastecidos de dados e algoritmos, deslocam tanto o papel da navegação, que fica em empresas como a Waze, quanto o papel de planejamento de rotas, que fica para as empresas de entrega. Ainda, um segundo elemento citado foi o de que, devido à não necessidade de conhecimento sobre navegação (a demonstração de conhecimentos de navegação nas ruas de uma região era comumente requerida, antes da difusão de aplicativos de navegação), houve um grande salto no número de pessoas que ingressaram na categoria de entregadores por moto, o que significou um maior acirramento pela oferta de trabalho e redução nos valores das entregas.

Buscando entender melhor as ocupações em termos de suas caracterizações,

e levando em consideração os potenciais impactos das novas tecnologias de automação nos empregos, motivamo-nos a investigar as bases de dados que tratam das qualificações dos trabalhos com descrições que situam cada ocupação em função das habilidades e conhecimentos requeridas para exercer suas atividades e tarefas. Um dos trabalhos de maior destaque nesta linha de investigação é o realizado por Frey e Osborne (FREY; OSBORNE, 2017), que busca quantificar os possíveis impactos da automação nos empregos nos Estados Unidos através de um modelo treinado por aprendizado supervisionado. O estudo, estendendo a ideia do trabalho de (AUTOR; LEVY; MURNANE, 2003), que partiu da distinção entre tarefas rotineiras e não-rotineiras para analisar as relações entre estas e a vulnerabilidade à automação por vias computacionais (computadorização), seleciona um conjunto de nove habilidades e conhecimentos que são denominadas variáveis indicadoras de *bottleneck* à computadorização de maneira geral, isto é, habilidades que indicam fatores de resistência ou impedimento contra a computadorização. A motivação desta escolha é a de que, apesar da intuitiva predição de que tarefas rotineiras no trabalho humano são mais vulneráveis à automação do que tarefas não-rotineiras, hoje, com a disponibilidade de *big data* e avanço nas tecnologias de robótica, passam a ser observadas tarefas não-rotineiras também passíveis de automação. A escolha das habilidades e conhecimentos descritores foi feita utilizando as variáveis da O*NET, uma base de dados desenvolvida pelo *Bureau of Labor Statistics* dos Estados Unidos, buscando curar uma seleção de variáveis que descrevessem níveis de percepção, manipulação, criatividade e inteligência social (O*NET, 2019). As variáveis da O*NET são variáveis numéricas de valores discretos de 0 a 100, que representam o grau de importância de cada um destes atributos ao exercício de uma determinada ocupação.

Para obter um conjunto de treinamento para o aprendizado supervisionado, esses autores realizaram uma oficina para a criação de uma base de referência da avaliação da vulnerabilidade de uma parcela das ocupações estadunidenses com dados disponíveis (70 ocupações de um total de 702). A anotação desses valores de referência foi feita de maneira binária, atribuindo-se "1" se a ocupação fosse considerada totalmente automatizável e "0" caso contrário. Os autores afirmam que as atribuições foram feitas apenas para ocupações em cuja avaliação os avaliadores tinham mais confiança. O treinamento supervisionado foi feito utilizando a estrutura de modelagem de um classificador binário $p(y = 1|f) = \frac{1}{1+exp(-f)}$ na qual, em vez de utilizar $f = \mathbf{w}.\mathbf{x}$ (regressão logística), foi utilizada uma modelagem via processos gaussianos para f.

Em linhas gerais, o trabalho de Frey e Osborne (FREY; OSBORNE, 2017) identifica grupos com diferentes níveis de risco de automação, estimando que 47% dos empregos dos Estados Unidos estão na categoria de alto risco, e faz a estimativa de duas ondas de automação. A primeira se vincula à automação devida à suscetibilidade das ocupações às tecnologias atuais e ao capital computacional, especialmente no âmbito de ocupações de transporte, logística, escritório, assistência administrativa e produção. Já

a segunda onda é condicionada à superação dos gargalos de engenharia relacionados às inteligências criativa e social. Também foi observado que os níveis de renda e de educação têm relações fortemente negativas em relação à probabilidade de computadorização, o que foi considerado pelos autores uma possível quebra da tendência de polarização do mercado de trabalho, com um foco na automação de ocupações de baixos níveis de renda e de habilidades, segundo as predições do modelo, sendo sugerido que habilidades criativas e sociais sejam fatores estratégicos para a manutenção da empregabilidade destes trabalhadores. Estas estimações realizadas por Frey e Osborne estão disponíveis para consulta (FREY; OSBORNE, 2017).

No contexto da força de trabalho brasileira, o trabalho publicado pelo Ipea (ALBUQUERQUE, 2019) se inspira no trabalho de Frey e Osborne para realizar predições da vulnerabilidade das ocupações brasileiras à automação utilizando as informações fornecidas pela Classificação Brasileira de Ocupações. O trabalho foi feito a partir de um questionário distribuído para diversos pesquisadores brasileiros cadastrados na plataforma Lattes do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) que tivessem atuado em projetos de automação relacionados à aprendizagem de máquina. Os pesquisadores recebiam um conjunto aleatoriamente escolhido de nomes das ocupações e descrições das atividades dessas ocupações, e, então, pedia-se que atribuíssem uma probabilidade entre 0 e 1 ao nível de automação. Diferentemente do estudo de Frey e Osborne, este estudo não possuía uma base de dados com descrições numéricas detalhando cada uma das habilidades que compunham as atividades das ocupações. Como alternativa, foram utilizados os textos descritivos de cada uma das ocupações, codificados em vetores de frequências de palavras (unigramas e bigramas). Utilizando uma modelagem matemática similar à utilizada por Frey e Osborne, foi treinado um modelo estimador das probabilidades de automação e, selecionando as ocupações consideradas de risco alto ou muito alto (ocupações com valores de risco pertencentes ao terceiro e quarto quartil da distribuição de riscos) e cruzando-as com as informações sobre trabalhadores formalmente empregados, disponibilizados pela Relação Anual de Informações Sociais (RAIS) de 2017, foi estimado que 54% da população formalmente empregada se encontra em situação de probabilidade alta ou muito alta de automação. Os valores individuais de risco de automação estimados para as ocupações brasileiras não foram publicados.

Recentemente, vem sendo desenvolvido, pelo Ministério da Economia, o Quadro Brasileiro de Qualificações (QBQ), que consiste de uma base de dados contendo descrições multidimensionais para cada uma das ocupações de acordo com a Classificação Brasileira de Ocupações (CBO), descritas em termos de habilidades, conhecimentos e atitudes e avaliadas em escalas numéricas (QBQ, 2020). Esta base de dados abre a oportunidade de investigarmos em mais detalhes as questões sobre os potenciais impactos dos recentes desenvolvimentos tecnológicos sobre os empregos no Brasil. Uma das motivações centrais deste trabalho é a de investigar os dados com a finalidade de encontrar perfis ocupacionais

e padrões em suas composições de habilidades, conhecimentos e atitudes de maneira a identificar fatores que possam ser úteis para entender e pensar as possibilidades econômicas da força de trabalho brasileira. Além disso, devido à falta de acesso a estimativas detalhadas das probabilidades de automação para as ocupações do Brasil, realizaremos, através de modelos de regressão, uma aproximação dos resultados divulgados por Frey e Osborne, conformando-os às ocupações brasileiras segundo a CBO. Assim como nos estudos citados, a utilização de métodos de aprendizagem de máquina se mostra bastante interessante, possibilitando o bom proveito de dados de alta dimensionalidade para as tarefas de análise de padrões (métodos não-supervisionados) e de criação de modelos de regressão (métodos supervisionados).

3 Técnicas de Análise de Dados e Extração de Informação

3.1 Introdução

Com a finalidade de investigar padrões e perfis ocupacionais do mercado de trabalho brasileiro através da análise de dados multidimensionais, decidimos utilizar métodos de análise de dados multivariados.

Nossa procura foi de métodos que nos possibilitassem compreender relações complexas existentes em bases de alta dimensionalidade, e nos ajudassem a perceber e visualizar padrões, tendências ou eventuais perfis específicos presentes nos dados. Métodos como estes exigem não só a capacidade de tratamento de grandes volumes de dados com alta dimensionalidade, mas também a flexibilidade de lidar com conjuntos de dados com estruturas e distribuições complexas, e ainda nos forneçam de resultados que sejam significativos e interpretáveis para nossa finalidade de investigar relações subjacentes. Estas relações podem, eventualmente, indicar fenômenos de ocorrência concreta.

É neste contexto que enfocamos a classe das técnicas de Aprendizado de Máquina (AM), que são particularmente úteis para a análise de dados em grande quantidade e com alta dimensionalidade. A classe abrange uma ampla gama de modelos e algoritmos capazes de lidar com as mais variadas situações de complexidade de dados. Questões acerca da relação entre a capacidade de complexidade destes modelos, bem como de sua interpretabilidade, são temas presentes na literatura recente (RUDIN, 2019; MURDOCH et al., 2019; MITTELSTADT; RUSSELL; WACHTER, 2019).

O aprendizado de máquina (AM) foi descrito de diferentes maneiras ao longo de sua história, com o destaque de diferentes aspectos. Em (SAMUEL, 1959), o AM é descrito como uma maneira de programar um computador sem a definição de instruções explícitas por parte de um programador. Nesta definição, há um foco na característica da flexibilidade que computadores passam a ter para realizar tarefas de modo adaptativo / baseado em dados com o AM.

Com maior ênfase na explicação dos mecanismos fundamentais do AM, (MITCHELL, 1997) descreve o AM em termos de um programa de computador que aprende a partir de uma experiência E, em relação a uma classe de tarefas T e medida de desempenho P. Nesse caso, considera-se que o programa aprende caso seu desempenho para realizar uma tarefa T, medido por P, melhore com uma experiência E. Esta definição apresenta paralelos explícitos com as formulações de otimização matemática iterativas comumente

utilizadas (ex.: bases de dados para E, funções custo para P e estruturas de modelos de classificação para T).

Já (GÉRON, 2017) descreve o AM como a ciência e arte de programar computadores de maneira que estes aprendam a partir de dados. Nesta descrição, podemos perceber uma centralidade atribuída aos dados no processo de AM. Aqui o AM pode, de certa forma, ser entendido como todo o conjunto de técnicas que nos ajuda a conseguir o máximo proveito das informações contidas dentro dos conjuntos de dados para a realização de tarefas.

Estas são descrições que se complementam. Os métodos de AM demonstram cada vez mais sua capacidade de realizar tarefas cada vez mais diversificadas. Grande parte dos desenvolvimentos teóricos e práticos gravitam ao redor de alguns dos mecanismos descritos por (MITCHELL, 1997), e hoje existe, de certa forma, expressiva centralidade nos dados. De fato, raramente tomamos decisões em relação aos métodos antes de construir ou obter um conjunto de dados e de realizar uma análise exploratória quanto ao conteúdo. Ainda, tipicamente, diversas combinações de modelos e métodos de otimização podem ser utilizadas para uma mesma tarefa (e.g. de regressão ou classificação), ainda que com eventuais diferenças em suas formas de representar os resultados ou ainda nas capacidades de otimizar os modelos em relação a uma determinada medida matemática de desempenho. Mas os resultados de quaisquer modelos estarão sempre condicionados pelas características dos dados disponíveis (a forma de construção de sua estrutura, a coleta de dados, vieses estatísticos etc.).

Neste capítulo, utilizamos as técnicas de AM para buscar a compreensão de padrões e perfis contidos nas bases de dados sobre as qualificações dos trabalhadores brasileiros pelo Quadro Brasileiro de Qualificações (QBQ), e a sua relação com os riscos de automação estimados por (FREY; OSBORNE, 2017). As técnicas estão divididas nas categorias de aprendizado não-supervisionado e aprendizado supervisionado. Para a análise dos dados multidimensionais do QBQ, são desejáveis diversas tarefas, como a redução de dimensionalidade, a descoberta de variáveis latentes e o agrupamento de dados (clustering). Estas fazem parte da categoria de modelos de aprendizado não-supervisionado. Além destas análises, realizaremos a aproximação dos resultados de (FREY; OSBORNE, 2017) através do treinamento de um modelo com estrutura de classificador binário. Esta tarefa, de otimizar um modelo segundo um conjunto de dados de saída de referência, pertence à categoria de aprendizado supervisionado.

3.2 Modelos de Aprendizado Não-Supervisionado

3.2.1 Redução de Dimensionalidade

Quando trabalhamos com análise de dados, deparamo-nos com dificuldades relacionadas tanto ao grande volume de dados quanto à alta dimensionalidade dos mesmos. Etapas de exploração e de redução de dimensionalidade dos dados são estratégias comuns na tentativa de transformar dados complexos em conjuntos mais tratáveis tanto computacionalmente quanto para fins de análise. Esta redução pode se dar de diferentes maneiras, tipicamente podendo ser relacionadas a uma das duas categorias: seleção de variáveis ou busca de novas variáveis. A manutenção apenas das variáveis mais relevantes do conjunto de dados original é chamada de seleção de variáveis, enquanto o proveito das redundâncias nos dados através do encontro de conjuntos menores de novas variáveis, cada uma destas sendo uma combinação das variáveis originais, contendo basicamente a mesma informação, é chamado de redução de dimensionalidade (SORZANO; VARGAS; MONTANO, 2014).

Do ponto de vista computacional, a redução de dimensionalidade é um fator benéfico, já que grande parte dos algoritmos possuem custo computacional bastante sensíveis à dimensionalidade dos dados. Em termos de análise, temos ainda outras vantagens: uma análise qualitativa mais proveitosa tem lugar quando tratamos de menos variáveis e, em alguns casos, as novas variáveis em si podem ser boas indicadoras para descoberta de padrões nos dados por variados motivos: seja pelas relações que possuem com os dados originais, seja pelas características próprias dessas novas variáveis.

Utilizaremos uma descrição generalizada de um método de redução de dimensionalidade similar à feita em (SORZANO; VARGAS; MONTANO, 2014). Seja uma base de dados representada por uma matriz $m \times n$ $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}]$, onde $\mathbf{x_i} \in \mathbb{R}^m$ são registros originais da base de dados. O objetivo da redução de dimensionalidade é encontrar uma outra representação para os registros da base de dados $\mathbf{w_i}$ com dimensão k, sendo $k < m$, tal que esta ainda retenha o máximo de informação possível do conjunto original de observações $\mathbf{X}$. Este processo envolve alguma transformação $\mathbf{w_i} = T(\mathbf{x_i})$ e, apesar de não necessariamente existir uma inversa, deve também existir uma transformação $\hat{\mathbf{x}}_\mathbf{i} = R(\mathbf{w_i})$, tal que $\hat{\mathbf{x}}_\mathbf{i} \approx \mathbf{x}$. Esta proximidade entre $\hat{\mathbf{x}}_\mathbf{i}$ e $\mathbf{x_i}$ pode ser calculada de diversas maneiras, tipicamente sendo calculada pela distância euclidiana ($\|\mathbf{x_i} - \hat{\mathbf{x}}_\mathbf{i}\|_2^2$).

Um dos métodos estatísticos mais amplamente utilizados de redução de dimensionalidade, segundo (Hart, Stork, and Duda 2000), é a Análise de Componentes Principais (Principal Component Analysis (PCA), do inglês *Principal Component Analysis*) (JOLLIFFE, 1986). A PCA realiza a redução através de combinações lineares das variáveis originais. Como definem (HART; STORK; DUDA, 2000), o objetivo da PCA é "encontrar uma representação de menor dimensões que considere a variância" das variáveis. Outro método bastante popular e, de certa forma, relacionado à PCA é a Análise de Componentes

Independentes (Independent Component Analysis (ICA), do inglês *Independent Component Analysis*) (HYVÄRINEN; OJA, 2000). A ICA é uma ferramenta amplamente utilizada no problema de separação cega de fontes (ROMANO et al., 2018).

Uma distinção entre métodos de redução de dimensionalidade é feita por (SORZANO; VARGAS; MONTANO, 2014). Segundo os autores, há duas categorias de modelos para esse propósito. A primeira se refere a métodos de redução de dimensionalidade baseados em estatística e na teoria da informação, na qual a PCA e a ICA se enquadram. A segunda categoria de métodos são os baseados em dicionários, como é o caso da Fatoração de Matrizes Não-negativas (NMF), método que também será utilizado neste trabalho.

Os métodos baseados em dicionários têm por base a escolha de um conjunto de vetores elementares, chamados de átomos, de maneira que o conjunto desses vetores forme uma transformação linear do conjunto de variáveis originais para um novo conjunto de variáveis, este denominado dicionário (RUBINSTEIN; BRUCKSTEIN; ELAD, 2010; SORZANO; VARGAS; MONTANO, 2014). Além da interpretação algébrica, da determinação de uma mudança de bases, buscando uma base que gere um subespaço capaz de representar os registros de uma base de dados com mínimos erros de reconstrução, temos também a noção geral de que átomos são considerados como partes elementares dos dados, através dos quais podemos compor qualquer registro de uma base. Isto é explicitamente ressaltado no trabalho de (LEE; SEUNG, 1999), que propõe a Fatorização de Matrizes Não-negativas (NMF) como um método capaz de identificar partes que se combinam para formar um todo. Além destas propriedades, em contraste com métodos como a PCA e a ICA, a NMF não exige hipóteses sobre as dependências estatísticas entre as novas variáveis aprendidas, a não ser sua não negatividade (LEE; SEUNG, 1999). A NMF será discutida em maiores detalhes a seguir na Seção 3.2.3.

3.2.2 Clustering

Uma maneira de obter uma compreensão coerente de um conjunto volumoso de dados multidimensionais é buscar o agrupamento destes dados. A tarefa de agrupamento de dados, chamada de *clustering* ou clusterização, tem finalidades que podem ser descritas de várias formas. Como levantado em (XU; WUNSCH, 2005), o objetivo geral da clusterização é separar um conjunto de dados originalmente não-rotulados e dividi-lo em um conjunto finito e discreto de estruturas ou grupos homogêneos, com base em alguma medida de similaridade. (XU; WUNSCH, 2005) ressalta ainda que, em última instância, o objetivo é prover aos usuários intuições significativas sobre os dados originais, compondo assim uma análise exploratória que antecede a especificação de modelos. Logo, análises posteriores são necessárias para validar o conhecimento extraído destes agrupamentos.

Em (ROKACH, 2010), as técnicas de clusterização são divididas em diferentes categorias: métodos hierárquicos, métodos baseados em particionamento, métodos baseados

em densidade e métodos baseados em modelo. Métodos hierárquicos (e.g. *single linkage*, *complete linkage*, *divisive analysis*) constroem clusters através de partições recursivas, de maneira a dividir um cluster inicial contendo todos os registros da base de dados (top-down), ou através da aglomeração de clusters a partir de um estado inicial em que cada registro representa um cluster (bottom-up). Métodos de particionamento (e.g. k-means e k-medoids) buscam, com base numa partição inicial dos clusters, mover os dados entre eles de maneira a encontrar a melhor configuração de atribuições possível. Métodos baseados em densidade (e.g. DBSCAN, AUTOCLASS) assumem que os registros pertencentes a cada cluster são sorteados de uma determinada distribuição probabilística e buscam identificar os clusters e os parâmetros dessas distribuições. Métodos baseados em modelos (e.g. árvores de decisão e mapas auto-organizáveis) tentam otimizar as atribuições de clusters dos dados a modelos matemáticos. Uma particularidade desses tipos de métodos é que estes, além de apenas identificarem agrupamentos de registros, conseguem também encontrar descrições características de cada grupo, sendo que cada grupo representa um conceito ou classe. Na seção 3.2.3, apresentamos um método de clustering via NMF que, aqui, consideramos um método baseado em modelo, visto que ele apresenta a existência de protótipos característicos para cada um dos clusters e que as atribuições de cluster para cada um dos registros é feita em função de sua pertinência às classes representadas por estes protótipos.

A maior parte dos algoritmos de clusterização possui uma quantidade de grupos ou categorias (comumente denotada por k) dada por um parâmetro de entrada pré-determinado (XU; WUNSCH, 2005; ROKACH, 2010). Segundo relata (XU; WUNSCH, 2005), existem casos onde a quantidade k de clusters já é conhecida a partir de informações sobre a base de dados, mas casos em que o valor de k precisa ser estimado são mais comuns. Foram propostos diversos métodos heurísticos como tentativas de estimar o valor apropriado para k. Dentro da grande variedade de heurísticas possíveis, destacamos os casos do índice de Davies-Bouldin (DAVIES; BOULDIN, 1979) e do coeficiente silhouette (ROUSSEEUW, 1987).

O índice Davies-Bouldin (DAVIES; BOULDIN, 1979) combina as ideias de minimização da dispersão interna de cada cluster e maximização da separação entre os clusters. A formulação do índice Davies-Bouldin para um conjunto de k clusters já determinados conta com os seguintes componentes: clusters denotados por C_i, quantidade de registros contidos em cada um dos cluster T_i, centróides de cada cluster $\mu_{\mathbf{i}}$ e $\mathbf{x_{i,j}}$ que representa o j-ésimo registro pertencente ao cluster C_i. É calculado

$$S_i = \left(\frac{1}{T_i} \sum_{i=1}^{T_i} \|\mathbf{x_{i,j}} - \mu_{\mathbf{i}}\|_p^q \right)^{\frac{1}{q}} \tag{3.1}$$

onde S_i representa, para $q = 1$, a medida de distância média de todos os registros contidos no cluster C_i ao centróide $\mu_{\mathbf{i}}$. Para o valor usual de $p = 2$ temos a utilização da distância

euclidiana. Além disso são calculados

$$M_{i,j} = \|\mu_{\mathbf{i}} - \mu_{\mathbf{j}}\|_p = \Big(\sum_{q=1}^{K} |\mathbf{x_{q,i}} - \mathbf{x_{q,j}}|^p\Big)^{\frac{1}{p}} \tag{3.2}$$

onde cada $M_{i,j}$ representa a medida de separação entre os clusters C_i e C_j. Com estes valores à mão, a medida $R_{i,j} = \frac{S_i+S_j}{M_{i,j}}$ expressa uma qualidade dos clusters C_i e C_j (valores menores de $M_{i,j}$ denotam uma configuração de clusters mais separados e compactos). Para cada cluster, o pior caso de interação entre clusters é denotado por $D_i = \max_{j,j\neq i} R_{i,j}$. Finalmente, o índice Davies-Bouldin é calculado por

$$IDB = \frac{1}{N}\sum_{i=1}^{N} D_i. \tag{3.3}$$

Já o coeficiente silhouette (ROUSSEEUW, 1987) busca uma consistência interna de cada cluster através da avaliação, para cada registro, de sua similaridade em relação ao cluster a que pertence (coesão ou compacticidade) e de sua diferença em relação aos outros clusters (separação). A ideia é encontrar uma medida que indique se um objeto pertence claramente a um cluster ou se está meramente numa região entre clusters. Idealmente, temos objetos que estão bastante próximos a todos os outros objetos do mesmo cluster (indicando clusters compactos) e que sua mínima distância a objetos de qualquer outro cluster seja grande (indicando clusters bastante separados).

Seja $d(m, n)$ a medida de dissimilaridade entre dois registros m e n quaisquer do conjunto de dados. Denotamos por $a(m)$ a dissimilaridade média de um registro qualquer $\mathbf{x_m}$, pertencente ao cluster C_i, em relação a todos os outros registros no mesmo cluster, calculada por

$$a(m) = \frac{1}{T_i - 1}\sum_{n\in C_i, m\neq n} d(m, n), \tag{3.4}$$

onde T_i indica a quantidade de registros contidos no cluster C_i.

Denotamos por $b(m)$ a dissimilaridade média de um registro qualquer $\mathbf{x_m}$, pertencente ao cluster C_i, em relação ao cluster C_j com o qual possua a menor dissimilaridade média, calculada por

$$b(m) = \min_{k\neq i} \frac{1}{T_k}\sum_{n\in C_k} d(m, n). \tag{3.5}$$

Sendo assim, a medida silhouette de um registro m pertencente ao cluster C_i é dada por

$$s(m) = \frac{b(m) - a(m)}{\max(a(m), b(m))}, \tag{3.6}$$

se $T_i > 1$ e $s(m) = 0$ se $T_i = 1$. Podemos ainda reescrever a função $s(m)$ como:

$$s(m) = \begin{cases} 1 - a(m)/b(m), & \text{se } a(m) < b(m) \\ 0, & \text{se } a(m) = b(m) \\ b(m)/a(m) - 1 & \text{se } a(m) > b(m). \end{cases} \tag{3.7}$$

Desta última definição, fica explícito que $-1 \leq s(m) \leq 1$. O melhor caso é observado quando um determinado registro é pouco dissimilar em relação a todos os registros pertencentes ao mesmo cluster e bastante dissimilar a todos os registros pertencentes aos outros clusters, resultando em valores baixos para $a(m)$ e valores altos para $b(m)$ e valores de $s(m)$ próximos a $+1$. Por outro lado, para os piores casos, $s(m)$ resulta em valores próximos a -1. Quando $s(m)$ é próximo de zero, significa que o dado está próximo à fronteira entre dois agrupamentos de dados. Finalmente, podemos calcular o valor médio do coeficiente silhouette como uma medida da qualidade da configuração de agrupamentos.

3.2.3 Fatoração de matrizes não-negativas (NMF)

Dada a contextualização dos métodos de redução de dimensionalidade e de clustering acima, aqui apresentamos a Fatoração de Matrizes Não-negativas (NMF) como um método de redução de dimensionalidade com capacidade de realizar clustering. Os trabalhos que deram início ao tópico da Fatoração de Matrizes Não-negativas como tópico de pesquisa foram (PAATERO; TAPPER, 1994), que introduziu a NMF através do conceito de Fatorização de Matrizes Positivas; e o trabalho de (LEE; SEUNG, 1999) que popularizou a NMF nos campos de aprendizado de máquina e mineração de dados (WANG; ZHANG, 2013; LI; al, 2014). Em (LEE; SEUNG, 1999), a NMF foi proposta como um algoritmo capaz de representar um conjunto de dados a partir da combinação de partes de um todo, fazendo distinção a uma análise baseada na busca de fatores de explicação holística dos dados.

Seja uma base de dados representada por uma matriz $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}]$, de dimensão $m \times n$ formada por n vetores-coluna $\mathbf{x}$ de m dimensões com apenas elementos não-negativos. O objetivo da NMF é decompor $\mathbf{X}$ em termos de uma matriz base $\mathbf{U} = [\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k}]$ de dimensão $m \times k$ e uma matriz de coeficientes $\mathbf{V} = [\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}]$ de dimensão $k \times n$, tal que $\mathbf{X} \approx \mathbf{UV}$. Equivalentemente, podemos entender esta reconstrução através fórmula vetorial $\mathbf{x_i} \approx \hat{\mathbf{x}}_\mathbf{i} = \mathbf{U}\mathbf{v_i}$. Será útil a interpretação de que a matriz $\mathbf{U}$ compõe um conjunto de vetores protótipos e que a reconstrução se dá pela combinação linear destes vetores através de um conjunto de coeficientes dados pelos elementos de $\mathbf{v_i}$.

O problema de obter as matrizes $\mathbf{U}$ e $\mathbf{V}$ é resolvido pela otimização de uma função objetivo que minimize a diferença entre a matriz aproximada $\hat{\mathbf{X}} = \mathbf{UV}$ e a matriz

original $\mathbf{X}$. A função objetivo mais comum é a de soma dos erros quadráticos, ou ainda o quadrado da distância euclidiana, representado pela formulação

$$\min_{\mathbf{U},\mathbf{V}\geq 0} J_{SEQ} = \|\mathbf{X} - \mathbf{UV}\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} (\mathbf{X}_{i,j} - [\mathbf{UV}]_{i,j})^2, \tag{3.8}$$

onde a norma $\|\cdot\|_F$ representa a norma Frobenius.

Outra possível função custo é a divergência de Kullback-Leibler generalizada, representada pela formulação

$$\min_{\mathbf{U},\mathbf{V}} J_{KL} = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{X}_{i,j} \ln \frac{\mathbf{X}_{i,j}}{[\mathbf{UV}]_{i,j}} - \mathbf{X}_{i,j} + [\mathbf{UV}]_{i,j}. \tag{3.9}$$

No contexto de decomposição por dicionários, os k vetores-coluna de $\mathbf{U}$ podem ser chamados de tópicos ou átomos do dicionário, com as n colunas de $\mathbf{V}$ representando os pesos de cada tópico na reconstrução de cada um dos n registros originais.

(WANG; ZHANG, 2013) avalia que existem duas consequências importantes que emergem a partir da não-negatividade das matrizes. A primeira é que, tanto para as observações originais em $\mathbf{X}$ quanto para os valores das matrizes $\mathbf{V}$ (coeficientes ou pesos) e $\mathbf{U}$ (protótipos átomos do dicionário) encontrados pela fatoração, é frequente que os valores sejam significativos apenas para valores não-negativos, como casos de dados físicos, químicos, genéticos ou imagens. A segunda consequência é que, devido à não-negatividade dos valores, a fatoração resulta em combinações puramente aditivas. Isto acaba promovendo naturalmente uma esparsidade (LEE; SEUNG, 1999), fazendo com que os protótipos sejam combinados como partes a serem incluídas ou não (valores nulos ou positivos) na recomposição dos registros.

Como mencionado anteriormente, a reconstrução aproximada dos vetores $\mathbf{x_j}$ se dá por $\hat{\mathbf{x}}_\mathbf{j} \approx \mathbf{x_j} = \mathbf{U}\mathbf{v_j}$. Uma maneira de entender as propriedades resultantes disso é pela perspectiva de que a matrizes $\mathbf{U}$ e $\mathbf{V}$, resultantes da aproximação de $\mathbf{X} \approx \mathbf{UV}$ formam uma matriz de mudança de base $\mathbf{U} = [\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k}]$ e uma matriz de coordenadas $\mathbf{V} = [\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}]$. Quando temos casos onde os valores de $\mathbf{v_i}$ (ou coeficientes de $\mathbf{v_i}$) são próximos de zero para todas as coordenadas exceto por uma determinada dimensão i, conseguimos identificar que aquele vetor protótipo u_i conseguiu capturar uma determinada característica comum a um grupo de registros de $\mathbf{X}$, já que é o protótipo que mais determina estes determinados registros.

Exemplos sintetizadores dessas propriedades são apresentados no trabalho seminal de (LEE; SEUNG, 1999). No trabalho, são ilustrados os protótipos encontrados para um caso de fatoração de uma base de dados de imagens de rostos e para outro caso de uma base de textos representados em codificação de frequências das palavras (*bag-of-words*). No caso da fatoração da base de imagens, é aparente a propriedade de composição por partes através da esparsidade. Os protótipos formam imagens que se assemelham a

partes do rosto, como olhos, bocas, cabelos etc. Na decomposição a partir de métodos como a PCA, os protótipos encontrados apresentam regiões tanto de valores negativos quanto positivos nas imagens, fazendo com que os protótipos fossem majoritariamente preenchidos de valores não-nulos e também que fosse necessária a utilização de maior parte dos protótipos (isto é, que maior parte dos coeficientes apresentassem valores não-nulos) para reconstruir qualquer um dos registros. No caso da fatoração da base de textos, fica claro que não existe a possibilidade de interpretar frequências negativas de palavras. A análise das coordenadas ou coeficientes de maior valor em cada protótipo, isto é, das palavras consideradas de maior importância para cada protótipo, revela conjuntos de palavras que puderam ser interpretadas, de acordo com conhecimento sobre os textos originais, como temas distintos uns dos outros, coesos internamente e coerentes com os conteúdos dos textos.

Como relatado nos trabalhos de (WANG; ZHANG, 2013) e (LI; al, 2014), a NMF possui a capacidade de realizar clusterização. Embora originalmente desenvolvida como estratégia para redução de dimensionalidade, o trabalho de (LEE; SEUNG, 1999) já apresentava propriedades de agrupamento, como no exemplo de fatoração da base de textos. No caso, foi observado o agrupamento de palavras relacionadas semanticamente em um mesmo protótipo. Por exemplo, palavras como “flowers”, “leaves” e “plant” apareciam com valores altos em um dos protótipos enquanto “disease”, “behaviour” e “glands” apareciam em outro protótipo, dando a impressão de que estes formavam grupos relacionados às temática de plantas e medicina, respectivamente.

Como exemplificado em (LI; al, 2014), a utilização dos valores da matriz $\mathbf{U}$ como coeficientes de atribuição fracional dos registros de um conjunto de dados a cada um dos clusters, tendo o papel de uma espécie de índice de pertinência aos clusters, resulta num processo de *soft-clustering*. Neste caso, da mesma maneira que (LEE; SEUNG, 1999) já apresentavam as propriedades de agrupamento, através da análise dos vetores-protótipo (vetores-coluna u_k da matriz $\mathbf{U}$), cada um dos k valores dos vetores-coluna $\mathbf{v_n}$ podem ser entendidos como coeficientes que indicam o pertencimento de um determinado vetor $\mathbf{x_n}$ a um dos k possíveis clusters.

Hard-clustering, isto é, a atribuição de cluster de maneira discreta e única a um cluster para cada registro, pode ser obtida implementando uma regra de atribuição, como $c_n = \arg\max \mathbf{v_n}$, onde $\mathbf{v_n}$ é um vetor-coluna dos coeficientes que representam um registro $\mathbf{x_n}$ em termos das combinações dos coeficientes aos protótipos $\mathbf{U}$ e c_n é o cluster ao qual o registro $\mathbf{x_n}$ é atribuído.

Além de a NMF possuir suas próprias capacidades de clustering, existem diversas variações da formulação do problema de NMF, que buscam aproximar os resultados deste processo de clusterização para os de outros métodos, como os de k-means (DING; HE; SIMON, 2005; LI; DING, 2006) (baseado na restrição de ortogonalidade de $\mathbf{V}$) e os da

análise probabilística de semântica latente (GAUSSIER; GOUTTE, 2005; DING; LI; PENG, 2006) (baseado na utilização da divergência generalizada de Kullback-Leibler como medida de dissimilaridade entre $\mathbf{X}$ e $\mathbf{UV}$).

3.3 Modelos de Aprendizado Supervisionado

O problema de aprendizado supervisionado pode, em geral, ser entendido como um problema de estimação ou predição. Podemos compreender os dados como sendo formados por um conjunto de observações que, por sua vez, são definidos por um conjunto de características ou atributos, e uma variável alvo que pode ser um rótulo ou um valor a ser estimado ou previsto (um valor de referência). O procedimento do aprendizado supervisionado consiste em, a partir dessas observações $\mathbf{x}$ e seus valores ou vetores alvo associados $\mathbf{y}$, aprender, ou seja, ajustar seus parâmnetros de maneira que a estimativa para $\mathbf{y}$ por ele fornecida a partir de $\mathbf{x}$, seja a mais acurada possível (GOODFELLOW; BENGIO; COURVILLE, 2016).

Nesta seção, descrevemos alguns modelos de aprendizado de máquina que serão utilizados na análise e estimação das probabilidades de risco de substituição de ocupações.

3.3.1 Modelos de Regressão

O objetivo dos modelos de regressão é predizer o valor de um ou mais valores contínuos de referência $\mathbf{y}$, dados os valores de um vetor de entrada $\mathbf{x}$ de d dimensões (BISHOP, 2006). A relação entre as referências $\mathbf{y}$ e os dados de entrada $\mathbf{x}$ é dada por uma função desconhecida, que buscaremos identificar utilizando um conjunto de dados de treinamento. A regressão busca a aproximação de uma função $\mathbf{y} = f(\mathbf{x}) + \epsilon$, onde ϵ é um ruído aleatório (ALPAYDIN, 2020). Se a relação buscada fosse diretamente $\mathbf{y} = f(\mathbf{x})$, a tarefa seria uma interpolação (ALPAYDIN, 2020). Podemos ainda dizer que existem fatores não observáveis z que afetam os resultados, dentre os quais o próprio ruído ϵ faz parte. Desta maneira, podemos ainda entender que o processo original que buscamos descobrir é da forma $\mathbf{y} = f^*(\mathbf{x}, \mathbf{z})$. Finalmente, buscamos um modelo aproximado g, tal que este se aproxime do processo descrito por f, das quais observamos apenas $\mathbf{x}$ e $\mathbf{y}$. Este procedimento de otimização em função da proximidade da saída do modelo g às observações $\mathbf{y}$ é descrito pela equação

$$g(\mathbf{x}) = \arg\min_{g} \frac{1}{n} \sum_{i=1}^{n} [\mathbf{y} - g(\mathbf{x_i})]^2, \tag{3.10}$$

onde $\mathbf{x_i}$ são cada um dos n vetores da base de dados $\mathbf{x}$. Neste caso, dependendo da estrutura de g, a minimização do erro empírico da aproximação pode ser encontrada de maneira analítica ou algorítmica.

Aqui, entendemos a regressão como um dos exemplos mais simples e ilustrativos do aprendizado supervisionado. Nas próximas seções descrevemos outras técnicas de aprendizado supervisionado, cada uma com suas especificidades em relação aos seus contextos de aplicação e suas formas de implementação matemática e computacional.

3.3.2 Classificação Binária e Regressão Logística

Outro exemplo importante de problema supervisionado é a classificação binária. Algumas das especificidades desse problema são o seu formato de saídas, as formas de aproximação dos resultados, as funções objetivo e os critérios de avaliação de desempenho do classificador.

Comparando com a formulação do modelo de regressão descrito anteriormente, podemos entender o problema da classificação como um problema que também busca mapear $\mathbf{y}$ (comumente chamadas de rótulo, no contexto de classificação) como uma função de vetores de entrada $\mathbf{x}$. A primeira diferença entre ambos os modelos vem do fato de que estes rótulos $\mathbf{y}$ são símbolos discretos e, tipicamente, não-ordinais. No caso da classificação binária, é comum codificar os dois possíveis valores das classes C_1, C_2 de $\mathbf{y}$ como $\{-1, 1\}$ (HART; STORK; DUDA, 2000) ou $\{0, 1\}$ (BISHOP, 2006; ALPAYDIN, 2020).

Sendo f o mapeamento original desconhecido $\mathbf{y} = f(\mathbf{x})$, buscamos um classificador g que se aproxime de f. Uma maneira de realizar este mapeamento é partir do desenvolvimento de funções discriminantes (HART; STORK; DUDA, 2000; BISHOP, 2006; ALPAYDIN, 2020). Uma função discriminante $\theta(\mathbf{x})$ é uma função construída de maneira que a classificação é realizada de acordo com a regra

$$g(\mathbf{x}) = \begin{cases} C_1, \text{ se } \theta(\mathbf{x}) < 0, \\ C_2, \text{ se } \theta(\mathbf{x}) > 0 \end{cases} . \tag{3.11}$$

A superfície $\theta(\mathbf{x}) = 0$ representa a fronteira de decisão (Bishop 2006; Alpaydin 2014). Neste caso, se necessário, alguma das classes pode ser atribuída à fronteira de maneira arbitrária. Se θ é da forma $\theta(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \mathbf{w_0}$, onde $\mathbf{w}$ é um vetor de parâmetros de mesma dimensão que $\mathbf{x}$ e $\mathbf{w_0}$ é um escalar independente, θ é considerado um discriminante linear (HART; STORK; DUDA, 2000; BISHOP, 2006; ALPAYDIN, 2020). Podemos ainda utilizar a notação em que $\mathbf{w_0}$ é adicionado como primeiro elemento do vetor $\mathbf{w}$ e $\mathbf{x}$ recebe um valor adicional igual a um em seu primeiro elemento ($\mathbf{x} = [1, x_1, x_2, \ldots, x_m]$), resultando na notação $\theta(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$. Se a base de dados inteira pode ser classificada exatamente por fronteiras de decisão lineares (formadas por discriminantes lineares), a base é considerada linearmente separável (BISHOP, 2006).

A otimização do classificador construído dessa maneira pode ser feita pela minimização da soma dos erros quadráticos, ou seja, a soma do quadrado das diferenças

entre o valor desejado $\mathbf{y}$ e o valor estimado pelo modelo $g(\mathbf{x})$, mas tipicamente as estimativas resultantes desta forma de aproximação apresentam desempenho bastante limitado, devido a problemas como a falta da capacidade do modelo linear limitar seus resultados a uma faixa de valores, como $(0, 1)$, e pela falta de robustez das soluções baseadas em mínimos quadrados a outliers (BISHOP, 2006). Um método capaz de superar essas limitações é a regressão logística.

A regressão logística (BISHOP, 2006) é o nome dado aos modelos classificação na forma

$$g(x) = \sigma(\mathbf{w}^T\mathbf{x}) = p(C_1|\mathbf{x}), \tag{3.12}$$

onde $\sigma(\cdot)$ é a função logística (ou função sigmóide) definida por

$$\sigma(\theta) = \frac{1}{1 + \exp(-\theta)}. \tag{3.13}$$

Enfatiza-se que, apesar de seu nome, este modelo é um modelo de classificação e não um de regressão. Temos descrita a estrutura de um modelo que limita os valores de saída entre $(0, 1)$, e que pode ser interpretada como uma chance de acerto de classificação — a probabilidade de que um determinado vetor de entrada $\mathbf{x}$ pertença à classe C_1. No caso da classificação binária, o resultado para a classe C_2 é dado por $p(C_2|\mathbf{x}) = 1 - p(C_1|\mathbf{x})$.

Para derivar os parâmetros do modelo de regressão logística $g(\mathbf{x}) = \sigma(\theta(\mathbf{x})) = \sigma(\mathbf{w}^T\mathbf{x})$, é utilizada a máxima verossimilhança. A estimação de máxima verossimilhança é o processo de encontrar o parâmetro $\mathbf{w}$ tal que este maximize

$$\mathbf{w_{MV}} = \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w}), \tag{3.14}$$

onde $\mathbf{y} = (y_1, y_2, \ldots, y_N)$.

Para os pares de dados (g_n, y_n), onde $g_n = p(y_n = C_1|\theta_n)$, com $n \in [1, N]$, a função de verossimilhança pode ser escrita (BISHOP, 2006) na forma

$$p(\mathbf{y}|\mathbf{w}) = \prod_{n=1}^{N} g_n^{y_n}(1 - g_n)^{1-y_n}. \tag{3.15}$$

Utiliza-se, ainda, o negativo do logaritmo da verossimilhança, que resulta na função de erro de entropia cruzada (BISHOP, 2006) na forma

$$EC(\mathbf{w}) = -\ln p(\mathbf{y}|\mathbf{w}) \tag{3.16}$$

$$= -\ln\left[\prod_{n=1}^{N} g_n^{y_n}(1 - g_n)^{1-y_n}\right] \tag{3.17}$$

$$= -\sum_{i=1}^{n}\Big[y_i \ln g_i + (1 - g_i)\ln(1 - g_i)\Big]. \tag{3.18}$$

O processo de otimização da regressão logística é dado através de métodos que computam o gradiente da entropia cruzada, buscando minimizá-la.

Nesta seção, foi discutido o problema geral de classificação binária e apresentada a solução da regressão logística, como uma maneira eficaz de realizar a tarefa da classificação. Na próxima seção, discutimos as redes neurais, que são classes de modelos capazes de mapear relações de dados mais complexas do que a regressão logística e que também são amplamente utilizadas para outras tarefas além da classificação.

3.3.3 Redes Neurais

Redes neurais compõem uma classe de modelos amplamente utilizados em aplicações de *big data* nos últimos anos. Um primeiro fator que explica essa popularidade é que as redes neurais possuem uma ampla flexibilidade em realizar tarefas de diversas naturezas — podem, de fato, realizar tanto tarefas de aprendizado supervisionado (regressão e classificação) quanto não supervisionado (autoencoders), por exemplo. Outro fator é que as redes neurais possuem, em conjunto com tecnologias comercialmente viabilizadas recentemente como as placas de processamento gráfico, capacidade de lidar com conjuntos de dados de grande volume e dimensionalidade.

Como levantado por (GOODFELLOW; BENGIO; COURVILLE, 2016), é possível entender o desenvolvimento das pesquisas em redes neurais artificiais em três ondas. A primeira onda da cibernética se passou durante os anos 1940 a 1960, influenciado pelos desenvolvimento de trabalhos da neurociência e sobre aprendizado biológico (MCCULLOCH; PITTS, 1943; HEBB, 2005) e contando com a implementação dos primeiros modelos de redes neurais como o perceptron (ROSENBLATT, 1958).

A segunda onda conexionista se passou durante o período de 1980 a 1995, e contou com o desenvolvimento da retropropagação (RUMELHART; HINTON; WILLIAMS, 1986), que possibilitou o treinamento de uma rede neural mais de uma camada oculta (perceptron multicamadas). Uma das ideias centrais do conexionismo é a de que a combinação de uma grande quantidade de unidades simples computacionais, como o perceptron, poderia alcançar comportamentos inteligentes quando conectadas em conjunto (GOODFELLOW; BENGIO; COURVILLE, 2016).

A terceira e atual onda do aprendizado profundo (no inglês, *deep learning*) começou em torno de 2006, marcado inicialmente pelos trabalhos de (HINTON; OSINDERO; TEH, 2006; BENGIO et al., 2007; RANZATO et al., 2007), que foram desenvolvimentos importantes para viabilizar o treino mais eficiente de redes com mais camadas, ou mais profundas.

As redes do tipo perceptron multicamadas são redes que podem ser descritas algebricamente como um mapeamento $g(\mathbf{x})$ que é resultante da composição de uma série de combinações lineares intercaladas de uma função chamada de função de ativação (tipicamente não-linear). Para uma determinada quantidade de camadas L, podemos

representar a rede neural do tipo perceptron multicamadas como

$$\begin{cases} l_0(\mathbf{x}) & = \mathbf{x} \\ l_i(\mathbf{x}) & = \sigma(\mathbf{W_i}^T l_{i-1} + \mathbf{b_i}) \\ g(\mathbf{x}) & = \sigma(\mathbf{W_L}^T l_L(\mathbf{x})) \end{cases} \tag{3.19}$$

onde cada camada $l_i(\mathbf{x})$ resulta numa dimensão (quantidade de neurônios) dependendo das dimensões da matriz de pesos $\mathbf{W_i}$ e vetor de *biases* $\mathbf{b_i}$. A função σ para o caso de argumento multi-dimensional é a generalização da função logística para múltiplas dimensões, também chamada de função *softmax* (GOODFELLOW; BENGIO; COURVILLE, 2016).

De uma maneira mais generalizada, redes neurais de alimentação direta podem ter diferentes funções de ativação tanto nas camadas internas quanto na última camada. Desta forma temos a forma mais generalizada

$$\begin{cases} l_0(\mathbf{x}) & = \mathbf{x} \\ l_i(\mathbf{x}) & = \phi_i(\mathbf{W_i}^T l_{i-1} + \mathbf{b_i}) \\ g(\mathbf{x}) & = \phi_L(\mathbf{W_L}^T l_L(\mathbf{x})). \end{cases} \tag{3.20}$$

onde cada função de ativação das camadas internas ϕ_i podem ser definidas separadamente (tipicamente a tangente hiperbólica ou a unidade linear retificada) (GOODFELLOW; BENGIO; COURVILLE, 2016) e a função de ativação da última camada ϕ_L é definida em função da tarefa realizada (comumente a função logística para classificação ou a função identidade para regressão) (BISHOP, 2006).

Similarmente à regressão logística, as redes neurais são tipicamente otimizadas através de métodos que computam o gradiente da função de custo (por exemplo a entropia cruzada, no caso de classificação) para atualizar os parâmetros do modelo. No caso das redes neurais, o processo é relativamente complicado pela existência de múltiplas camadas, algo com que se lida através da utilização da técnica de retropropagação (BISHOP, 2006; GOODFELLOW; BENGIO; COURVILLE, 2016).

4 Dados, Estratégia Empírica e Pré-Processamento

Neste capítulo, apresentamos os dados e as etapas da estratégia de processamento e análise dos dados. Conforme ilustra a Figura 4.1, a estratégia consiste numa etapa de processamento dos dados do Quadro Brasileiro de Qualificações (QBQ) que resultará numa seleção de configuração final de clusters, uma etapa de treinamento de um modelo para extrapolar os dados de probabilidade de automação das ocupações publicados por Frey e Osborne e uma etapa final de análise de padrões, que será realizada pela combinação destes dois resultados intermediários com os dados da Classificação Brasileira de Ocupações (CBO).

Primeiramente, na Seção 4.1, apresentamos as bases de dados utilizadas, suas estruturas e características. A etapa de processamento de dados do QBQ que resultará nos clusters será dividida em duas seções: a Seção 4.2 apresenta o processo de escolha do parâmetro da quantidade de clusters através do uso de métricas de desempenho para clusters e; a Seção 4.3 apresenta um processo alternativo para a mesma escolha do mesmo parâmetro, através da construção de resumos de clusters. A etapa de treinamento do modelo que extrapolará as probabilidades de automação para as ocupações brasileiras é descrita na Seção 4.4. A última etapa, da combinação dos resultados e análise de padrões é o objeto de discussão do Capítulo 5.

4.1 Apresentação das Bases de Dados

A estratégia empírica é baseada principalmente em três bases de dados: a Classificação Brasileira de Ocupações de 2002 (CBO, 2002a), o Quadro Brasileiro de Qualificações (QBQ) (QBQ, 2020) e as probabilidades (risco) de substituição de ocupações estimadas por Frey e Osborne (FREY; OSBORNE, 2017), descritas a seguir.

4.1.1 Classificação Brasileira de Ocupações

A Classificação Brasileira de Ocupações (CBO), desenvolvida pelo então existente Ministério do Trabalho e Emprego, é um documento que tem a função de normatizar o reconhecimento, a nomeação e a codificação de títulos e conteúdos das ocupações dos mercado de trabalho brasileiro (CBO, 2002a). A CBO teve suas primeiras estruturas básicas elaboradas em 1977, e marca um esforço para a criação de uma codificação unificada das classificações de ocupação utilizadas pelos órgãos públicos brasileiros. Para estruturar

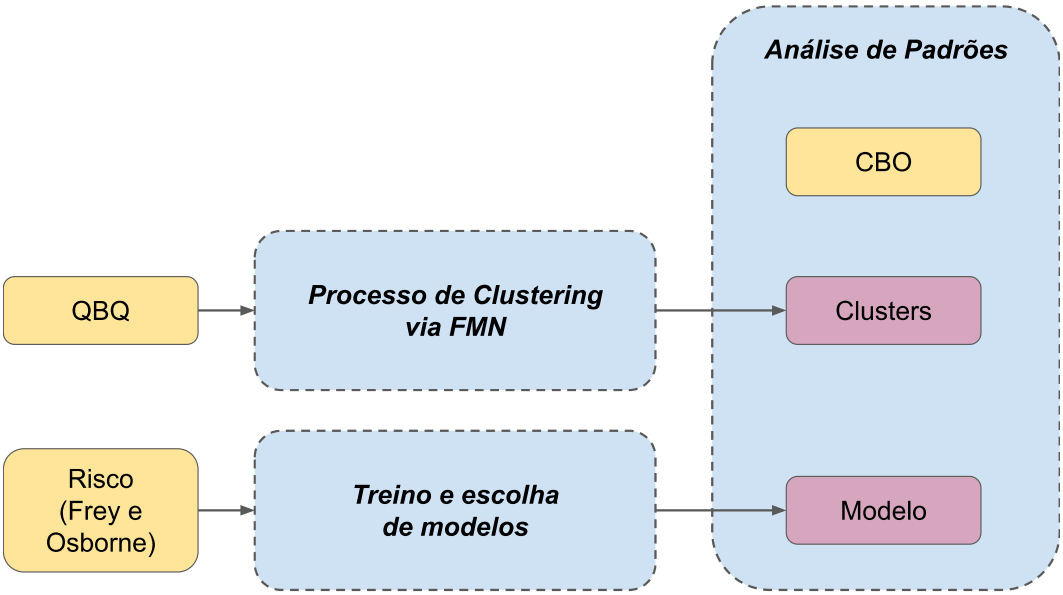

Figura 4.1 – Ilustração em diagrama da estratégia de preparo dos dados e análise em etapas. Bases de dados originais em amarelo, procedimentos em azul e dados de resultados em vermelho.

conceitualmente sua metodologia de descrição e classificação, a CBO estabelece, para seu próprio contexto, as definições para os conceitos de ocupação e de emprego:

> "Ocupação é um conceito sintético não natural, artificialmente construído pelos analistas ocupacionais. O que existe no mundo concreto são as atividades exercidas pelo cidadão em um emprego ou outro tipo de relação de trabalho (autônomo, por exemplo)" (CBO, 2002a).

> "Ocupação é a agregação de empregos ou situações de trabalho similares quanto às atividades realizadas" (CBO, 2002a).

> "Emprego ou situação de trabalho: definido como um conjunto de atividades desempenhadas por uma pessoa, com ou sem vínculo empregatício. Esta é a unidade estatística da CBO" (CBO, 2002a).

A CBO existe como uma base de dados que atua sobre dois tipos de funções: uma de classificação enumerativa e outra de classificação descritiva.

As informações enumerativas são as codificações dos empregos e outras situações de trabalho, incluindo essencialmente os códigos (que chamaremos de códigos CBO)

e títulos ocupacionais. Esta função enumerativa é tipicamente utilizada para registros administrativos e censos populacionais, como os realizados pela Relação Anual de Informações Sociais (Rais), Cadastro Geral de Empregados e Desempregados (CAGED) e outras pesquisas do Instituto Brasileiro de Geografia e Estatística (IBGE).

As informações descritivas são compostas por fichas de descrição que detalham as atividades realizadas no trabalho, os requisitos de formação e experiência profissionais e as condições de trabalho. Esta função descritiva é utilizada por órgãos como o Sistema Nacional de Empregos (SINE), na elaboração de currículos e na avaliação de formação profissional.

Destas duas funções da CBO, a enumerativa será a mais relevante para este trabalho, pois utilizaremos estas informações para analisar em conjunto e em comparação com os resultados de clusterização trazidos mais adiante.

As classificações da CBO são hierarquizadas em cinco níveis (Tabela 4.1): Grande grupo (GG), Subgrupo principal (SGP), Subgrupo (SG), Grupo de base ou família (GB) e Ocupações (O).

Nível	Sigla	Quantidade
Grandes Grupos	GG	10
Subgrupos Principais	SGP	47
Subgrupos	SG	192
Grupos de base ou Famílias	GB	596
Ocupações	O	2422

Tabela 4.1 – Tabela de composição hierárquica da CBO.

Na CBO, os grandes grupos representam o nível mais agregado da classificação, composto de dez conjuntos que são formados pelo agrupamento de ocupações similares. Abaixo listamos descrições de cada um dos Grandes Grupos da CBO 2002 e seus respectivos títulos segundo a própria CBO (CBO, 2002a).

- **GG 0 - Forças Armadas, Policiais e Bombeiros Militares:** Compreende as ocupações vinculadas às Forças Armadas (exército, marinha, aeronáutica e policiais e bombeiros militares).
- **GG 1 - Membros superiores do poder público, dirigentes de organizações de interesse público e de empresas e gerentes:** Compreende profissões cujas atividades consistem em definir e formular políticas de governo, leis e regulamentos, fiscalizar a aplicação dos mesmos e representar as diversas esferas de governo e atuar em seu nome.
- **GG 2 - Profissionais das ciências e das artes:** Compreende as ocupações cujas atividades principais requerem, para seu desempenho, profissionais de alto nível e

experiência em matéria de ciências físicas, biológicas, sociais e humanas. Também está incluído neste grande grupo o pessoal das artes e desporto de alto nível de competência.

- **GG 3 - Técnicos de nível médio:** Compreende as ocupações cujas atividades principais requerem, para seu desempenho, conhecimentos técnicos e experiência de uma ou várias disciplinas das ciências físicas e biológicas ou das ciências sociais e humanas (educação de nível médio).

- **GG 4 - Trabalhadores de serviços administrativos:** Compreende dois subtipos. Aqueles que realizam trabalhos burocráticos, sem contato constante com o público e aqueles que realizam trabalho administrativo de atendimento ao público. O primeiro subtipo compreende atividades principais que requerem conhecimentos e experiência para ordenar, armazenar, computar e recuperar informações. O segundo subtipo compreende atividades de fornecimento de serviços a clientes como os realizados por auxiliares de biblioteca, documentação e correios, operadores de caixa, atendentes, etc.

- **GG 5 - Trabalhadores dos serviços, vendedores do comércio em lojas e mercados:** Compreende as ocupações cujas tarefas principais requerem, para seu desempenho, os conhecimentos e a experiência necessários para as prestações de serviços às pessoas, serviços de proteção e segurança ou a venda de mercadorias em comércio e mercados. Tais atividades consistem em serviços relacionados a viagens, trabalhos domésticos, restaurantes e cuidados pessoais, proteção às pessoas e bens e a manutenção da ordem pública, venda de mercadorias em comércio e mercados.

- **GG 6 - Trabalhadores agropecuários, florestais, da caça e pesca:** Compreende as ocupações cujas atividades principais requerem, para seu desempenho, os conhecimentos e a experiência necessários para a obtenção de produtos da agricultura, da silvicultura e da pesca.

- **GG 7 - Trabalhadores da produção de bens e serviços industriais:** Compreende os trabalhadores de produção extrativa, da construção civil e da produção industrial de processos discretos, que mobilizam habilidades psicomotoras e mentais voltadas primordialmente à forma dos produtos.

- **GG 8 - Trabalhadores da produção de bens e serviços industriais:** Compreende os trabalhadores que operam processos industriais contínuos, que demandam habilidades mentais de controle de variáveis físico-químicas de processos.

- **GG 9 - Trabalhadores de manutenção e reparação:** Compreende as ocupações cujas atividades principais requerem, para seu desempenho, os conhecimentos e as

atividades necessários para reparar e manter toda a sorte de bens e equipamentos, seja para uso pessoal, de instituições, empresas e do governo.

Os GGs 7 e 8 possuem, de fato, o mesmo título, mas representam setores produtivos distintos dentro destas mesmas áreas.

Por fim, a codificação completa das ocupações da CBO é realizada através de seis dígitos numéricos, como exemplificado na Tabela 4.2.

Hierarquia	Código	Título
Grande Grupo	2	Profissionais nas ciências e artes
Subgrupo Principal	23	Profissionais da educação
Subgrupo	232	Professores do ensino médio
Família	2321	Professores do ensino médio
Ocupação	232130	Professores de física do ensino médio

Tabela 4.2 – Exemplo da codificação hierárquica da CBO para o caso de professores de física do ensino médio. Nota-se que, para este campo em particular, a descrição de "professores do ensino médio" é utilizada repetidamente nos níveis de subgrupo e família.

4.1.2 Classificação Brasileira de Ocupações

4.1.2.1 Sobre o QBQ

O Quadro Brasileiro de Qualificações (QBQ) é uma base de dados (atualmente em conclusão), desenvolvida pelo Ministério da Economia, que fornece caracterizações para as ocupações formais brasileiras em forma numérica (QBQ, 2020). O QBQ é baseado na documentação de ocupação da CBO e é inspirada em bases de dados similares como o Quadro Europeu de Qualificações (QEQ) (QEQ, 2017).

A Comissão Europeia desenvolveu o Quadro Europeu de Qualificações (QEQ) para viabilizar a comparação de sistemas de qualificações entre vários países, especialmente os de países europeus. A partir do QEQ, cada país da Comunidade Europeia desenvolveu seu próprio Quadro Nacional de Qualificações (QNQ), utilizado como referência para a classificação de todas as qualificações no âmbito do sistema educativo e formativo nacional.

Com inspiração nesses desenvolvimentos, foi criado o Quadro Brasileiro de Qualificações. Em contraste com o QEQ e o QNQ europeus, o QBQ foi desenvolvido em relação às descrições das ocupações do mercado de trabalho (CBO) em vez dos sistemas de educação e formação profissional. Além da CBO, servem de referência, informações sobre as mudanças tecnológicas ou na organização do trabalho, no âmbito da ocupação (QBQ, 2020).

Para entender a razão de ser do conteúdo do QBQ, podemos recorrer à página da base de dados:

> "O Quadro Brasileiro de Qualificações pode ser definido como um conjunto de informações — perfis e indicadores associados aos conhecimentos, às habilidades e às atitudes — que, de forma articulada, permitem delinear o preparo necessário ao desempenho do trabalhador em cada ocupação oficialmente reconhecida no mercado de trabalho brasileiro" (QBQ, 2020).

Neste trabalho, referimo-nos comumente aos valores numéricos registrados para os indicadores de competências que qualificam as ocupações.

4.1.2.2 Estrutura dos Dados do QBQ

O QBQ busca criar um sistema de descrição das qualificações profissionais de cada ocupação em termos dos conhecimentos, habilidades e atitudes relacionadas às atividades realizadas no trabalho. Cada ocupação tem um conjunto de características relevantes, cada uma quantificada de acordo com três tipos de medidas diferentes: profundidade, frequência e importância. As quantidades numéricas são representadas por números inteiros de 1 a 5, com valores maiores indicando graus maiores de profundidade, frequência ou importância. A Figura 4.2 ilustra a estrutura geral dos dados. Até o momento de escrita deste trabalho, o QBQ possuía 1000 ocupações documentadas, com um total de 1399 propriedades (1301 conhecimentos, 85 habilidades e 13 atitudes). Destas propriedades, uma pré-seleção foi realizada para remover habilidades com mesmo código e nomes/descrições distintas, para evitar possíveis inconsistências devidas à disparidade de significados. Após esta pré-seleção, temos os dados do QBQ compostos de 1221 propriedades, nas quais temos 1126 conhecimentos, 82 habilidades e 13 atitudes.

Para facilitar o uso das informações quantitativas, reorganizamos os dados do QBQ em uma única tabela agregada (Figura 4.2), onde cada linha representa uma ocupação e cada coluna representa uma propriedade.

Para facilitar as formulações matemáticas no resto deste trabalho, podemos ainda entender esta base como uma matriz X de dimensões $m \times n$

$$\mathbf{X} = \begin{bmatrix} C_{1,1} & \dots & C_{1,M_k} & H_{1,1} & \dots & H_{1,M_a} & A_{1,1} & \dots & A_{1,M_b} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ C_{N,1} & \dots & C_{N,M_k} & H_{N,1} & \dots & H_{N,M_a} & A_{N,1} & \dots & A_{N,M_b} \end{bmatrix}^T, \tag{4.1}$$

onde a base possui n ocupações, M_c conhecimentos, M_h habilidades, M_a atitudes e uma quantidade total de propriedades $M = M_c + M_h + M_a$. Cada um dos elementos desta matriz possui valor entre $[0, 5]$, devido à adição do valor zero onde as propriedades não possuíam valor atribuído originalmente.

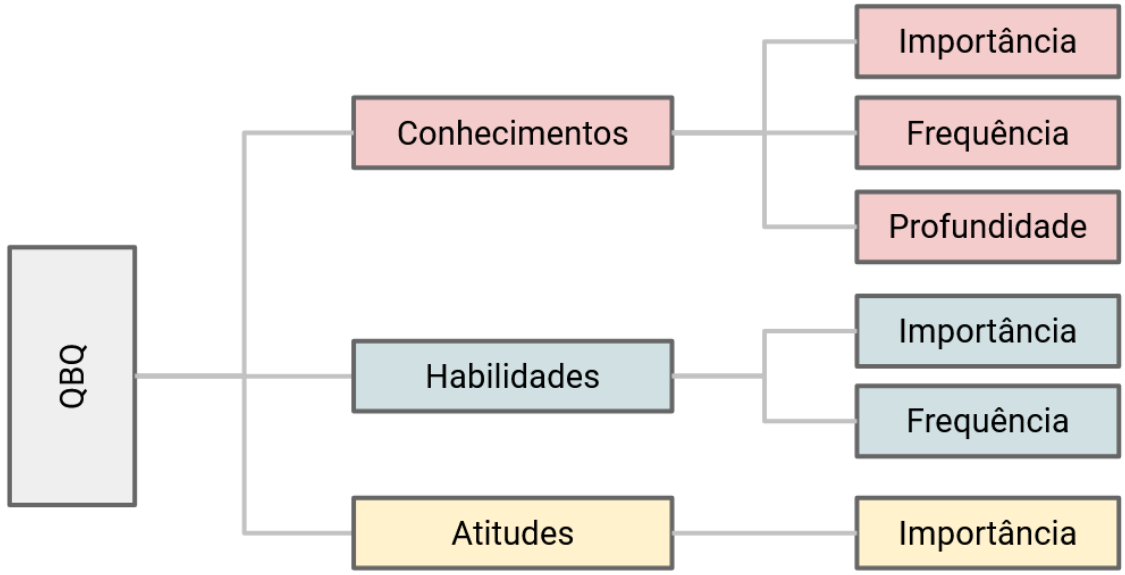

Figura 4.2 – Diagrama da estrutura dos dados do QBQ. Cada uma das três bases de dados (conhecimentos, habilidades e atitudes) possui entradas representando cada uma das 1000 ocupações, descrevendo perfis ocupacionais através da atribuição de valores inteiros de 1 a 5 para as propriedades relevantes.

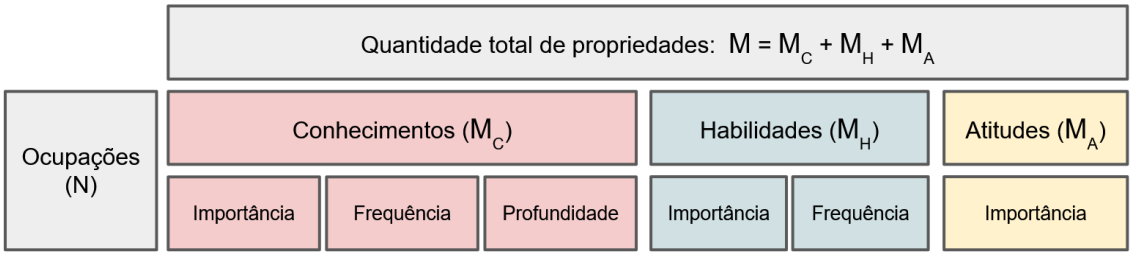

Figura 4.3 – Estrutura de dados resultante da reorganização dos dados. Cada linha corresponde a uma ocupação e possui valores para todas as propriedades existentes no QBQ. Quando uma determinada propriedade não possui nenhum valor atribuído originalmente para uma ocupação, isto é, a propriedade não é relevante à ocupação, o valor zero é atribuído a ela.

4.1.3 Dados sobre o Risco de Substituição devido à Computadorização e Automação de Processos

A base de dados que utilizaremos como referência para desenvolver estas informações é a disponibilizada por Frey e Osborne (FREY; OSBORNE, 2017), que desenvolveram um modelo de estimação dos riscos de automação devido à computadorização das ocupações e atividades nos Estados Unidos.

4.1.3.1 O Modelo de Frey e Osborne

O trabalho de Frey e Osborne (FREY; OSBORNE, 2017) fornece uma lista de estimações dos riscos de automação devidos à computadorização para as ocupações dos Estados Unidos.

O procedimento que os autores adotaram para obter essas estimativas começa pelo uso de duas bases de dados: um conjunto $\mathbf{X_o}$ de dados extraídos da O*NET e outro conjunto $\mathbf{y_{ws}}$, desenvolvido em um workshop realizado no Oxford University Engineering Sciences Department, de rótulos binários anotados manualmente que avaliam, de acordo com a opinião dos participantes do workshop, se as ocupações podem ser consideradas automatizáveis ou não.

O conjunto de dados $\mathbf{X}$ é composto de variáveis derivadas de questionários com respostas em escalas ordinais (como "importância" ou "nível") que mensuram a importância ou necessidade de habilidades, atividades e capacidades como "destreza manual" ou "percepção social", para cada ocupação no país, tendo os valores padronizados em uma escala de 0 a 100 (BAKHSHI et al., 2017; FREY; OSBORNE, 2017).

Durante o workshop realizado em Oxford, os participantes foram questionados sobre quais tipos de tarefas representariam *bottlenecks*, ou limitadores, à computadorização. Combinando os resultados dessas questões à literatura de aprendizado de máquina e robótica, foram selecionadas nove variáveis da base da O*NET, representando tarefas relacionadas às categorias de percepção e manipulação, inteligência criativa e inteligência social.

O $\mathbf{y_{ws}}$ foi construído como uma série de atribuições binárias, obtidas durante o workshop realizado em Oxford, através de respostas à questão "esta tarefa pode ser suficientemente especificada, condicionada na disponibilidade de *big data*, para ser realizada por equipamentos controlados por computadores no estado da arte?" (FREY; OSBORNE, 2017). Foram, ao todo, anotadas 70 ocupações de um total de 702 ocupações existentes na O*NET.

Na formulação de seu modelo, Frey e Osborne (FREY; OSBORNE, 2017) definem os elementos $\mathbf{x_o}$ da base de dados $\mathbf{X_{onet}}$ como vetores reais de 9 dimensões ($\mathbf{x} \in \mathbb{R}^9$), enquanto $\mathbf{y_{ws}}$ é composta por elementos de valor binário ($y \in \{0, 1\}$). O modelo $\mathbf{y_{fo}} = f(\mathbf{x})$ foi treinado para aproximar a probabilidade

$$p\Big(y_i = 1|\theta(\mathbf{x_i})\Big) = \frac{1}{1 + \exp(-\theta(\mathbf{x_i}))}. \tag{4.2}$$

Se a função latente $\theta(\mathbf{x})$ possui a forma $\theta(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, isto é, é uma combinação linear de $\mathbf{x}$ com coeficientes $\mathbf{x}$, a forma final de $f(\mathbf{x})$ configura um modelo de regressão logística. No trabalho de Frey e Osborne, $\theta(\mathbf{x})$ é modelado com um processo gaussiano (WILLIAMS; RASMUSSEN, 2006; FREY; OSBORNE, 2017).

4.1.3.2 Equivalências das ocupações

As ocupações nos Estados Unidos seguem a Classificação Ocupacional Padrão (COP) de 2010 (SOC, 2018). Atualmente não existe uma tradução oficial direta das

ocupações entre a COP 2010 e a CBO 2002. Neste trabalho, nós utilizamos a Classificação Internacional Uniforme de Ocupações (CIUO) (ISCO, 2008) para intermediar uma tradução entre elas, seguindo a sequência: CBO 2002 – CIUO 1998 – CIUO 2008 – COP 2010 (CBO, 2002b; ISCO, 2008; SOC, 2018).

4.2 Escolha da Clusterização via Métricas de Desempenho

A seguir, iniciamos a busca de uma das configurações centrais para a realização das análises de padrões: a quantidade k de clusters. A quantidade de clusters influencia diretamente a qualidade dos resultados. Se a quantidade de clusters é demasiado baixa, podem ser gerados clusters muito abrangentes e pouco claros. Caso contrário, uma quantidade muito alta de clusters pode tornar os clusters muito esparsos e específicos, levando a resultados que perdem a capacidade de redução do volume de dados e a interpretação e explicação dos grupos gerados.

Nesta seção, buscamos encontrar uma boa escolha para o valor de k a partir da avaliação das realizações da NMF em função de métricas de desempenho próprias para tarefas de clusterização. Na Seção 4.3, discutimos maneiras alternativas de determinar as configurações de clusterização mais úteis aos nossos objetivos de análise dos dados através de procedimentos que consideram mais a natureza qualitativa dos dados.

4.2.1 Métricas de Desempenho

A técnica NMF utilizada no agrupamento dos perfis ocupacionais para o caso brasileiro, a partir dos dados do QBQ e as probabilidades de Frey e Osborne, teve os seus parâmetros otimizados pelo algoritmo de Coordenadas Descendentes, implementado pela biblioteca *scikit-learn* (PEDREGOSA et al., 2011) para uma variedade de valores de k.

Avaliamos o desempenho dos agrupamentos resultantes através de duas métricas: o índice Davies-Bouldin (DAVIES; BOULDIN, 1979) (Figura 4.4) e a pontuação silhouette (ROUSSEEUW, 1987) (Figura 4.5).

A NMF possui um fator de aleatoriedade na inicialização das matrizes $\mathbf{U}$ e $\mathbf{V}$, portanto desejamos também visualizar a variabilidade dos resultados causada pela aleatoriedade da condição inicial. Para tal, o procedimento de otimização dos parâmetros da NMF foi repetido cinquenta vezes, com cada realização (configuração de k) limitada a 5000 iterações, com tolerância de $1e-4$.

Na Figura 4.4, verificamos uma tendência à melhoria do índice Davies-Bouldin (valores menores são considerados melhores) com o aumento da quantidade k. No entanto, a configuração $k = 3$ representou uma quebra desta tendência, com um desempenho relativamente melhor do que configurações próximas. Neste contexto, concluímos que $k = 3$

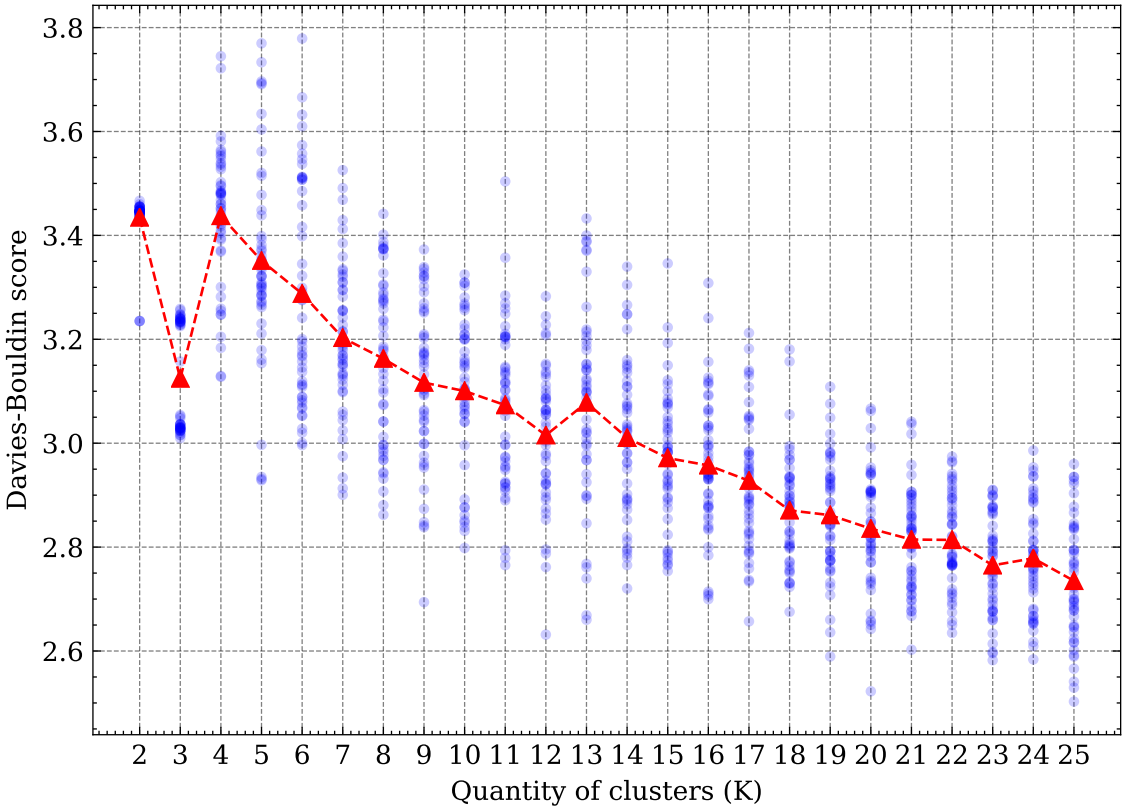

Figura 4.4 – Gráfico das distribuições do índice Davies-Bouldin para a clusterização via NMF na base de dados do QBQ. Cada execução da clusterização NMF corresponde a um ponto azul. Triângulos vermelhos correspondem ao valor médio para uma dada quantidade de cluster. Valores baixos do índice Davies-Bouldin são considerados melhores.

seja uma das melhores configurações de acordo com o índice Davies-Bouldin.

Na figura 4.5, verificamos uma desempenho relativamente melhor para os valores iniciais de $k = 2$ e $k = 3$ (valores maiores são melhores, com a pontuação máxima igual a 1), com médias de pontuação silhouette acima de $0,075$ e o restante das outras configurações resultando em pontuações médias abaixo de $0,025$.

Tendo estes resultados em consideração, uma boa escolha, de acordo com a análise quantitativa, seria a de $k = 3$. É válido notar que, ainda que estas configurações tenham apresentado as melhores pontuações nos gráficos das figuras, estas ainda são pontuações aquém do desejado, considerando que o valor ótimo do índice Davies-Bouldin é o mínimo de zero e o valor ótimo da pontuação silhouette é o máximo de 1. Particularmente, através da avaliação das pontuações de silhouette, o desempenho de uma maneira geral nos indica que os clusters apresentam bastante sobreposição, seja devido ao fato de os pontos de cada cluster estarem muito espalhados, seja pelo fato de os centroides de cada cluster estarem muito próximos uns dos outros. Apesar de isso poder estar relacionado a uma limitação na capacidade de clusterização da NMF, também se deve atentar para o fato de que a base de dados não foi construída necessariamente para ser fortemente

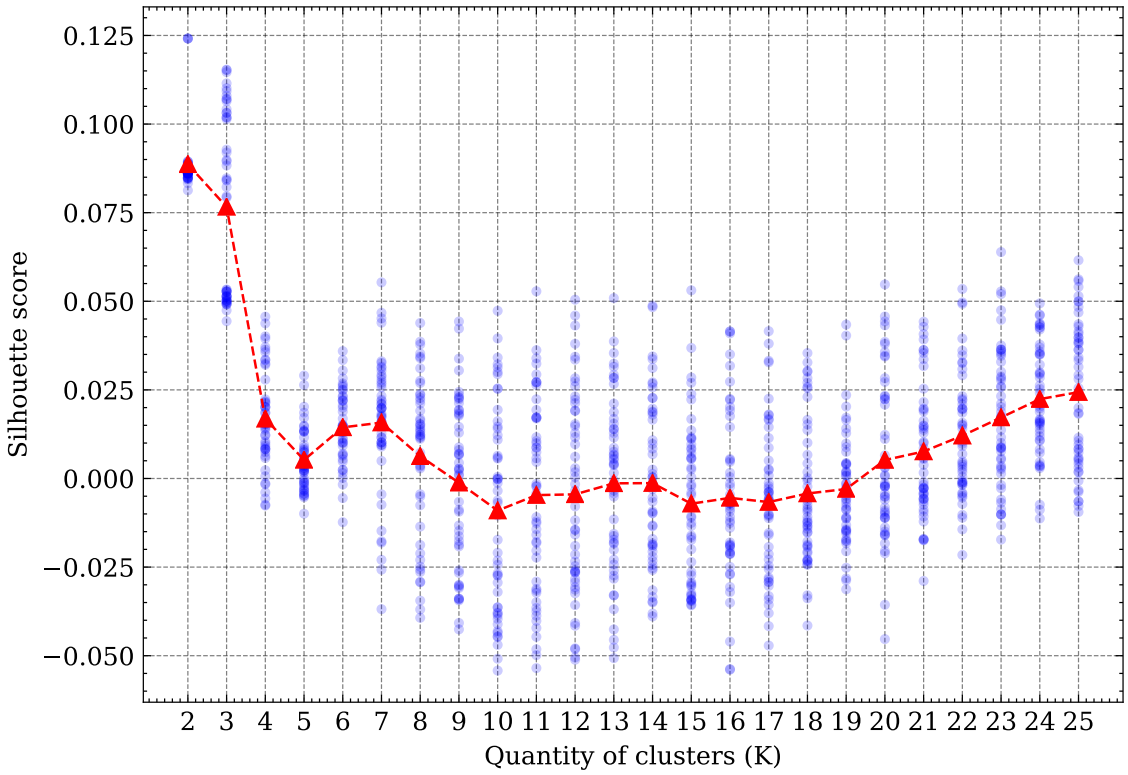

Figura 4.5 – Gráfico das distribuições da pontuação silhouette para a clusterização via NMF na base de dados do QBQ. Cada execução da clusterização NMF corresponde a um ponto azul. Triângulos vermelhos correspondem ao valor médio para uma dada quantidade de cluster. Valores altos, próximos ao valor máximo de 1, são considerados melhores.

separável por nenhuma categoria externa ou propriedades internas. Por exemplo, a maioria das ocupações possui um grande número de propriedades comuns e com valores em faixas próximas. Particularmente, 93,1% das ocupações possuem o conhecimento "Internet" com um valor de importância dentro da faixa de importância média-alta ($2 \leq x_{\text{Internet},n} \leq 5$).

4.3 Escolha da Clusterização via Resumos de Clusters

Nesta seção, desenvolvemos procedimentos para a avaliação de desempenho dos clusters de forma qualitativa. Um das principais motivações para isso é o de que a avaliação por meios quantitativos não é capaz de fornecer resultados que nos permitam escolher uma configuração de maneira completamente satisfatória: as métricas não indicam um ponto ótimo claro e também não são capazes de dar indicação inequívoca da razão pela qual escolher uma configuração sobre outra.

Para realizar uma análise qualitativa, avaliamos cada partição em termos de seus componentes mais marcantes (entre ocupações e propriedades). Em seguida, fazemos uso desta capacidade de visualizar e interpretar resumos dos clusters para avaliar a qualidade dos clusters em diferentes configurações (quantidades k de cluster), e, por fim, escolhemos

o parâmetro k.

4.3.1 Procedimento de Resumo dos Clusters

Buscamos resumir cada um dos clusters na forma de listas das propriedades e ocupações que melhor os representem. Para encontrar estas listas, buscamos construir critérios que nos indiquem quais são as ocupações e propriedades mais distintivas e características de um determinado cluster.

Essencialmente, realizaremos um procedimento de sumarização da matriz $\mathbf{U}$, que é a matriz composta dos protótipos dos clusters, o que nos retornará, para cada cluster, um conjunto de propriedades distintivas daquele cluster. Por outro lado, a sumarização da matriz $\mathbf{V}$, que é a matriz que codifica as pertinências para cada cluster, retornará os conjuntos de ocupações mais distintivos para cada cluster.

4.3.1.1 Selecionando Propriedades

Para cada realização da NMF, condicionada a um parâmetro de quantidade de clusters k, temos k vetores-coluna $\mathbf{u}_i$ de m dimensões constituindo a matriz $\mathbf{U}$ (que foi normalizada e limitada ao mesmo espaço $[0, 5]$ dos dados originais), os quais podem ser interpretados como uma representação característica das ocupações pertencentes àquele cluster. Com o objetivo de resumir cada um destes vetores $\mathbf{u}_i$, definimos uma maneira sistemática de selecionar um conjunto de propriedades que sejam as mais distintivas de cada cluster.

Através dos k clusters, eventualmente existem propriedades que ou aparecem com valores altos em poucos clusters (ou até mesmo apenas um deles) ou aparecem com valores distribuídos de uma maneira similar em vários clusters.

Idealmente, queremos encontrar os atributos que apareçam apenas em um cluster, o que indicaria que tal atributo é característico daquele cluster em específico. Para encontrar uma maneira de distinguir entre esses diferentes casos, utilizamos a função entropia discreta.

A função entropia discreta foi originalmente projetada para quantificar incerteza em distribuições probabilísticas, mas sua estrutura matemática nos permite usá-la para quantificar a concentração de valores num vetor. Aqui, em vez de olharmos para $\mathbf{U}$ como uma matriz composta de k protótipos $\mathbf{u}_{\text{col}}$ (vetores-coluna de m dimensões), consideramos $\mathbf{U}$ como uma matriz composta de m vetores-linha de k dimensões $\mathbf{u}_{\text{lin}}$ (um vetor para cada uma das m propriedades, percorrendo todos os n registros), de tal maneira que visualizaremos o quão concentrados ou dispersos estão os valores de cada propriedade ao longo dos k clusters. Em seguida, selecionamos apenas aqueles considerados concentrados

o suficiente em um cluster e consideramos cada uma dessas propriedades uma candidata à lista de propriedades características daquele cluster.

A função entropia para variáveis aleatórias discretas é definida como

$$H(\mathbf{w}) = -\sum_{i=1}^{n} \mathrm{P}(\mathbf{w} = w_i) \log \mathrm{P}(\mathbf{w} = w_i), \tag{4.3}$$

onde $\mathbf{w}$ é alguma variável aleatória discreta com resultados possíveis $\{w_1, w_2, \ldots, w_N\}$.

Em vez de utilizarmos a distribuição de probabilidades de uma variável discreta $\mathbf{w}$, usamos os valores normalizados de um dos m vetores-linha $\mathbf{u}_{\text{lin}} = [u_1, u_2, \ldots, u_k]$. Esta normalização é feita de maneira que os valores de $\mathbf{u}_{\text{lin}}$ somados resultem em 1, assim como numa distribuição de probabilidades. Para tal, atualizamos os valores de cada vetor $\mathbf{u}_{\text{lin}}$ de acordo com $\mathbf{u}^*_{\text{lin}} = \frac{1}{\sum_{i=1}^{k} u_i}\mathbf{u}_{\text{lin}}$.

Aplicando isto ao longo da matriz $\mathbf{U}$ inteira, obtemos um novo vetor de m dimensões, que apresenta os valores de entropia para cada propriedade. Chamaremos estes valores de entropia de propriedade. A medida de entropia possui um valor mínimo de zero e um valor máximo que depende do número de resultados possíveis de uma variável aleatória (mais possibilidades resultam em maiores valores, ou seja, maior incerteza). Neste caso, o valor máximo de entropia será determinado pela quantidade de clusters k. Valores maiores de entropia de propriedade significam que a propriedade possui presença similar na maioria dos clusters, enquanto valores baixos de entropia de propriedade significam que a propriedade está concentrada em poucos (ou apenas um) cluster.

Para determinar um limiar mínimo, calculamos a entropia máxima para a quantidade k de clusters e utilizamos 5% deste valor como o valor de corte (selecionamos apenas as propriedades com entropia menor que este valor).

Podemos visualizar a distribuição de entropia de propriedade de todas as propriedades da matriz $\mathbf{U}$ encontrada através da clusterização dos dados do QBQ para $k = 3$ na Figura 4.6.

Mesmo para as propriedades com baixa entropia, é desejável garantir que selecionemos apenas as propriedades que apresentem um valor mínimo, para que tenhamos uma indicação de que elas sejam minimamente determinantes para o processo de agrupamento dos clusters. Com isso em mente, analisamos a distribuição de valores em $\mathbf{U}$ e determinamos o valor mínimo de corte, de acordo com o percentil de 95% dos valores de $\mathbf{U}$. A Figura 4.7 ilustra, para o caso da configuração $k = 3$, os valores de importância de cada um dos vetores protótipos de cada cluster, indicando em linha tracejada preta o valor mínimo de corte que servirá como um dos critérios para a seleção das propriedades características de cada cluster.

Combinando ambos critérios de seleção, obtemos um ranking final de propriedades únicas e determinantes para cada cluster e formamos uma lista final das propriedades

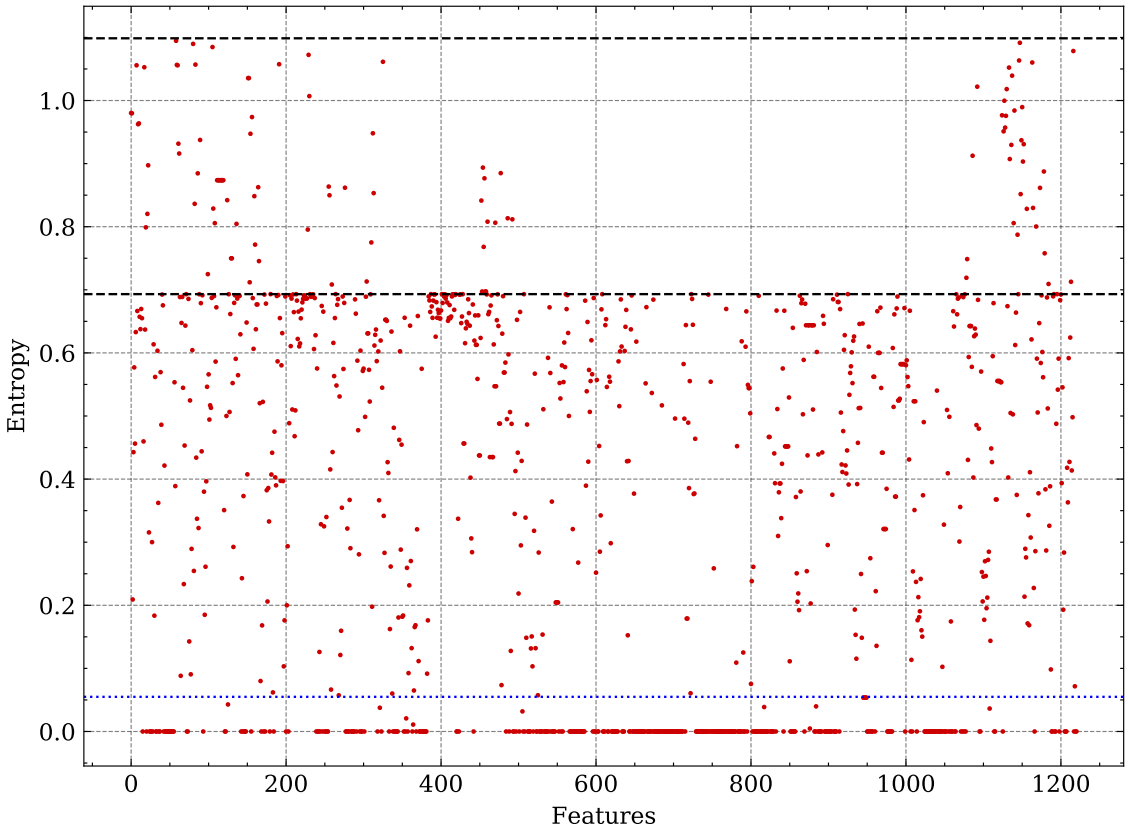

Figura 4.6 – Distribuição da entropia de propriedade em $\mathbf{U}$ para o caso de $k = 3$. A linha pontilhada azul indica o valor de corte calculado por 5% da entropia máxima. A linha tracejada preta ilustra os valores de entropia máxima teórica para $k = 2$ e $k = 3$.

com as menores entropias e que atendem ao corte de importância mínima. Em situações onde existem muitas ocupações com entropia igual a zero, o desempate é feito através da ordenação das propriedades em função de seus valores originais em $\mathbf{U}$ (escolhendo a de maior valor entre elas).

4.3.1.2 Selecionando Ocupações

A fim de encontrar uma seleção resumida para as ocupações em cada cluster, seguimos uma estratégia similar à realizada para a seleção das propriedades. Olhamos para as ocupações através da matriz $\mathbf{V}$, de pontuações de pertinência de cluster. Aqui, passamos a enxergar a matriz $\mathbf{V}$ como sendo composta de n vetores-coluna $\mathbf{v}_{\text{col}} = [v_1, v_2, \ldots, v_k]$. Neste caso, interessamo-nos em encontrar um conjunto de ocupações representativas daquele cluster. Em estratégia similar à seleção das propriedades, a função entropia discreta se torna uma medida útil para mensurar concentração nos valores dos vetores-coluna v_n, que entendemos como graus de pertinência aos clusters. Normalizamos os valores da matriz $\mathbf{V}$ de acordo com $\mathbf{v}^*_{\text{col}} = \frac{1}{\sum_{i=1}^{k} v_i}\mathbf{v_{col}}$, produzindo um vetor de entropia de ocupação $H(\mathbf{v}_{\text{col}})$ de n dimensões. A Figura 4.8 ilustra a distribuição de $H(\mathbf{v}_{\text{col}})$ para a configuração de $k = 3$

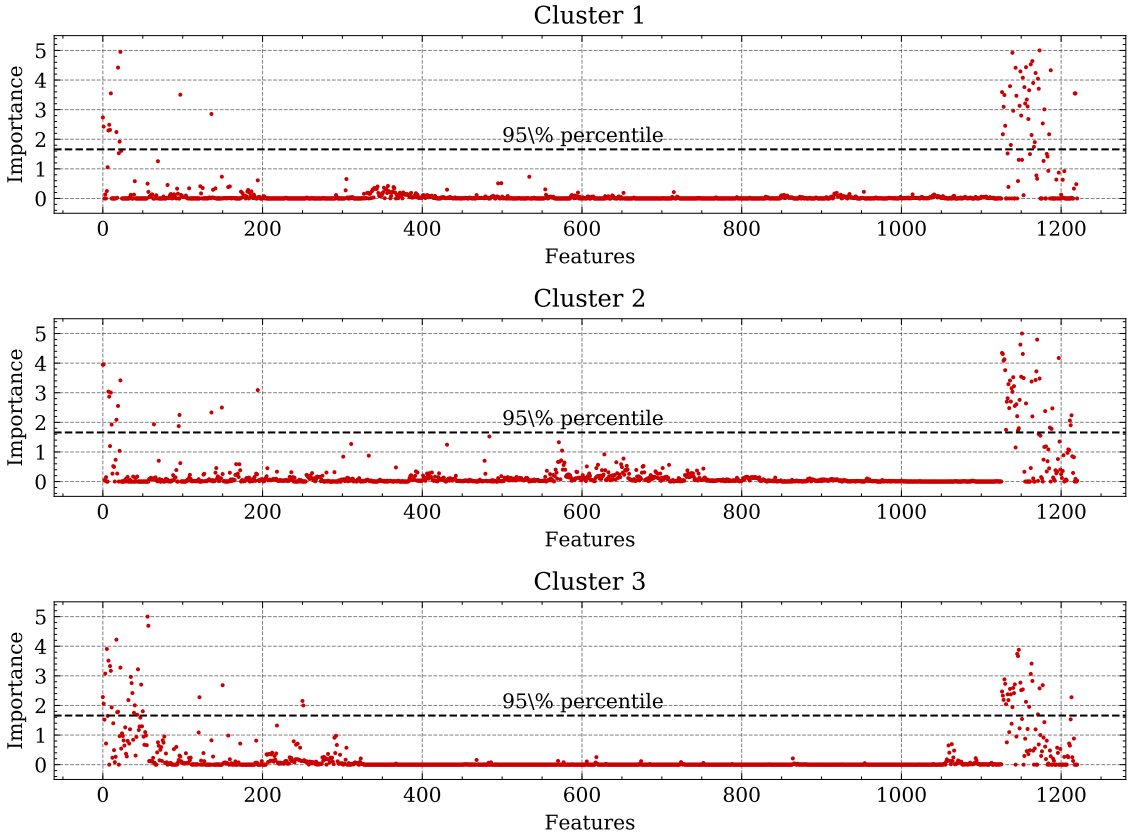

Figura 4.7 – Distribuição das importâncias as propriedades para cada protótipo. Linhas tracejadas pretas indicam o valor de corte de importância mínima, definido como o percentil de 95% da distribuição dos valores de $\mathbf{U}$.

como exemplo. Na figura, a linha tracejada azul indica o valor de corte de $H(\mathbf{v}_{\text{col}})$, que atua como um corte para a seleção de ocupações que possuam a distribuição de pertinências aos clusters suficientemente concentrada.

As linhas tracejadas pretas indicam valores de entropia máxima para os casos de $k = 2$ e $k = 3$. É observado que existem concentrações de vários pontos próximos às linhas pretas, o que indica que muitos dos casos tiveram pertinências similarmente distribuídas entre todos os 3 clusters (linha tracejada preta na posição mais alta) ou entre apenas 2 dos clusters com pertinência próxima a 0 para um dos clusters (linha tracejada preta em altura intermediária).

Após uma etapa preliminar de *hard-clustering*, através do cálculo de $C_i = \arg\max_i \mathbf{v}_{\text{col}}(i)$, onde $\mathbf{v}_{\text{col}}(i) = v_i$, ordenamos as ocupações, dentro de cada cluster, por seus valores de entropia de ocupação $H(\mathbf{v}_{\text{col}})$ em ordem crescente. A lista final é formada através da seleção das ocupações com menores valores de entropia. Como na seleção de features, quaisquer empates entre ocupações, indexadas na matriz $\mathbf{X}$ por j, candidatas dentro de um mesmo cluster C_i são quebrados através da utilização dos valores originais dos elementos $\mathbf{V}_{i,j}$ como pontuação de pertinência, escolhendo os de maior valor.

Utilizando o critério de ranking de propriedades e ocupações descritas acima,

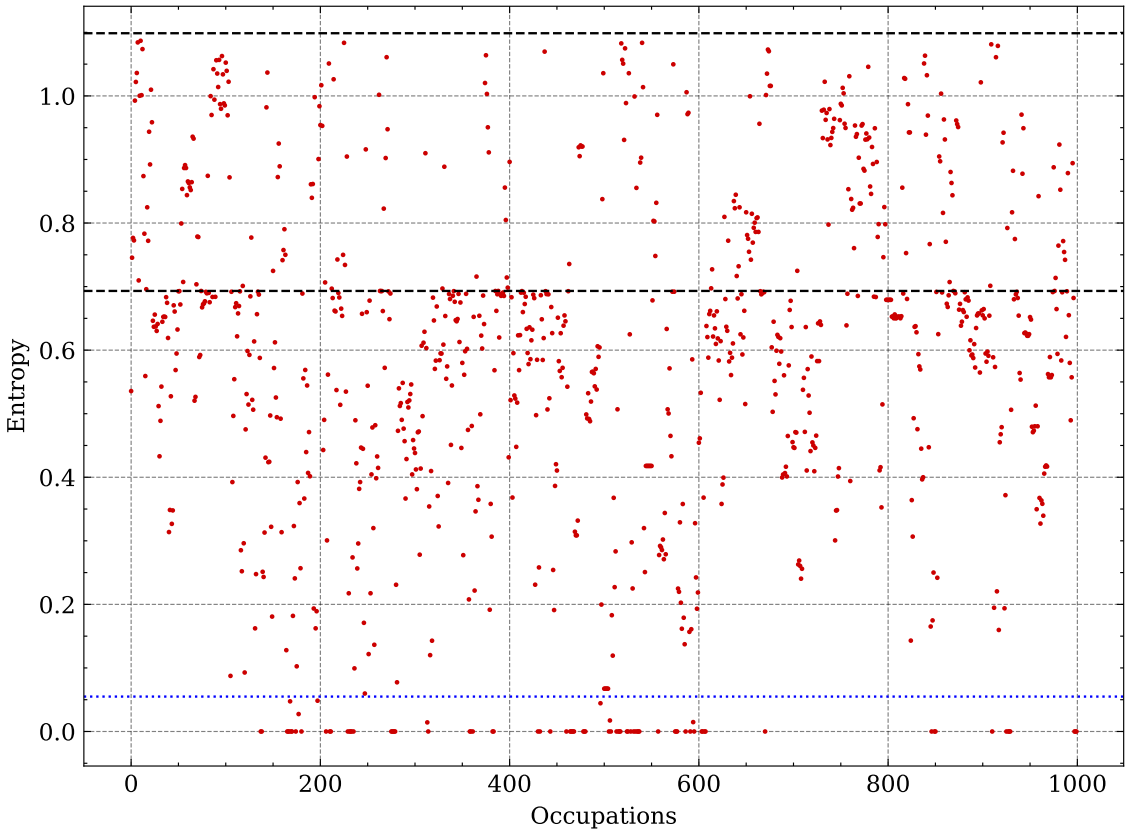

Figura 4.8 – Distribuição de entropia de ocupação para o caso de $k = 3$. A linha pontilhada azul indica o valor de corte calculado por 5% da entropia máxima. As linhas tracejadas pretas ilustram os valores de máxima entropia para os casos de $k = 2$ e $k - 3$.

selecionamos as 6 propriedades mais distintivas e as 10 ocupações mais representativas para cada cluster. Um exemplo desta tabela de resumo e cluster é mostrado na Tabela 4.3.

As quantidades de 6 propriedades e 10 ocupações foram escolhidas de maneira que, para as configurações com valores mais altos de k, existissem quantidades de propriedades e ocupações suficientes que atendessem aos critérios de seleção. Nestes casos, devido ao número maior de clusters, a quantidade de propriedades que atendem aos requisitos para cada cluster se torna mais esparsa. Realizando este procedimento da criação de listas para cada cluster em valores de $k = 3$ até $k = 10$, as quantidades de 6 propriedades e de 10 ocupações foram as que puderam ser alcançada comumente por todas as configurações.

Estes formatos de tabelas de resumo serão úteis não apenas para analisar as relações entre clusters, para uma dada configuração de k clusters, mas também para comparar entre diferentes configurações de clusters.

4.3.2 Compatibilidade entre propriedades e ocupações

Utilizando as tabelas de resumo descritas acima, analisamos os resultados da configuração de $k = 3$ até $k = 10$. Buscaremos respostas para os 3 seguintes grupos de

Ocupações		Propriedades	
Código CBO	Título da ocupação	#	Nome da propriedade
311720	Preparador de tintas (fábrica de tecidos)	1	Tecnologia na área de gastronomia e alimentação
311725	Tingidor de couros e peles	2	Corte de calçados
766305	Acabador de embalagens (flexíveis e cartotécnicas)	3	Processos de acabamento gráfico
841408	Cozinhador (conservação de alimentos)	4	Matérias-primas e aditivos
841416	Cozinhador de carnes	5	Polímeros sintéticos
841420	Cozinhador de frutas e legumes	6	Moldagem e matrizes de polímeros
841428	Cozinhador de pescado		
841432	Desidratador de alimentos		
841440	Esterilizador de alimentos		
848325	Trabalhador de fabricação de sorvete		

Tabela 4.3 – Um exemplo de tabela de resumo de cluster. As primeiras duas colunas indicam as 10 ocupações mais características deste cluster e seus respectívos códigos da CBO. A terceira e a quarta coluna indicam as 6 propriedades mais características deste cluster. Como consequência do critério de seleção, ocupações e propriedades nunca se repetem ao em nenhuma outra tabela de resumo.

perguntas:

- Existe uma relação consistente entre os clusters encontrados para valores diferentes de k?
- Qual é a cobertura de áreas de trabalho dos clusters? Existem clusters redundantes?
- Todas as ocupações são representadas adequadamente pelos clusters em algum nível. Existe uma escolha ótima clara para o valor de k através da análise qualitativa?

Os principais fatores que utilizamos são as listas de ocupações e propriedades encontradas para cada cluster. Buscamos comparar, para cada cluster, em cada configuração, uma compatibilidade temática entre os conjuntos selecionados de propriedades e de ocupações. O objetivo de realizar este procedimento de comparação entre as temáticas constituídas pelas ocupações e pelas propriedades, é a de determinar uma maneira de avaliar se o cluster expressa uma temática clara e consistente, de acordo com a compatibilidade entre ocupações e propriedades.

Aqui, utilizamos apenas uma realização para cada valor de k, pois as variações de resultado da NMF devidas à aleatoriedade na inicialização das matrizes $\mathbf{U}$ e $\mathbf{V}$ não aparentaram refletir em mudanças nas tabelas de resumo de cluster.

Feitas estas observações para cada cluster para valores diferentes de k, o primeiro ponto a notar é que as propriedades, de fato, correspondem tematicamente às ocupações selecionadas de $k = 3$ até $k = 8$. Para $k = 9$ e $k = 10$, por outro lado, há ocorrências de clusters com propriedades tematicamente distantes da maior parte das ocupações correspondentes. Por exemplo, em $k = 9$, um dos clusters tem parte de suas ocupações relacionadas à saúde (ex.: radiologia, hemoterapia, citotecnologista), mas a lista de atributos não tem nenhuma propriedade relacionada à saúde ou ciências médicas. Elas correspondem, na verdade, a propriedades como "meteorologia", "big data", "video game" e "lógica matemática". Em outro caso, ainda mais interessante, um cluster representa principalmente ocupações relacionadas a "motorista" (ex.: motorista de ônibus, taxista, instrutor de direção, motorista de ambulância), mas suas propriedades correspondentes são relacionadas majoritariamente à saúde, como "cuidados de enfermagem", "farmacologia" e "imunologia celular", mas também relacionada a "direção defensiva". Isso nos leva a inferir de que, nesse caso em particular, uma única ocupação, notadamente a de "motorista de ambulância", foi a determinante majoritária daquele cluster.

Estes resultados nos dão uma indicação de que existe um ponto em que a correspondência entre as ocupações e as propriedades selecionadas começa a se deteriorar. Isto pode ser por causa da crescente especificidade das ocupações em cada cluster, que leva a alguns dos protótipos dos clusters a apresentarem composição de valores das propriedades bastante próximas à composição de uma ocupação em particular. Neste contexto, a configuração de $k = 8$ é a de maior valor de k que atende às condições de compatibilidade temática entre ocupações e propriedades para todos seus clusters.

4.3.3 Expansão e Especialização dos Clusters

Analisando os tipos de cluster encontrados através dos diferentes valores de k, foi observado um padrão consistente: para cada incremento em k, houve ou uma divisão de um cluster em múltiplos clusters (tipicamente dois) ou a emergência de um novo cluster. Embora não estritamente consistente, uma sequência geral de expansões e adições de clusters pôde ser observada. Os grupos resultantes para $k = 3 \ldots 8$ foram rotulados como se mostra na lista a seguir:

- **K=3**: "trabalho fisicamente intenso", "operação em máquinas" e "serviços administrativos".
- **K=4**: "construção, esportes e agropecuária", "trabalho em fábrica e agropecuária", "operação em máquinas" e "serviços administrativos".
- **K=5**: "construção, esportes e agropecuária", "trabalho em fábrica (alimentos, têxtil)", "operação em máquinas", "atendente/telemarketing" e "supervisor (construção, TI, eletricista)".

- **K=6**: “construção e esportes”, “trabalho em fábrica (alimentos, têxtil)”, “manufatura mecânica”, “atendente/telemarketing”, “supervisor (tesouraria, vendas)” e “TI e eletricista”.
- **K=7**: “construção e agropecuária”, “trabalho em fábrica (alimentos, têxtil)”, “manufatura mecânica”, “atendente/telemarketing”, “supervisor (tesouraria, vendas)”, “eletricista” e “saúde”.
- **K=8**: “construção e esportes”, “agropecuária”, “trabalho em fábrica (alimentos, têxtil)”, “manufatura mecânica”, “atendente/telemarketing”, “supervisor (tesouraria, construção, têxtil)”, “TI e eletricista” e “saúde”.

Ainda que alguns dos clusters tenham alguma sobreposição, podemos dizer que não existem clusters redundantes na configuração de $k = 8$ clusters, mas não é possível afirmar que os clusters cobrem todos os tipos de ocupações do conjunto completo de dados. Até $k = 8$, observamos um refino progressivo dos temas cobertos por cada cluster, porem, de $k = 9$ em diante, observamos uma deterioração da correspondência entre as seleções de propriedades e de ocupações.

4.3.4 Escolha da Quantidade k de Clusters

Existem diferentes métricas que consideram noções como compacticidade e separação de cluster ou o quão similar um determinado ponto é similar aos outros pontos no mesmo cluster. Experimentamos a otimização em relação a estas métricas e também desenvolvemos um procedimento para avaliação subjetiva da qualidade dos clusters, tanto como forma alternativa de avaliar a utilidade das decisões baseadas nas métricas mas também como uma nova forma de buscar uma resposta para a quantidade de clusters k.

Considerando a análise qualitativa dessas diferentes configurações, temos $k = 8$ como a melhor escolha para análise, superando a escolha anterior de $k = 3$ que, em termos qualitativos, apresentou clusters de composição ambígua e difícil interpretação. Conforme visto nas Seções 4.3.2 e 4.3.3, $k = 8$ foi o valor que conseguiu tornar os clusters maximamente específicos atendendo a requisição de uma noção de coesão interna, avaliada qualitativamente pela compatibilidade temática entre os conjuntos selecionados de ocupações e propriedades nas tabelas de resumo de cada cluster.

Os clusters resultantes da configuração $k = 8$ serão objetos centrais nas análises deste trabalho, que serão apresentadas no Capítulo 5.

4.4 Treinamendo do modelo de probabilidades de automação das ocupações

Neste trabalho, usaremos as estimações resultantes do modelo treinado por Frey e Osborne $\mathbf{y}_{\text{fo}} = [y_{\text{fo},1}, y_{\text{fo},2}, \ldots, y_{\text{fo},n}]$ como o conjunto de treinamento para nosso próprio modelo. Nosso conjunto de dados de propriedades $\mathbf{X}_q$ é composto pelos dados do QBQ, com cada ocupação representada por um vetor-coluna de valores reais positivos $\mathbf{x_i} \in \mathbb{R}^L$, onde L é o número total de propriedades usadas do QBQ. Cada ocupação x_i possui uma avaliação de probabilidade de automação $y_{\text{fo},i}$, formando o mapeamento $y_{\text{fo},i} = f_{\text{fo}}(x_i)$ ou ainda $\mathbf{y}_{\text{fo}} = f_{\text{fo}}(\mathbf{X}_q)$. Estes dados serão utilizados para treinarmos um modelo f_{a} que aproxime o mapeamento f_{fo} tal que $f_{\text{a}}(\mathbf{X}_q) = \mathbf{y}_{\text{a}} \approx \mathbf{y}_{\text{fo}}$. Este mapeamento f_{a} será utilizado para extrapolarmos a avaliação das probabilidades de automação das ocupações para os casos brasileiros que ainda não possuem avaliações.

Evidentemente, usar esses dados como referência necessariamente tem as suas limitações. Em primeiro lugar, o modelo ajustado incorporará os vieses presentes no conjunto de treinamento. Em segundo lugar, as definições para as ocupações nos Estados Unidos não necessariamente são as mesmas que as dadas para o caso brasileiro, ou seja, há especificidades da estrutura ocupacional brasileira que não se encontram incorporadas nas probabilidades estimadas para outras economias. Ainda, um terceiro fator a ser considerado é que o conjunto de treino, formado pelos resultados de Frey e Osborne, são probabilidades de risco de substituição das ocupações (valores no intervalo unitário), ao invés dos valores binários utilizados como referência no treinamento do modelo original, o que altera a configuração matemática do treinamento de um modelo de classificação binária para um modelo que efetua uma espécie de regressão dentro do intervalo unitário.

Tendo isso em consideração, os dados resultantes da estimação serão utilizados majoritariamente com o propósito de visualizar relações mais qualitativas dentro dos dados, como contrastes e tendências gerais através dos grupos.

4.4.1 Seleção de Variáveis

Como mencionado antes, na base de dados do QBQ, cada ocupação é descrita apenas por algumas poucas propriedades (neste contexto, vistas como variáveis), o que levou à introdução de zeros na representação matricial dos dados. Além disso, a base de dados segue sendo construída incrementalmente, o que produz novas colunas (propriedades) e novas linhas (ocupações), contribuindo com isso para uma maior esparsidade da matriz. Também podemos enxergar esta esparsidade em termos de frequência de ocorrência de cada propriedade. Sendo Z a quantidade de vezes que uma propriedade possui valor zero na base de dados e N a quantidade total de ocupações, calculamos a frequência de ocorrência pela razão $\frac{N-Z}{N}$.

Um dos principais problemas que essa esparsidade pode causar é que se, quando dividirmos os dados em conjuntos de treinamento e de validação, as frequências de ocorrência das propriedades ficarem desbalanceadas, o desempenho no treinamento do modelo pode ser comprometido. Para mitigar esses possíveis problemas devidos à divisão dos conjuntos de dados entre treino e validação, é necessário efetuar uma seleção de variáveis, de forma a manter apenas aquelas que atinjam uma frequência mínima em ambos os conjuntos ao mesmo tempo. Idealmente, buscamos uma seleção que aproveite o máximo de variáveis possíveis, mas que garanta uma frequência de ocorrência mínima. O valor de corte escolhido para a frequência de ocorrência mínima foi o de 10%, porque resultou na seleção de quantidades moderadas (entre 110 e 128) de variáveis e representa uma frequência absoluta de 100 ocorrências para cada uma das variáveis selecionadas.

Outra questão a ser considerada é que temos a divisão entre dados rotulados (dados de referência) e não-rotulados. Os dados rotulados são compostos de 622 ocupações, enquanto o conjunto de dados não-rotulados possui 378 ocupações. Isto faz com que seja possível que uma determinada seleção de variáveis feita desconsiderando esta divisão acabe selecionando variáveis com desbalanceamento nas frequências de ocorrência.

Do total de 1221 propriedades, foram selecionadas 110 para o conjunto rotulado e 128 para o conjunto não-rotulado. Tomando a interseção entre estes conjuntos de seleções, isto é, escolhendo apenas as propriedades que atendem ao critério de frequência mínima para ambos os conjuntos ao mesmo tempo, resulta numa seleção final de 100 variáveis.

4.4.2 Considerações sobre a Entropia Cruzada

Para garantir que $y : \mathbb{R} \to [0, 1]$, utilizamos a função logística

$$f(x) = \frac{1}{1 + \exp\left(-\theta(\mathbf{x})\right)} \tag{4.4}$$

usando alguma função latente $\theta : \mathcal{X} \to \mathbb{R}$, onde $\mathcal{X}$ é o conjunto ao qual pertencem os vetores de entrada $\mathbf{x}$. Para a otimização do modelo, utilizamos a entropia cruzada

$$L(y, \hat{y}) = -\Big(y \log\left(\hat{y}\right) + (1 - y) \log\left(1 - \hat{y}\right)\Big) \tag{4.5}$$

como função custo a ser minimizada, onde y são os valores de referência e $\hat{y}$ são os valores estimados pelo modelo. No contexto de classificação, esta função é tipicamente avaliada apenas para os valores de $y = 0$ e $y = 1$. Estas curvas de custo são ilustradas na Figura 4.9. Verifica-se que para cada caso, o valor na curva tende a zero na medida em que $\hat{y}$ se aproxima de y.

Mas, devido ao fato de que usaremos o intervalo completo de $y \in [0, 1]$, obteremos uma gama completa de curvas que dependem do valor de y. Na Figura 4.10, pode ser percebido que, quando y não é nem 0 nem 1, o valor mínimo da curva não é

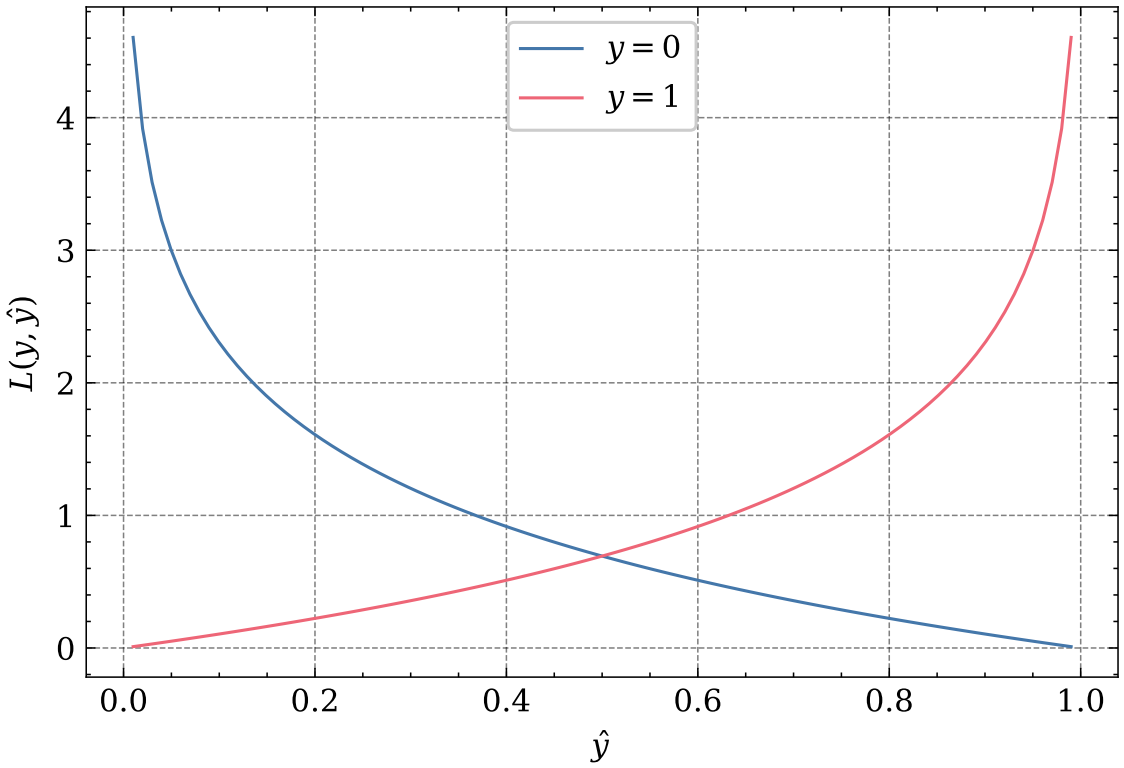

Figura 4.9 – Curvas de entropia cruzada para $y = 0$ e $y = 1$.

mais zero. Isto pois cada valor de y diferente de 0 e 1 resultará um valor mínimo de perda teórico necessariamente maior que zero. Para referência, calculando o custo ótimo médio para $\hat{y} = y$, isto é, o custo médio gerado por um modelo hipotético que replica exatamente todas os resultados do conjunto de treinamento, nós obtemos um custo médio de $0,4339$.

Adicionalmente, para $y = 0,5$, a função possui um valor mínimo de $0,6931$, tem um perfil simétrico e é aproximadamente plana em torno de seu ponto mínimo. Uma possível consequência é que, para ocupações onde o valor de referência seja próximo a $0,5$, qualquer estimador poderia ter uma margem considerável de erro para qualquer direção sem qualquer incremento significativo na função de perda. Isto traz a interpretação de que estimações em torno da faixa de valores próximos a $0,5$ possuem um tipo de incerteza intrínseco devido à forma da função custo. Focamos em realizar interpretações através da análise relações nos dados que indiquem noções menos sensíveis a vieses dos valores, como relações que indiquem contraste entre valores baixos e altos ou correlações.

4.4.3 Configurações para o Ajuste dos Modelos e Resultados

Sendo $f(\mathbf{x}) = \frac{1}{1+\exp(-\theta(\mathbf{x}))}$ a estrutura básica a ser utilizada para a aproximação dos resultados de Frey e Osborne, resta-nos definir a forma funcional de θ. Com esse intuito, avaliamos diferentes tipos de modelos para o mapeamento $f : \mathbb{R}^{100} \to \mathbb{R}$:

- Um modelo linear para $\theta(\mathbf{x})$, que resulta na configuração de um modelo final $f(\mathbf{x})$

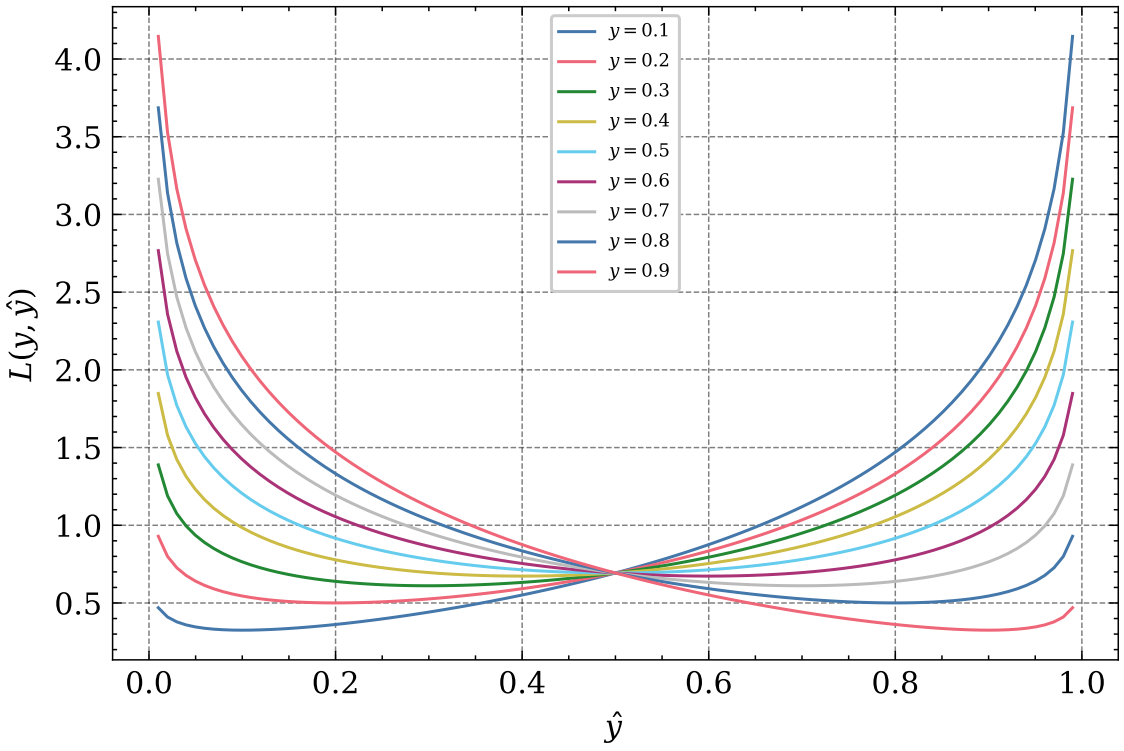

Figura 4.10 – Curvas de entropia cruzada para $y \in [0, 1]$. Nota-se que o valor mínimo para cada curva não é mais zero.

de estrutura idêntica à regressão logística, por ser um modelo bastante usual e;

- Um modelo de rede neural, para possibilitar a criação de mapeamentos não-lineares e com maior poder de generalização.

Regressão Logística: A primeira forma funcional é dada por:

$$\theta(x) = \mathbf{w}^T\mathbf{x} + \mathbf{b}. \tag{4.6}$$

Esta configuração resulta na forma funcional final de $f(\mathbf{x})$ basicamente idêntica à regressão logística, exceto pelo fato de que não usamos rótulos binários para a supervisão do modelo.

Rede Neural: A segunda forma funcional avaliada para θ é dada por uma rede neural, com uma função logística como função de ativação em sua camada final. Se trata de uma rede totalmente conexa com unidades lineares retificadas (em inglês, *Rectified Linear Unit* (ReLU)), $\text{ReLU}(t) = \max(0, t)$, como suas funções de ativação, formando uma rede com P de camadas ocultas. Matematicamente, a rede neural proposta é expressa pelo sistema de equações

$$\begin{cases} l_0(\mathbf{x}) & = \mathbf{x} \\ l_i(\mathbf{x}) & = \text{ReLU}(\mathbf{W_i}^T l_{i-1}(\mathbf{x}) + \mathbf{b_i}),\ i \in [1, N] \\ y(\mathbf{x}) & = \frac{1}{1+\exp\left(-l_N(\mathbf{x})\right)}. \end{cases} \tag{4.7}$$

A rede foi configurada com $P = 5$ camadas ocultas e quantidades de neurônios de $\{100, 50, 25, 10, 5\}$ em suas camadas, respectivamente.

Ambos modelos tiveram o processo de ajuste ou treinamento implementado utilizando a biblioteca TensorFlow (ABADI et al., 2016). Para cada modelo, um conjunto de 100 realizações do modelo foi treinado, cada uma com diferentes divisões dos conjuntos de treino e validação. O modelo final para cada tipo será do tipo ensemble, formado pelo resultante da média de todas as 100 realizações do modelo. A divisão entre treino e validação foi realizada aleatoriamente com a regra de 75% dos dados separados para o ajuste dos modelos e os 25% restantes para validação. Cada instância é configurada para sempre retornar os parâmetros que alcançarem os melhores desempenhos de validação dentro do histórico de validação cruzada. O desempenho de cada ensemble (um para cada tipo de modelo) é estimado através da média dos custos médios de validação de todas as 100 realizações. O desempenho resultante para cada tipo de modelo é exibido na Tabela 4.4. Ambos modelos alcançaram um nível de perda razoavelmente bom, com a rede neural sendo a melhor entre as duas.

Modelo	Custo médio
Regressão logística	0,5313
Rede neural	0,4939

Tabela 4.4 – Perda média calculada através das 100 instâncias para cada tipo de modelo.

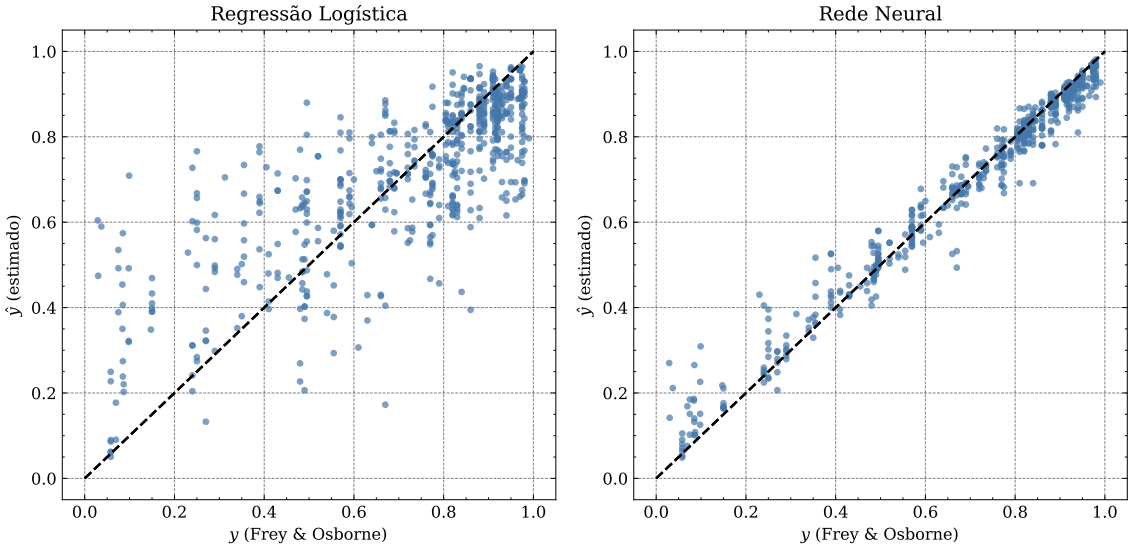

Figura 4.11 – Visualização dos resultados das predições comparadas aos resultados de Frey e Osborne. Resultados estimados no eixo vertical e resultados de Frey e Osborne no eixo horizontal.

Pela visualização comparativa na Figura 4.11, entre os resultados estimados (eixo vertical) e os resultados de Frey e Osborne (eixo horizontal), podemos verificar que ambos os modelos tiveram uma tendência a errar para valores mais próximos de 0,5, ou

seja, superestimar o risco para referências nas faixas inferiores de risco e subestimar o risco para referências nas faixas superiores de risco.

Em termos de custo médio, não há diferença significativa do modelo de regressão logística em relação à rede neural. Já pela visualização dos resultados, nota-se que o modelo de regressão logística apresentou maiores desvios de uma maneira geral, especialmente para as referências nas faixas inferiores de risco (com muitas das estimações passando de 0,5).

Sendo assim, o capítulo seguinte se centrará na discussão dos resultados usando a rede neural, devido ao seu melhor desempenho e pela possibilidade que este modelo tenha refletido, implicitamente, numa variedade de relações não-lineares entre as variáveis não incluídas no processo de treino e o restante das variáveis utilizadas nesta seção, que possam ser de interesse para a análise de padrões.

5 Resultados e Análises

Neste capítulo exibimos os resultados principais dos processos de clusterização e analisamos os padrões observados nos clusters. Seguimos com o objetivo de utilizar os resultados da clusterização e dos riscos de automação para nos ajudar na visualização de relações implícitas nos dados, que podem ser refletidas, entre outras maneiras, na forma de amplas tendências na associação de dados a clusters ou valores de risco de automação, padrões recorrentes no agrupamento de dados ou ainda relações díspares com os clusters ou dados de risco de automação, vindas de ocupações aparentemente similares.

Primeiramente analisamos, na Seção 5.1, os significados que podem ser entendidos dos clusters na configuração $k = 8$, encontrados no Capítulo 3. Descrevemos um procedimento de inspeção dos clusters e estabelecemos breves descrições e títulos para cada um dos clusters.

Em seguida, cruzamos os dados de clusterização e das classificações da CBO com os dados de risco de automação, utilizando a última como uma nova camada de informação para observar tanto associações de grupos e ocupações individuais aos diferentes níveis de riscos obtidos pelo modelo de aproximação (apresentado na Seção 4.4), quanto às relações de similaridade e diferenças entre grupos e ocupações utilizando as associações aos níveis de risco como um fator de comparação.

5.1 Avaliando os Temas dos Clusters

Na Seção 4.3, chegamos à escolha da quantidade de clusters $k = 8$ através de uma análise qualitativa dos resumos dos clusters. Esta foi a configuração que atingiu o maior nível de especificidade dos clusters sem deteriorar a coesão entre ocupações e propriedades selecionadas.

Nesta seção, começamos a analisar em mais detalhes a composição de cada um desses clusters. Buscamos entender quais são os fatores que podem ser considerados como indicadores de uma unidade interna de cada cluster. Queremos identificar temáticas capturadas pelos recortes gerados pelos clusters. Estas temáticas não são fatores diretamente derivados das representações numéricas, mas ainda assim entendemos que é possível compreender relações entre os registros agrupados de uma maneira geral.

As condições de agrupamento no momento de clusterização são originadas do fato de que a NMF busca uma representação dos dados originais através de duas matrizes: uma de protótipos e uma de codificações. Isto faz com que de um lado tenhamos vetores-protótipo dos quais extraímos as propriedades mais marcantes de um cluster e do

outro lado tenhamos codificações que indicam, em alguns casos, que algumas ocupações são muito facilmente representadas por apenas um dos protótipos, indicando que essas ocupações sejam, instâncias concretas que ilustram bem as características centrais do cluster.

Em parte, uma noção de temática pode ser derivada desta percepção de compatibilidade entre os nomes dessas ocupações e propriedades selecionadas. Aqui, formulamos mais uma forma de descrição sumária para os clusters, na forma de uma tabela que nos informe sobre a composição de cada um dos clusters em termos das classificações da CBO.

Cada cluster é composto por ocupações de diferentes GGs da CBO e dentro de cada GG existe uma subdivisão entre diferentes SGPs. Realizamos sumarizações dos clusters (disponíveis no Apêndice B) com base nessas hierarquias da CBO e as utilizamos em conjunto com as tabelas das seleções de ocupações e propriedades do QBQ (disponíveis no Apêndice A) para determinar os temas de cada cluster. Para cada cluster, é analisada sua composição completa de ocupações, propriedades e classificações CBO, é realizada uma descrição sintetizante e coesiva de todos estes componentes e então é dado um título ao cluster. Os resultados são os apresentados na lista abaixo.

1. **Trabalho em fábrica**: Majoritariamente composto de ocupações relacionadas a produção industrial, especialmente àquelas relacionadas às indústrias têxtil e de alimentos. Inclui ocupações como “tingidor de couro e pele”, “esterilizador de alimentos”, “cozinhador de carne”, “desidratador de alimentos”. Entre as propriedades para este cluster estão “tecnologias nas áreas da gastronomia e alimentos”, “corte de sapatos” e “materiais e aditivos”.

2. **Agropecuária**: Este cluster é composto somente de ocupações no GG 6, de trabalhadores da agropecuária, florestagem e pescaria. Algumas das ocupações incluídas neste cluster estão trabalhadores na cultivação de vários produtos agriculturais, com árvores frutíferas, arroz e café. Entre as propriedades para este cluster estão “operação de equipamento e maquinário agricultural”, “nutrição vegetal, fertilizantes e corretivos” e “processos de armazenamento de produto”.

3. **Manufatura mecânica**: Majoritariamente composto por trabalhadores no SGP 72 “metalurgia e compósitos”, mas também incluindo trabalhadores em manutenção mecânica (SGP 91) e siderúrgicas (SGP 82). Exemplos de ocupações neste cluster são “ajustador de ferramentas” e “operador de laminadora”. Entre as propriedades para este cluster estão “usinagem”, “solda” e “controle dimensional”.

4. **Vendas e escritório**: Este cluster pode ser descrto como principalmente composto

de dois blocos principais: um primeiro de trabalhadores no comércio (GG 5) e um segundo compreendendo várias ocupações relacionadas a serviços administrativos (GG 4 e SGPs 35 e 33). Algumas ocupações neste cluster são "assistente de vendas", "atendente de banco", "operador de televendas" e "vendedor de loja". Entre as propriedades selecionadas para este cluster estão "serviços e operações bancárias", "processo de vendas" e "gerenciamento de relacionamento com cliente (CRM)".

5. **Construção e industria extrativa**: Este cluster é majoritariamente composto de ocupações relacionadas tanto a construção (SGP 71) quanto as indústrias de extração e processamento de recursos (SGPs 77, 72 e 82 e GG 6). Exemplos de ocupações neste cluster são "guincheiro", "assentador de canalização", "destroçador de pedras" e "operador de máquina perfuradora". Entre as propriedades selecionadas para este cluster estão "condifionamento físico" e "operação de maquinas de construção".

6. **Saúde**: Este cluster pode ser entendido como majoritariamente composto de dois grupos maiores: ocupações relacionadas a saúde e ocupações que são relacionadas a normas de saúde. O primeiro grupo é claramente exemplificado por ocupações no SGP 32, como "técnico de radiologia" e "citotécnico" e ocupações com CBO começando com o SG 515 (um subgrupo do SGP 51, relacionado a serviços na saúde), como "auxiliar de farmácia de manipulação" e "assistente de laboratório de imunobiologia". O segundo grupo é exemplificado por ocupações em uma grande variedade de grupos CBO, desde o GG 6 até o GG 9. Entre as propriedades selecionadas para este cluster estão "boas práticas na área da sáude", "normas regulamentadoras", "protocolo de trabalho em saúde" e "fisiologia".

7. **Supervisor**: Apesar deste cluster conter a composição mais diversa em termos de grupos CBO, a maioria das ocupações copartilha uma característica comum de que elas são, em sua maioria, posições de gerenciamento ou supervisão dentro de determinado campo de atuação. Exemplos disso são "supervisor de depósito" (SGP 41), "mestre de construção civil" (SGP 71) e "contramestre de tecelagem". Entre as propriedades selecionadas para este cluster estão "dinâmica das relações interpessoais", "liderança' e "planejamento de produção".

8. **Eletricista e TI**: Este cluster é majoritariamente composto de ocupações relacionadas à eletrotécnica, tecnologia da informação e telecomunicação. Exemplo de ocupações neste cluster incluem "técnico de manutenção em máquinas elétricas", "técnico em eletrônica", "técnico em manutenção de equipamento de informática" e "técnico de comunicação de dados". Entre as propriedades para este cluster estão "circuitos lógicos digitais", "systemas de potência elétricos" e "circuitos magnéticos".

Nesta seção, foram identificados temas e atribuídos nomes para cada um dos clusters, o que nos dá uma visão mais clara das relações internas dos dados capturadas

pela clusterização. Nota-se que a maioria dos clusters não coincidiram com as classificações da CBO. O cluster 2, que identificamos como de temática de agropecuária, é o que mais se aproximou de uma das classificações originais da CBO (o GG 6, de agropecuária, silvicultura e pesca). Temos então a situação de que o processo de clusterização encontrou recortes, em sua maioria, distintos dos da CBO.

No restante deste capítulo, introduzimos os dados sobre risco de automação, obtidos na Seção 3.4, e os utilizaremos para buscar por visualizações de relações e padrões neles contidos.

5.2 Incorporando os Dados de Risco de Substituição

Nesta seção, utilizamos os dados de risco de automação estimados para o mercado de trabalho brasileiro, os quais resultam da aproximação e extrapolação dos resultados publicados por Frey e Osborne (FREY; OSBORNE, 2017). Como discutido anteriormente, o processo de aproximação e extrapolação traz consigo possíveis vieses ou distorções, que podem ser devidas a uma gama de fatores como erro de aproximação, limitações do modelo construído, a utilização dos dados probabilísticos (valores contínuos) numa estrutura idealmente desenvolvida para dados binários ou ainda possíveis vieses ou distorções oriundas do próprio modelo desenvolvido por Frey e Osborne.

Portanto, para os propósitos deste trabalho, não centramos nossas análises nos valores estimados em si, mas sim na utilização deles como um fator adicional para visualizar relações entre os dados da QBQ, da CBO e dos clusters encontrados. Os valores resultantes do modelo de risco de automação, ainda que carreguem tais vieses ou distorções, têm o potencial de ilustrar tendências mais amplas ou relações de contraste entre diferentes partes das bases de dados, por exemplo.

Começamos por uma análise da distribuição dos riscos visualizada em diferentes perspectivas e, em seguida, buscamos observar diferentes relações emergentes entre esses valores de risco de automação e as ocupações, classificações CBO e clusters.

5.2.1 Caracterização das Distribuições dos Riscos de Automação

Começamos visualizando as distribuições completas dos riscos de automação estimados por Frey e Osborne e as aproximações feitas pelo nosso modelo, para as 1000 ocupações da QBQ. Numa primeira vista, pode ser notado que a estimação de risco tem um viés geral para valores mais altos de risco (Figura 5.1), visto que mais de 30% das ocupações possuem valor de risco de automação entre $0,8$ e $1,0$ para ambos os modelos. Como discutido previamente, ao invés de analisarmos diretamente em função dos valores estimados de risco de automação, optamos em analisá-las em termos de diferenças relativas.

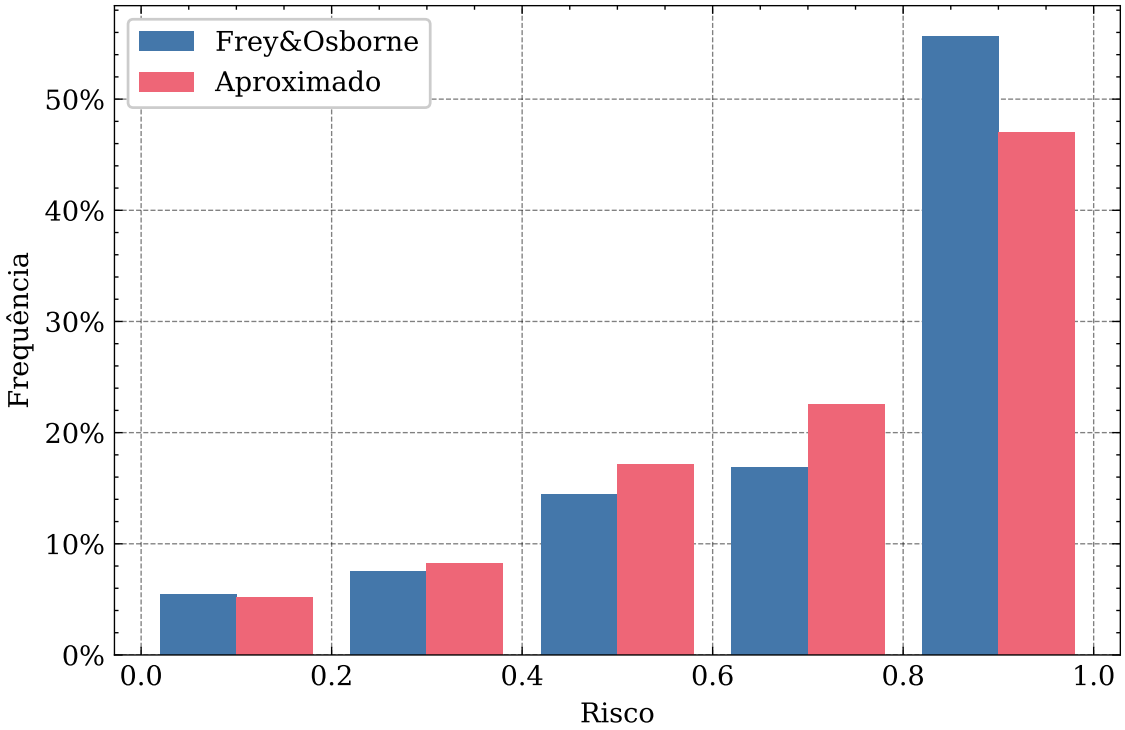

Figura 5.1 – Comparação entre as distribuições de risco de automação encontradas por Frey e Osborne e pelo nosso modelo. Houve uma tendência geral para valores altos de risco. Nosso modelo resultou em uma distribuição similar.

A partir daqui, olhamos para as distribuições de risco sob diferentes perspectivas. Agrupamos os dados em função de suas atribuições de clusters e visualizamos as distribuições de risco condicionadas por cluster através de boxplots (Figura 5.2). Aqui vemos que, apesar de existir uma tendência a valores elevados de risco de automação (todas as médias estão acima de $0,6$), ainda podemos distinguir alguns destes clusters em termos dos valores médios de risco ou ainda pelas amplitudes entre quartis apresentadas pelo boxplot.

Neste contexto, podemos identificar algumas diferentes faixas de risco, baseadas nas diferentes distribuições observadas de cada cluster. Uma primeira faixa, associada às maiores medianas dos valores de risco de automação e com pequenas amplitudes interquartil, é composta pelos clusters 1 (trabalho em fábrica) e 3 (manufatura mecânica). Uma segunda faixa, de valores medianos de risco de automação altos mas com com amplitudes relativamente grandes, é composta pelos clusters 2 (agropecuária), 4 (vendas e escritório) e 5 (construção e esportes) e pode ser considerada uma faixa de médio risco. Por último, uma faixa com os menores valores medianos de risco e com as maiores amplitudes interquartil é composta pelos clusters 6 (saúde), 7 (supervisor) e 8 (eletricista e TI).

Numa segunda perspectiva, agrupamos os dados em função das classificações em Grandes Grupos (GGs) da CBO, e visualizamos as distribuições em boxplot (Figura 5.3). Similarmente, conseguimos observar diferentes faixas de risco, de acordo com as

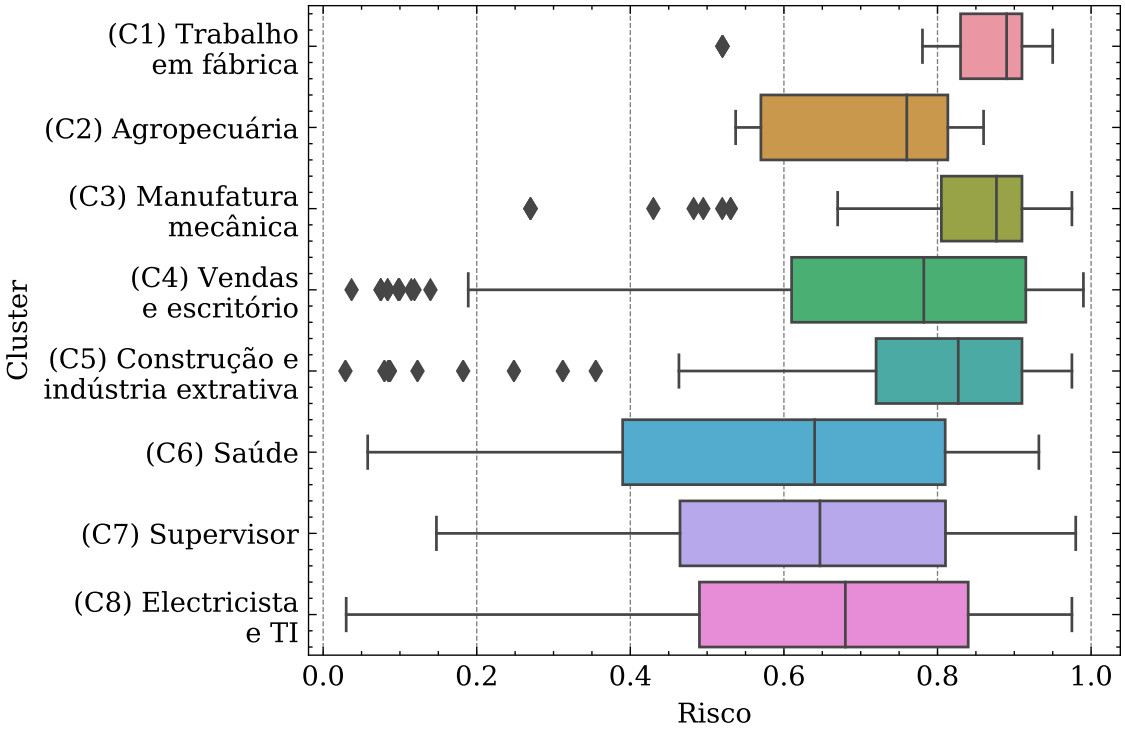

Figura 5.2 – Distribuição de riscos estimados por cluster. Apesar de haver um viés geral para valores altos de risco, nota-se que alguns clusters apresentam valores médios mais altos com menor variância.

características das distribuições, segundo visualizadas pelo boxplot.

Percebemos uma primeira faixa apresentando distribuições com as maiores medianas de risco de automação e amplitudes interquartil pequenas, composta pelos GG 4 (serviços administrativos) e GG 8 (produção industrial). A segunda faixa, composta pelos GG 5 (serviços de comércio), GG 6 (agropecuária) e GG 7 (produção industrial), apresentou distribuições com medianas de risco de automação intermediárias e amplitudes interquartil relativamente grandes. Por fim, os GG 3 (técnicos de nível médio) e GG 9 (manutenção industrial) foram os que apresentaram as menores medianas de risco de automação (ambas menores que $0,6$), com o GG 3 (técnicos de nível médio) apresentando uma amplitude interquartil bastante grande, o que indica que seja uma área que abrange muitos tipos diferentes de ocupações, pelo menos no que se refere à probabilidade de automação.

Nota-se que, ainda que os GG 7 e 8 tenham o mesmo nome pela definição da CBO (produção industrial), eles apresentaram distribuições de risco consideravelmente distintas, com o GG 7 tendo uma grande variância e um valor médio de risco relativamente alto e o GG 8 apresentando um risco médio um pouco mais alto e variância bastante reduzida. A documentação da CBO distingue estes dois GGs determinando que o GG 7 tende às ocupações relacionadas aos processos discretos e o GG 8 aos processos contínuos. Considerando isso, uma possível interpretação destes dados é a de que as ocupações ligadas

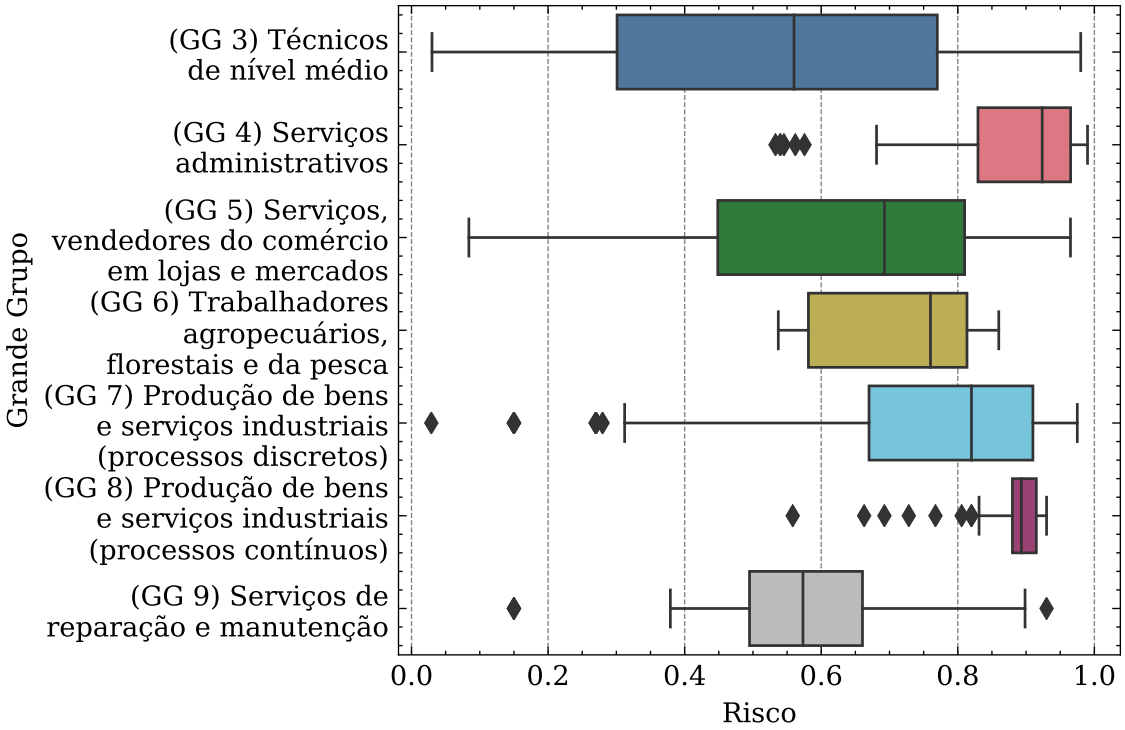

Figura 5.3 – Distribuição de risco estimado de grandes grupos CBO. Similarmente ao agrupamento por clusters, existem concentrações de altos riscos com baixa variância em alguns grupos, notavelmente o GG 4 (serviços administrativos) e GG 8 (produção industrial).

aos processos discretos de produção abrangem uma variedade relativamente grande de condições de trabalho, com parte dessas ocupações tendo menores níveis de vulnerabilidade à automação, enquanto as ocupações relacionadas aos processos contínuos de produção estão majoritariamente associadas a maiores riscos de automação.

Analisando mais a fundo as relações dos resultados das tabelas de composição dos cluster em função das classificações da CBO (Apêndice B) em relação aos temas dos clusters encontrados, percebemos que o processo de clusterização acaba dividindo os grandes grupos em vários clusters. Descrevemos brevemente estas relações entre as classificações CBO e os clusters na lista abaixo:

1. O GG 3 (técnicos de nível médio) foi dividido entre os cluster 4 (vendas e escritório), 6 (saúde), 7 (supervisor) e 8 (eletricista e TI), sendo o cluster 4 o associado aos maiores níveis de risco de automação entre eles, potencialmente explicando os valores mais altos da distribuição de riscos do GG 3.

2. Os GGs 4 (trabalhadores de serviços administrativos) e 5 (trabalhadores dos serviços, vendedores do comércio em lojas e mercados) foram divididos entre os clusters 4 (vendas e escritório), 5 (construção e indústria extrativa) e 7 (supervisor).

3. O GG 6 (trabalhadores agropecuários, florestais e da pesca) basicamente correspondeu ao cluster 2 (agropecuária).

4. Os GGs 7 (trabalhadores da produção de bens e serviços industriais - processos discretos), 8 (trabalhadores da produção de bens e serviços industriais - processos contínuos) e 9 (trabalhadores em serviços de reparação e manutenção) foram divididos entre os clusters 1 (trabalho em fábrica), 3 (manufatura mecânica), 5 (construção e indústria extrativa) e 8 (eletricista e TI).

Os clusters, na maior parte, não apresentaram correspondência direta com nenhum dos grandes grupos da CBO (exceto pelo cluster 2 com o GG 6). Ainda assim verificamos um certo nível de correspondência entre conjuntos de clusters e conjuntos de GGs da CBO de maneira geral. Uma das maiores diferenças foi a divisão do GG 3 (técnicos de nível médio), cujas ocupações foram divididas em áreas mais específicas de atuação. Outro fator diferencial dos clusters em relação às classificações da CBO, foi a criação de um cluster que agrupa ocupações de variados GGs, geralmente associados ao cluster devido aos requisitos de conhecimentos relacionados à saúde ou a normas de saúde.

5.2.2 Análise pela Correlação entre as Propriedades e os Riscos de Automação

Na seção anterior, visualizamos diretamente as distribuições dos valores resultantes dos modelos de risco de automação das ocupações. Conseguimos identificar, de maneira geral, algumas tendências presentes nos dados quando visualizamos os riscos de automação através de recortes baseados nas classificações da CBO e nos clusters.

Nesta seção, buscamos visualizar mais um tipo de relação entre os dados: a relação entre as propriedades que descrevem as atividades de trabalho (habilidades, conhecimentos e atitudes) e os riscos de automação. Numa estratégia similar à seção anterior, começamos tentando visualizar características gerais dessas relações entre as variáveis e em seguida especificamos uma análise similar baseada nos recortes das classificações da CBO e dos clusters.

Buscando por uma maneira de estabelecer visualizações destas correlações entre os valores dos riscos de automação e os valores das importâncias das propriedades, partimos utilizando o identicamente nomeado coeficiente de correlação. Calculamos o coeficiente de correlação $r_{Y,X}$ entre o risco (Y) e cada uma das propriedades (X_m) através de $r_{Y,X_m} = \frac{\mathbb{E}[(Y-\mu_Y)(X_m-\mu_{X_m})]}{\sigma_Y \sigma_{X_m}}$, onde μ é o valor médio de sua variável correspondente e σ é o desvio padrão de sua variável correspondente.

Analisando as propriedades que resultaram nos maiores e menores valores de correlação (dez de cada) para o conjunto de dados completo (Tabela 5.1), encontramos que as propriedades com os valores de correlação mais positivos, isto é, as propriedades

geralmente correlacionadas com o aumento no risco de automação, são majoritariamente associadas às tarefas rotineiras ou procedurais, enquanto parte delas são relacionadas ao planejamento e monitoramento. As outras dez propriedades com correlação mais negativas podem ser resumidas como compreendendo áreas como comunicação, pensamento analítico e crítico e saúde.

Correlações mais positivas		Correlações mais negativas	
Correlação	Propriedade	Correlação	Propriedade
+0,24	Aplicação de instruções simples e rotineiras	-0,33	Reconhecimento de fala
+0,23	Trabalho sob supervisão direta, com alguma autonomia	-0,31	Ações preventivas
+0,22	Planejamento do trabalho	-0,29	Clareza de fala
+0,19	Normas regulamentadoras para área industrial	-0,29	Raciocínio analítico
+0,17	Rotinas contábeis	-0,29	Normas regulamentadoras
+0,17	Avaliação do próprio desempenho, com alguma orientação	-0,28	Ações emergenciais e de urgência
+0,16	Rotinas financeiras	-0,28	Anatomia
+0,16	Processos administrativos	-0,28	Visão noturna
+0,16	Painéis de controle	-0,26	Suporte básico à vida
+0,16	Controle da qualidade	-0,26	Medidas de higiene e assepsia

Tabela 5.1 – Seleção das propriedades com os valores mais positivos e negativos de correlação com o risco de automação, calculados na base de dados completa. Valores positivos indicam um aumento no risco de automação enquanto valores negativos indicam uma diminuição do valor de automação.

Notavelmente, as propriedades selecionadas para cada cluster através do valor de correlação, apesar de não bater exatamente com os conjuntos de propriedades selecionados pelo processo de resumo dos protótipos dos clusters, tem uma compatibilidade temática com os resultados encontrados pelos resumos, reforçando um senso de consistência no processo geral da análise de clusters realizado até agora.

Repetindo as mesmas análises de correlação separadamente para cada cluster, encontramos que os resultados variam significativamente. Apesar de cada um dos clusters apresentar seu conjunto particular de propriedades, existem comunalidades gerais entre conjuntos de propriedades associadas com correlações tanto positivas quanto negativas. A lista completa das propriedades com correlações mais positivas e mais negativas para cada cluster estão disponíveis no Apêndice C.

Houve uma grande quantidade de propriedades contendo uma variação de "processo" ou "rotina" em seus títulos, o que nos permite caracterizá-los como pertencentes

a um grupo de propriedades relacionadas a tarefas procedurais e/ou rotineiras. Estas tarefas estavam presentes em quase todos os clusters e foram consistentemente associadas com os valores mais positivos de correlação com risco de automação. Similarmente, outro conjunto de propriedades relacionadas com controle de processo, todas com "controle" em seus títulos, foram consistentemente correlacionadas positivamente com risco de automação. Outro grupo que pôde ser identificado com correlações consistentemente positivas foi o grupo de propriedades relacionadas a conhecimentos em normas regulamentadoras e técnicas. Distinguindo-se da tendência de correlações positivas, propriedades relacionadas às interações sociais, como "clareza na fala" ou "cuidar de pessoas", foram consistentemente negativamente relacionadas ao risco de automação.

Além dessas relações formadas por comunalidades e consistências, também houve grupos de propriedades que, dependendo do contexto, eram relacionadas a tanto correlações negativas ou positivas. Propriedades relacionadas à percepção física foram frequentemente dependentes em contexto, como exemplificado pela propriedade "escuta ativa" sendo positivamente correlacionada ao risco de automação no cluster 3 (manufatura mecânica) mas negativamente correlacionada no cluster 1 (trabalho em fábricas). Um exemplo adicional é "sensibilidade auditiva", que foi positivamente correlacionada ao risco de automação no cluster 8 (eletricista e TI) mas negativamente correlacionada no cluster 4 (vendas e escritório) e 7 (supervisor). Apesar de serem propriedades distintas, tanto a "visão periférica" quanto a "visão a distância" são relacionadas a percepção física, mas a primeira foi negativamente correlacionada ao risco de automação no cluster 7 (supervisor) e a segunda foi positivamente correlacionada no cluster 3 (manufatura mecânica).

Num caso similar, algumas propriedades relacionadas a habilidades motoras tiveram variadas correlações com os valores de risco dependendo do contexto: "coordenação multimembros" teve correlação positiva no cluster 3 (manufatura mecânica) e negativa no cluster 8 (eletricista e TI), enquanto "velocidade de movimento de membros" teve correlação positiva no cluster 2 (agropecuária) e negativa no cluster 1 (trabalho em fábrica). Novamente, apesar de serem propriedades distintas, "força estática" e "força dinâmica" são ambas relacionadas a habilidades motoras, mas a primeira foi positivamente correlacionada no cluster 1 (trabalho em fábrica) e a segunda foi negativamente correlacionada no cluster 8 (eletricista e TI).

Uma diferença similar ocorreu entre "informática" e "informática médica", com "informática" sendo positivamente correlacionada com o risco de automação no cluster 4 (vendas e escritório) e "informática médica" sendo negativamente correlacionada no cluster 6 (saúde). Outro caso de inconsistência foi encontrado no cluster 6 (saúde), onde a habilidade "comunicação oral e escrita" foi positivamente correlacionada com o risco de automação, enquanto "comunicação e expressão" apresentou correlação negativa.

Realizando o mesmo tipo de análise para os recortes baseados nas classificações

da CBO (grandes grupos), obtemos outras sete tabelas (GGs 3 a 9) (tabelas disponíveis no Apêndice D).

Nestas tabelas podem ser observadas algumas propriedades que são consistentemente relacionadas ao aumento no risco de automatização (correlações positivas), como aquelas associadas às atividades rotineiras (que contém "rotina" em seu título), presentes nos GGs 3 (técnicos de ensino médio) e 4 (serviços administrativos) na forma de "rotinas financeiras", "rotinas contábeis" e "rotinas de operações e serviços bancários". A habilidade de "aplicação de instruções simples e rotineiras" também foi consistentemente associada ao aumento do risco de automação e foi encontrada em todos os GGs exceto o 3 (técnicos de ensino médio) e o 7 (produção de bens e serviços industriais - processos discretos). Os conhecimentos de "informática" e "internet" e a habilidade de "trabalho sob supervisão" também apresentaram correlações positivas consistentemente.

As propriedades consistentemente associadas à diminuição do risco de automação em diferentes GGs foram relativamente mais raras. Uma delas foi a habilidade de "ações preventivas" e outra foi a de "tempo de reação", ambas presentes nos GGs 3 (técnico de ensino médio) e 5 (trabalhadores de serviços, vendedores do comércio e lojas e mercados).

Similarmente ao observado nos clusters, foram observados casos de propriedades sendo correlacionadas tanto positivamente quanto negativamente em diferentes GGs. Podemos identificar um primeiro grupo de propriedades associadas à comunicação por fala, "clareza de fala" e "reconhecimento de fala", que foram associadas à diminuição do risco de automação nos GGs 6 (agropecuários, florestais e pesca) e 7 (produção de bens e serviços industriais), mas associadas ao aumento do risco no GG 4 (serviços administrativos). Num segundo grupo, associado à percepção auditiva, estão as propriedades de "localização de som" e "atenção auditiva", que foram associados à diminuição do risco nos clusters 4 (serviços administrativos) e 5 (vendedores) mas foram associados ao aumento do risco no cluster 9 (reparação e manutenção). Por último, um terceiro grupo associado à capacidade de força física dos trabalhadores, incluindo as habilidades de "força de tronco", "força dinâmica" e "força estática", teve associação à diminuição do risco para o GG 3 (técnico de ensino médio) e ao aumento do risco para os GGs 8 (produção de bens e serviços industriais) e 9 (reparação e manutenção).

A existência de vários exemplos onde ocorreram correlações entre propriedades e o risco de automação com valores contraditórios dependendo do contexto, dão-nos indicações de que, ainda que a base de dados tenha sido construída para atribuir significados padronizados a cada uma dessas propriedades (conhecimentos, habilidades e atitudes), elas ainda podem refletir significados diferentes dependendo da realidade concreta de cada trabalho. Propriedades que apresentaram este comportamento foram tipicamente relacionadas a capacidades humanas físicas (percepção e força). Particularmente no caso

da percepção auditiva, é possível que o contexto das ocupações pertencentes ao GG 9 (reparação e manutenção) seja, por exemplo, um contexto onde o trabalhador precisa perceber ruídos em máquinas com a finalidade de detectar problemas para as atividades de reparação e manutenção. Por outro lado, essas mesmas habilidades de percepção auditiva provavelmente possuem utilização bastante distinta quando pensamos em possíveis situações de trabalho das ocupações pertencentes ao GG 5 (vendedores), que são tipicamente ocupações que trabalham com pessoas em espaços de lojas e mercados.

6 Conclusão

Neste trabalho, realizamos a análise dos dados do Quadro Brasileiro de Qualificações em conjunto com as informações sobre a estrutura dessas ocupações fornecidas pela Classificação Brasileira de Ocupações e as estimativas de risco de substituição de ocupações originalmente realizadas por Frey e Osborne. Realizamos as análises desses dados sem pretensão de realizar predições sobre a realidade futura das ocupações e do mercado de trabalho brasileiro, nem de identificar os padrões encontrados nos dados como representações diretas de processos ou fenômenos socio-econômicos. Procuramos encontrar relações e padrões expressadas nos dados e buscar por explicações, sempre que possível, também contidas nos próprios dados.

Começamos pela análise de clusters via FMN da base de dados do Quadro Brasileiro de Qualificações. A escolha de quantidade de clusters K foi realizada em primeiro instante em função de métricas de qualidade de cluster (índice Davies-Bouldin e pontuação silhouette), mas, após a avaliação das configurações de clustering através de uma comparação subjetiva da compatibilidade temática entre as ocupações e as propriedades mais características de cada cluster, obtivemos uma visão mais clara da utilidade que resultava de cada configuração de cluster. Com o aumento da quantidade K, houve a maior especificidade nas temáticas das ocupações de cada cluster, mas, para $K > 8$, essa compatibilidade temática começou a deteriorar, nos indicando um início de perda da capacidade de agrupamento de propriedades e ocupações em torno de temáticas mais amplas e passando a agrupar os clusters em torno de ocupações particulares, como o caso da ocupação de "motorista de ambulância", que relacionou diversas ocupações relacionadas a diferentes tipos de motoristas com habilidades e conhecimentos relacionados à área da saúde.

Notavelmente, os clusters não coincidiram com a hierarquia de grupos da Classificação Brasileira de Ocupações, mas tiveram sucesso em agregar ocupações em torno de temas comuns. Conseguimos entender os resultados numéricos do clustering FMN através da interpretação dos K protótipos, cada um representando um ponto principal entre todas as ocupações de um dado cluster. Isto esclarece porque os clusters conseguem agrupar ocupações de grandes grupos distintos da CBO e ao mesmo tempo manter uma coesão temática. Os protótipos frequentemente têm, entre suas propriedades mais importantes, algumas propriedades que são bastante relevantes em vários grandes grupos distintos da CBO. Também ajuda a clarificar alguns resultados inesperados, como o caso da associação entre ocupações nos campos da construção civil e dos esportes, que ocorreu porque estes campos compartilham propriedades relacionadas à coordenação motora; Ou ainda o caso da associação entre ocupações no campo da saúde e diversas outras ocupações em várias

outras áreas não relacionadas entre si nem à saúde, apenas porque todas apresentavam propriedades relacionadas a normas de saúde.

Os resultados do modelo de risco de automação foram analisados em combinação com os clusters do QBQ e as classificações hierárquicas da CBO. Analisando as distribuições dos riscos de automação por cluster, ainda que a distribuição geral seja enviesada para valores altos de risco (mais de 50% de risco de automação), após uma análise das distribuições condicionadas por clusters ou por grupos da CBO, pudemos distinguir perfis consistentemente associados aos maiores valores relativos de risco de substituição — ocupações geralmente relacionadas às áreas de produção industrial e serviços administrativos.

Analisando as correlações entre as propriedades e os valores de risco de automação, vários padrões puderam ser observados. Alguns grupos de propriedades foram consistentemente correlacionados com valores mais altos de automação, como as propriedades relacionadas a tarefas rotineiras, procedurais de planejamento e de monitoramento, enquanto outros grupos foram consistentemente correlacionados a valores mais baixos de risco de automação, como as habilidades relacionadas à comunicação social. Interessantemente, houve casos em que certas propriedades inverteram sua correlação com o risco de automação dependendo do contexto. Um exemplo disso foi que algumas propriedades relacionadas a habilidades físicas, como escuta e habilidades motoras, apresentaram correlação positiva com o risco de automação no cluster 3 (associado a metalurgia) mas isto inverte numa correlação negativa no cluster 1 (associado a trabalho em fábricas).

Concluímos que estes resultados podem trazer novas nuances para a ideia recorrente de que algumas habilidades ou conhecimentos possam ser consideradas, de uma maneira rígida, fatores limitadores do processo de substituição do trabalhador devido à automatização de tarefas. Talvez um exemplo que contradiga esta noção de uma propriedade como um limitador estrito da substituição dos trabalhadores seja a comunicação social, que é geralmente considerada como uma habilidade inerentemente humana e até mesmo apresentou uma correlação consistentemente negativa em relação às probabilidades de automação mas, apesar disso, ocupações como atendentes e televendas, que apresentam valores altos de importância em habilidades relacionadas à comunicação social, foram consistentemente relacionadas aos valores mais altos de risco de automação. Outro aspecto a ser considerado é que as propriedades em si podem ter significados diferentes dependendo do contexto de cada trabalho, como ilustrado pelos exemplos de percepção auditiva, que mostraram correlações positivas com os riscos de automação para ocupações ligadas à indústria e correlações negativas para as ocupações relacionadas ao comércio.

Como uma consideração final, deve-se notar que a base de dados do QBQ ainda não está completa, o que trará a necessidade de uma análise atualizada destes dados no futuro. É esperado que, com a adição de mais ocupações em diferentes campos de atuação

e mais propriedades sendo introduzidas, o número de clusters que pode ser identificável, mantendo-se as mesmas condições de coesão temática entre propriedades e ocupações, deve aumentar. Nesse caso, os procedimentos de seleção da quantidade de K e de avaliação de clusters ainda devem ser úteis para encontrar o equilíbrio entre a especificidade e a coesão interna de cada cluster.

Referências

ABADI, M. et al. Tensorflow: A system for large-scale machine learning. In: *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. [S.l.: s.n.], 2016. p. 265–283. ISBN 1-931971-33-1. Citado na página 57.

AI for Health. 2021. Https://www.microsoft.com/en-us/ai/ai-for-health. Citado na página 16.

ALBUQUERQUE, P. H. M. NA ERA DAS MÁQUINAS, O EMPREGO É DE QUEM? ESTIMAÇÃO DA PROBABILIDADE DE AUTOMAÇÃO DE OCUPAÇÕES NO BRASIL. *IPEA*, p. 40, 2019. Citado 2 vezes nas páginas 13 e 18.

ALPAYDIN, E. *Introduction to Machine Learning*. [S.l.]: MIT press, 2020. ISBN 0-262-04379-3. Citado 2 vezes nas páginas 29 e 30.

AUTOR, D. H.; LEVY, F.; MURNANE, R. J. The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics*, MIT Press, v. 118, n. 4, p. 1279–1333, 2003. Citado na página 17.

BAKHSHI, H. et al. *The Future of Skills: Employment in 2030*. United Kingdom: Pearson, 2017. ISBN 978-0-9924259-5-1. Citado 2 vezes nas páginas 13 e 41.

BENGIO, Y. et al. Greedy layer-wise training of deep networks. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2007. p. 153–160. Citado na página 32.

BISHOP, C. M. Pattern recognition. *Machine learning*, v. 128, n. 9, 2006. Citado 4 vezes nas páginas 29, 30, 31 e 33.

CBO. *Classificação Brasileira de Ocupações. Ministério Do Trabalho*. 2002. Http://www.mtecbo.gov.br/cbosite/pages/home.jsf. Citado 4 vezes nas páginas 13, 34, 35 e 36.

CBO. *Tábua de Conversão CBO2002 - CBO94 - CIUO88 - 5.1.2*. 2002. Http://www.mtecbo.gov.br/cbosite/pages/tabua/FiltroConversao_CBO2002_CBO94_CIUO88.jsf. Citado na página 42.

CEPAL. *Mudança estrutural para a igualdade: uma visão integrada do desenvolvimento*. [S.l.]: CEPAL, 2014. Citado 2 vezes nas páginas 13 e 15.

Chatbot Realiza 70% Dos Atendimentos Do BB Em Rede Social. 2018. Https://bb.com.br/pbb/pagina-inicial/imprensa/n/57066/#/. Citado na página 16.

DAVIES, D. L.; BOULDIN, D. W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, abr. 1979. ISSN 1939-3539. Citado 2 vezes nas páginas 24 e 42.

DeepMind health team joins Google Health. 2019. Https://deepmind.com/blog/announcements/deepmind-health-joins-google-health. Citado na página 16.

DING, C.; HE, X.; SIMON, H. D. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. In: *Proceedings of the 2005 SIAM International Conference on Data Mining (SDM)*. [S.l.]: Society for Industrial and Applied Mathematics, 2005, (Proceedings). p. 606–610. ISBN 978-0-89871-593-4. Citado na página 28.

DING, C.; LI, T.; PENG, W. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method. In: *AAAI*. [S.l.: s.n.], 2006. v. 42, p. 137–43. Citado na página 29.

FREY, C. B.; OSBORNE, M. A. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, v. 114, p. 254–280, jan. 2017. ISSN 00401625. Citado 8 vezes nas páginas 13, 17, 18, 21, 34, 40, 41 e 62.

GAUSSIER, E.; GOUTTE, C. Relation between PLSA and NMF and implications. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2005. (SIGIR '05), p. 601–602. ISBN 978-1-59593-034-7. Citado na página 29.

GÉRON, A. Hands-on machine learning with scikit-learn and tensorflow: Concepts. *Tools, and Techniques to build intelligent systems*, 2017. Citado na página 21.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT press, 2016. ISBN 0-262-33737-1. Citado 3 vezes nas páginas 29, 32 e 33.

HART, P. E.; STORK, D. G.; DUDA, R. O. *Pattern Classification*. [S.l.]: Wiley Hoboken, 2000. ISBN 0-471-05669-3. Citado 2 vezes nas páginas 22 e 30.

HEBB, D. O. *The Organization of Behavior: A Neuropsychological Theory*. [S.l.]: Psychology Press, 2005. ISBN 978-1-135-63190-1. Citado na página 32.

HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A fast learning algorithm for deep belief nets. *Neural computation*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 18, n. 7, p. 1527–1554, 2006. Citado na página 32.

HYVÄRINEN, A.; OJA, E. Independent component analysis: Algorithms and applications. *Neural Networks*, v. 13, n. 4, p. 411–430, jun. 2000. ISSN 0893-6080. Citado na página 23.

Instituto de Economia da Unicamp. *O Mundo Do Trabalho Na Pandemia: Precarização Do Emprego e Banalização Da Vida*. 2020. Citado na página 16.

ISCO. *ISCO - International Standard Classification of Occupations*. 2008. Https://www.ilo.org/public/english/bureau/stat/isco/isco08/. Citado na página 42.

JOLLIFFE, I. T. Principal components in regression analysis. In: *Principal Component Analysis*. [S.l.]: Springer, 1986. p. 129–155. Citado na página 22.

JUMPER, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, Nature Publishing Group, v. 596, n. 7873, p. 583–589, ago. 2021. ISSN 1476-4687. Citado na página 16.

LASI, H. et al. Industry 4.0. *Business & Information Systems Engineering*, v. 6, n. 4, p. 239–242, ago. 2014. ISSN 1867-0202. Citado na página 15.

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, Nature Publishing Group, v. 401, n. 6755, p. 788–791, out. 1999. ISSN 1476-4687. Citado 4 vezes nas páginas 23, 26, 27 e 28.

LI, T.; al, e. *Non-Negative Matrix Factorizations for Clustering: A Survey*. [S.l.: s.n.], 2014. Citado 2 vezes nas páginas 26 e 28.

LI, T.; DING, C. The relationships among various nonnegative matrix factorization methods for clustering. In: *Sixth International Conference on Data Mining (ICDM'06)*. [S.l.]: IEEE, 2006. p. 362–371. ISBN 0-7695-2701-9. Citado na página 28.

LUNA, I. *Transformaciones Productivas e Impactos Potenciales En Las Ocupaciones: Análisis de Escenarios Para El Caso Brasileño*. [S.l.], 2019. Citado na página 15.

MARR, B. *The Future of Lawyers: Legal Tech, AI, Big Data And Online Courts*. 2020. Https://www.forbes.com/sites/bernardmarr/2020/01/17/the-future-of-lawyers-legal-tech-ai-big-data-and-online-courts/. Citado na página 16.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943. Citado na página 32.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill, 1997. ISBN 978-0-07-115467-3. Citado 2 vezes nas páginas 20 e 21.

MITTELSTADT, B.; RUSSELL, C.; WACHTER, S. Explaining Explanations in AI. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, 2019. (FAT* '19), p. 279–288. ISBN 978-1-4503-6125-5. Citado na página 20.

MORGAN, B. *How Amazon Has Reorganized Around Artificial Intelligence And Machine Learning*. 2018. Https://www.forbes.com/sites/blakemorgan/2018/07/16/how-amazon-has-re-organized-around-artificial-intelligence-and-machine-learning/. Citado na página 16.

MURDOCH, W. J. et al. Interpretable machine learning: Definitions, methods, and applications. *Proceedings of the National Academy of Sciences*, v. 116, n. 44, p. 22071–22080, out. 2019. ISSN 0027-8424, 1091-6490. Citado na página 20.

OLICK, D. *Artificial Intelligence Is Taking over Real Estate – Here's What That Means for Homebuyers*. 2021. Https://www.cnbc.com/2021/09/17/what-artificial-intelligence-means-for-homebuyers-real-estate-market.html. Citado na página 16.

O*NET. *O*NET | U.S. Department of Labor*. 2019. Https://www.dol.gov/agencies/eta/onet. Disponível em: <https://www.dol.gov/agencies/eta/onet>. Citado 2 vezes nas páginas 13 e 17.

PAATERO, P.; TAPPER, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, v. 5, n. 2, p. 111–126, 1994. ISSN 1099-095X. Citado na página 26.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, n. 85, p. 2825–2830, 2011. ISSN 1533-7928. Citado na página 42.

QBQ. *Quadro Brasileiro de Qualificações. Ministério Da Economia.* 2020. Http://qbqconsulta.fipe.org.br/. Citado 5 vezes nas páginas 13, 18, 34, 38 e 39.

QEQ. *O Quadro Europeu de Qualificações (QEQ) | Europass.* 2017. Https://europa.eu/europass/pt/european-qualifications-framework-eqf. Citado na página 38.

RANZATO, M. et al. Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, MIT; 1998, v. 19, p. 1137, 2007. Citado na página 32.

ROKACH, L. A survey of Clustering Algorithms. In: MAIMON, O.; ROKACH, L. (Ed.). *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, 2010. p. 269–298. ISBN 978-0-387-09823-4. Citado 2 vezes nas páginas 23 e 24.

ROMANO, J. M. T. et al. *Unsupervised Signal Processing: Channel Equalization and Source Separation*. [S.l.]: CRC Press, 2018. ISBN 1-4200-1946-5. Citado na página 23.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958. Citado na página 32.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, nov. 1987. ISSN 0377-0427. Citado 3 vezes nas páginas 24, 25 e 42.

RUBINSTEIN, R.; BRUCKSTEIN, A. M.; ELAD, M. Dictionaries for Sparse Representation Modeling. *Proceedings of the IEEE*, v. 98, n. 6, p. 1045–1057, jun. 2010. ISSN 1558-2256. Citado na página 23.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, Nature Publishing Group, v. 1, n. 5, p. 206–215, maio 2019. ISSN 2522-5839. Citado na página 20.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *nature*, Nature Publishing Group, v. 323, n. 6088, p. 533–536, 1986. Citado na página 32.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, IBM, v. 3, n. 3, p. 210–229, 1959. Citado na página 20.

SELF-DRIVING Vehicles | One Hundred Year Study on Artificial Intelligence (AI100). 2016. Https://ai100.stanford.edu/2016-report/section-ii-ai-domain/transportation/self-driving-vehicles. Citado na página 16.

SENIOR, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature*, Nature Publishing Group, v. 577, n. 7792, p. 706–710, jan. 2020. ISSN 1476-4687. Citado na página 16.

SOC. *Standard Occupational Classification (SOC) System*. 2018. Https://www.bls.gov/soc/. Citado 2 vezes nas páginas 41 e 42.

SORZANO, C. O. S.; VARGAS, J.; MONTANO, A. P. A survey of dimensionality reduction techniques. *arXiv:1403.2877 [cs, q-bio, stat]*, mar. 2014. Citado 2 vezes nas páginas 22 e 23.

WALCH, K. *AI's Increasing Role In Customer Service*. 2019. Https://www.forbes.com/sites/cognitiveworld/2019/07/02/ais-increasing-role-in-customer-service/. Citado na página 16.

WANG, Y.-X.; ZHANG, Y.-J. Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge and Data Engineering*, v. 25, n. 6, p. 1336–1353, jun. 2013. ISSN 1558-2191. Citado 3 vezes nas páginas 26, 27 e 28.

WILLIAMS, C. K.; RASMUSSEN, C. E. *Gaussian Processes for Machine Learning*. [S.l.]: MIT press Cambridge, MA, 2006. v. 2. Citado na página 41.

XU, R.; WUNSCH, D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, v. 16, n. 3, p. 645–678, maio 2005. ISSN 1941-0093. Citado 2 vezes nas páginas 23 e 24.

Apêndices

APÊNDICE A – Tabelas de resumo dos clusters

Ocupações		Propriedades	
Código CBO	Título da ocupação	#	Nome da propriedade
311720	Preparador de tintas (fábrica de tecidos)	1	Tecnologia na área de gastronomia e alimentação
311725	Tingidor de couros e peles	2	Corte de calçados
766305	Acabador de embalagens (flexíveis e cartotécnicas)	3	Processos de acabamento gráfico
841408	Cozinhador (conservação de alimentos)	4	Matérias-primas e aditivos
841416	Cozinhador de carnes	5	Polímeros sintéticos
841420	Cozinhador de frutas e legumes	6	Moldagem e matrizes de polímeros
841428	Cozinhador de pescado		
841432	Desidratador de alimentos		
841440	Esterilizador de alimentos		
848325	Trabalhador de fabricação de sorvete		

Tabela A.1 – Cluster 1 (trabalho em fábrica).

Ocupações		Propriedades	
Código CBO	Título da ocupação	#	Nome da propriedade
621005	Trabalhador agropecuário em geral	1	Manutenção de equipamentos
622020	Trabalhador volante da agricultura	2	Máquinas e implementos agrícolas
622105	Trabalhador da cultura de arroz	3	Operação de máquinas e equipamentos agrícolas
622305	Trabalhador na olericultura (frutos e sementes)	4	Fitossanidade e proteção de plantas
622310	Trabalhador na olericultura (legumes)	5	Nutrição vegetal, adubos e corretivos
622315	Trabalhador na olericultura (raízes, bulbos e tubérculos)	6	Processos de armazenagem de produtos
622320	Trabalhador na olericultura (talos, folhas e flores)		
622505	Trabalhador no cultivo de árvores frutíferas		
622510	Trabalhador no cultivo de espécies frutíferas rasteiras		
622610	Trabalhador da cultura de café		

Tabela A.2 – Cluster 2 (agropecuária).

Ocupações		Propriedades	
Código CBO	Título da ocupação	#	Nome da propriedade
724220	Preparador de estruturas metálicas	1	Controle dimensional
724230	Rebitador, a mão	2	Processos de união mecânica
724415	Chapeador	3	Usinagem convencional
725005	Ajustador ferramenteiro	4	Soldagem
725010	Ajustador mecânico	5	Conformação à frio
725015	Ajustador mecânico (usinagem em bancada e em máquinas-ferramentas)	6	Ajustagem
725020	Ajustador mecânico em bancada		
725025	Ajustador naval (reparo e construção)		
821305	Operador de laminador		
821310	Operador de laminador de barras a frio		

Tabela A.3 – Cluster 3 (manufatura mecânica).

Ocupações		Propriedades	
Código CBO	Título da ocupação	#	Nome da propriedade
354125	Assistente de vendas	1	Serviços de atendimento ao cliente
354145	Vendedor pracista	2	Perfil de clientes
411005	Auxiliar de escritório	3	Rotinas de operações e serviços bancários
413205	Atendente de agência	4	Processo de venda
413225	Escriturário de banco	5	Gestão de relacionamento com o cliente (crm)
413230	Operador de cobrança bancária	6	Divulgação de produtos e serviços
422305	Operador de telemarketing ativo		
422310	Operador de telemarketing ativo e receptivo		
521105	Vendedor em comércio atacadista		
521110	Vendedor de comércio varejista		

Tabela A.4 – Cluster 4 (vendas e escritório).

Ocupações		Propriedades	
Código CBO	Título da ocupação	#	Nome da propriedade
711110	Canteiro	1	Condicionamento físico
711115	Destroçador de pedra	2	Organização esportiva
711215	Operador de máquina cortadora (minas e pedreiras)	3	Operação de máquinas de construção civil
711225	Operador de máquina perfuradora (minas e pedreiras)	4	Desmonte de rochas
715410	Operador de bomba de concreto	5	Monitoramento de estabilidade de rochas
717010	Operador de martelete	6	Completação de poços
717020	Servente de obras		
724105	Assentador de canalização (edificações)		
782205	Guincheiro (construção civil)		
783230	Bloqueiro (trabalhador portuário)		

Tabela A.5 – Cluster 5 (construção e indústria extrativa).

Ocupações		Propriedades	
Código CBO	Título da ocupação	#	Nome da propriedade
324105	Técnico em métodos eletrográficos em encefalografia	1	Anatomia
324110	Técnico em métodos gráficos em cardiologia	2	Boas práticas na área de saúde
324115	Técnico em radiologia e imagenologia	3	Normas regulamentadoras
324120	Tecnólogo em radiologia	4	Protocolos de trabalho em saúde
324130	Técnico em espirometria	5	Registros e informações em saúde
324135	Técnico em polissonografia	6	Fisiologia
324215	Citotécnico		
324220	Técnico em hemoterapia		
515210	Auxiliar de farmácia de manipulação		
515220	Auxiliar de laboratório de imunobiológicos		

Tabela A.6 – Cluster 6 (saúde).

Ocupações		Propriedades	
Código CBO	Título da ocupação	#	Nome da propriedade
410205	Supervisor de almoxarifado	1	Dinâmica das relações interpessoais
410220	Supervisor de controle patrimonial	2	Liderança
410230	Supervisor de orçamento	3	Planejamento da produção.1
710205	Mestre (construção civil)	4	Sistemas integrados de gestão empresarial (erp)
710210	Mestre de linhas (ferrovias)	5	Robotização nos serviços de entrega
710215	Inspetor de terraplenagem	6	Gerenciamento de recursos humanos
760105	Contramestre de acabamento (indústria têxtil)		
760110	Contramestre de fiação (indústria têxtil)		
760115	Contramestre de malharia (indústria têxtil)		
760120	Contramestre de tecelagem (indústria têxtil)		

Tabela A.7 – Cluster 7 (supervisor).

Ocupações		Propriedades	
Código CBO	Título da ocupação	#	Nome da propriedade
313105	Eletrotécnico	1	Acionamento elétrico
313110	Eletrotécnico (produção de energia)	2	Circuitos lógicos digitais
313125	Técnico de manutenção elétrica de máquina	3	Circuitos magnéticos
313205	Técnico de manutenção eletrônica	4	Conversão de energia
313210	Técnico de manutenção eletrônica (circuitos de máquinas com comando numérico)	5	Magnetismo
313215	Técnico eletrônico	6	Sistemas elétricos de potência
313220	Técnico em manutenção de equipamentos de informática		
313305	Técnico de comunicação de dados		
514310	Auxiliar de manutenção predial		
731180	Operador de linha de montagem (aparelhos eletrônicos)		

Tabela A.8 – Cluster 8 (eletricista e TI).

APÊNDICE B – Composição dos clusters em grupos da CBO

Grande Grupo		Subgrupo Principal	
Frequência	Código e título	Frequência	Código e título
20	7: Trabalhadores da produção de bens e serviços industriais	16	76: Trabalhadores nas indústrias têxtil, do curtimento, do ve...
		2	78: Trabalhadores de funções transversais
		2	71: Trabalhadores da indústria extrativa e da construção civil
13	8: Trabalhadores da produção de bens e serviços industriais	11	84: Trabalhadores da fabricação de alimentos, bebidas e fumo
		2	81: Trabalhadores em indústrias de processos contínuos e outr...
3	3: Técnicos de nivel médio	3	31: Técnicos de nível médio das ciências físicas, químicas, e...
1	9: Trabalhadores em serviços de reparação e manutenção	1	99: Outros trabalhadores da conservação, manutenção e reparação
1	5: Trabalhadores dos serviços, vendedores do comércio em loj...	1	51: Trabalhadores dos serviços

Tabela B.1 – Cluster 1 (trabalho em fábrica). 38 amostras.

Grande Grupo		Subgrupo Principal	
Frequência	Código e título	Frequência	Name
47	6: Trabalhadores agropecuários, florestais e da pesca	46	62: Trabalhadores na exploração agropecuária
		1	64: Trabalhadores da mecanização agropecuária e florestal

Tabela B.2 – Cluster 2 (agropecuária). 47 amostras.

Grande Grupo		Subgrupo Principal	
Frequência	Código e título	Frequência	Código e título
53	7: Trabalhadores da produção de bens e serviços industriais	45	72: Trabalhadores da transformação de metais e de compósitos
		3	77: Trabalhadores das indústrias de madeira e do mobiliário
		3	74: Montadores de aparelhos e instrumentos de precisão e musi...
		1	78: Trabalhadores de funções transversais
		1	71: Trabalhadores da indústria extrativa e da construção civil
7	9: Trabalhadores em serviços de reparação e manutenção	6	91: Trabalhadores em serviços de reparação e manutenção mecânica
		1	99: Outros trabalhadores da conservação, manutenção e reparação
7	8: Trabalhadores da produção de bens e serviços industriais	7	82: Trabalhadores de instalações siderúrgicas e de materiais ...
1	3: Técnicos de nivel médio	1	31: Técnicos de nível médio das ciências físicas, químicas, e...

Tabela B.3 – Cluster 3 (manufatura mecânica). 68 amostras.

Grande Grupo		Subgrupo Principal	
Frequência	Código e título	Frequência	Código e título
93	5: Trabalhadores dos serviços, vendedores do comércio em loj...	82	51: Trabalhadores dos serviços
		11	52: Vendedores e prestadores de serviços do comércio
65	4: Trabalhadores de serviços administrativos	34	41: Escriturários
		31	42: Trabalhadores de atendimento ao público
58	3: Técnicos de nivel médio	33	35: Técnicos de nivel médio nas ciências administrativas
		10	33: Professores leigos e de nível médio
		9	37: Técnicos em nivel médio dos serviços culturai²s, das comun...
		3	34: Técnicos de nível médio em serviços de transportes
		2	31: Técnicos de nível médio das ciências físicas, químicas, e...
		1	39: Outros técnicos de nível médio
10	7: Trabalhadores da produção de bens e serviços industriais	3	78: Trabalhadores de funções transversais
		3	76: Trabalhadores nas indústrias têxtil, do curtimento, do ve...
		3	71: Trabalhadores da indústria extrativa e da construção civil
		1	72: Trabalhadores da transformação de metais e de compósitos
2	9: Trabalhadores em serviços de reparação e manutenção	1	95: Polimantenedores
		1	91: Trabalhadores em serviços de reparação e manutenção mecânica
1	8: Trabalhadores da produção de bens e serviços industriais	1	82: Trabalhadores de instalações siderúrgicas e de materiais ...

Tabela B.4 – Cluster 4 (vendas e escritório). 229 amostras.

Grande Grupo		Subgrupo Principal	
Frequência	Código e título	Frequência	Código e título
127	7: Trabalhadores da produção de bens e serviços industriais	66	71: Trabalhadores da indústria extrativa e da construção civil
		21	77: Trabalhadores das indústrias de madeira e do mobiliário
		19	78: Trabalhadores de funções transversais
		16	72: Trabalhadores da transformação de metais e de compósitos
		5	76: Trabalhadores nas indústrias têxtil, do curtimento, do ve...
38	8: Trabalhadores da produção de bens e serviços industriais	17	82: Trabalhadores de instalações siderúrgicas e de materiais ...
		13	84: Trabalhadores da fabricação de alimentos, bebidas e fumo
		4	86: Operadores de produção, captação, tratamento e distribuiç...
		4	81: Trabalhadores em indústrias de processos contínuos e outr...
24	5: Trabalhadores dos serviços, vendedores do comércio em loj...	24	51: Trabalhadores dos serviços
10	9: Trabalhadores em serviços de reparação e manutenção	8	99: Outros trabalhadores da conservação, manutenção e reparação
		2	91: Trabalhadores em serviços de reparação e manutenção mecânica
7	6: Trabalhadores agropecuários, florestais e da pesca	4	63: Pescadores e extrativistas florestais
		2	64: Trabalhadores da mecanização agropecuária e florestal
		1	62: Trabalhadores na exploração agropecuária
6	3: Técnicos de nivel médio	6	37: Técnicos em nivel médio dos serviços culturais, das comun...

Tabela B.5 – Cluster 5 (construção e indústria extrativa). 212 amostras.

Grande Grupo		Subgrupo Principal	
Frequência	Código e título	Frequência	Código e título
68	3: Técnicos de nivel médio	29	32: Técnicos de nível médio das ciências biológicas, bioquími...
		12	31: Técnicos de nível médio das ciências físicas, químicas, e...
		7	34: Técnicos de nível médio em serviços de transportes
		6	39: Outros técnicos de nível médio
		6	35: Técnicos de nivel médio nas ciências administrativas
		4	37: Técnicos em nivel médio dos serviços culturais, das comun...
		3	30: Técnicos polivalentes
		1	33: Professores leigos e de nível médio
57	7: Trabalhadores da produção de bens e serviços industriais	21	72: Trabalhadores da transformação de metais e de compósitos
		18	78: Trabalhadores de funções transversais
		10	71: Trabalhadores da indústria extrativa e da construção civil
		5	76: Trabalhadores nas indústrias têxtil, do curtimento, do ve...
		2	74: Montadores de aparelhos e instrumentos de precisão e musi...
		1	73: Trabalhadores da fabricação e instalação eletroeletrônica
17	8: Trabalhadores da produção de bens e serviços industriais	8	81: Trabalhadores em indústrias de processos contínuos e outr...
		4	86: Operadores de produção, captação, tratamento e distribuiç...
		3	82: Trabalhadores de instalações siderúrgicas e de materiais ...
		2	84: Trabalhadores da fabricação de alimentos, bebidas e fumo
12	5: Trabalhadores dos serviços, vendedores do comércio em loj...	12	51: Trabalhadores dos serviços
6	9: Trabalhadores em serviços de reparação e manutenção	5	91: Trabalhadores em serviços de reparação e manutenção mecânica
		1	95: Polimantenedores
5	6: Trabalhadores agropecuários, florestais e da pesca	4	62: Trabalhadores na exploração agropecuária
		1	63: Pescadores e extrativistas florestais

Tabela B.6 – Cluster 6 (saúde). 165 amostras.

Grande Grupo		Subgrupo Principal	
Frequência	Código e título	Frequência	Código e título
58	3: Técnicos de nivel médio	18	31: Técnicos de nível médio das ciências físicas, químicas, e...
		11	35: Técnicos de nivel médio nas ciências administrativas
		10	34: Técnicos de nível médio em serviços de transportes
		8	39: Outros técnicos de nível médio
		7	37: Técnicos em nivel médio dos serviços culturais, das comun...
		4	32: Técnicos de nível médio das ciências biológicas, bioquími...
36	7: Trabalhadores da produção de bens e serviços industriais	17	72: Trabalhadores da transformação de metais e de compósitos
		11	76: Trabalhadores nas indústrias têxtil, do curtimento, do ve...
		5	71: Trabalhadores da indústria extrativa e da construção civil
		2	77: Trabalhadores das indústrias de madeira e do mobiliário
		1	73: Trabalhadores da fabricação e instalação eletroeletrônica
19	4: Trabalhadores de serviços administrativos	12	41: Escriturários
		7	42: Trabalhadores de atendimento ao público
12	5: Trabalhadores dos serviços, vendedores do comércio em loj...	12	51: Trabalhadores dos serviços
10	9: Trabalhadores em serviços de reparação e manutenção	8	91: Trabalhadores em serviços de reparação e manutenção mecânica
		2	99: Outros trabalhadores da conservação, manutenção e reparação
3	8: Trabalhadores da produção de bens e serviços industriais	2	86: Operadores de produção, captação, tratamento e distribuiç...
		1	84: Trabalhadores da fabricação de alimentos, bebidas e fumo
2	6: Trabalhadores agropecuários, florestais e da pesca	2	62: Trabalhadores na exploração agropecuária

Tabela B.7 – Cluster 7 (supervisor). 140 amostras.

Grande Grupo		Subgrupo Principal	
Frequência	Código e título	Frequência	Código e título
38	7: Trabalhadores da produção de bens e serviços industriais	29	73: Trabalhadores da fabricação e instalação eletroeletrônica
		5	72: Trabalhadores da transformação de metais e de compósitos
		3	71: Trabalhadores da indústria extrativa e da construção civil
		1	77: Trabalhadores das indústrias de madeira e do mobiliário
35	3: Técnicos de nivel médio	32	31: Técnicos de nível médio das ciências físicas, químicas, e...
		2	39: Outros técnicos de nível médio
		1	30: Técnicos polivalentes
20	9: Trabalhadores em serviços de reparação e manutenção	12	91: Trabalhadores em serviços de reparação e manutenção mecânica
		8	95: Polimantenedores
6	8: Trabalhadores da produção de bens e serviços industriais	4	86: Operadores de produção, captação, tratamento e distribuiç...
		2	81: Trabalhadores em indústrias de processos contínuos e outr...
2	5: Trabalhadores dos serviços, vendedores do comércio em loj...	2	51: Trabalhadores dos serviços

Tabela B.8 – Cluster 8 (eletricista e TI). 101 amostras.

APÊNDICE C – Correlação entre propriedades e riscos de automação por cluster

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.83	Aplicação de instruções simples e rotineiras	-0.73	Software - cad-cam
0.55	Coordenação corporal bruta	-0.73	Corte de calçados
0.48	Percepção de profundidade	-0.64	Estabilidade (firmeza) braço-mão
0.40	Controle da qualidade	-0.55	Precisão de controle
0.36	Processos de embalagem e etiquetagem	-0.51	Processos de preparação de confecção de calçados
0.36	Controle da produção	-0.50	Troca de informações
0.36	Alimentação de máquinas e equipamentos	-0.44	Velocidade de movimento dos membros
0.36	Produção em batelada	-0.44	Escuta ativa
0.32	Força estática	-0.42	Aplicação de princípios tecnológicos de baixa complexidade
0.30	Legislação aplicada à área ocupacional	-0.41	Comparação de dados

Tabela C.1 – Cluster 1 (trabalho em fábrica).

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.65	Aplicação de instruções simples e rotineiras	-0.84	Atendimento a solicitações e pedidos das pessoas...
0.62	Velocidade de movimento dos membros	-0.71	Manejo agrícola
0.57	Nutrição animal	-0.69	Irrigação
0.57	Produção animal confinada	-0.63	Georreferenciamento rural
0.56	Manejo e tratamento animal	-0.63	Reconhecimento de fala
0.56	Zootecnia	-0.63	Clareza de fala
0.55	Bem-estar animal	-0.60	Sistemas de parceria da produção
0.55	Zoologia	-0.57	Sistemas de integração da produção
0.51	Produção animal extensiva	-0.55	Mecanização agrícola
0.51	Reprodução animal	-0.53	Drenagem

Tabela C.2 – Cluster 2 (agropecuária).

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.56	Escuta ativa	-0.78	Medidas elétricas
0.39	Visão a distância	-0.68	Estatística
0.37	Saúde ocupacional	-0.68	Probabilidade
0.30	Controle da qualidade	-0.68	Ótica
0.29	Coordenação multimembros	-0.63	Testes
0.28	Raciocínio concreto para seguir instruções	-0.59	Mecânica de precisão
0.27	Tempo de reação (ou de resposta)	-0.58	Classificação de dados
0.25	Taxa de controle	-0.57	Garantia de qualidade
0.25	Velocidade de pulso-dedos	-0.51	Elementos de máquinas
0.24	Compreensão oral	-0.49	Programação da manutenção

Tabela C.3 – Cluster 3 (manufatura mecânica).

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.35	Processos administrativos	-0.35	Análise de riscos (habilidade)
0.32	Planejamento do trabalho	-0.33	Cuidado às pessoas (care)
0.29	Rotinas financeiras	-0.31	Limpeza e conservação doméstica
0.28	Rotinas contábeis	-0.31	Tempo de reação (ou de resposta)
0.25	Matemática financeira	-0.30	Sensibilidade auditiva
0.24	Rotinas de operações e serviços bancários	-0.30	Tecnologias de defesa e segurança
0.24	Informática	-0.29	Segurança no trabalho
0.21	Normas regulamentadoras em serviços financeiros	-0.29	Segurança e ordem pública
0.21	Normas técnicas em serviços financeiros	-0.29	Sensibilidade ao brilho
0.20	Processos organizacionais	-0.29	Apoio às atividades da vida diária

Tabela C.4 – Cluster 4 (vendas e escritório).

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.41	Análise de garantia de qualidade	-0.59	Reconhecimento de fala
0.36	Raciocínio concreto para seguir instruções	-0.57	Exercícios físicos
0.35	Normas regulamentadoras para área industrial	-0.56	Anatomia
0.32	Planejamento do trabalho	-0.55	Clareza de fala
0.32	Normas técnicas para área industrial	-0.55	Condicionamento físico
0.31	Controle da qualidade	-0.53	Estatística
0.28	Mecânica	-0.52	Rendimento esportivo
0.27	Eletricidade	-0.52	Esportes e jogos
0.27	Controle dimensional	-0.51	Regras esportivas
0.26	Aplicação de instruções simples e rotineiras	-0.47	Resposta de orientação

Tabela C.5 – Cluster 5 (construção e indústria extrativa).

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.47	Análise de garantia de qualidade	-0.49	Reconhecimento de fala
0.42	Segurança no trabalho	-0.47	Clareza de fala
0.36	Normas regulamentadoras para área industrial	-0.44	Informática médica
0.34	Aplicação de instruções simples e rotineiras	-0.42	Boas práticas na área de saúde
0.33	Normas técnicas para área industrial	-0.41	Imunologia celular
0.33	Português: comunicação oral e escrita	-0.41	Medidas de higiene e assepsia
0.32	Controle estatístico de processo (cep)	-0.40	Anatomia
0.31	Controle da produção	-0.40	Assistência de enfermagem em ambientes especializados
0.30	Tratamento de superfície	-0.40	Autonomia em contextos de trabalho com previsão de mudanças
0.29	Materiais metálicos ferrosos	-0.39	Português: comunicação e expressão

Tabela C.6 – Cluster 6 (saúde).

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.50	Português: leitura e interpretação de textos	-0.42	Raciocínio concreto para seguir instruções
0.45	Processo decisório	-0.40	Visão periférica
0.40	Sistemas de informação	-0.38	Elaboração de projeto
0.40	Rotinas contábeis	-0.36	Primeiros socorros
0.35	Rotinas financeiras	-0.35	Discriminação de cor visual
0.35	Normas regulamentadoras relativas à saúde, segurança e higiene no trabalho	-0.34	Sensibilidade auditiva
0.32	Matemática financeira	-0.34	Análise de riscos (habilidade)
0.32	Gestão da qualidade	-0.33	Espaços confinados
0.31	Normas técnicas em serviços financeiros	-0.33	Avaliação de projeto
0.31	Normas regulamentadoras em serviços financeiros	-0.32	Trabalho em altura

Tabela C.7 – Cluster 7 (supervisor).

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.59	Montagem eletrônica	-0.48	Reconhecimento de fala
0.46	Circuitos eletrônicos	-0.46	Força dinâmica
0.43	Calibração	-0.46	Orçamento
0.39	Normas regulamentadoras para área industrial	-0.44	Clareza de fala
0.38	Desenho técnico geral	-0.38	Comparação de dados
0.35	Normas técnicas para área industrial	-0.38	Banco de dados
0.34	Sensibilidade auditiva	-0.37	Lógica matemática
0.33	Aplicação de instruções simples e rotineiras	-0.37	Inteligência artificial
0.30	Eletricidade	-0.37	Coordenação multimembros
0.29	Eletrônica	-0.36	Big data

Tabela C.8 – Cluster 8 (eletricista e TI).

APÊNDICE D – Correlação entre propriedades e riscos de automação por grande grupo da CBO

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.32	Rotinas financeiras	-0.40	Equilíbrio corporal bruto
0.32	Matemática financeira	-0.39	Força dinâmica
0.31	Internet	-0.38	Força de tronco
0.29	Seguros	-0.36	Coordenação multimembros
0.29	Valores	-0.34	Tempo de reação (ou de resposta)
0.28	Informática	-0.33	Serviços de vacinação
0.28	Planejamento do trabalho	-0.33	Medidas de higiene e assepsia
0.28	Rotinas contábeis	-0.33	Vigor (estamina)
0.28	Rotinas de operações e serviços bancários	-0.33	Ações preventivas
0.27	Medidas elétricas	-0.32	Precisão de controle

Tabela D.1 – Grande grupo 3: Técnicos de nivel médio

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.48	Planejamento do trabalho	-0.51	Estabilidade (firmeza) braço-mão
0.41	Rotinas contábeis	-0.42	Discriminação de cor visual
0.31	Internet	-0.41	Aplicação de questionários
0.30	Português: comunicação e expressão	-0.40	Técnicas de pesquisa
0.30	Análise de dados	-0.39	Elaboração de roteiro de entrevista
0.30	Aplicação de instruções simples e rotineiras	-0.37	Entrevista
0.30	Processos administrativos	-0.37	Transcrição de dados
0.29	Clareza de fala	-0.36	Atenção auditiva
0.28	Informática	-0.36	Elaboração de questionário
0.28	Português: leitura e interpretação de textos	-0.31	Relações públicas

Tabela D.2 – Grande grupo 4: Trabalhadores de serviços administrativos

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.34	Aplicação de instruções simples e rotineiras	-0.53	Tempo de reação (ou de resposta)
0.33	Análise de garantia de qualidade	-0.46	Sensibilidade auditiva
0.29	Trabalho sob supervisão direta	-0.44	Sensibilidade ao brilho
0.26	Seleção de materiais (utensílios e equipamentos) de alimentação	-0.43	Desenvolvimento comportamental
0.25	Procedimentos de mise em place	-0.41	Ações emergenciais e de urgência
0.25	Técnicas de armazenamento de alimentos e bebidas	-0.41	Localização de som
0.24	Segurança alimentar	-0.38	Segurança e ordem pública
0.24	Divulgação de produtos e serviços	-0.35	Tecnologias de defesa e segurança
0.23	Processo de venda	-0.34	Análise de riscos (habilidade)
0.23	Informática	-0.33	Ações preventivas

Tabela D.3 – Grande grupo 5: Trabalhadores dos serviços, vendedores do comércio em lojas e...

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.56	Aplicação de instruções simples e rotineiras	-0.76	Atendimento a solicitações e pedidos das pessoas...
0.56	Velocidade de movimento dos membros	-0.71	Manejo agrícola
0.53	Zootecnia	-0.70	Irrigação
0.52	Zoologia	-0.58	Georreferenciamento rural
0.52	Manejo e tratamento animal	-0.57	Sistemas de integração da produção
0.51	Bem-estar animal	-0.56	Sistemas de parceria da produção
0.49	Nutrição animal	-0.56	Mecanização agrícola
0.48	Seleção e melhoramento dos animais	-0.55	Drenagem
0.48	Genética	-0.55	Reconhecimento de fala
0.48	Bioética	-0.55	Clareza de fala

Tabela D.4 – Grande grupo 6: Trabalhadores agropecuários, florestais e da pesca

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.41	Normas regulamentadoras para área industrial	-0.45	Clareza de fala
0.36	Normas técnicas para área industrial	-0.38	Reconhecimento de fala
0.30	Mecânica	-0.38	Trabalho em altura
0.28	Desenho técnico geral	-0.35	Instalações elétricas
0.28	Controle dimensional	-0.34	Segurança no trânsito
0.24	Transcrição de dados	-0.34	Visão noturna
0.23	Materiais metálicos ferrosos	-0.34	Normas técnicas em energia
0.23	Análise de garantia de qualidade	-0.33	Automação na área de transportes
0.23	Montagem eletrônica	-0.33	Geografia relacionada à área ocupacional
0.22	Lubrificação	-0.33	Direção defensiva

Tabela D.5 – Grande grupo 7: Trabalhadores da produção de bens e serviços industriais

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.38	Segurança no trabalho	-0.60	Classificação de dados
0.36	Compreensão numérica	-0.60	Terminais aeroviários
0.36	Força estática	-0.53	Raciocínio sintético (capacidade de diagnóstico de problemas ou troubleshooting)
0.31	Destreza manual	-0.53	Inglês instrumental
0.28	Expressão numérica	-0.53	Planejamento da produção
0.28	Ergonomia	-0.53	Química física
0.26	Raciocínio concreto para seguir instruções	-0.53	Química geral
0.25	Compreensão oral	-0.53	Cinética química
0.24	Trabalho sob supervisão direta, com alguma autonomia	-0.53	Química analítica
0.24	Avaliação do próprio desempenho, com alguma orientação	-0.53	Eletroeletrônica

Tabela D.6 – Grande grupo 8: Trabalhadores da produção de bens e serviços industriais

Correlação mais positiva		Correlação mais negativa	
Correlação	Propriedade	Correlação	Propriedade
0.45	Visão a distância	-0.64	Montagem elétrica
0.43	Trabalho sob supervisão direta	-0.63	Inspeção de manutenção
0.37	Velocidade de pulso-dedos	-0.55	Circuitos elétricos
0.37	Aplicação de instruções simples e rotineiras	-0.53	Medidas elétricas
0.35	Aritmética	-0.52	Tipos de manutenção
0.32	Localização de som	-0.50	Instrumentação
0.31	Flexibilidade dinâmica	-0.50	Eletroeletrônica
0.31	Atenção auditiva	-0.47	Circuitos lógicos digitais
0.31	Tratamento de superfície	-0.47	Proteção de sistemas elétricos
0.27	Força de tronco	-0.47	Medidas eletrônicas

Tabela D.7 – Grande grupo 9: Trabalhadores em serviços de reparação e manutenção