UNIVERSIDADE FEDERÁL DO CEARÁ

CAMPUS DE QUIXADÁ

CURSO DE GRADUAÇÃO EM ENGENHARIA DE SOFTWARE

ÍTALO LIMA DANTAS

SISTEMA DE RECOMENDAÇÃO DE BENEFÍCIOS PARA FAMÍLIAS INSCRITAS EM PROGRAMAS SOCIOASSISTENCIAIS NO ESTADO DO CEARÁ

QUIXADÁ

2022

ÍTALO LIMA DANTAS

SISTEMA DE RECOMENDAÇÃO DE BENEFÍCIOS PARA FAMÍLIAS INSCRITAS EM PROGRAMAS SOCIOASSISTENCIAIS NO ESTADO DO CEARÁ

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em ENGENHARIA DE SOFTWARE do CAMPUS DE QUIXADÁ da UNIVERSIDADE FEDERÁL DO CEARÁ, como requisito parcial à obtenção do grau de bacharel em ENGENHARIA DE SOFTWARE.

Orientador: Prof. Dr. Criston Pereira de Souza.

QUIXADÁ

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

D216s Dantas, Ítalo Lima.
Sistema de recomendação de benefícios para famílias inscritas em programas socioassistenciais no Estado do Ceará / Ítalo Lima Dantas. – 2022.
60 f. : il.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Engenharia de Software, Quixadá, 2022.
Orientação: Prof. Dr. Criston Pereira de Souza.

1. Inteligência Artificial. 2. Sistemas de recomendação (filtragem de informações). 3. Programas Sociais. 4. Política pública. I. Título.

CDD 005.1

ÍTALO LIMA DANTAS

SISTEMA DE RECOMENDAÇÃO DE BENEFÍCIOS PARA FAMÍLIAS INSCRITAS EM PROGRAMAS SOCIOASSISTENCIAIS NO ESTADO DO CEARÁ

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em ENGENHARIA DE SOFTWARE do CAMPUS DE QUIXADÁ da UNIVERSIDADE FEDERÁL DO CEARÁ, como requisito parcial à obtenção do grau de bacharel em ENGENHARIA DE SOFTWARE.

Aprovada em: ____/ ____/ ____.

BANCA EXAMINADORA

Prof. Dr. Criston Pereira de Souza (Orientador)
Universidade Federal do Ceará (UFC)

Profa. Dra. Ticiana Linhares Coelho da Silva
Universidade Federal do Ceará (UFC)

Prof. Dr. Regis Pires Magalhães
Universidade Federal do Ceará (UFC)

Agradeço primeiramente a Deus, por me abençoar e proteger, sinto que a minha luz brilha mais forte porque você olhou diferente para mim, aí de cima. Muitíssimo obrigado a minha família, namorada e amigos. A minha mãe, em especial, toda a gratidão e reconhecimento, por ser a maior responsável e incentivadora de todas às etapas educacionais da minha vida.

AGRADECIMENTOS

Inicialmente, agradeço a Deus, pela saúde, sabedoria e discernimento. Durante o período que frequentei a faculdade, coincidentemente ou não, fui menos frequente na Igreja, mas nunca me senti distante de ti, nunca me senti sozinho e sempre que conversei com o senhor, você me acalmou ou me atendeu.

Posteriormente, agradeço aos meus pais, Josefa Rejane de Lima e Francisco Willamy de Castro Dantas, vocês sempre me incentivaram e me proporcionaram as melhores condições de vida e estudo, esse é o mínimo que posso oferecer de retorno. Em especial, assim como na dedicatória, agradeço a minha mãe, sem ela nada disso seria possível, nem muito do que aconteceu antes disso. Mando lembranças ao meu irmão, João Otávio e aos meus avós, maternos e paternos, amo vocês.

Agradeço também a mim, por sempre acreditar e não desistir daquilo que almejo, neste período, assim como na vida, fui o meu maior crítico e o meu maior fã. Passei por momentos extremamente desgastantes e estressantes, que exigiram tudo de mim, mas quase sempre fui recompensado na mesma proporção. Esta foi uma trajetória repleta de sucesso, desde o seu início, até o final, onde tive o prazer de vivenciar experiências voltadas à academia e pesquisa, assim como no mercado de trabalho. O meu esforço incondicional foi um diferencial marcante para isso.

Agradeço a minha namorada, Dara, por toda a paciência que teve comigo, imagino o quão difícil foi aguentar todas as reclamações, mas você sempre me orientou da melhor forma. Obrigado por me acalmar nos momentos mais aflitos, e me fazer enxergar a situação com outros olhos, isso foi essencial para sair de problemas que pareciam sem solução.

Meus agradecimentos aos meus amigos mais próximos de Solonópole, cujos tenho amizade desde a infância, sei que vocês torcem bastante por mim. Agradeço também aos meus colegas de curso, em especial, ao Marcos Gênesis, Fabrício Pinheiro, Lucas Nascimento, Gustavo Colombo, Jeferson Gonçalves e Eric Rodrigues. Uma lembrança a outros colegas cujos dividi o início dessa jornada, Andson Silva, Matheus Felipe, Michel Sales, Cristiano Júnior, Miguel Neto, João Ygo, etc. Agradeço também ao meu líder, Felipe Alves, por toda compreensão e apoio nos momentos em que as rotinas mais conflitaram. Por último, agradeço aos meus amigos da PTF, aprendi muito com vocês, obrigado pelos ensinamentos.

Agradeço ao Prof. Dr. Criston Pereira de Souza por ter me aceitado como orientando no Trabalho de Conclusão de Curso (TCC) e me guiado ao longo de dois semestres. Agradeço

aos professores Prof. Dr. Régis Pires Magalhães e Profa. Dra. Ticiana Linhares Coelho da Silva por aceitarem participar da banca avaliadora deste trabalho e por colaborarem também durante o meu período de bolsista no *Insight Lab*, o meu tema é fruto de curiosidade e vontade de colaborar com ações que apoiam os programas socioassistenciais do Ceará.

“O mundo está cheio de pessoas talentosas que fracassaram. Talento não é tão importante. O que realmente importa é o quão dedicado você é ao seu ofício. Ser disciplinado é uma coisa, mas ser dedicado é um jogo diferente.”
(Cristiano Ronaldo)

RESUMO

Atualmente, os programas socioassistenciais no estado do Ceará realizam a consulta na base de dados do Cadastro Único (CadÚnico) como o instrumento exclusivo para verificação da aptidão de uma família para a inscrição em programas de transferência de renda. A busca por otimização no fornecimento de políticas públicas para o cidadão é uma temática altamente relevante, dentro do contexto de um governo digital. No entanto, a imensidão e concentração dos dados dos cidadãos é algo que dificulta a identificação das informações certas para as pessoas certas, e consequentemente, a assertividade dos serviços. Para tentar sanar tais problemas, a construção de sistemas de recomendação é uma estratégia amplamente usada. Diante disso, este trabalho objetivou a criação de um sistema de recomendação de benefícios, para famílias inseridas em programas socioassistenciais no Ceará. A partir dos dados coletados por uma ação do Programa Cartão Mais Infância Ceará (CMIC), realizou-se a construção de um modelo de aprendizado de máquina, que utilizou as informações de 30.889 famílias e é capaz fornecer a probabilidade da família possuir determinado benefício, o que pode ser utilizado como guia para a decisão sobre a concessão dele. Através do presente trabalho, foi possível obter resultados em seis seções distintas: pré-processamento dos dados, engenharia de recursos, seleção de características, avaliação de desempenho dos modelos no conjunto de treino, avaliação de desempenho dos modelos no conjunto de teste e avaliação do sistema de recomendação. Conclui-se através deste, que existe um grande potencial na utilização dos dados, extraídos dos formulários CMIC, para otimizar a entrega de benefícios às famílias inseridas nos programas. Além disso, o modelo criado consegue recomendar um benefício específico, para novos registros, ou registros desconhecidos por ele. Essa recomendação, no cenário atual, serve, inclusive, para comparar as concessões já existentes com as sugestões de recomendação.

Palavras-chave: aprendizado de máquina; sistema de recomendação; programas socioassistenciais; políticas públicas

ABSTRACT

Currently, social assistance programs in the state of Ceará consult the Unified Registry (CadÚnico) database the Cadastro Único (Single Registry) database as the sole instrument to verify the suitability of a family for enrollment in cash transfer programs. The search for optimization in the provision of public policies for the citizen is a highly relevant theme within the context of a digital government. However, the immensity and concentration of citizen data is something that makes it difficult to identify the right information for the right people, and consequently, the assertiveness of services. To try to solve these problems, the construction of recommendation systems is a widely used strategy. Therefore, this work aimed to create a benefit recommendation system for families enrolled in social assistance programs in the state of Ceará. From the data collected by an action of the Mais Infância Ceará Card Program (CMIC), a machine learning model was built, which used the information from 30,889 families and is able to provide the probability of the family having a certain benefit, which This can be used as a guide for the decision about granting it. Through this work, it was possible to obtain results in six distinct sections: data preprocessing, feature engineering, feature selection, performance evaluation of the models on the training set, performance evaluation of the models on the test set, and evaluation of the recommendation system. It is concluded through this, that there is a great potential in using the data, extracted from the CMIC forms, to optimize the delivery of benefits to families in the programs. Furthermore, the model created is able to recommend a specific benefit, for new records, or records unknown to it. This recommendation, in the current scenario, even serves to compare the existing concessions with the suggested recommendations.

Keywords: machine learning; recommendation system; social assistance programs; public policies

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Vivemos na época do *Big Data*, denominação utilizada para externar a imensa quantidade de dados produzida com grande velocidade e em volumes crescentes. Isto pode ser atribuído a vários fatores, mas um crucial é o aumento substancial do poderio computacional, que promove uma melhor infraestrutura para o armazenamento e o processamento destes dados. Além disso, há a hiperconectividade, o estado de alta disponibilidade digital, e a integração entre sistemas, plataformas e serviços que também contribuem significativamente para a interconexão e consequentemente, com o crescente volume de dados produzidos e armazenados.

Porém, o processo não se limita somente a produção e armazenamento dos dados, o *Data Driven*, termo que adjetiva processos orientados por dados, está se tornando cada vez mais presente em empresas de todos os portes, corporações e instituições governamentais. Além disso, várias *startups* tem sido criadas para prover a aplicação do *Data Driven*, como a empresa brasileira *AI ROBOTS* [1], que apesar de recente no mercado, é um modelo no gerenciamento de projetos industriais, e, além de aplicar a cultura para a gerir as informações de forma inteligente, oferece isso como um de seus serviços. Essa cultura citada promove um aumento na utilização potencial dos dados, assim, a tomada de decisões é moldada pela exploração dessas informações coletadas.

Os benefícios do *Data Driven*, no entanto, não se restringem à indústria. A pesquisa científica também é beneficiada, como visto em Lusher *et al.* (2014), onde há uma discussão sobre como a química medicinal pode usufruir das técnicas citadas e que os métodos e abordagens desenvolvidos em um campo podem ter aplicações potenciais em outros que não se relacionam diretamente. Baseando-se nessa utilização dos dados em diversos âmbitos para potencializar as informações produzidas e transformá-las em decisões, mais aplicações vão sendo desenvolvidas e novos setores explorados. Um exemplo considerado recente é o setor público, que pode utilizar esses *insights* obtidos para transformar o relacionamento com os seus cidadãos. Alguns desses métodos são o direcionamento de políticas públicas para o atendimento das necessidades populares e a otimização no uso dos recursos em geral.

Uma maneira de abranger às duas metodologias anteriormente citadas, é realizar recomendações otimizadas, seja de ações e políticas públicas, ou de recursos. Isso pode ser feito com a construção de sistemas de recomendação, cujos foram criados ainda na última década do século XX. Atualmente, estes sistemas oferecem grandes oportunidades e desafios

[1] https://airobots.com.br/

para negócios, governos, educação e outros domínios, e podem ser agrupados em 8 categorias: *e-government*, *e-business*, *e-commerce*/*e-shopping*, *e-library*, *e-learning*, *e-tourism*, serviços de recursos eletrônicos e atividades de grupos eletrônicos (LU *et al.*, 2015). Neste trabalho, o foco é no cenário de *E-government*, expressão usada para denotar um governo digital e definido como serviços ao cidadão, reengenharia com tecnologia ou buscas na Internet (TAMBOURIS *et al.*, 2001). Na prática, é um governo que disponibiliza o acesso a serviços e informações utilizando soluções tecnológicas e busca melhorar a interação do cidadão com o Estado, além de otimizar o fornecimento de políticas públicas para o cidadão.

Todo esse conjunto de ações, atores e infraestrutura moderna caracterizam o *GovTech*, um fenômeno mundial envolto de inovação e transformação digital que vem se consolidando, como aponta o estudo completo produzido pela Organização das Nações Unidas (ONU). Como um dos resultados desse movimento, em 2021, o Brasil foi reconhecido pelo Grupo Banco Mundial (WBG) como o 7º país com a mais alta maturidade em Governo Digital, dentre 198 países. Para esta maturidade, foi utilizado o Índice de Maturidade GovTech (GTMI), que mede os principais aspectos de quatro áreas de foco de um governo tecnológico: suporte aos principais sistemas governamentais, melhoria da prestação de serviços, integração do envolvimento do cidadão e promoção de capacitadores da *GovTech* (DENER HUBERT NII-APONSAH; JOHNS, 2021).

Atualmente, os programas socioassistenciais no estado do Ceará realizam a consulta na base de dados do programa Cadastro Único (CadÚnico) como o instrumento exclusivo para a verificação da aptidão de uma família para a inscrição em um destes programas. O CadÚnico é um instrumento que visa identificar todas as famílias brasileiras em situação de pobreza e extrema pobreza existentes no país, para fins de inclusão em programas socioassistenciais e de transferência de renda. Assim, é possível afirmar que a tecnologia da informação está presente nessa esfera governamental, com ênfase no registro de informações de seus cidadãos. No entanto, no cenário de governos eletrônicos, a imensidão e concentração destes dados foram vistos como prejudiciais para a eficácia dos serviços, trazendo dificuldades na identificação das informações certas para os usuários certos. Para tentar sanar tais problemas, os governos adotam os sistemas de recomendação (GUO; LU, 2007).

Dessa forma, os dados dos cidadãos podem ser ainda melhores aproveitados pelo próprio CadÚnico, onde, ao invés de ser possível apenas consultá-los individualmente, em um cenário de análise para aprovação de um benefício, pode-se utilizá-los como entrada para

modelos de recomendação, que não serão responsáveis por automatizar o processo, mas, servir como ferramenta de apoio às decisões dos gestores, proporcionando maior eficácia na busca e inclusão de famílias, assim como a concessão de benefícios e na verificação de fraudes.

Baseando-se no que foi supracitado e nas lacunas observadas pela subutilização dos dados dos cidadãos, o objetivo deste trabalho é criar um sistema de recomendação de benefícios para famílias inseridas em programas socioassistenciais no Ceará, com base nos dados dessas famílias. O sistema fornecerá a probabilidade da família possuir determinado benefício, o que poderá ser utilizado como guia para a decisão sobre a concessão dele. Tal proposta não pretende automatizar todo o fluxo atual, devido ao entendimento crítico e burocrático do âmbito, por se tratar de recursos públicos. Entretanto, as recomendações podem apoiar o direcionamento de políticas públicas e promover melhores e mais rápidas decisões baseadas nos dados.

1.1 Organização

Este trabalho está organizado da seguinte forma: no Capítulo 2 são apresentados os termos e conceitos necessários para o entendimento do trabalho. No Capítulo 3 são apresentados trabalhos relacionados com o presente trabalho. O Capítulo 4 contém os procedimentos metodológicos que serão seguidos durante a realização deste. O Capítulo 5 apresenta os resultados e a discussão deles, a partir dos experimentos realizados. Por fim, o capítulo 6 conclui o presente trabalho e elicita pontos que podem ser melhorados ou realizados em cenários futuros.

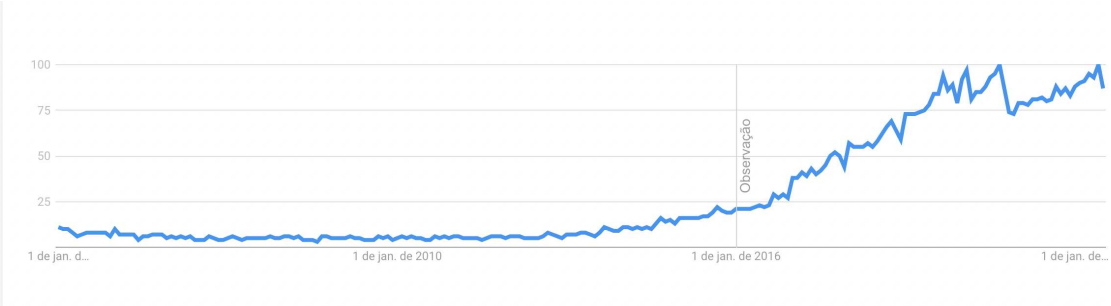
2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os conceitos presentes no embasamento teórico que possuem maior importância para entender o estudo e a disposição das etapas que compõem o procedimento metodológico. Na Seção 2.1 será introduzida a área de Ciência de dados e exemplificado como sua aplicação tem impacto nesta categoria de estudo. Na Seção 2.2 será apresentado os conceitos e etapas importantes de *Machine Learning*. Na Seção 2.3, os sistemas de recomendação são definidos e exemplos de aplicações e técnicas introduzidos. Em 2.4 e 2.5 serão exibidas técnicas e práticas usadas na construção do sistema de recomendação, como etapas de engenharia de recursos e o modelo de probabilidades. Por último, na Seção 2.7 será definido o que é *E-government* e exibido qual o cenário atual, no qual a área está inserida.

2.1 *Ciência de dados*

A Ciência de Dados pode ser definida primordialmente como a transformação de dados, usando matemática e estatística, em *insights*, decisões e produtos valiosos (FOREMAN *et al.*, 2014). No entanto, esta pode ser vista como uma área interdisciplinar entre a estatística e a ciência da computação, com foco aplicável de métodos científicos voltados para análise de dados em diversos âmbitos distintos. Nos últimos anos, esse termo sofreu uma grande ascensão, o que pode ser evidenciado pelo fato de que as empresas procuram cada vez mais por cientistas de dados e as instituições acadêmicas estão se esforçando para montar programas de ciência de dados, além das próprias publicações colocarem a profissão como *sexy* (PROVOST; FAWCETT, 2013). O gráfico exibido na Figura 1 torna explícita a crescente na busca pelo termo *Data Science*, nos últimos 6 anos, o que pode significar um dos indícios para o aumento do número de estudos e pesquisas acadêmicas em torno da área. Tais fatos servem como motivação para o trabalho proposto.

Figura 1 – Resultado do Google trends para o termo “Data Science”.



Fonte: elaborado pelo autor

2.2 Aprendizado de Máquina

Machine Learning (ML) é uma subárea da Inteligência Artificial (IA) que reúne "um conjunto de métodos que permitem aos computadores aprender com os dados para fazer e melhorar predições" (ALPAYDIN, 2020). Esta área tem como fundamentação a própria definição do problema de aprendizagem geral, proposta por Tom Mitchell, em que ele diz que

> "Um programa de computador aprende com a experiência E em relação a alguma classe de tarefas T e medida de desempenho P, se seu desempenho na tarefa T, medido por P, melhora com a experiência E."

Apesar de complexa, tal definição é continuamente listada nos livros, documentários e estudos que envolvem dados e algoritmos capazes de aprender com eles, essa também é uma fundamentação para as argumentações de que os dados são o novo petróleo, frase dita pelo matemático britânico Clive Humby, ainda em 2006.

Existem diferentes formas de aprender, tanto para humanos quanto para máquinas. Por isso, o aprendizado de máquina foi inicialmente subdivido em duas categorias de aprendizado: supervisionado e não supervisionado, que possuem como característica definidora a disponibilidade de dados de treinamento rotulados (CUNNINGHAM *et al.*, 2008), ou seja, na primeira abordagem, os algoritmos aprendem a partir de uma fundamentação ou de um conhecimento prévio sobre determinado contexto, já a segunda, pode ter como base a neurociência e se fundamentar no uso que o cérebro pode efetuar do fluxo maciço de informações sensoriais que ocorre, sem nenhuma recompensa ou punição associada (BARLOW, 1989). Atualmente, já podemos verificar na literatura algumas novas variações das supracitadas, como o aprendizado semi-supervisionado (ZHU; GOLDBERG, 2009), o aprendizado por reforço (WIERING; OTTERLO, 2012) e um dos mais recentes, o fracamente supervisionado (ZHOU, 2018).

Neste trabalho, o foco será no aprendizado supervisionado, visto que há a disponibilidade de uma base de dados com rótulos previamente definidos. A resolução de problemas desse tipo geralmente pode ser fragmentada e estruturada nas seguintes etapas: coleta e preparação dos dados, engenharia de recursos, seleção das *features*, que pode ser considerada uma sub etapa do processo anterior, escolha do algoritmo de aprendizado de máquina, seleção dos hiperparâmetros e construção do modelo, e por fim, avaliação (MARSLAND, 2015). Abaixo, será detalhada a definição de cada uma das etapas e apresentado o que cada passo propõe, em sua execução.

1. **Coleta e preparação dos dados**. Nesta etapa os dados são coletados e então o *dataset* (conjunto de dados) é formado. Geralmente, esses dados são uma amostra para a população que o contexto do estudo representa. Posteriormente este conjunto deve ser dividido para o treinamento e teste dos modelos de aprendizado de máquina. Antes dessa divisão ser efetuada, deve-se realizar uma limpeza nos dados, onde técnicas e métodos estatísticos tem sua aplicação vista como crucial para o tratamento de possíveis cenários, como: informações faltantes, inválidas ou em escalas desproporcionais com as unidades relacionadas com os dados.
2. **Engenharia de recursos**. A engenharia de recursos é um componente essencial, mas trabalhoso, das aplicações de aprendizado de máquina Bengio *et al.* (2013). Este termo é usado para definir um conjunto de técnicas utilizado na criação e manipulação de *features*(recursos), tendo como objetivo desenvolver um bom modelo de aprendizado de máquina, visto que grande parte do desempenho dele depende muito da representação do vetor de recursos.
3. **Seleção de características**. Nesta etapa são selecionadas os recursos de maior importância, a partir das considerações de relevância para a resolução do problema. Com isso, algumas serão removidas, conforme não seja identificada a necessidade. O conjunto dessas atividades forma o processo hoje conhecido como seleção de características. As decisões tomadas nessa fase são fundamentadas pela análise estatística.
4. **Escolha dos algoritmos de aprendizado de máquina**. A depender do aprendizado e das características do domínio do problema em questão, existem diferentes algoritmos que podem ser utilizados. Essa etapa é destinada para a distinção e definição de quais são úteis e serão usados na construção dos modelos.
5. **Seleção dos melhores hiperparâmetros**. Para um melhor desempenho dos modelos temos os hiperparâmetros, que basicamente são parâmetros configuráveis que modificam como o modelo funciona. Os algoritmos utilizados na construção do modelo já possuem um valor padrão, porém, é possível otimizar o seu funcionamento, isso pode ser feito a partir de buscas exaustivas que avaliam os valores para os parâmetros cujo modelo proporciona melhores resultados. As denominadas *técnicas de ajustes* são aplicadas nessa fase.
6. **Avaliação**. Antes de avaliar um modelo utilizando os dados de teste, deve-se treiná-lo e avaliá-lo e validar seus resultados com a porção de dados de treino. Apenas no último estágio da avaliação os dados de teste deverão ser utilizados. Essa é uma restrição

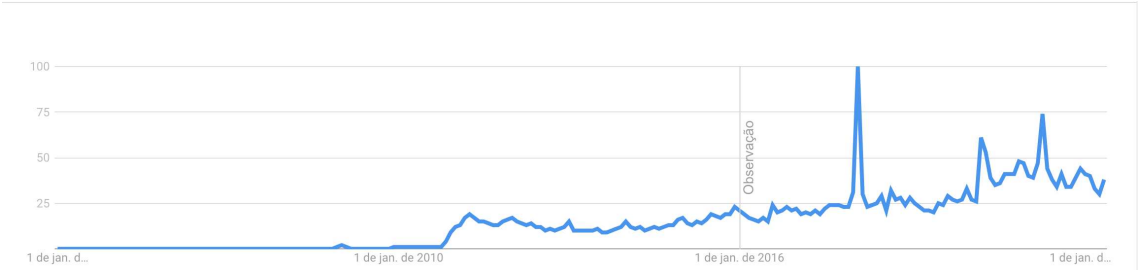
importante para evitar problemas de ajuste nos dados. A avaliação é feita a partir das métricas selecionadas para o tipo de problema ou com o auxílio de especialistas do domínio.

2.3 Sistemas de recomendação

Os sistemas de recomendação podem ser definidos como programas que tentam recomendar os itens mais adequados (produtos ou serviços) a usuários específicos (indivíduos ou empresas), prevendo o interesse de um usuário em um item com base em informações relacionadas aos itens, aos usuários e às interações entre os dois (BOBADILLA *et al.*, 2013). Esses sistemas podem utilizar métodos explícitos ou implícitos para a coleta de dados, onde a diferenciação é feita baseada na forma de interação com o ator externo (GUO; LU, 2007). Existem algumas categorias de aplicações conhecidas onde os sistemas de recomendação podem estar inseridos, como, por exemplo, os *E-commerces* e os *E-resources*, aplicações utilizadas, respectivamente, para a recomendação de produtos e serviços de recursos eletrônicos. Nos últimos anos, *E-Government* foi inserido como um novo âmbito que engloba um leque de aplicações relacionadas a recomendação.

O gráfico presente na Figura 2 exibe os resultados de busca para o termo *Recommender Systems* nos últimos 18 anos. Apesar de o tema ser conhecido desde o século passado, sua popularidade aumentou consideravelmente a partir da década passada, se mantendo em alta até os dias atuais.

Figura 2 – Resultado do Google trends para o termo “Recommender Systems”.



Fonte: elaborado pelo autor

Para entender como funcionam os sistemas de recomendação se faz necessário conhecer as possíveis técnicas da literatura para aplicar recomendações a um conjunto de dados. Deve-se então entender o domínio da aplicação e conceber o problema do trabalho, para então decidir quais métodos serão usados. As principais técnicas de recomendação incluem

métodos baseados em filtragem colaborativa, baseados em conteúdo, baseadas em conhecimento e híbridos, além de técnicas mais modernas, como: abordagens baseadas em conjuntos difusos, apoiadas em redes sociais, fundamentadas em conscientização de contexto, recomendações de grupo, dentre outras (LU *et al.*, 2015). Este trabalho se baseará nos métodos tradicionais. Logo abaixo serão descritas e detalhadas às técnicas de filtragem colaborativa, baseadas em conteúdo e híbridas.

2.3.1 Técnicas baseadas em conteúdo

Os sistemas de recomendação construídos a partir da aplicação de técnicas baseadas em conteúdo analisam as descrições dos itens para identificar aqueles que são de interesse particular para o usuário (PAZZANI; BILLSUS, 2007). Esta abordagem ainda se subdivide em duas vertentes: a centrada no item e a centrada no usuário. A primeira faz novas recomendações de itens (serviços, produtos, etc.) baseando-se na semelhança dos itens anteriormente recomendados ou utilizados pelo usuário. Já a segunda vertente dá ênfase maior ao usuário e faz recomendações a partir de preferências coletadas dele, de forma explícita. Essa coleta pode acontecer na própria plataforma que oferta o serviço ou produto, como quando a Netflix permite que um usuário avalie um filme e o inclua na sua lista de preferências, ou através de formulários e pesquisas externas que requerem a avaliação de satisfação. Esta abordagem não será explicitamente adotada neste trabalho porque essa interação com os objetos de estudos não acontece de forma contínua.

2.3.2 Técnicas de filtragem colaborativa

As técnicas que se baseiam no usuário foram amplamente utilizadas com sucesso em sistemas de recomendação. No entanto, além da complexidade exigida para lidar com o número de usuários da base, as interações contínuas (explícitas ou implícitas) com os usuários podem ser difíceis em inúmeros casos de uso. As técnicas de filtragem colaborativa constroem o modelo de recomendação analisando as semelhanças entre os vários itens e, em seguida, usam os seus semelhantes para identificar o conjunto a ser recomendado. Os principais passos dos algoritmos desta classe são (i) calcular a semelhança entre os objetos e (ii) combinar essas semelhanças para calcular a semelhança entre uma cesta de itens e um item candidato à recomendação (DESHPANDE; KARYPIS, 2004).

Essas técnicas analisam a matriz usuário-item para descobrir relações entre os diferentes itens e usam essas relações para calcular a lista de recomendações e, em seguida,

usá-los para identificar o conjunto de itens a serem recomendados.

2.3.3 Técnicas híbridas

Um sistema de recomendação híbrido é aquele que utiliza mais de uma técnica de recomendação. Mais precisamente, é uma combinação de métodos de recomendações baseadas em conteúdo com a filtragem colaborativa. Essa alternativa é indicada para cenários em que se deseja trabalhar com as principais vantagens de cada metodologia e superar as adversidades encontradas em cada uma. As recomendações baseadas em conteúdo não necessitam de dados de outros usuários e conseguem realizar recomendações de itens para nichos específicos. No entanto, sua utilização requer que as preferências dos usuários sejam explicitamente conhecidas e haja um elevado conhecimento do domínio e de suas respectivas regras do negócio. Ao se referir à filtragem colaborativa, temos as vantagens de ajudar os usuários a descobrirem novos interesses, sem necessariamente conhecer bastante do domínio. Todavia, essa abordagem encontra dificuldades ao lidar com itens desconhecidos (novos). Dessa forma, os sistemas híbridos usam várias fontes de informação e combinam métodos distintos, como os supracitados, para tentar sanar os problemas. Estes sistemas usam dados de uso de usuários e dados de conteúdo de itens. Assim, além de capturar as semelhanças de conteúdo entre os itens, esses sistemas conseguem revelar outras relações, como associações e co-ocorrências entre eles (PARRA; SAHEBI, 2013).

2.4 Engenharia de recursos

A engenharia de recursos surgiu do desejo de transformar entradas de regressão linear não distribuídas normalmente. Tal transformação pode ser útil para regressão linear Heaton (2016), mas não é o único modelo de aprendizado de máquina que pode se beneficiar da aplicação dessa metodologia e de outras possíveis transformações. Em 1999, pesquisadores demonstraram que a engenharia de recursos pode melhorar o desempenho de modelos além da regressão linear, quando, inclusive, usaram técnicas pertencentes à metodologia para aprendizado de regras para classificação de texto Scott e Matwin (1999). Identificar os recursos de entrada apropriados é uma das etapas fundamentais e desafiadoras para a aplicação de métodos de aprendizado de máquina.

2.5 Modelo para previsão de probabilidades

A regressão logística é um dos algoritmos que podem ser usados para classificação, no entanto, ele trabalha com as probabilidades. Por exemplo, dado um conjunto de dados fictício, em que a resposta é uma saída binária, sim ou não, ao invés de diretamente predizer uma resposta para uma das possibilidades, a regressão logística modela a probabilidade de que uma entrada de dados qualquer pertença a uma dessas categorias em específico.

Exemplificando, para um conjunto de dados com algumas características sobre uma amostra de pessoas da população do Texas, cujas características estão relacionadas com o fato da pessoa ser fumante ou não. Neste caso, utiliza-se a regressão logística não para prever se uma pessoa é fumante ou não, mas para fornecer probabilidades, a depender da variável alvo (ser fumante ou não). Assim, se o exemplo for reproduzido seguindo a lógica booliana, o 0 pode representar o não fumante e o 1 o fumante. Dessa forma, a probabilidade (P) de X, uma entrada qualquer, ser fumante, é descrita a partir da fórmula a seguir:

$$P(X) = Pr(Y = 1|X) \tag{2.1}$$

2.5.1 Regressão Polinomial

Historicamente, a maneira padrão de estender a regressão linear para configurações em que a relação entre os preditores e a resposta é não linear tem sido substituir o modelo linear padrão. Nessa abordagem, existe um grau d, que representa o grau de um termo de uma variável em um polinômio e corresponde ao expoente dessa variável nesse termo. Para um grau d suficientemente grande, uma regressão polinomial permite a produção de uma curva extremamente não linear.

2.6 Avaliação de Desempenho

Existem duas formas complementares para avaliar os sistemas de recomendação. A primeira consiste em avaliar o desempenho de previsão do sistema usando conjuntos de dados e validação cruzada. A segunda forma consiste em pedir a usuários reais sua satisfação testando o sistema (CANDILLIER *et al.*, 2009). Neste trabalho será avaliada apenas a primeira forma, visto que o contexto onde o trabalho está inserido impossibilita a interação supracitada, o que pode ser considerada uma limitação. Dessa forma, o trabalho proposto se concentra em avaliar as previsões, porém, também focará em validar etapas intermediárias do estudo. Abaixo são

descritos alguns tipos de avaliação e validação dos resultados e como estas são utilizadas.

2.6.1 Validação cruzada

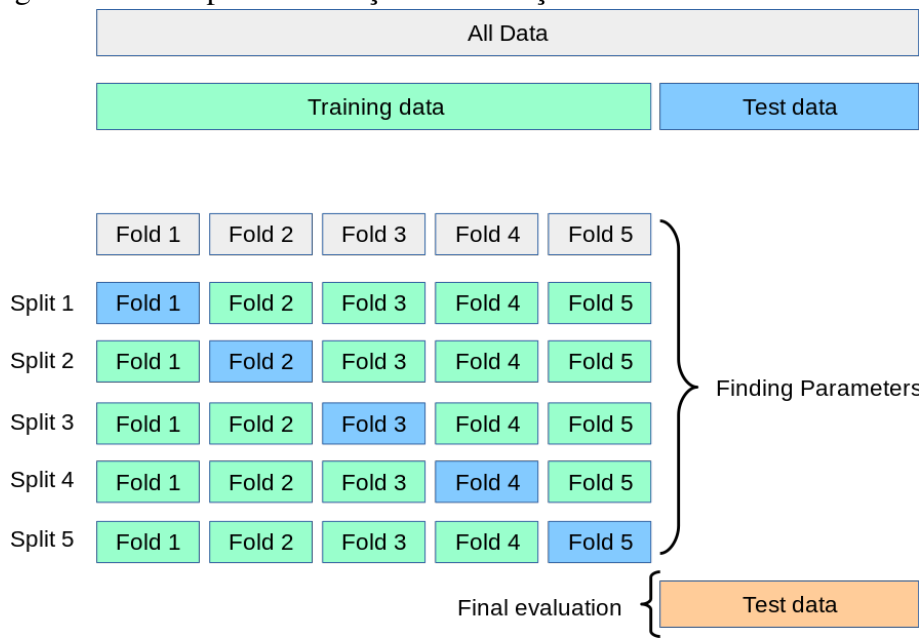
A validação cruzada é um método estatístico de avaliação e comparação de algoritmos de aprendizado, dividindo os dados em dois segmentos: um usado para aprender ou treinar um modelo e o outro usado para validar o modelo. Nesse tipo de validação, os subconjuntos de treinamento e validação devem ser cruzados em execuções sucessivas, de modo que cada ponto de dados tenha a chance de ser validado. A aplicação mais comum dessa validação é a *k-fold*, na qual os dados são particionados inicialmente em k segmentos, ou dobras, de tamanho que tende a ser igual. Subsequentemente, k iterações de treinamento e validação são realizadas de tal forma que dentro de cada iteração uma dobra diferente dos dados é mantida para validação, enquanto as restantes k-1 dobras são usadas para aprendizado. (REFAEILZADEH *et al.*, 2009)

A Imagem 3 ilustra um exemplo de execução da validação cruzada com a técnica *k-fold*, com 5 segmentos e consequentemente 5 iterações. Na primeira execução, em que o k é igual a 1, a primeira dobra é mantida para validação, enquanto às outras são utilizadas no treinamento, para aprendizado. Na segunda execução, em que o k tem o seu valor igual a 2, a segunda dobra é mantida para validação, enquanto às outras, incluindo a primeira dobra, são utilizadas, e assim sucessivamente.

Para avaliar este processo, utiliza-se uma métrica que estime o erro em cada execução, como a *Mean Absolute Error* (MSE). Assim, os resultados obtidos são k estimativas de erro, MSE1, MSE2..., MSEk. A estimativa da validação cruzada com *k-fold* é então calculada a partir da média desses valores. (JAMES *et al.*, 2013)

Repare que ao lado há a descrição de um passo denominado como busca de parâmetros, essa é uma abordagem complementar usada para testar exaustivamente os valores de parâmetros dos algoritmos, essa execução utiliza a validação cruzada sobre configurações de parâmetros. Existem duas aplicações para essa abordagem, o *Grid Search* e o *Randomized Search*. A diferença entre as duas está no número de valores testados, para os parâmetros selecionados. A primeira realiza esse teste exaustivo por cada um dos valores, para cada parâmetros. Já a segunda, realiza o processo para um número fixo de configurações de parâmetro, cujo é definido como uma amostra das distribuições especificadas.

Figura 3 – Exemplo de execução da validação cruzada com técnica K-Fold.



Fonte: *Scikit-learn*

2.6.2 Mean Absolute Error (MSE)

As diferentes etapas até chegar na construção de um modelo de recomendação podem ser avaliadas de formas distintas, variando conforme o domínio do problema e o resultado que se espera obter. Com essa validação intermediária, espera-se que as predições tenham melhores resultados. Um exemplo de métrica que pode ser usada para avaliar sub-etapas da construção de um modelo de recomendação é o Erro quadrático médio, em tradução livre do inglês *Mean squared error* (MSE). Nesta estratégia, busca-se selecionar as alternativas que fornecem o menor de MSE.

A fórmula 2.2 representa o cálculo para a métrica:

$$MSE = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{2.2}$$

Em outras palavras, o MSE é o somatório $\sum_{i=1}^{D}$, para um conjunto de D entradas, dos quadrados dos erros $(x_i - y_i)^2$, onde (x_i) representa o valor real e (y_i) representa o valor predito. Essa métrica mede o quanto alguns resultados se afastam da média obtida inicialmente. O MSE dá um maior peso aos maiores erros, já que, ao ser calculada, tem os erros elevados ao quadrado individualmente e, após isso, a média destes erros quadráticos é calculada.

2.6.3 Avaliação das predições

Para avaliar as predições realizadas pelo modelo de recomendação, também existem diferentes estratégias e medidas. Para alguns modelos, pode-se usar medidas preditivas, cujas medem o quão próximas às classificações dos sistemas de recomendação estão das classificações já obtidas, costumam ser usadas em conjuntos de dados cuja saída é não binária. O *Mean absolute error* (MAE) e o *Root mean squared error* (RMSE) são as métricas preditivas mais populares. Além dessa, existem as métricas de classificação, que avaliam a capacidade de decisão dos sistemas de recomendação. É comum ver sua utilização em conjuntos de dados que se quer identificar, produtos ou objetos, relevantes ou irrelevantes para o usuário ou outro tipo de registro da entrada dos dados. Exemplos para esse último contexto são *Recall*, *Precision* e *F1-Score*.

2.6.3.1 Matriz de Confusão

Para compreender e interpretar as medidas supracitadas, pode-se obter os seus resultados a partir da visualização de uma matriz, denominada como Matriz de Confusão. Uma matriz de confusão de tamanho n x n associada a um classificador mostra a classificação prevista e real, onde *n* é o número de classes diferentes, como visto em Visa *et al.* (2011). Nela, existem quatro tipos de valores que podem ser medidos, são eles: verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

Os verdadeiros positivos são aqueles valores cujo modelo prevê corretamente a classe positiva, ou seja, o resultado obtido é positivo, assim como o esperado. Os verdadeiros negativos ocorrem quando, no conjunto real, a classe que se busca prever é prevista incorretamente, ou seja, o resultado obtido é positivo, diferente do esperado. Os falsos positivos são denominados a partir da previsão incorreta para a classe positiva, isto é, o modelo prevê que sim, quando deve ser não. Já os falsos negativos são aqueles valores que o modelo prevê de forma incorreta a classe negativa, isto significa que a previsão deve ser negativa, mas a predição é positiva.

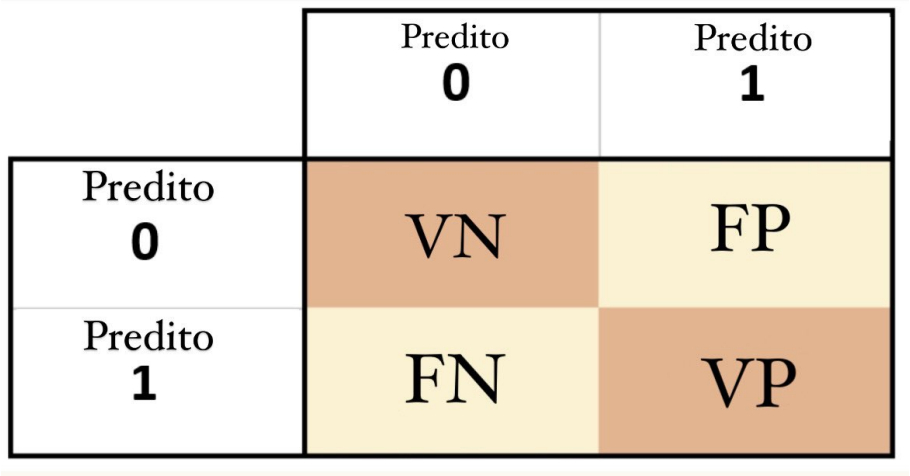
A exemplificação destes valores pode ser vista a partir da Imagem 4.

2.6.3.2 Precision

A fórmula 2.3 representa o cálculo para métrica e pode ser vista a seguir:

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

Figura 4 – Exemplo de modelo para a Matriz de Confusão



Fonte: autor

Em que *VP* representa os verdadeiros positivos e *FP* representa os falsos positivos.

Em suma, a métrica *precision* visa responder a seguinte pergunta: "qual a proporção de previsões positivas está correta?", ou seja, das previsões cuja resposta é "sim", qual a proporção das que foram identificadas corretamente?

2.6.3.3 Revocação

A fórmula 2.4 representa o cálculo para obter a métrica e está detalhada a seguir:

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

Onde, *VP*, como foi supracitado, representa os verdadeiros positivos, e *FN* representa os falsos negativos.

Em suma, a métrica *recall* visa responder a seguinte pergunta: "quando a resposta é sim, com que frequência prevê sim?", ou seja, delimitando o espaço amostral para os verdadeiros positivos, o interesse é saber qual proporção foi identificada corretamente pelo modelo.

2.6.3.4 F1 Score

Esta métrica é uma média harmônica das métricas apresentadas acima: *recall* e *precision*.

Seu cálculo é realizado a partir da Fórmula 2.5, apresentada a seguir:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{2.5}$$

O cálculo e consequentemente o valor desta métrica pode diferir, conforme a técnica aplicada. As três principais variações são a avaliação Macro, Micro e Ponderada. A primeira, é a média não ponderada das pontuações de F1 calculadas por classe. Já a segunda, segue a mesma fórmula da métrica, porém, é calculada usando o número total de Verdadeiros Positivos (TP), Falsos Positivos (FP) e Falsos Negativos (FN). Ela agrupa as classificações por amostra entre as classes e, em seguida, calcula a pontuação geral de F1. (TAKAHASHI *et al.*, 2022). Por último, o cálculo da pontuação F1 média ponderada é realizado a partir da média de todas as pontuações F1 por classe, considerando o suporte de cada classe. O suporte, neste caso, refere-se ao número de ocorrências reais da classe no conjunto de dados. Esses cálculos também podem ser utilizados para as métricas de *recall* e *precision*.

As Equações 2.6 e 2.7 exibem as respectivas fórmulas para o cálculo das duas abordagens supracitadas.

$$F1 = \frac{TP}{TP + \frac{1}{2} * (FP + FN)} \tag{2.6}$$

$$F1 = \frac{sum(F1scores)}{n} \tag{2.7}$$

Onde n corresponde ao número de classes utilizadas para o cálculo e o *sum (F1 scores)* ao somatório dos valores de *f1* para todas as classes.

2.6.4 Label Ranking Average Precision

A precisão média de classificação do rótulo, em tradução livre *Label Ranking Average Precision* (LRAP) é a média sobre cada rótulo atribuído como verdadeiro, em cada amostra, da proporção de rótulos verdadeiros em comparação com os totais de pontuação mais baixa. Essa métrica é utilizada em problemas de ranqueamento com múltiplos rótulos, onde o objetivo é fornecer o melhor ranqueamento aos rótulos associados a cada amostra. A pontuação obtida é sempre estritamente maior que 0 e o melhor valor é 1. Essa métrica visa responder a seguinte questão: para cada rótulo atribuído como positivo, que fração de rótulos com classificação mais alta eram rótulos realmente verdadeiros? Essa medida de desempenho será maior se os rótulos associados a cada amostra forem bem classificados.

A pontuação obtida é maior que 0, e o melhor valor possível é 1. Para se aproximar do melhor valor, não basta apenas que uma grande proporção de classificações esteja correta, mas, que essas classificações estejam corretas de forma pelo menos balanceada, em todos os rótulos associados.

2.7 *E-government*

O *E-government* tem como definição base o uso das Tecnologias da Informação (TIC) para melhorar o processo de governo (KOLACHALAM, 2012). Na prática, isso impacta diretamente como o governo se relaciona com cidadãos que são por ele governados. Estes governos eletrônicos emergiram a partir da grande transformação que a tecnologia causa em diversos âmbitos e aspectos na vida das pessoas, e consequentemente em tudo que elas estão envolvidas, seja no trabalho, na escola, nos negócios, ou até mesmo em atividades no cotidiano. Assim, surge uma nova oportunidade de como reinventar os governos e como que eles servem ao seu povo. Depois disso, está sendo reconhecida a necessidade de mudar a forma de fazer negócios, de prestar serviços e até mesmo de disponibilizar as informações centrados no cidadão (SILCOCK, 2001). A governança eletrônica é definida como a transformação de processos governamentais resultantes da introdução contínua e exponencial nas tecnologias digitais mais avançadas (BILL, 2002).

A popularidade dos governos eletrônicos emergiu ainda nos primeiros anos do século XXI e se manteve em alta até o período atual. Nesse meio-termo, muitos trabalhos foram realizados com o intuito de explorar novas abordagens e inovações para esse tipo de governo, como Chun *et al.* (2012), que introduz os tipos de modelos de colaboração e como eles podem ser facilitados pelo uso da tecnologia e Bas *et al.* (), que traz uma nova perspectiva de governo eletrônico, baseando-se na proatividade dos cidadãos com relação aos serviços disponíveis.

3 TRABALHOS RELACIONADOS

A partir de uma revisão na literatura, foram identificados alguns estudos que se relacionam com este projeto de pesquisa proposto, seja de forma técnica e replicável, onde há a construção de uma sistema de recomendação e pontos comuns no delineamento experimental, ou através do modelo conceitual, cuja abordagem envolve os principais componentes e compartilha contexto compatível ou semelhante com este trabalho. Nesta seção, estes estudos são detalhados, suas contribuições são descritas e o modo como se relacionam ou diferem do presente trabalho é apresentado.

3.0.1 Recommender systems for smart cities

Em Quijano-Sánchez *et al.* (2020), os autores apresentam uma taxonomia de características, dimensões, ações e objetivos para *smart cities* e introduzem conceitos relevantes no âmbito de sistema de recomendação. Feito isso, às duas áreas citadas são trabalhadas de forma conjunta, visando identificar as principais tendências de pesquisa e exibir oportunidades e desafios atuais, onde pode-se explorar soluções de recomendações personalizadas para cidadãos, empresas e administrações públicas. Na explicação sobre os principais conceitos, os autores definem que os problemas de recomendação tem basicamente três tarefas: a primeira é coletar informações sobre o objeto de estudo - ator - em questão. A segunda, é aprender com as informações coletadas e prever as preferências destes atores para novos itens. E, por último, aplicar uma função ou construir um modelo que selecione e recomende aqueles com maior probabilidade de serem preferidos. Estes três passos são detalhados e cada uma das sub-atividades que os compõem são descritos. A etapa da coleta dos dados e criação de um perfil, pode acontecer de forma explícita ou implícita, a depender de como os dados de preferência desses usuários são coletados. A coleta explícita remete a declarações de preferência diretas feitas por eles sobre itens que eles já conhecem, enquanto a implícita, ocorre quando as preferências são modeladas a partir de *logs*, registros e consultas referentes às ações destes usuários.

Ademais, ainda nessa primeira etapa há a seleção de características usadas pelos modelos, e o possível cruzamento de dados através da integração com outras bases. Na segunda tarefa, para problemas de recomendação, deve-se definir o tipo de sistema que será utilizado. Os autores citam dois dos principais tipos conhecidos no estado da arte: os sistemas baseados em conteúdo e os de filtragem colaborativa. Em suma, aqueles baseados em conteúdo usam

as informações para representar usuários e itens, e sugerir aqueles cujos perfis sejam mais semelhantes aos requisitos do usuário que se deseja fazer a recomendação. Já os de filtragem colaborativa se baseiam nas opiniões já atribuídas pelos usuários aos elementos existentes. Assim, a partir de padrões de classificação identificados ou de fatores latentes, esses sistemas fazem recomendações de usuário de itens preferidos por pessoas com gostos e interesses semelhantes.

Para a parte final do processo, o trabalho sugere algoritmos de aprendizado de máquina que podem ser usados para cada uma das abordagens e elicita três formas de avaliar os sistemas de recomendação construídos: avaliação online, estudo de usuários e avaliação *offline*. O trabalho conduz de forma exploratória parte do contexto que envolve os sistemas de recomendação, desde a coleta dos dados, passando por algoritmos que podem ser usados nos modelos, até à avaliação, de fato. No entanto, sua principal contribuição é relacionar esse tema com as *smart cities*, exibindo as oportunidades, desafios, e como os sistemas podem se relacionar com subtópicos como *smart governance* e *smart people*. Isso se mostrou como inovador no quesito. Essa discussão fornece um embasamento que pode ser reaproveitado no trabalho aqui proposto, visto que há também o intuito de utilizar recomendações para uma área de aplicação, neste caso, o âmbito de governo eletrônico. Entretanto, as semelhanças se limitam ao campo teórico, visto que este trabalho propõe de fato a implementação de um sistema de recomendação com dados coletados de cidadãos.

3.0.2 A decision support system for designing new services tailored to citizen profiles in a complex and distributed e-government scenario

Meo *et al.* (2008) propõe um sistema de recomendação visando apoiar os tomadores de decisão de agências governamentais na projeção de novos serviços adaptados aos perfis dos cidadãos em um cenário de governo eletrônico complexo e distribuído. De modo mais específico, o sistema se dispõe a auxiliar os gerentes de agências governamentais, cujos planos são de liberar novos serviços para os cidadãos. Dessa forma, o foco é identificar os cidadãos que podem extrair o maior benefício de cada um dos serviços. Ou seja, baseado nas necessidades das pessoas, os gestores devem decidir quais serviços serão fornecidos e como estes podem se adaptar melhor a cada realidade, de modo que seja garantido que a indicação realizada será a de maior potencial assistencial.

Para acessar e gerenciar as necessidades dos cidadãos, o sistema cria uma associação entre o cidadão no papel de usuário e um perfil adequado, no momento do uso do programa. Estes

perfis são estruturados a partir das informações que o governo armazena, sobre estes usuários, sejam informações do domínio técnico, como comportamentos anteriores e preferências, ou de aspecto geral, como os dados demográficos. Traçar esses perfis é uma decisão que na opinião dos autores pode melhorar a eficácia e precisão do sistema, visto que se torna possível contribuir com as duas vias: os cidadãos só receberão recomendações de serviços necessários e os governos poderão escolher aqueles que extrairão maior benefício de um serviço específico. Além disso, o sistema proposto se baseia nos recursos financeiros disponíveis para a execução de tais ações. No presente trabalho, a análise otimizada a partir dos recursos financeiros disponíveis não será realizada, primeiramente, porque o objetivo desse sistema proposto não é interagir diretamente com os usuários finais, e segundo porque essas informações não são de domínio público. Por isso, o intuito é prover um embasamento otimizado para as autoridades correspondentes, que ficam então responsáveis por alocar os benefícios recomendados, mediante as limitações existentes e conhecidas por elas e pelo cenário público.

3.0.3 State-of-the-art recommender systems

Em Candillier *et al.* (2009), os autores introduzem conceitos relevantes no âmbito de sistema de recomendação. Primeiro são apresentadas algumas técnicas que podem ser utilizados por eles, tais como a filtragem colaborativa, baseada em conteúdo e híbrida. Após introduzir e exemplificar quais são as especificidades de cada uma das técnicas, os autores listam casos de uso aplicáveis e diferenciam formas de avaliar distintos sistemas de recomendação. Ademais, um grupo de experimentos é realizado, comparando diferentes recomendadores, implementados e avaliados para dois conjuntos de dados de classificação reais que estão disponíveis publicamente: *MovieLens* (www.grouplens.org) e *Netflix* (www.netflixprize.com). Os sistemas construídos utilizam a filtragem colaborativa e os três recomendadores foram criados, baseando-se em três abordagens: a baseada no usuário, no item e no modelo. A avaliação dos resultados foi realizada através da validação cruzada, com 90% dos dados para o conjunto de treino e 10% para o de teste. Esta avaliação é aplicada aos três recomendadores, para os dois conjuntos de dados, a partir das métricas *Mean Squared Error* (MAE), *Root Mean Squared Error* (RMSE) e *precision*.

Ademais, os autores também realizam validações no experimento relacionadas com a decisão tomada para os seguintes passos: escolha da métrica de similaridade, número de vizinhos K utilizados no agrupamento dos dados, abordagem para calcular as predições, os *clusters* baseados em itens ou em usuários e qual será o número de *clusters* desejado. Estas

etapas são medidas e avaliadas com base nas métricas de avaliação, para os dois conjuntos de dados. Ao final do processo, os resultados comparativos são sumarizados e apresentados através de gráficos e tabelas. Este trabalho também utilizará as técnicas de filtragem colaborativa, que conforme o autor do trabalho supracitado, são utilizadas com maior frequência, se comparada as outras duas. Além disso, a validação cruzada será aplicada, utilizando o conjunto de dados de treino e validação.

3.0.4 Intelligent e-government services with personalized recommendation techniques

O trabalho proposto em Guo e Lu (2007) pretende aumentar a eficácia dos serviços de governo eletrônico. Para isso, os autores propõem uma nova abordagem com a finalidade de lidar com questões de recomendação de itens únicos, para os serviços do governo. Esta abordagem integra as técnicas de similaridade semântica e aplica uma das técnicas de filtragem colaborativa: a filtragem baseada em itens. O sistema Smart Trade Exhibition Finder foi desenvolvido utilizando uma amostra de dados com 300 feiras de exposições comerciais, coletada entre janeiro de 2004 e janeiro de 2005 e obtida a partir dos seguintes *datasets*: Australian Trade Commission [1], Australia Exhibition and Conference [2], Global Sources [3] e Tradeshow Week [4].

A validação do estudo se dividiu em duas etapas. Na primeira é realizada a avaliação dos sistemas de recomendação, utilizando as métricas *MAE*, *recall*, *precision* e *F1*. Já no segundo passo, os autores avaliam o experimento a partir da observação dos resultados de *MAE* e *F1* com relação à alteração do número de vizinhos, algo relatado como de fundamental importância para a qualidade da recomendação. Os resultados obtidos no estudo mostram que a abordagem proposta alcança um melhor desempenho na recomendação de novos itens adicionados e melhoram os valores para a precisão. Além disso, a contribuição chave é que antes desse estudo não havia uma abordagem consolidada para a recomendação de itens únicos.

A Tabela 1 realiza um comparativo entre os trabalhos relacionados e o trabalho aqui proposto, essa comparação é feita a partir das seguintes questões: domínio em que os dados estão inseridos, implementação de um sistema de recomendação, métricas de avaliação e as técnicas de recomendação abordadas no trabalho. O domínio é importante para diferenciar em que cenário o trabalho e os dados coletados estão inseridos, no caso do trabalho proposto, os dados utilizados

[1] http://www.austrade.gov.au
[2] http://www.aec.net.au
[3] http://www.globalsources.com
[4] http://directory.tradeshowweek.com/directory/index.asp

são de cidadãos e o cenário é o de *E-Government*. A implementação ou não de um sistema de recomendação também é crucial na diferenciação dos trabalhos e contribuição deste, visto que alguns estudos apenas modelam ou designam como seria esse sistema, sem implementá-lo, de fato. As métricas de avaliação e as técnicas de recomendação são critérios técnicos relacionados com o domínio, dessa forma, cada problema a ser resolvido pode necessitar de diferentes técnicas e formas de avaliar suas aplicações, através dos modelos construídos.

Tabela 1 – Tabela comparativa entre o trabalho proposto e os seus relacionados

Trabalhos	Domínio dos dados	Implementa um sistema de recomendação	Métricas de avaliação	Técnicas de recomendação
Trabalho proposto	Governo eletrônico	Sim	*Precision*, *Recall*, *F1-Score* e *Average Precision*	Filtragem colaborativa, recomendação baseada em conteúdo e híbridas
MEO et al., 2008	Governo eletrônico	Sim	*Precision* e *Recall*	Técnica proposta pelos autores
CANDILLIER et al., 2009	Filmes	Sim	*Average Precision* e *Granularity*	Filtragem colaborativa, recomendação baseada em conteúdo, baseada em conhecimento, baseada em inteligência computacional e híbridas
QUIJANO-SÁNCHEZ, 2020	Cidades inteligentes	Não	*RMSE, MAE, MRR, Precision, Recall e nDCG*	Filtragem colaborativa e recomendação baseada em conteúdo
GUO; LU, 2007	Governo eletrônico	Sim	*MAE*, *Precision*, *Recall* e *F1*	Filtragem colaborativa

Fonte: elaborada pelo autor.

4 PROCEDIMENTOS METODOLÓGICOS

Nesta seção são apresentadas as etapas necessárias para a realização do presente trabalho. As subseções a seguir descrevem cada uma dessas etapas, realizadas de forma sequencial.

4.1 Coleta da base de dados

A coleta dos dados é uma das ações itinerantes do Programa Cartão Mais Infância Ceará (CMIC), lançado em 2015 e que atualmente conta com 150.000 beneficiários, como pode ser visto na plataforma do Big Data Social[1]. A plataforma apresenta esta e outras informações referentes a proteção social do Ceará. O conjunto de dados utilizado é composto de dados coletados pelo Levantamento da Situação Sócio Familiar da Família CMIC, um formulário criado como uma das ações do programa supracitado. Essa base é denominada de base dos Agentes e contém perguntas distintas das que estão na base do Cadastro Único. A pesquisa ainda está em curso, por isso um recorte temporal será realizado e terá como parâmetro final o mês de junho, do ano de 2022, pois foi quando se iniciou a parte prática do trabalho aqui proposto.

A pesquisa foi respondida por famílias em mais de 167 municípios cearenses em todas as regiões do estado, e esses dados foram coletados a partir de visitas realizadas pelos agentes sociais. Cada entrevista respondida corresponde a uma linha no conjunto de dados, as perguntas correspondem às colunas. No momento da ação, estes agentes orientavam o responsável familiar no que fosse requerido pelas perguntas, que estavam disponíveis de forma online através da plataforma *Jotform*. Este formulário é composto por quase 200 perguntas e as famílias inclusas no espaço amostral devem estar inclusas no CMIC. As questões foram agrupadas logicamente em seções, cuja divisão se dá a partir dos indicadores de interesse a serem observados em: condições de moradia, educação, saúde, assistência social, renda e qualificação, insegurança nutricional, crianças, grávidas e adolescentes, além de informações sobre à aplicação do questionário e uma seção com informações gerais sobre o entrevistado (chefe familiar).

Esses dados são objetos de diversos estudos do *Insight Data Science Lab*[2], um laboratório de pesquisa aplicada em Ciência de Dados e Inteligência Artificial do Departamento de Computação da Universidade Federal do Ceará. Os projetos Plataforma Big Data para Acelerar a Transformação Digital do Estado do Ceará (FUNCAP nº 04772551/2020), Plataforma Governo Digital do Estado do Ceará (FUNCAP nº 04772314/2020) e Plataforma para Trans-

[1] https://bigdatasocial.irislab.ce.gov.br
[2] https://insightlab.ufc.br

formação Digital do Estado do Ceará (FUNCAP nº 04772420/2020) são casos de uso ativos que utilizam os dados citados com algum objetivo de pesquisa. A extração, a primeira etapa de pré-processamento, a modelagem e o armazenamento são exemplos de atividades realizadas por participantes dessas iniciativas. O presente trabalho pode ser considerado uma extensão a esses programas.

4.2 Pré-processamento

A etapa de pré-processamento é responsável por preparar os dados que serão utilizados como entrada para os modelos. Essa preparação inclui uma série de tratamentos que podem ser aplicados, a depender do formato e qualidade das informações. Nesse passo dos procedimentos metodológicos existem atividades relacionadas à limpeza, transformação e redução dos dados, além de algumas sub etapas pertencentes à engenharia de recursos. A limpeza deve lidar com os dados ausentes, tentar reduzir os ruídos, identificar e remover valores inconsistentes ou que fogem completamente da escala apresentada. Para a transformação, atividades de normalização e discretização podem ser realizadas, servindo para transformar os dados em valores ou estados limitados por intervalos e quantidades. A redução é usada na maioria das vezes para lidar com um grande volume de dados, mas também pode acontecer quando há a presença de recursos que não considerados importantes, portanto, não serão utilizados pelos modelos, estes, devem então ser removidos.

Neste trabalho, o pré-processamento é subdividido em: limpeza, divisão dos dados em conjuntos de treino e teste, que deve ser realizado somente após as etapas de engenharia de recursos. Esse tópico aborda atividades como criação, remoção e seleção de recursos, como detalhado na Seção 4.3.

4.2.1 Limpeza dos dados

. A etapa de limpeza é crucial para o sucesso de um estudo ciência de dados. É neste momento do processo que os dados inválidos serão removidos e também quando há a avaliação sobre o método de substituição das informações nulas e seleção dos recursos de interesse. Todos esses procedimentos são realizados a partir da aplicação de técnicas estatísticas, combinadas com ferramentas de aprendizado de máquina. Os dados faltantes para perguntas obrigatórias ou aqueles com valores de resposta identificados como fora da escala considerada válida serão

removidos do conjunto, pois não é possível identificar qual momento originou inconsistências ocorreram, podendo ser uma falha na própria coleta, no momento de exportar os dados da ferramenta de formulários e modelá-los em um banco de dados de armazenamento ou outro tipo de falha de segurança, como o vazamento de dados de teste para o ambiente de produção.

A decisão sobre substituição ou remoção das informações nulas foi tomada a partir dos fatores supracitados. Dessa forma, em caso de existir um valor considerado inválido, para uma coluna do conjunto, todo o registro foi removido, dada a impossibilidade de rastreio da inconsistência. Ou seja, não há como saber precisamente se a falha aconteceu somente para o dado dessa coluna ou o porquê e como ela foi produzida. Para realizar esse processo, todos os intervalos e possíveis valores para todas as colunas do conjunto foram examinados. Inclusive, existem colunas em que os valores pertenciam a uma escala válida para o subconjunto que o dado pertence, mas que ao relacionado com a pergunta, se tornava inválido. Por exemplo, uma coluna que representa a quantidade de meses de gravidez de alguma mulher do domicílio. O número 15 pertence ao subconjunto dos números inteiros positivos, o esperado para perguntas que se referem a meses, no entanto, no contexto da gestação isso é algo inviável, então esse tipo de dado foi removido.

4.2.2 Benefícios

No conjunto de dados inicial haviam sete benefícios, distribuídos entre as famílias: programa do cartão mais infância, vale-gás, cesta básica, isenção da tarifa de energia, isenção da tarifa de água, alimentos *in natura* e virando o jogo. No entanto, os registros do primeiro e do último benefícios listados acima foram removidos, primeiramente, porque estar inserido no Programa Cartão Mais Infância é um pré-requisito para estar nessa base de dados, então essa informação não contribui para o modelo, visto que todos os registros devem possuir. Já para o benefício denominado de "virando o jogo"existem apenas cinco registros no conjunto de dados, e por isso ele não pode ser usado como parâmetro para as características e predição.

4.3 Engenharia de recursos

Etapas de transformação, seleção e remoção fazem parte dos processos de engenharia de recursos, utilizadas na busca da definição dos melhores recursos que servem como entrada para um modelo preditivo. Este processo também abrange a transformação matemática de recursos já

existentes e a normalização dos valores contidos na base de dados, em casos, por exemplo, de conversão das medidas usadas ou para reduzir a redundância de dados, como introduzido em na Seção 4.2. A criação de novos recursos a partir de outros já existentes é também um passo fundamental dentre as etapas, visto que é possível criar dados a partir dos disponíveis, e estes podem ser necessários para a resolução de um problema.

Nesta seção, é apresentado um detalhamento das atividades de criação e remoção de colunas do conjunto de dados, no entanto, o trabalho proposto também foca na seleção de características, que poderá ser explorada com maior clareza na seção 4.4.

O resultado de uma pesquisa de aprendizado de máquina é um cenário final resumido das decisões tomadas durante o projeto, das mais simples às mais complexas. No panorama, há decisões de nível macro, mas em cada sub etapa desenvolvida há também outras tantas menores. Já o processo de criação de novas colunas foi realizado para as características categóricas, desde as que representavam valores inclusos na lógica booliana (verdadeiro ou falso, sim ou não) até aquelas que possuíam um máximo 8 valores distintos.

Para ambos os casos, novas colunas foram criadas para representar cada um desses valores, e os dados, a partir das condições de compatibilidade com a categoria, foram preenchidos com *True* ou *False*. Para tornar mais palpável, tomemos como exemplo uma coluna que representa a seguinte pergunta no formulário "Há alguma grávida no domicílio?", os valores de resposta só podem ser sim ou não, então foi criada uma coluna para representar a resposta positiva à pergunta, em que os registros que possuírem grávidas ficarão com valor *True* e todos os que não, com o valor *False* preenchido. Um exemplo para o segundo caso é a coluna que representa a pergunta "O carro de lixo da Prefeitura busca o seu lixo quantas vezes por semana?", para ela, existem três possíveis respostas, delimitadas pelo próprio formulário: "nenhuma vez"; "de uma a duas vezes"; "mais de duas vezes". Dessa forma, foram criadas três novas colunas, uma para cada resposta possível, e o preenchimento do valor seguiu a mesma lógica do exemplo anterior.

4.3.1 Particionamento dos dados

Após às etapas de criação, transformação e seleção de recursos terem sido concluídas e antes da execução da seleção de características, se faz necessário realizar o particionamento dos dados. Neste processo, o conjunto de dados é dividido em uma proporção para treino e outra para teste. A porção de treino é utilizada durante todas as etapas de treinamento dos modelos e de validação cruzada dos resultados. O subconjunto de treino corresponderá a uma amostra

aleatória com 90% dos dados do conjunto principal, enquanto o subconjunto de teste tem os 10% restantes. Essa divisão é realizada a partir da implementação de uma função específica para dividir conjuntos de dados, sua nomenclatura pode variar conforme a ferramenta utilizada. Neste trabalho, a implementação usada é mantida pelo *Scikit*[3] e nomeada de *train test split*, além do fator aleatório, uma semente é definida para que o usuário obtenha a mesma divisão entre os conjuntos. Este procedimento é muito importante, apesar de simples, e deve acontecer nesse estágio da pesquisa, pois se o conjunto de dados completo for usado para executar a etapa de seleção de características, os erros do conjunto de validação e os erros de validação cruzada obtidos não serão estimativas precisas do erro de teste.

O conjunto de dados de teste deve ser utilizado somente na etapa final do estudo, onde o modelo de aprendizado de máquina, construído configurado com os melhores hiperparâmetros validados de forma cruzada, possa ser executado a partir desses dados. Tal abordagem é usada para evitar problemas de ajustes nos dados, como o *Overfitting*, que ocorre quando o modelo se ajusta demais ao conjunto utilizado em seu treinamento, mas se mostra ineficaz ou pouco eficiente na predição de novos dados, neste caso, isso acabaria prejudicando a recomendação de novos itens.

4.4 Seleção de características

Para a seleção das características avaliadas nesse procedimento, existem três técnicas, *Forward and backward stepwise selection* (FBS), *Lasso* e *Partial least squares* (PLS). A técnica *Lasso* foi selecionada pela facilidade e viabilidade de implementação, assim como, pelos testes realizados, se mostrar um modelo mais preciso e mais interpretável, com ênfase para o segundo fator. Ou seja, os resultados obtidos eram mais fáceis de ser interpretados, podendo agregar mais para as etapas posteriores. Por exemplo, na técnica PLS é necessário entender a variação na variável que se deseja prever, para isso, se faz necessária a interpretação dos componentes gerados para cada benefício, como no *Principal component analysis* (PCA), e isso pode não ser trivial.

O modelo *Lasso* seleciona o subconjunto de recursos mais relevantes para a construção do modelo, dentre todos os recursos do conjunto de dados completo. Esse modelo atribui um coeficiente zero para as variáveis que possuem pouca importância para a predição, a implementação utilizada neste trabalho é a disponibilizada e mantida pelo *Scikit*. Identificadas as colunas de

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

maior relevância pelo modelo, o conjunto de dados deve ser atualizado, deixando somente as colunas selecionadas.

4.4.1 Escolha dos melhores hiperparâmetros

O hiper-parâmetro para aplicação da validação cruzada no *Lasso* é o *alpha*, que funciona para controlar a penalidade *l1* do modelo, visto que estas variáveis possuem valor inversamente proporcionais. Ou seja, quanto menor o valor de *alpha*, maior será a penalidade aplicada. Em outras palavras, o modelo ficará mais criterioso, e menos características serão selecionadas para o subconjunto.

Para validar de forma cruzada os resultados obtidos a partir da variação do valor do hiper-parâmetro *alpha*, são aplicadas duas técnicas de ajuste: *GridSearchCV* e *RandomizedSearchCV*, as quais testam exaustivamente o valor hiper-parâmetro, baseando-se no resultado do MSE. A seguinte variação de valores de *alpha* foram testados: 0,05, 0,1, 0,5, 1, 0,0001, 0.0005. O menor resultado de MSE obtido indica o valor de *alpha* que deve ser usado como hiper-parâmetro, na execução do modelo *Lasso*, utilizado para selecionar as características principais do conjunto. Em ambas aplicações das técnicas de ajuste, o valor de k para a técnica do *k-fold* foi definido como 10, que segundo James *et al.* (2013) fornece uma boa aproximação para a taxa de erro do teste.

Dessa forma, para cada valor de *alpha*, dentre os apresentados, 10 execuções foram realizadas. Em cada uma delas, um segmento de dados foi usado para validação e os outros k-1 para aprendizado.

4.5 Implementação do modelo de aprendizado de máquina

Atualmente, existem várias maneiras de implementar algoritmos e construir modelos visando a recomendação de itens, diversas bibliotecas e inclusive *frameworks* fornecem a implementação pronta para uso, como é caso do *Turi Create*[4], biblioteca mantida pela Apple que simplifica o desenvolvimento de modelos personalizados de aprendizado de máquina e pode ser utilizada para problemas de recomendação, detecção de objetos, classificação de imagens, etc. No entanto, devido ao escopo do atual trabalho, onde o objetivo não é comparar diferentes modelos e seus resultados, apenas um algoritmo será utilizado na construção do modelo de aprendizado de máquina, o de regressão logística. A comparação de modelos distintos e seus

[4] https://github.com/apple/turicreate

resultados pode ser algo realizado em um trabalho futuro.

A implementação desse algoritmo já existe e é de código aberto, mantida e disponibilizada pelo *Scikit*.

4.5.1 Validação cruzada para definir o grau do polinômio na regressão polinomial

As aplicações das técnicas de validação cruzada são comumente vistas no cenário de uma regressão linear múltipla, como citado em Browne (2000). Dessa forma, nesta etapa do trabalho a validação cruzada é utilizada para definir o melhor grau do polinômio.

De modo geral, é incomum usar um grau d maior que 3 ou 4, porque para grandes valores de d, a curva polinomial pode se tornar excessivamente flexível e pode assumir algumas formas estranhas e difíceis de serem interpretadas. Isso é especialmente verdadeiro perto do limite da variável X (JAMES *et al.*, 2013). Isto posto, para definir o melhor grau do polinômio, foi definido um espaço amostral de três variações de d, iniciando em 2 e indo até 4. A execução é realizada com 1/3 do conjunto de dados, devido à limitação na capacidade de processamento e utiliza a métrica MSE como avaliador do desempenho. Cada execução também utilizou a técnica *k-fold* na validação cruzada, definindo o k igual a 3. Neste caso, o valor de k foi diferente do apresentado anteriormente, por conta de limitações da máquina que realizou o processamento dos dados.

4.6 Treinamento do modelo

No cenário deste trabalho proposto, o treinamento do modelo de aprendizado de máquina ocorre para cada benefício. O objetivo é fornecer a probabilidade da família ter o benefício em questão. Cada execução deve usar o modelo criado, com os hiperparâmetros definidos, no conjunto de treino. Os dados são novamente divididos em subconjuntos de treino e teste, seguindo a proporção de 70% e 30% respectivamente. Os dados de treino são todas as colunas selecionadas na saída do processo de seleção de características, selecionado na Seção 4.4. Já os dados de teste são os valores para a coluna que representa se a família possui o benefício em questão. Ou seja, o mesmo modelo é executado para os 6 benefícios distintos.

4.7 Execução e avaliação do modelo de aprendizado de máquina com os dados de teste

Para cada benefício, o modelo criado e treinado utilizando toda a base de treino e com os hiperparâmetros determinados na validação cruzada é executado para o subconjunto de dados de teste. Os resultados obtidos serão os resultados apresentados no trabalho. A saída de cada execução é um vetor com duas probabilidades, a primeira, representa a classe 0, ou seja, a probabilidade da família não possuir aquele benefício. Já a segunda, representa a classe 1, que indica a probabilidade da família possuir o benefício. Após essa etapa, em todas as saídas, será executado um processamento, para coletar somente as probabilidades para a classe 1, e em cada saída, caso a probabilidade seja maior que 0,5, o benefício a recomendação será marcada como verdadeira, em uma nova coluna de dados. Caso seja menor que o valor citado, a recomendação será negativa, ou seja, para não conceder o benefício. Após isso, os dados preditos são avaliados a partir das seguintes métricas: *F1-Score*, *precision* e *recall*. Esses resultados serão sumarizados e representados por gráficos, tabelas e por uma matriz de confusão, apresentado em 2.6.3.1.

5 RESULTADOS

Este Capítulo apresenta e discute os resultados obtidos com a execução dos passos presentes na metodologia do trabalho proposto.

5.1 Dados coletados

O conjunto de dados inicial utilizado pelo trabalho é composto por 42.966 entrevistas respondidas e corresponde a um recorte temporal final realizado para o mês de junho de 2022. O início deste recorte temporal é de 2018. Uma possível limitação identificada nesses resultados é que o conjunto de entrevistados não é aleatório, ou seja, o processo das entrevistas acontece conforme a divisão de localidades e segue critérios pré-definidos pelos próprios municípios. Portanto, não ocorre de maneira aleatória, o que pode implicar em algum viés para os dados já coletados. É importante reiterar que cada registro é uma linha no conjunto de dados, e esta corresponde as informações de uma família.

Os resultados exibidos pela Tabela 2 mostram a quantidade de benefícios concedidos, divididos por tipo do benefício. O benefício de cesta básica é o mais concedido, seguido do vale-gás. Ambos se encaixam no contexto de alimentação, o que pode significar que este setor é uma das maiores necessidades das famílias.

A Figura 5 exibe graficamente a distribuição apresentada na tabela anterior. Nesta representação é possível observar visualmente a disparidade entre o benefício mais concedido e os demais.

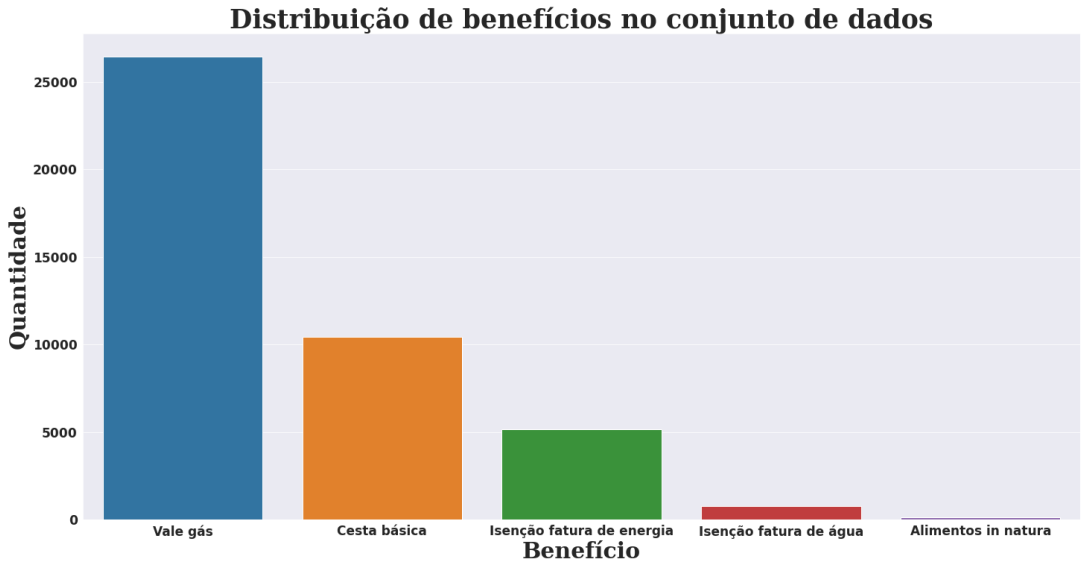
A Figura 6 exibe outra visualização a respeito da distribuição dos benefícios. Nesta representação há a distribuição para a quantidade de famílias que recebem 1, 2, 3, 4 ou 5 benefícios. É importante destacar que cada família pode receber mais de um benefício. É possível verificar que aproximadamente 49% das famílias recebe somente 1 benefício, seguido pelas famílias que recebem dois (34%). As famílias que recebem 3 ou 4 benefícios aparecem

Tabela 2 – Distribuição quantitativa dos benefícios concedidos

Benefício	Número de famílias na amostra
Vale-gás	26444
Cesta básica	10438
Isenção da tarifa de energia	5154
Isenção da tarifa de água	745
Alimentos in natura	147

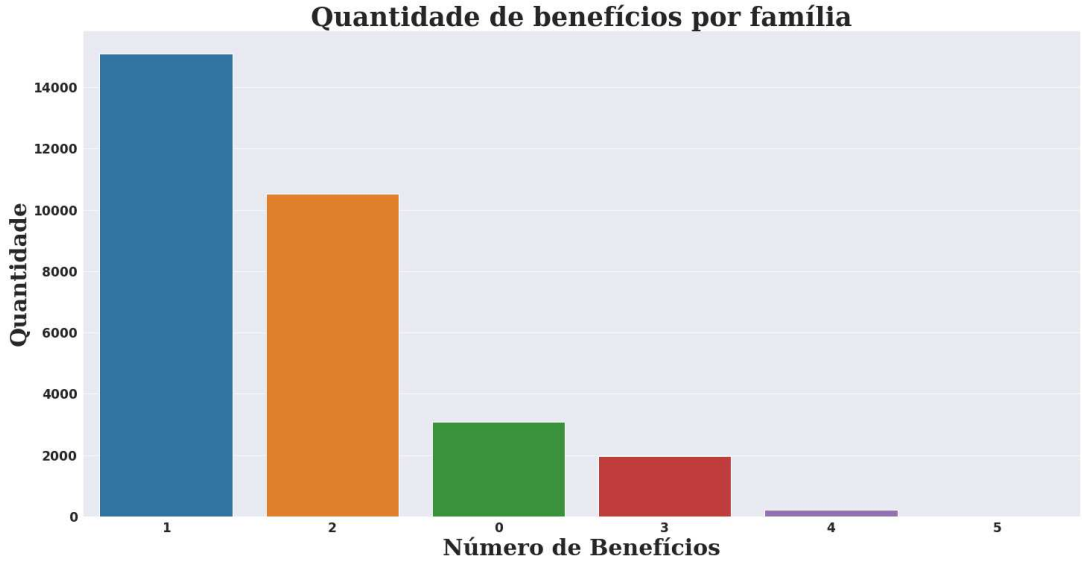
Fonte: elaborada pelo autor.

Figura 5 – Distribuição quantitativa dos benefícios



Fonte: elaborado pelo autor

Figura 6 – Distribuição quantitativa dos benefícios por família



Fonte: elaborado pelo autor

com menor frequência, e somente duas, das 30889, recebem todos os benefícios.

5.2 Pré-processamento

Após a conclusão de todas as etapas de pré-processamento, seguindo os critérios definidos em 4.2.1, foram removidos 12077 registros, representando aproximadamente 28% do conjunto de dados inicial. O conjunto de dados utilizado posteriormente ficou com 30889

registros.

5.3 Engenharia de recursos

Inicialmente, haviam 213 colunas que representavam as características, no conjunto de dados. Após as etapas de engenharia de recursos, como remoção e criação de novas colunas, o conjunto ficou com 312 colunas. O primeiro processo executado foi a remoção de características, onde 9 colunas foram excluídas do conjunto de dados, ou por representarem recursos gerados de forma automática, ou por não agregarem valor para a recomendação, partindo da visão de como elas se relacionam com os benefícios. A Lista 5.3 apresenta as colunas removidas.

- *timer*
- emailAgente
- telefoneContato
- telefoneContatoOutro
- perguntaAnteriorFoiRespondida
- telefoneContatoOutro
- dataDeHoje
- seSimQuem
- agenteAplicouQuestionario
- formularioAplicadoNaCasaDaFamilia

Após isso, como descrito em 4.3, o processo de criação de novas colunas foi realizado para as características categóricas, desde que possuíssem no máximo 8 valores distintos. Dessa forma, para cada coluna que se enquadra no critério citado, novas colunas foram criadas, uma para cada valor possível nos dados, e a coluna inicial foi removida. Dessa forma, houve um acréscimo de 109 colunas.

5.4 Seleção de características

Antes de executar o modelo *Lasso* para selecionar as características mais importantes, as técnicas *Grid Search* e *Randomized Search* foram aplicadas, como descrito em 4.4.1. O melhor valor de *alpha* obtido em 10 execuções foi o de 0,0005. Este valor para o hiperpârametro é baseado nos valores testados exaustivamente e validados, a partir da validação cruzada, para a métrica MSE.

Tabela 3 – Recorte para a tabela-resumo das execuções da técnica *Randomized Search* para o modelo Lasso

Mean fit time (s)	*Alpha*	*Mean test score*	*Rank test score*
30,76	0,0005	-0,0764	1
32,39	0,0001	-0,0765	2
19,83	0,05	-0,0834	3
17,21	0,1	-0,0834	4
7,13	0,5	-0,0837	5
4,18	1	-0,0840	6

Fonte: elaborada pelo autor.

A Tabela 3 apresenta o recorte para o resumo de cada uma das 5 execuções da aplicação da técnica *Randomized Search*, para cada valor de *Alpha* distinto. Neste exemplo, para cada uma das execuções é exibido alguns valores: o *Mean fit time* representa o tempo médio necessário para realizar o treinamento, o *Alpha* representa o valor do hiperparâmetro do respectivo *Alpha*. Já o *Mean test score* representa o resultado médio das execuções para a métrica MSE e *Rank test score* representa qual a posição no *ranking* daquela respectiva execução. É importante realçar que em cada série citada, dez execuções são realizadas, configurando o valor de *k*, da técnica *k-fold*. É possível verificar que o melhor resultado de MSE obtido com a aplicação do *Randomized Search* foi de aproximadamente -0,0764. Este foi o melhor resultado obtido, entre todas as execuções, para o *Randomized Search* e o *Grid Search*.

A tabela completa retorna os resultados de cada execução, a partir de cada segmento de dados, para treino e para teste. Além da média, há também os valores de desvio padrão para cada unidade, como o tempo total da execução, tempo de treinamento e teste, entre outros valores.

Após configurar os hiperparâmetros do modelo, ele foi executado com todos os dados, exceto as colunas que representavam cada um dos benefícios, visto que estes dados foram usados como *label* em passos posteriores a esse. O modelo selecionou 161 colunas e a partir dela um novo conjunto de dados foi criado, mantendo todos os registros (linhas), mas agora somente com as características selecionadas (colunas). Além disso, cinco novos conjuntos de dados foram criados, um para cada benefício. Nestes conjuntos, além das características selecionadas pelo *Lasso*, há a coluna de *jotformId*, que representa um *ID* para a entrevista respondida por cada família, e a coluna com que informa se a família possui o respectivo benefício. A lista completa de colunas selecionadas pode ser verificada no Anexo A.

5.5 Avaliação de desempenho dos modelos no conjunto de treino

Para o conjunto de dados de treino, que corresponde a 90% do conjunto original, os dados foram subdivididos em conjuntos de treino e validação, seguindo a proporção de 70% e 30% , como apresentado em 4.6. O treinamento do modelo de regressão foi realizado e testes de desempenho para as métricas de classificação foram executados. O relatório de testes exibe os resultados para as principais métricas de classificação, que também são utilizadas no presente trabalho: *precision*, *recall* e *f1-score*. Além disso, há os resultados para os métodos de média, apresentados em 2.6.3.4. O objetivo é validar qual método deve ser utilizado para treinar o modelo, que será executado com os dados de teste, posteriormente.

A Tabela 4 exibe a divisão dos dados dos benefícios a partir do rótulo, o que, neste caso, representa se a família possui ou não possui determinado benefício.

Tabela 4 – Tabela-resumo da divisão dos dados dos benefícios a partir do rótulo

Benefício	Não possui o benefício (0)	Possui o benefício (1)
Vale-gás	23.789	4.011
Cesta básica	18.392	9.408
Alimentos *in natura*	27.668	132
Isenção da tarifa de energia	23.154	4.646
Isenção da tarifa de água	27.132	668

Fonte: elaborada pelo autor.

Essa divisão é importante para definir o método de média a ser utilizado pelo modelo. Neste caso, pelo fato dos dados estarem bastante desbalanceados entre as duas classes, é indicado que a média ponderada seja o método selecionado. Todavia, técnicas de balanceamento poderiam ser utilizadas, visando uma melhoria dos resultados. Isso pode se encaixar no escopo de um trabalho futuro.

Os resultados apresentados nesta seção são para o subconjunto de testes, após a aplicação da *train test split*, com 30% (8340 registros) dos dados. As tabelas 5, 6, 7, 8 e 9 representam os valores obtidos para as métricas já citadas anteriormente.

É possível observar que para todos os benefícios, a média ponderada apresenta os melhores resultados, para as métricas selecionadas. Isto justifica o que foi citado no início desta Seção, dado que as classes estão desbalanceadas. No entanto, é possível observar que à medida que a diferença na quantidade de benefícios entre as duas classes aumenta, os resultados se aproximam de 1, como nos casos em que o resultado das métricas chega a 0,99. Este valor não necessariamente indica um ótimo resultado, podendo significar um *overfitting* no modelo.

Tabela 5 – Desempenho dos dados de treino para o benefício Vale-gás

	Precision	*Recall*	*F1-Score*	Número de registros
Não possui o benefício Vale-gás	0,42	0,00	0,01	1.210
Possui o benefício Vale-gás	0,86	1,00	0,92	7.130
Média micro	*	*	0,85	8.340
Média macro	0,64	0,50	0,46	8.340
Média ponderada (*weighted*)	0,79	0,85	0,79	8.340

Fonte: elaborada pelo autor.

Tabela 6 – Desempenho dos dados de treino para o benefício Cesta básica

	Precision	*Recall*	*F1-Score*	Número de registros
Não possui o benefício Cesta básica	0,66	0,97	0,79	5.496
Possui o benefício Cesta básica	0,47	0,07	0,08	2.844
Média micro	*	*	0,66	8.340
Média macro	0,57	0,51	0,44	8.340
Média ponderada (*weighted*)	0,60	0,66	0,55	8.340

Fonte: elaborada pelo autor.

Tabela 7 – Desempenho dos dados de treino para o benefício Isenção da tarifa de energia

	Precision	*Recall*	*F1-Score*	Número de registros
Não possui o benefício Isenção da tarifa de energia	0,84	1,00	0,91	6.965
Possui o benefício Isenção da tarifa de energia	0,14	0.00	0.00	1.375
Média micro	*	*	0,83	8.340
Média macro	0,49	0,50	0,46	8.340
Média ponderada (*weighted*)	0,72	0,83	0,76	8.340

Fonte: elaborada pelo autor.

Tabela 8 – Desempenho dos dados de treino para o benefício Isenção da tarifa de água

	Precision	*Recall*	*F1-Score*	Número de registros
Não possui o benefício Isenção da tarifa de água	0,98	1,00	0,99	8.149
Possui o benefício Isenção da tarifa de água	0,00	0,00	0,00	191
Média micro	*	*	0,98	8.340
Média macro	0,49	0,50	0,49	8.340
Média ponderada (*weighted*)	0,95	0,98	0,97	8.340

Fonte: elaborada pelo autor.

5.6 Avaliação de desempenho dos modelos no conjunto de teste

Após a etapa de treinamento do modelo de aprendizado de máquina, para cada um dos benefícios foi executada a predição, utilizando o subconjunto de dados de teste como entrada. É importante realçar que estes dados nunca haviam sido utilizados pelo modelo, como previsto em 4.3.1 e 4.7. Para cada um dos benefícios, o modelo foi treinado com o conjunto de treino

Tabela 9 – Desempenho dos dados de treino para o benefício Alimentos *in natura*

	Precision	*Recall*	*F1-Score*	Número de registros
Não possui o benefício Alimentos *in natura*	1,00	1,00	1,00	8.296
Possui o benefício Alimentos *in natura*	0,30	0,07	0,11	44
Média micro (*weighted*)	*	*	0,99	8.340
Média macro (*weighted*)	0,65	0,53	0,55	8.340
Média ponderada (*weighted*)	0,99	0,99	0,99	8.340

Fonte: elaborada pelo autor.

Tabela 10 – Avaliação de desempenho dos dados de teste para o benefício Vale-gás

	Precision	*Recall*	*F1-Score*	Número de registros
Não possui o benefício Vale-gás	0,80	0,01	0,02	434
Possui o benefício Vale-gás	0,86	1,00	0,92	2.655
Média ponderada (*weighted*)	0,85	0,86	0,80	3.089

Fonte: elaborada pelo autor.

e validação, os dados de teste, que não possuem rótulos, foram usados como entrada para as predições de probabilidade.

Os resultados obtidos após a execução, em cada um dos benefícios, estão resumidos nas Tabelas 10, 11, 12, 13 e 14. Os valores para as métricas selecionadas foram apresentados e diferenciados pelas classes. Além disso, a última linha de cada tabela traz o resultado de cada métrica utilizando o cálculo a partir da média ponderada. Além disso, as tabelas citadas contêm a informação do número de registros, também divididos por classe e na perspectiva total.

Tabela 11 – Avaliação de desempenho dos dados de teste para o benefício Cesta básica

	Precision	*Recall*	*F1-Score*	Número de registros
Não possui o benefício Cesta básica	0,67	0,98	0,80	2.059
Possui o benefício Cesta básica	0,50	0,05	0,09	1.030
Média ponderada (*weighted*)	0,61	0,67	0,56	3.089

Fonte: elaborada pelo autor.

O benefício Isenção de água de energia obteve os melhores resultados para *precision*,

Tabela 12 – Avaliação de desempenho dos dados de teste para o benefício Isenção da fatura de energia

	Precision	*Recall*	*F1-Score*	Número de registros
Não possui o benefício Isenção da fatura de energia	0,84	1,00	0,91	2.581
Possui o benefício Isenção da fatura de energia	0,30	0,01	0,01	508
Média ponderada (*weighted*)	0,75	0,83	0,76	3.089

Fonte: elaborada pelo autor.

Tabela 13 – Avaliação de desempenho dos dados de teste para o benefício Isenção da fatura de água

	Precision	*Recall*	*F1-Score*	Número de registros
Não possui o benefício Isenção da fatura de água	0,98	1,00	0,99	3.012
Possui o benefício Isenção da fatura de água	0,00	0,00	0,00	77
Média ponderada (*weighted*)	0,95	0,97	0,96	3.089

Fonte: elaborada pelo autor.

Tabela 14 – Avaliação de desempenho dos dados de teste para o benefício Alimentos *in natura*

	Precision	*Recall*	*F1-Score*	Número de registros
Não possui o benefício Alimentos *in natura*	1,00	1,00	1,00	3.074
Possui o benefício Alimentos *in natura*	0,00	0,00	0,00	15
Média ponderada (*weighted*)	0,75	0,86	0,80	3.089

Fonte: elaborada pelo autor.

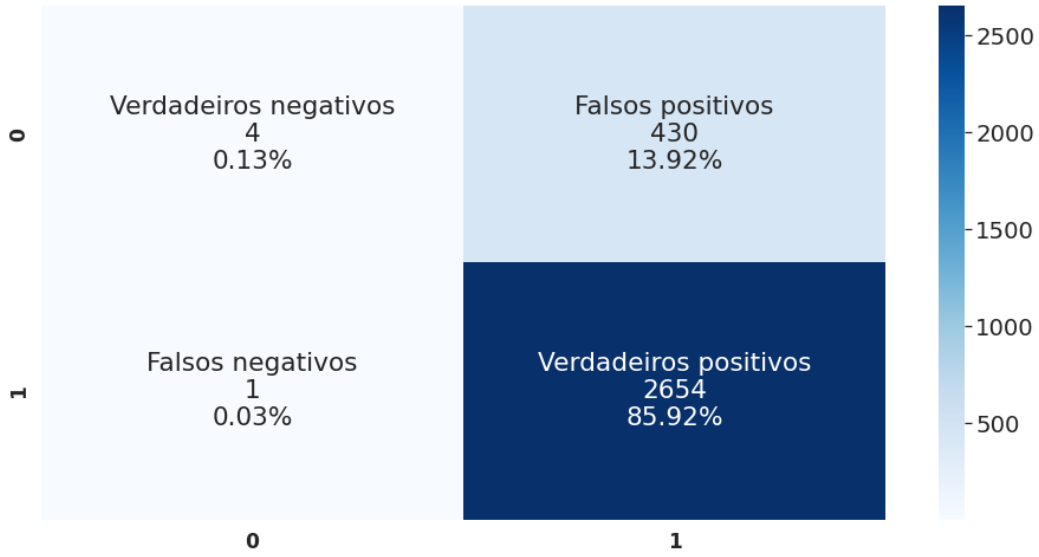
recall e *f1-score*. Já o benefício de Cesta básica apresentou resultados medianos, com um *f1* de 0,56. Além disso, é possível verificar que os resultados para as duas classes se mostraram desequilibrados. Pelo que foi observado, no caso em que as classes não estão tão desbalanceadas, essa média pode não ser a melhor opção a ser escolhida, visto que a média não considera as classes individualmente. Uma sugestão para trabalhos futuros é que cada benefício possa usar uma estratégia específica para calcular cada uma das métricas, baseado inclusive na distribuição dos valores entre as classes. Por exemplo, em um caso, a média *micro*, em outro a média *macro*, etc. Isso será possível, mediante o fato que o fluxo de treino, validação e teste de cada benefício ocorre de forma individual, já que para usar os dados de teste para determinado benefício, é necessário e interessante que ele tenha sido treinado com os dados do mesmo.

Outra forma de visualizar os resultados é através da matriz de confusão. Para cada benefício, há uma matriz que compara os valores reais e os valores preditos.

Na Figura 7 a matriz de confusão para o benefício Vale-gás traz algumas informações que devem ser destacadas: 2654 registros se caracterizam como verdadeiros positivos, neste caso, as famílias possuíam o benefício e o modelo o recomendou. Além disso, 430 registros foram falsos positivos, o que ocorre quando o modelo realiza a recomendação, mas a família não possuía o benefício. Este último deve ser um dos pontos de atenção deste trabalho, visto que pode ser algo utilizado para incluir mais famílias nos programas socioassistenciais e conceder a elas outros benefícios.

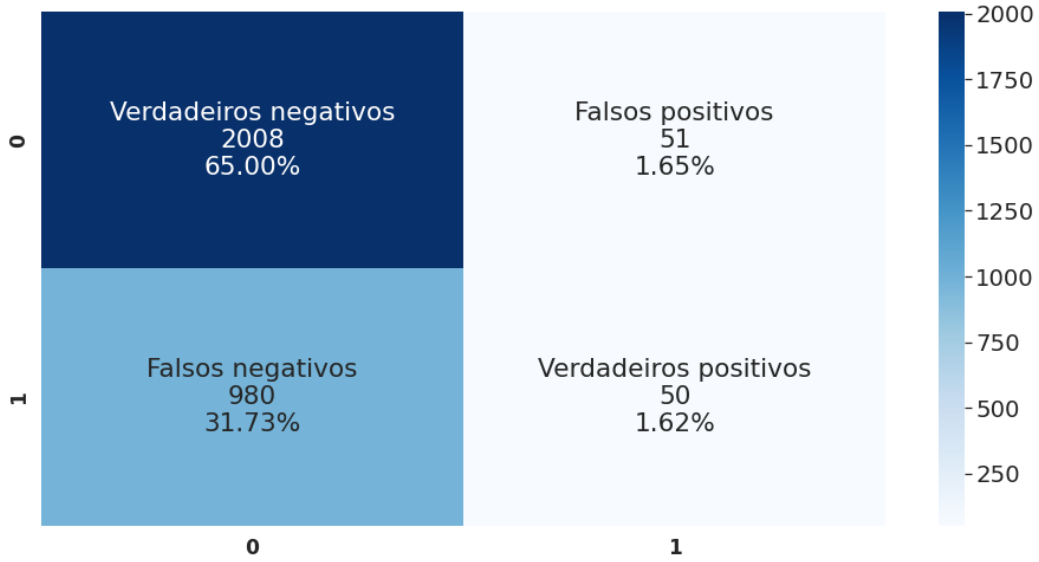
Nesta Figura 8, os resultados mostram 51 registros como falsos positivos, o que

Figura 7 – Matriz de confusão para predição do benefício Vale-gás



Fonte: elaborado pelo autor

Figura 8 – Matriz de confusão para predição do benefício Cesta básica



Fonte: elaborado pelo autor

indica que o valor real dessa recomendação é *false*, mas o modelo a recomendou como *true* para esses registros. Esses dados correspondem a apenas 1,65% dos dados de teste. Ademais, 2008 registros foram indicados como verdadeiros negativos. Ou seja, a família não possui o benefício e o modelo não o recomendou, esse tipo de informação pode servir como validação das concessões não realizadas, em caso de automatização. Por último, houveram 980 registros

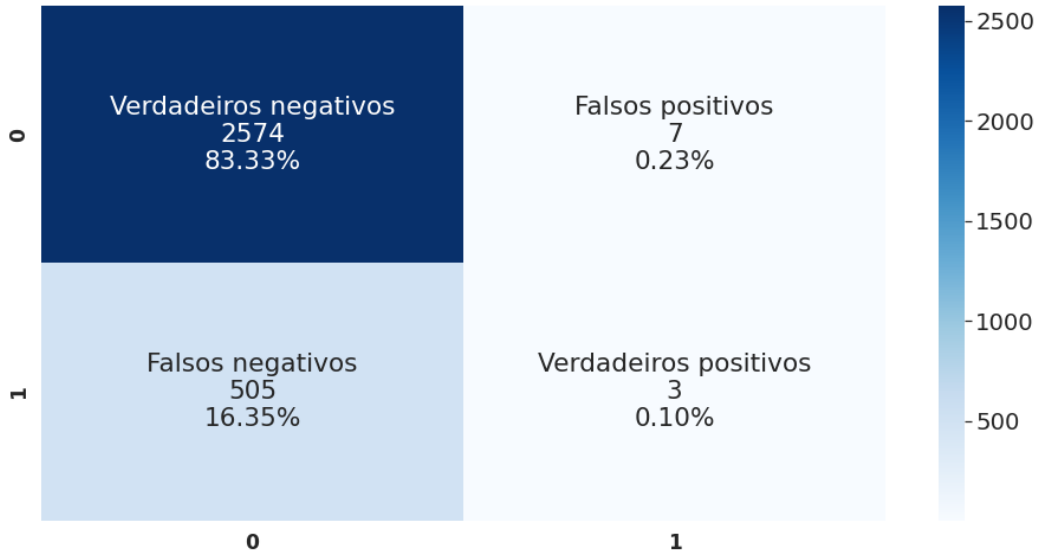
Tabela 15 – Recomendações dos modelos para cada benefício

Benefício	Recomendações positivas (P > 0,5)	Recomendações negativas (P < 0,5)
Vale-gás	3087	2
Cesta básica	101	2988
Alimentos in natura	5	3085
Isenção da tarifa de energia	10	3079
Isenção da tarifa de água	3	3086
Total	3.206	12.240

Fonte: elaborada pelo autor.

considerados falsos negativos, cujos modelo não realizou a recomendação, mas os seus valores reais são positivos. Este é outro ponto de atenção para o trabalho, pois, em caso de utilização do modelo como ferramenta de apoio para novos registros, este poderia ser um cenário onde um benefício está sendo negado, quando na verdade a família tem características semelhantes com outras que recebem o mesmo benefício.

Figura 9 – Matriz de confusão para predição do benefício Isenção da fatura de energia

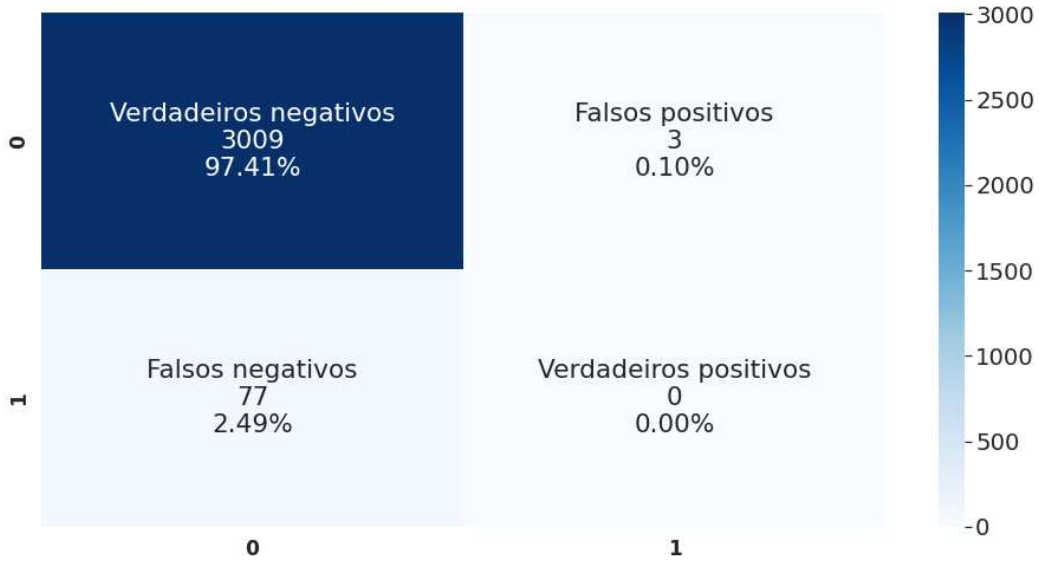


Fonte: elaborado pelo autor

5.7 Avaliação do sistema de recomendação

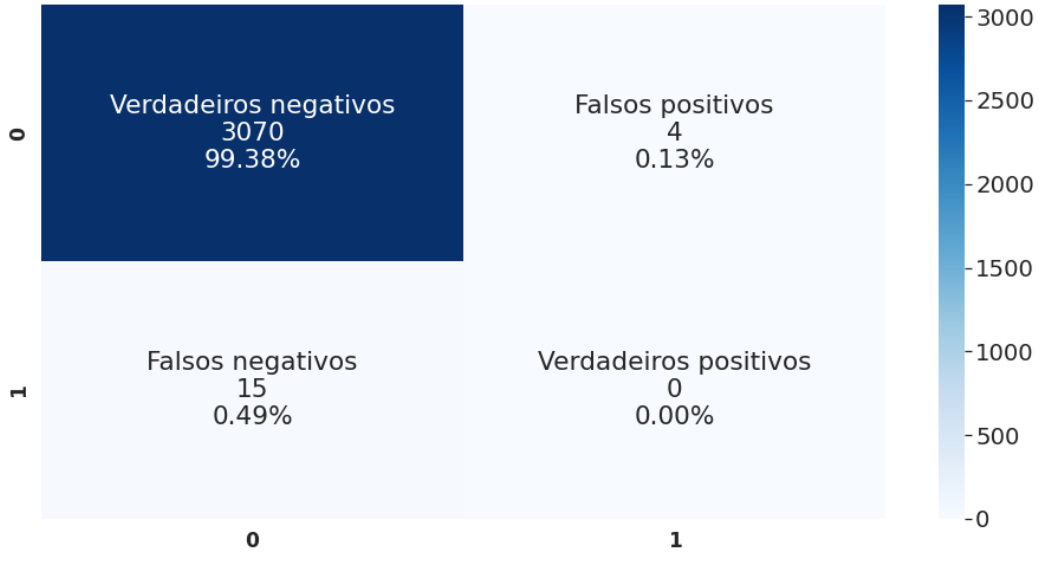
A Tabela 15 exibe uma comparação entre as decisões tomadas por cada modelo criado, executado no conjunto de testes, a partir de cada benefício. A decisão baseada na probabilidade define se o benefício deve ou não ser recomendado para cada família.

Figura 10 – Matriz de confusão para predição do benefício Isenção da fatura de água



Fonte: elaborado pelo autor

Figura 11 – Matriz de confusão para predição do benefício Alimentos *in natura*



Fonte: elaborado pelo autor

Pelos resultados, é possível verificar que aproximadamente 80% das saídas probabilísticas possuíam valor menor que 0,5, e por isso, a recomendação do benefício foi negativa. Já a proporção de recomendações positivas foi de aproximadamente 20% . O benefício Cesta básica foi o único cujo número de recomendações positivas foi maior que o de negativas. Ou seja, em uma interpretação real e baseando-se nas famílias do conjunto de testes, este benefício

está sendo concedido de maneira assertiva. Outra observação a respeito é que as recomendações negativas não significam erros do modelo, mas sim, em grande maioria, benefícios que não estão sendo concedidos, quando de fato, não deveriam. Essa visualização é melhor explorada de forma individual, nas matrizes de confusão apresentadas na Seção 5.6.

Os resultados obtidos podem possuir uma variância considerável, porque foram obtidos a partir de uma pequena amostra, de testes. Ademais, os dados das classes cujo modelo foi treinado estavam desbalanceadas de duas formas: entre as classes, e na própria classe. Isso significa que a proporção de dados é bem diferente, entre os benefícios, e que, além disso, em um benefício há um desequilíbrio relevante entre os valores. Tais fatos podem ser considerados limitações para o trabalho proposto.

Para avaliar o sistema de recomendação construído, dois novos subconjuntos de dados foram criados, o primeiro com todas as predições realizadas para os dados de teste, para cada um dos benefícios, e o segundo com os valores reais. Dessa forma, tem-se um conjunto de várias predições, para todos os benefícios, em uma situação que se caracteriza como multi-rótulos. Além disso, os valores reais servem para identificar a assertividade do sistema de recomendação. A métrica utilizada foi a *Label ranking average precision* e a pontuação obtida foi de 0,28. Isto pode significar que para cada rótulo atribuído como positivo, a fração de rótulos com classificação diferente de positivo, assim como os resultados entre as amostras, pode estar desbalanceado, como citado em 2.6.4. Os dados retornadas pelo sistema de recomendação representam as probabilidades, em ordem respectiva, para os benefícios cesta básica, vale-gás, isenção da tarifa de energia, isenção da tarifa de água e alimentos *in-natura*. Uma propriedade *rows* indica o número de registros, que corresponde com o conjunto de testes, cujo corresponde a 10% do conjunto inicial. Os registros não ficaram divididos em *arrays* de 5 posições (uma probabilidade por benefício), o que pode ser considerado uma limitação para a visualização. No entanto, seguindo a ordem lógica dos registros e a ordem dos dados de entrada dos benefícios, citada acima, a primeira probabilidade representa o benefício cesta básica, para a primeira família do conjunto, a segunda probabilidade representa o benefício vale-gás, para a primeira família, e assim por diante.

6 CONCLUSÃO E TRABALHOS FUTUROS

A partir deste trabalho, foi possível realizar novas etapas de pré-processamento, que tratavam os dados inválidos, para todas as colunas do conjunto de dados original. Essa etapa demandou grande esforço, primeiro na identificação de todos os tipos de irregularidades, seja no formato, nas unidades ou nas escalas apresentadas pelas informações. Outro fator importante é que não é possível obter uma rastreabilidade confiável desses dados inválidos, o que tornou necessário a remoção de todo o registro, mesmo que a informação seja em uma única coluna.

Com base nos resultados obtidos a partir dos experimentos exibidos no Capítulo 5, principalmente os distribuídos entre as Seções 5.2, 5.4, 5.6, 5.6 e 5.7, é possível verificar o potencial dos dados extraídos dos formulários CMIC, com o intuito de otimizar a entrega dos benefícios, para as famílias que participam dos programas socioassistenciais. Um problema observável é a imensidão dos dados, que se apresentam em muitas dimensões, podendo configurar inclusive o cenário de problema comum na computação, conhecido como a maldição da dimensionalidade. No entanto, como foi mostrado, é possível reduzir essas dimensões, mesmo sem aplicar técnicas de redução de dimensionalidade, algo que pode ser feito em um trabalho futuro, através de processos relacionados com a seleção de características, onde a partir de um modelo, as principais características são identificadas e mantidas, já as outras, são excluídas do conjunto principal.

Também é possível identificar algumas limitações, que podem ser tratadas em trabalhos futuros. Com relação ao particionamento dos dados, podem ser testadas novas proporções, que aumentem a quantidade de dados de teste, como, por exemplo, 70% e 30% . Além disso, um balanceamento entre as classes é essencial, pois, como foi observado, isso pode influenciar o modelo de recomendação, em caso de classes que possuem poucos registros ou registros concentrados em uma única classe. Para este problema, trabalhos futuros devem considerar a utilização de técnicas de balanceamento, como *Smote* e *Tomek links*.

Ainda sobre o conjunto de dados, é necessário realçar suas limitações não técnicas. O conjunto utilizado para treinamento, cujo rótulo representa se a família possui ou não determinado benefício, para cada um dos benefícios, não é visto como completamente íntegro e verídico. Isto se dá pelo fato de como ocorre a concessão dos benefícios, atualmente, podendo haver casos de fraudes, manipulações e manobras políticas no processo de concessão.

Como maior contribuição, neste trabalho foi criado um modelo capaz de recomendar um benefício específico, para novos registros, ou registros desconhecidos por ele. Essa reco-

mendação, no cenário atual, serviu para verificar e comparar as classes reais com as classes preditas, através da matriz de confusão e dos seus valores: verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Cada um desses atributos pode ser utilizado e analisado de forma distinta pelos responsáveis, seja no sentido de recomendar novos benefícios para as famílias ou de identificar erros em benefícios já concedidos.

REFERÊNCIAS

ALPAYDIN, E. **Introduction to machine learning**. [*S. l.*]: MIT press, 2020.

BARLOW, H. B. Unsupervised learning. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 1, n. 3, p. 295–311, 1989.

BAS, S.; BHAROSA, N. N.; SPOELSTRA, F. F.; VOORT, H. H. van der; JANSSEN, M. M. **Inclusion through proactive public services:** findings from the netherlands. [S.l: s.n], 2021.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 35, n. 8, p. 1798–1828, 2013.

BILL, M. e-governance: Towards a practioner's definition. **American Society for Public Administration**, v. 13, 2002.

BOBADILLA, J.; ORTEGA, F.; HERNANDO, A.; GUTIÉRREZ, A. Recommender systems survey. **Knowledge-based systems**, Elsevier, v. 46, p. 109–132, 2013.

BROWNE, M. W. Cross-validation methods. **Journal of mathematical psychology**, Elsevier, v. 44, n. 1, p. 108–132, 2000.

CANDILLIER, L.; JACK, K.; FESSANT, F.; MEYER, F. State-of-the-art recommender systems. In: **Collaborative and Social Information Retrieval and Access**: Techniques for improved user modeling. [*S. l.*]: IGI Global, 2009. p. 1–22.

CHUN, S. A.; LUNA-REYES, L. F.; SANDOVAL-ALMAZÁN, R. Collaborative e-government. **Transforming Government:** People, Process and Policy, Emerald Group Publishing Limited. [S.l], 2012.

CUNNINGHAM, P.; CORD, M.; DELANY, S. J. Supervised learning. In: **Machine learning techniques for multimedia**. [*S. l.*]: Springer, 2008. p. 21–49.

DENER HUBERT NII-APONSAH, L. E. G. C.; JOHNS, K. D. **GovTech Maturity Index**: the state of public sector digital transformation. [*S. l.*]: World Bank, 2021.

DESHPANDE, M.; KARYPIS, G. **Item-based top-n recommendation algorithms**. **ACM Transactions on Information Systems (TOIS)**, ACM New York, NY, USA, v. 22, n. 1, p. 143–177, 2004.

FOREMAN, J. W.; JENNINGS, G.; MILLER, E. **Data smart**: Using data science to transform information into insight. [*S. l.*]: Wiley Indianapolis, 2014. v. 1.

GUO, X.; LU, J. Intelligent e-government services with personalized recommendation techniques. **International journal of intelligent systems**, Wiley Online Library, v. 22, n. 5, p. 401–417, 2007.

HEATON, J. An empirical analysis of feature engineering for predictive modeling. In: **SoutheastCon 2016**. [*S. l.*]: IEEE, 2016. p. 1–6.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. [*S. l.*]: Springer, 2013. v. 112.

KOLACHALAM, S. An overview of e-government. **Economia Aziendale Online**. [S.l], n. 1, p. 1–12, 2012.

LU, J.; WU, D.; MAO, M.; WANG, W.; ZHANG, G. Recommender system application developments: a survey. **Decision Support Systems**, Elsevier, v. 74, p. 12–32, 2015.

LUSHER, S. J.; MCGUIRE, R.; SCHAIK, R. C. van; NICHOLSON, C. D.; VLIEG, J. de. Data-driven medicinal chemistry in the era of big data. **Drug discovery today**, Elsevier, v. 19, n. 7, p. 859–868, 2014.

MARSLAND, S. **Machine learning**: an algorithmic perspective. [*S. l.*]: CRC press, 2015.

MEO, P. D.; QUATTRONE, G.; URSINO, D. A decision support system for designing new services tailored to citizen profiles in a complex and distributed e-government scenario. **Data & Knowledge Engineering**, Elsevier, v. 67, n. 1, p. 161–184, 2008.

PARRA, D.; SAHEBI, S. Recommender systems: Sources of knowledge and evaluation metrics. In: **Advanced techniques in web intelligence-2**. [*S. l.*]: Springer, 2013. p. 149–175.

PAZZANI, M. J.; BILLSUS, D. Content-based recommendation systems. In: **The adaptive web**. [*S. l.*]: Springer, 2007. p. 325–341.

PROVOST, F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. **Big data**, **Mary Ann Liebert, Inc.** 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 1, n. 1, p. 51–59, 2013.

QUIJANO-SÁNCHEZ, L.; CANTADOR, I.; CORTÉS-CEDIEL, M. E.; GIL, O. Recommender systems for smart cities. **Information systems**, Elsevier, v. 92, p. 101545, 2020.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. Encyclopedia of database systems. **Springer**. [S.l], v. 5, p. 532–538, 2009.

SCOTT, S.; MATWIN, S. Feature engineering for text classification. In: **ICML**. [*S. l.*: *s. n.*], 1999. v. 99, p. 379–388.

SILCOCK, R. What is e-government. **Parliamentary affairs**. Oxford University Press, [S.l], v. 54, n. 1, p. 88–101, 2001.

TAKAHASHI, K.; YAMAMOTO, K.; KUCHIBA, A.; KOYAMA, T. Confidence interval for micro-averaged f1 and macro-averaged f1 scores. **Applied Intelligence**, Springer, v. 52, n. 5, p. 4961–4972, 2022.

TAMBOURIS, E.; GORILAS, S.; BOUKIS, G. **Investigation of electronic government**. [S.l]: Citeseer, 2001.

VISA, S.; RAMSAY, B.; RALESCU, A. L.; KNAAP, E. V. D. Confusion matrix-based feature selection. **MAICS**. [S.l], v. 710, n. 1, p. 120–127, 2011.

WIERING, M. A.; OTTERLO, M. V. Reinforcement learning. **Adaptation, learning, and optimization**. **Springer**. [S.l], v. 12, n. 3, p. 729, 2012.

ZHOU, Z.-H. A brief introduction to weakly supervised learning. **National science review**. Oxford University Press. [S.l], v. 5, n. 1, p. 44–53, 2018.

ZHU, X.; GOLDBERG, A. B. Introduction to semi-supervised learning. **Synthesis lectures on artificial intelligence and machine learning**, Morgan & Claypool Publishers. [S.l], v. 3, n. 1, p. 1–130, 2009.

APÊNDICE A – EXEMPLOS DE RESULTADOS

Figura 12 – Exemplo de 5 famílias aleatórias: comparação da concessão e recomendação

index	jotform_id	vale_gas_pred	beneficios_vale_gas	cesta_basica_pred	beneficios_cesta_basica
860	14962	1	1	0	0
1294	11655	1	1	0	1
1130	2320	1	1	0	1
1095	20372	1	1	0	1
1638	13372	1	1	0	1

Fonte: elaborado pelo autor

Figura 13 – Exemplo de 5 famílias aleatórias: comparação da concessão e recomendação

index	jotform_id	isencao_tarifa_energia_pred	beneficios_isencao_tarifa_energia	isencao_tarifa_agua_pred	beneficios_isencao_tarifa_agua
382	8466	0	0	0	0
2441	30487	1	1	0	0
1177	5394	0	0	0	0
1556	4969	0	0	0	0
2984	13211	0	0	0	0

Fonte: elaborado pelo autor