Faculdades Integradas de Taquara - Faccat
Av. Oscar Martins Rangel, 4.500
Taquara, RS, CEP 95600-000

Curso de Sistemas de Informação

CENTRAL DE AVALIAÇÕES DE PRODUTOS: APLICAÇÃO DA ANÁLISE DE AGRUPAMENTO POR APRENDIZADO DE MÁQUINA PARA RECOMENDAÇÃO DE AVALIAÇÕES DE PRODUTOS[1]

Eduardo Arndt[2]

Marcelo Cunha de Azambuja[3]

RESUMO

Este projeto apresenta um sistema de avaliação de produtos que utiliza práticas de aprendizado de máquina para facilitar a decisão de compra das pessoas que buscam mercadorias na internet. Um sistema análogo é o site Reclame Aqui, que age como ponto central de reclamações de empresas. Tal-qualmente, o projeto tem o objetivo de expor as opiniões das pessoas a respeito de produtos, centralizando as diversas avaliações realizadas em um único local. Além disso, outra finalidade do projeto é indicar ao usuário quais avaliações são relevantes para ele, dentre a grande quantidade de opiniões compartilhadas. Para agrupar os usuários cadastrados por semelhança de características, no decorrer do artigo, são expostos estudos realizados sobre aprendizado de máquina e algoritmos de agrupamento. Além disso, o artigo apresenta tecnologias e práticas para implantação de um sistema web integrado ao aprendizado de máquina, utilizando serviços do AWS para gerenciar produtos, avaliações e usuários. A integração do sistema de aprendizado de máquina e aplicação web formam a Central de Avaliações, sistema obtido como resultado deste projeto.

Palavras-chave: Aprendizado de Máquina. Agrupamento. Avaliações. Desenvolvimento de Software.

[1] Trabalho de Conclusão de Curso - 2022.

[2] Acadêmico do curso de Sistemas de Informação das Faculdades Integradas de Taquara - Faccat/RS. E-mail: eduardoarndt@sou.faccat.br.

[3] Orientador – Doutor em Ciência da Computação. E-mail: marcelo.azambuja@gmail.com.

1 INTRODUÇÃO

O presente artigo apresenta um trabalho de pesquisa e desenvolvimento com o objetivo de propor uma nova forma de lidar com as avaliações de produtos em portais na internet, realizando a indicação de avaliações relevantes para o usuário que procura resenhas de mercadorias.

Atualmente, as avaliações de produtos se encontram descentralizadas em diversos canais pela internet, como e-commerces, redes sociais, blogs e plataformas de vídeo. Isso implica que, para encontrar um conteúdo relevante e útil, uma pessoa pode navegar em dezenas de sites na internet e, mesmo assim, ficar limitada a avaliações somente em estrelas de 1 a 5, ou acabar por encontrar avaliações realizadas por pessoas que adquiriram o produto com expectativas e/ou intuitos diferentes.

Através de relatos de experiência espontâneos recebidos ao longo do desenvolvimento deste trabalho, diversos usuários de e-commerces afirmaram encontrar dificuldades no processo de decisão de compra. Ao buscarem, por exemplo, um novo smartphone, a falta de conhecimento técnico dificultou a escolha do produto, por não entenderem o que é uma boa quantidade de memória, resolução de câmera ou capacidade de processamento. É comum que os usuários apliquem a regra do “quanto mais, melhor” e acabem pagando por recursos dos quais nunca tirarão total proveito, enquanto um equipamento inferior atenderia perfeitamente as suas necessidades.

A Amazon, maior empresa de comércio eletrônico do mundo, através de uma publicação em seu blog[4], afirmou que avaliações falsas ou “‘incentivadas” são um problema, e que a companhia removeu 200 milhões de avaliações falsas do site, no ano de 2020.

Outra situação que justifica a implantação de um sistema aberto de avaliações é o poder que se há sobre a informação. É comum a prática de deletar avaliações negativas que os usuários escrevem em alguns portais, como, por exemplo, na loja de aplicativos Google Play, da empresa Google, onde centenas de milhares de avaliações negativas foram removidas do aplicativo Robinhood, pertencente à

[4] https://www.aboutamazon.com/news/how-amazon-works/creating-a-trustworthy-reviews-experience

plataforma de investimentos de mesmo nome, conforme reportado pelo jornalista Noah Manskar, do jornal The New York Post[5], em janeiro de 2021. Circunstância semelhante ocorreu na loja virtual da empresa Apple, onde as avaliações na página de um produto foram removidas e a função de escrever avaliações foi desabilitada por completo após o produto atingir nota de avaliação baixíssima, de acordo com o portal de notícias Variety[6].

Considerando soluções para os problemas apresentados, surge a ideia de um sistema que permite que uma pessoa, ao pesquisar sobre um produto de seu interesse, encontre avaliações mais profundas e significativas, escritas por pessoas semelhantes a ela – semelhança em termos de características que podem ser coletadas do usuário, como idade, profissão, sexo, etc.

O presente trabalho tem o objetivo geral de estudar, implementar e validar a metodologia de análise de agrupamentos, com intuito de melhorar a experiência de busca por avaliações de produtos na internet, tendo em vista os problemas citados anteriormente. Desenvolver um sistema de indicação de avaliações relevantes para o usuário a partir da análise de seu perfil, sinalizando como relevantes as avaliações de pessoas semelhantes ao usuário, assim servindo como um facilitador à decisão de compra. Como objetivos específicos, apresentam-se: (i) pesquisar sobre métodos de classificação e agrupamento e como estes podem ser úteis para o presente cenário; (ii) determinar o método de agrupamento mais adequado para o problema apresentado; (iii) implementar um sistema de aprendizado de máquina que agrupe pessoas baseado na semelhança de características; (iv) implementar um sistema web com estratégias de desenvolvimento cloud e integrado ao sistema aprendizado de máquina.

Para alcançar os objetivos descritos acima, foram feitas pesquisas na área de análise de dados e aprendizado de máquina, mais especificamente de aprendizado não supervisionado. A partir dessas pesquisas foi possível constatar que um dos métodos possíveis para resolução do problema foco desta pesquisa é o uso de algoritmos de agrupamento.

---

[5] https://nypost.com/2021/01/29/google-deletes-thousands-of-negative-robinhood-app-reviews
[6] https://variety.com/2019/digital/news/apple-deletes-customer-reviews-ratings-online-store-1203412128/

Como resultado, foi desenvolvido o sistema intitulado “Central de Avaliações de Produtos”, que identifica avaliações relevantes para um usuário dentro de um protótipo de central de avaliações, através da utilização de práticas de aprendizado de máquina e desenvolvimento cloud. Este protótipo evidencia que a análise de agrupamento oferece capabilidade de melhorar o setor de avaliações de produtos na internet, predisposto da quantidade e qualidade de dados de entrada.

2 REFERENCIAL TEÓRICO

2.1 Aprendizado de Máquina

De acordo com IBM Cloud Education (2020a), o aprendizado de máquina é um ramo da inteligência artificial (IA) e da ciência da computação que se concentra no uso de dados e algoritmos para imitar a maneira como os humanos aprendem, melhorando gradualmente sua precisão.

Para Brown (2021), o aprendizado de máquina começa com dados – números, fotos ou texto, como transações bancárias, fotos de pessoas ou até registros de uma padaria, séries de dados de sensores ou relatórios de vendas. Após os dados serem coletados, eles são separados em duas coleções, uma para treinamento e outra para teste. Com a coleção de treinamento, o modelo de aprendizado de máquina é treinado, isto é, aprende os padrões da coleção.

Brown (2021) complementa que, os programadores escolhem um modelo de aprendizado de máquina para usar, fornecem os dados e desenvolvem o modelo de aprendizado de máquina que se treine para encontrar padrões ou fazer previsões. O programador também pode ajustar o modelo, alterando seus parâmetros, para ajudá-lo a obter resultados mais precisos.

Entende-se então que o aprendizado de máquina, neste trabalho, pode ser utilizado para classificar as pessoas em diferentes grupos, com o objetivo de relacionar as experiências que um indivíduo tem com os produtos que adquire à sua classe identificada. Para atingir esta meta, é preciso entender os tipos de aprendizado de máquina que existem.

2.1.1 Aprendizado Supervisionado

Sobre os modelos de aprendizado de máquina supervisionados, Brown (2021) diz que estes são treinados com conjuntos de dados rotulados, que permitem que os modelos aprendam e se tornem mais precisos ao longo do tempo. Por exemplo, um algoritmo seria treinado com fotos de cães e outras fotos diversas, todas rotuladas por humanos indicando quais imagens são realmente de cães, e a máquina aprenderia maneiras de identificar fotos de cães por conta própria após compreender os padrões existentes. O aprendizado de máquina supervisionado é o tipo mais comumente usados hoje.

Entende-se que a aprendizagem de máquina supervisionada não é a solução ideal para o presente trabalho, pois, os usuários cadastrados no sistema não são rotulados, isto é, não se há o entendimento de qual grupo um usuário pertence.

Mesmo que fossem utilizadas bases de dados com rótulos, estas normalmente são orientadas a problemas específicos, diferentes do proposto no presente artigo, e acabam por apresentar dados enviesados que afetam a eficácia do algoritmo.

2.1.2 Aprendizado Não Supervisionado

A respeito do aprendizado de máquina não supervisionado, Brown (2021) discorre que são utilizados algoritmos que procuram padrões em dados não rotulados.

Esses algoritmos descobrem padrões ocultos em dados sem a necessidade de intervenção humana. Sua capacidade de descobrir semelhanças e diferenças nas informações o torna a solução ideal para análise exploratória de dados, estratégias de venda cruzada, segmentação de clientes e reconhecimento de imagens (IBM CLOUD EDUCATION, 2020b).

Além disso, há a técnica de agrupamento, que é uma técnica de mineração de dados que agrupa dados não rotulados com base em suas semelhanças ou diferenças. Os algoritmos de agrupamento são usados para processar objetos de dados não classificados em grupos representados por estruturas ou padrões nas informações (IBM CLOUD EDUCATION, 2020b).

Destaca-se que os algoritmos de agrupamento, baseados em aprendizagem não supervisionada, podem então identificar padrões em características dos usuários

do sistema sem intervenção humana, assim sendo uma solução possível, alinhada com o problema do presente artigo, e que, com sua utilização, alcança-se o objetivo de indicar as avaliações de uma pessoa para outras semelhantes.

2.2 Dataset

Um dataset é um conjunto de dados organizado e relacionado a um tópico específico. Geralmente organizados em arquivos CSV, os datasets são utilizados em aprendizado de máquina como a base de dados para o desenvolvimento do algoritmo proposto.

A gama de datasets públicos que contenham as informações pessoais desejadas para o agrupamento de usuários é limitada, porém, há a possibilidade da geração de um dataset, com dados fictícios, a partir da biblioteca Mimesis, para Python.

2.2.1 Dataset Adult

O dataset intitulado “Adult” é um dataset que consiste em dados pessoais derivados do censo demográfico de 1994 dos EUA. O dataset contém informações de mais de 32000 pessoas, dentre os dados presentes estão características como: idade, sexo, nível de educação, estado civil, nacionalidade, profissão, raça, faixa salarial, entre outros.

A tarefa típica aplicada neste conjunto de dados é prever se um indivíduo está em certa faixa salarial, porém, seus dados, extraídos da população dos EUA, têm grande valor para outros tipos de análise. Este dataset é um dos poucos que apresenta dados de pessoas reais vindos de uma fonte confiável (o governo americano) e outros datasets semelhantes não oferecem a mesma quantidade de observações e/ou estão direcionados a outros problemas.

Opta-se então pela utilização do dataset Adult para a tarefa de agrupamento, onde os indivíduos descritos no dataset serão agrupados por semelhança de suas características.

2.2.2 Escolha e Classificação de Atributos

A seguir, os atributos presentes no dataset Adult que se referem a informações do indivíduo e a classificação do tipo de dado que representam.

Tabela 1 – Atributos escolhidos do dataset Adult

Atributo	**Classificação**
Idade	Quantitativo, discreto
Sexo	Qualitativo, binário
Grau de educação	Qualitativo, ordinal
Estado civil	Qualitativo, nominal
Profissão	Qualitativo, nominal

Fonte: O autor (2022).

De acordo com UCLA: Statistical Consulting Group, uma variável qualitativa é aquela que possui duas ou mais categorias. No tipo ordinal, há ordem entre os valores, por exemplo, no atributo educação, podemos dizer que ensino superior é maior que ensino médio. No tipo nominal, não há uma maneira acordada de ordená-los do maior para o menor, por exemplo, o atributo profissão (eletricista, engenheiro, cabeleireiro etc.).

Na classificação qualitativa binária, apresenta-se somente duas categorias, no caso do atributo sexo, masculino ou feminino, fala-se binário então pois podem ser representados por masculino = 0 e feminino = 1.

O tipo quantitativo discreto aponta dados numéricos onde os intervalos entre os valores da variável são igualmente espaçados, por exemplo, o atributo idade, espaçado de um em um ano.

Cálculos e análises estatísticas assumem que as variáveis têm níveis específicos de medição. Por exemplo, não faria sentido calcular uma profissão média. Suponha, por exemplo, que você tenha alguma variável categórica chamada "profissão" que possa assumir os valores advogado, médico ou professor. Se simplesmente os codificarmos numericamente como 1, 2 e 3, respectivamente, nosso algoritmo pensará que o advogado (1) está mais próximo do médico (2) do que do professor (3), onde não há proximidade mensurável entre estes atributos. Uma média

de uma variável nominal não faz muito sentido porque não há ordenação intrínseca dos níveis das categorias (UCLA: STATISTICAL CONSULTING GROUP).

É de suma importância que o algoritmo de agrupamento aplicado esteja alinhado com os tipos de dados presentes no dataset, somente assim seu resultado terá sentido.

2.3 Algoritmos de Agrupamentos

Para Bai et al (2012), agrupamento é o processo de agrupar um conjunto de objetos em grupos (clusters) para que os objetos no mesmo grupo tenham alta similaridade, mas sejam muito diferentes de objetos em outros grupos.

O principal uso de algoritmos de agrupamento é descobrir as estruturas de agrupamento inerentes a dados (GUHA, RASTOGI e SHIM, 1999).

Quase todos os algoritmos de agrupamento são explicitamente ou implicitamente conectados a alguma definição de medida de proximidade. No entanto, não existe um algoritmo de agrupamento que possa ser universalmente usado para resolver todos os problemas (XU e WUNSCH, 2005).

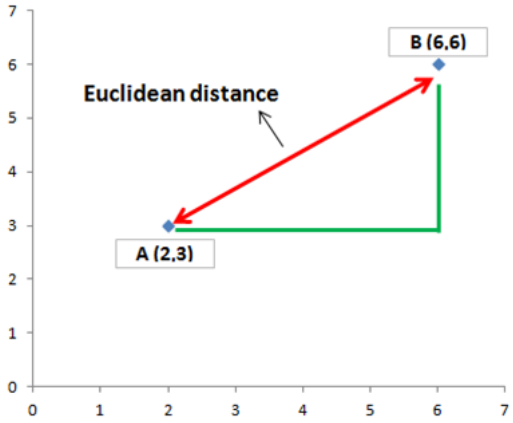
Um dos principais problemas encontrados em algoritmos de agrupamento é a definição de quantos grupos dividir os dados. Sinaga e Yang (2020), dizem que os algoritmos de agrupamento geralmente são afetados por parâmetros de inicialização precisam receber um número de grupos a priori.

2.3.1 K-Means

O algoritmo k-means é o método de agrupamento mais conhecido e utilizado, e existem várias extensões do algoritmo propostas na literatura acadêmica. É um algoritmo de reconhecimento de padrões e agrupamento em aprendizado de máquina não supervisionado (SINAGA e YANG, 2020).

O agrupamento K-means visa particionar dados em k clusters de forma que os pontos de dados no mesmo cluster sejam semelhantes. A semelhança entre dois pontos é determinada pela distância entre eles. Existem muitos métodos para medir a distância entre dois pontos, porém, a distância euclidiana é uma das medidas mais usadas.

Figura 1 – Distância Euclidiana



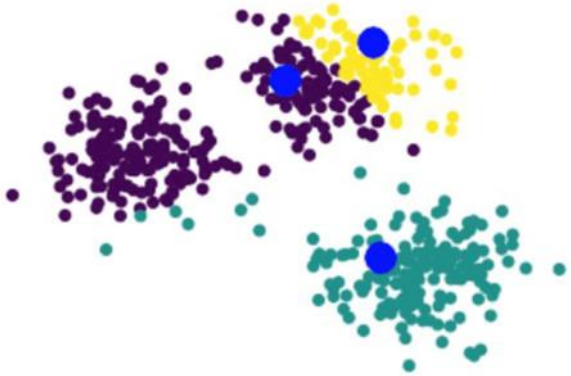
$$Euclidean\ distance\ (a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

Fonte: SONER (2020).

De acordo com Camargo (2020), o funcionamento do algoritmo k-means consiste em duas etapas: (i) calcular a quantidade aproximada de grupos e o ponto central estimado de cada um deles; (ii) realizar iterações de cálculos sobre o conjunto de dados, corrigir e aumentar a precisão de classificação dos grupos.

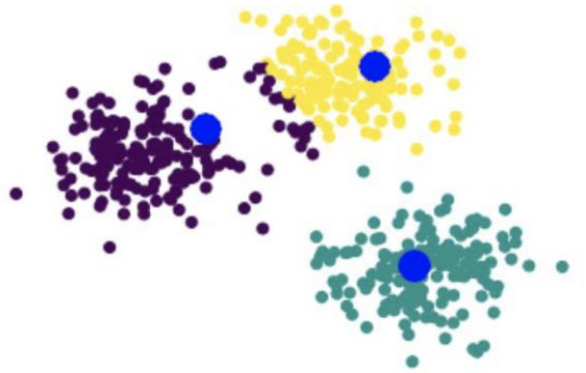
Camargo (2020), também demonstra visualmente o funcionamento do algoritmo k-means. Em azul, podemos ver os objetos que representam os centroides de cada grupo. Em roxo, amarelo e verde, os grupos em que os objetos foram divididos.

Figura 2 - Primeira iteração do algoritmo k-means



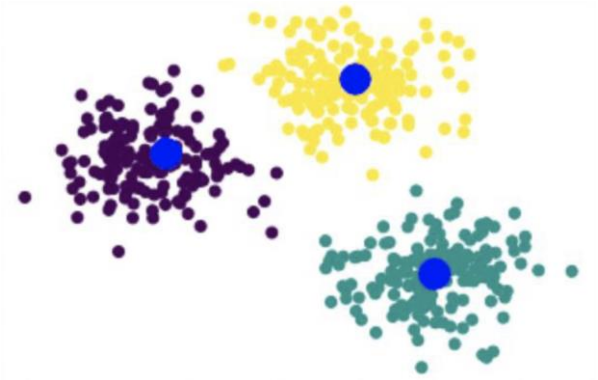
Fonte: CAMARGO (2020).

Figura 3 - N+1 iteração do algoritmo k-means



Fonte: CAMARGO (2020).

Figura 4 - Centroides definidos



Fonte: CAMARGO (2020).

Apesar de sua popularidade, o algoritmo tem algumas limitações: (i) para alcançar eficácia na tarefa de agrupamento necessita da implementação de outros algoritmos, para atribuição de centroides e para definição do número de clusters ideal; (ii) não suporta lidar com dados que não são quantitativos (AHMED, SERAJ e ISLAM, 2020).

Ao lidar com diferentes tipos de dados, Ahmed, Seraj e Islam (2020), dizem que o os métodos tradicionais de cálculo de distância não são adequados para conjuntos de dados que consistem em atributos mistos, por exemplo, categóricos e binários.

Para elucidar a afirmação descrita anteriormente, pode-se exemplificar um cenário: a distância entre idades de indivíduos do dataset Adult pode ser representada numericamente - a diferença de 20 anos para 35 anos é 15 anos. Ao implantar a mesma lógica para um atributo categórico, percebemos que uma representação numérica não faz sentido, por exemplo, no atributo profissão – qual a distância entre

advogado e fotografo? O advogado estaria mais próximo de um fotografo ou psicólogo?

De acordo com Huang (1998), o fato do algoritmo k-means funcionar apenas em dados numéricos limita seu uso em muitas aplicações de mineração de dados devido ao envolvimento de dados categóricos.

Em vista que o conjunto de dados do dataset Adult contém dados categóricos e binários, não é correta a implantação do algoritmo k-means para realizar a tarefa de agrupamento, sendo assim necessária outra solução que possibilite o cálculo de similaridade em atributos mistos.

2.3.2 K-Modes e K-Prototypes

Proposto por Huang (1998), o algoritmo k-modes é uma extensão do algoritmo k-means que visa endereçar os problemas ao lidar com dados categóricos. Além disso, Huang (1998) também propõe o algoritmo k-prototypes, que une as funcionalidades dos algoritmos k-means e k-modes, para lidar com datasets que contenham dados mistos.

O algoritmo k-modes usa uma medida de dissimilaridade simples para lidar com objetos categóricos, substitui as médias dos grupos por modos[7] e usa um método baseado em frequência para atualizar os modos no processo de agrupamento para minimizar o custo da função de agrupamento. Com essas extensões, o algoritmo k-modes permite o agrupamento de dados categóricos de maneira semelhante ao k-means. O algoritmo k-prototypes, através da definição de uma medida de dissimilaridade combinada, integra ainda os algoritmos k-means e k-modes para permitir o agrupamento de objetos descritos por atributos numéricos e categóricos mistos (HUANG, 1998).

Huang (1998) explica que a medida de dissimilaridade simples pode ser dada como: A distância entre dois pontos X e Y é o número de observações[8] em X e Y as quais os valores são diferentes, formalmente definida como:

$$d(X,Y) = \sum_{i=1}^{n} f\,(x_i, y_i)$$

[7] Modo é o valor de dado que mais ocorre para cada observação.
[8] Observação também pode ser entendida como característica de um dado, como idade, sexo, etc.

$$f(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & x_i \neq y_i \end{cases}$$

Onde x_i é o valor da i observação do dado X, e y_i é o valor da i observação do dado Y, e n é o número de observações.

APRILLIANT (2021a), demonstra as etapas para o agrupamento k-modes:

1. Selecionar o modo k inicial;
2. Alocar uma observação ao grupo mais próximo baseado na medida de dissimilaridade simples; atualizar o modo de cada grupo após cada alocação;
3. Após todas as observações serem alocadas a algum grupo, validar o valor de dissimilaridade de cada observação ao modo de todos os grupos; caso a observação esteja mais próxima de um grupo diferente do atual, mover a observação para este; atualizar o modo dos grupos;
4. Repetir o passo 3 até que nenhuma observação mude de grupo.

Resumidamente, o algoritmo k-modes define grupos com base no número de categorias correspondentes entre os pontos de dados. O algoritmo k-prototypes combina k-modes e k-means e é capaz de agrupar dados numéricos/categóricos mistos (VOS, 2022).

A propriedade mais atraente do algoritmo k-means na mineração de dados é sua eficiência em agrupar grandes conjuntos de dados. No entanto, o fato de funcionar apenas em dados numéricos limita seu uso em muitas aplicações de mineração de dados devido ao envolvimento de dados categóricos. Os algoritmos k-modes e k-prototypes removeram essa limitação e estenderam o paradigma k-means para uma operação genérica de particionamento de dados para mineração de dados (HUANG, 1998).

Sobre acurácia, Huang (1998) demonstra com testes em datasets rotulados que, na maioria dos casos, a precisão será superior a 70%. Cao, Liang e Bai (2009), manifestam que um dos principais fatores que implicam na acurácia do agrupamento é a escolha dos k-centroides iniciais, os autores propõem um método de inicialização para o algoritmo k-prototypes que supera a precisão de 90% em testes em datasets rotulados.

Posto isto, a implantação do algoritmo k-prototypes está alinhada com o problema lidado pelo presente artigo e possibilitará a criação de grupos de usuários com características quantitativas e qualitativas semelhantes.

3 DESENVOLVIMENTO

3.1 Jupyter Notebooks

Para o desenvolvimento do sistema de agrupamento, utilizou-se da linguagem de programação Python, juntamente com a ferramenta Jupyter para produção de notebooks. Documentos de notebook (ou somente “notebooks”) são documentos produzidos pelo Jupyter Notebook, que contêm código de computador (por exemplo, Python) quanto elementos de texto (parágrafo, equações, figuras, links etc.). Notebooks são documentos legíveis para humanos contendo a descrição da análise e os resultados (figuras, tabelas etc.), bem como trechos executáveis de código que podem ser executados individualmente.

3.2 Amazon SageMaker

O Amazon SageMaker é um serviço totalmente gerenciado que permite que cientistas de dados e desenvolvedores criem, treinem e implantem modelos de aprendizado de máquina de maneira rápida e fácil em qualquer escala. O Amazon SageMaker inclui módulos que podem ser usados juntos ou de forma independente para criar, treinar e implantar modelos (AMAZON WEB SERVICES, INC., 2017).

Além de outras vantagens, o principal benefício que o Amazon SageMaker traz para o presente trabalho é a conexão nativa com outros componentes da infraestrutura da AWS.

3.3 Pré-processamento de Dados

3.3.1 Seleção de Atributos

A partir do dataset Adult foram selecionadas as seguintes observações para análise de agrupamento:

- Idade;
- grau de educação;
- estado civil;
- profissão;
- sexo.

As seguintes observações foram desconsideradas:

- classe de trabalho – desconsiderada por similaridade a profissão;
- raça – desconsiderada pois mais de 90% dos registros apresentavam o mesmo valor;
- número grau de educação – desconsiderada pois somente representa numericamente a observação grau de educação;
- relacionamento – desconsiderada por similaridade à estado civil;
- ganho capital – desconsiderada por não remeter a característica;
- perda capital - desconsiderada por não remeter a característica;
- horas de trabalho semanais - desconsiderada por correlação a profissão;
- país nativo - desconsiderada pois mais de 98% dos registros apresentavam o mesmo valor.

3.3.2 Valores Nulos

O dataset Adult contém valores nulos/inválidos que necessitam ser corrigidos antes de qualquer análise para melhores resultados. Das observações selecionadas, a profissão contém 1843 registros com o valor ‘?’, estes serão substituídos pelo valor mais comum de profissão no dataset.

3.3.3 Redução de Dimensionalidade

O número de variáveis de entrada para um conjunto de dados é chamado de dimensionalidade. A redução de dimensionalidade refere-se a técnicas que reduzem o número de variáveis de entrada em um conjunto de dados.

Quando os dados são esparsos, as observações ou amostras no conjunto de dados de treinamento são difíceis de agrupar, pois dados de alta dimensão fazem com

que todas as observações no conjunto de dados pareçam equidistantes umas das outras.

Para facilitar o agrupamento de observações semelhantes, é aplicada a redução de dimensionalidade nas observações educação e estado civil, reduzindo a quantidade de variáveis existentes.

3.3.4 Padronização de Dados

Ao padronizar as variáveis contínuas (no caso, somente a idade), tornam-se todas igualmente importantes para a análise. Caso houver grandes diferenças entre o intervalo das variáveis numéricas, as variáveis com maior intervalo de valores dominarão sobre aquelas com intervalos menores. A padronização redimensiona os dados para ter uma média (μ) de 0 e desvio padrão (σ) de 1.

3.3.5 Modelagem

É crucial determinar o número ideal de clusters para a qualidade do agrupamento na análise (ZHOU, XU e LIU, 2017).

Os autores Bholowalia e Kumar (2014) explicam que o método do cotovelo é um método que analisa a porcentagem de variância em função do número de clusters. Este método baseia-se na ideia de que se deve escolher um número de clusters o qual a adição de outro cluster não forneça uma modelagem muito melhor dos dados. A porcentagem de variância pelos clusters é então plotada, os primeiros clusters adicionarão muitas informações, mas em algum momento o ganho marginal cairá drasticamente e surgirá um ângulo no gráfico - o 'k' ideal, ou seja, o número de clusters é escolhido neste ponto, eis o "método do cotovelo".

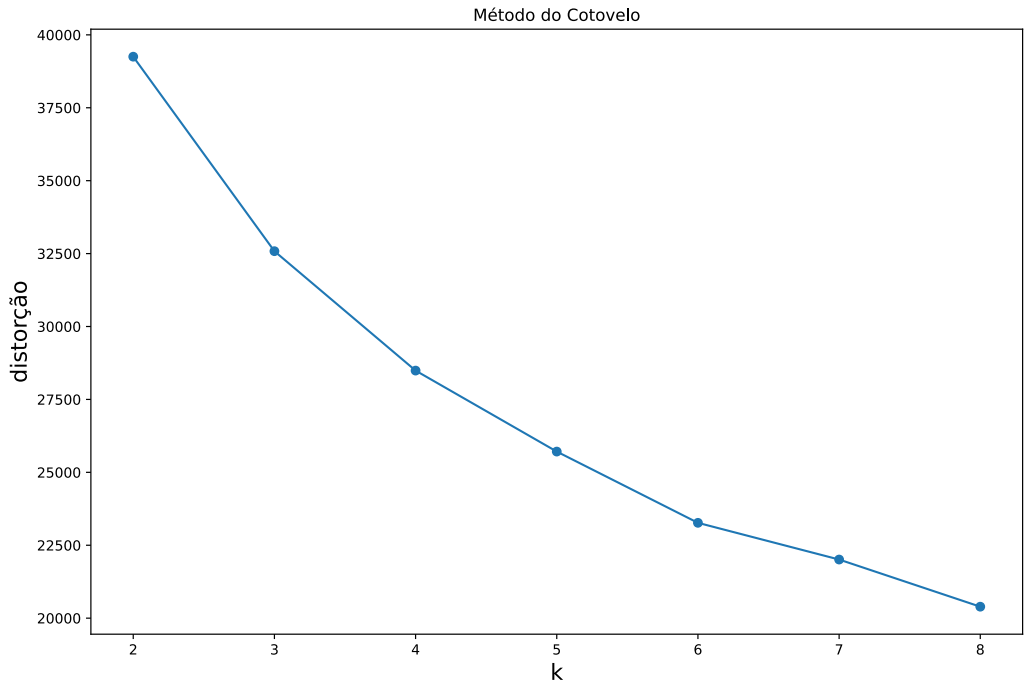
O algoritmo K-Prototypes fornece a função de custo combinada em variáveis numéricas e categóricas. Podemos olhar o cotovelo no gráfico para determinar o número ideal de grupos (APRILLIANT, 2021b).

Figura 5 – Encontrando custo/distorção para k grupos

```
K = range(2,9)
cost = []
for k in K:
    kproto = KPrototypes(n_clusters=k, init='Cao')
    # 1 = education, 3 = occupation, 4 = race
    kproto.fit_predict(df_final, categorical=[1, 3, 4])
    cost.append(kproto.cost_)
```

Fonte: O autor (2022).

Figura 6 – Visualização do método do cotovelo



Fonte: O autor (2022).

Pode-se perceber que, a partir de 4 grupos, a redução de distorção passa a ser linear até 6 grupos, e depois até 8 grupos. Entende-se que a redução linear da distorção se dá somente pela maior quantidade de grupos, e não pela diferenciação do caráter de cada grupo. Portanto, o valor ideal de grupos é 4.

Esta é a etapa mais custosa do modelo de aprendizado de máquina, pois são realizados os cálculos para agrupamentos em 2 a 8 grupos, levando cerca de 40 minutos para execução no ambiente Amazon SageMaker.

A implementação do algoritmo de agrupamento K-Prototypes, bem como K-Modes é publicamente disponível a partir de bibliotecas Python. No presente trabalho, utiliza-se da biblioteca k-modes.

Como visto anteriormente, o algoritmo k-prototypes funciona para dados numéricos e categóricos, na utilização da biblioteca k-modes é necessária a indicação de quais atributos do dataset são categóricos, estes terão o algoritmo k-modes aplicado, os dados numéricos serão tratados pelo algoritmo k-means. Também é informado a quantidade de grupos a serem gerados, a partir do valor ideal encontrado pelo método do cotovelo.

Figura 7 - Implantação do algoritmo K-Prototypes com biblioteca Python

```
from kmodes.kprototypes import KPrototypes

kproto = KPrototypes(n_clusters=4, init="Cao")
# categorical data: 1 = education, 3 = occupation, 4 = race
clusters = kproto.fit_predict(df_final, categorical=[1, 3, 4])

# merging original data with clusters
df_clusters = pd.concat([df_original, pd.DataFrame({"cluster": clusters})], axis=1)
```

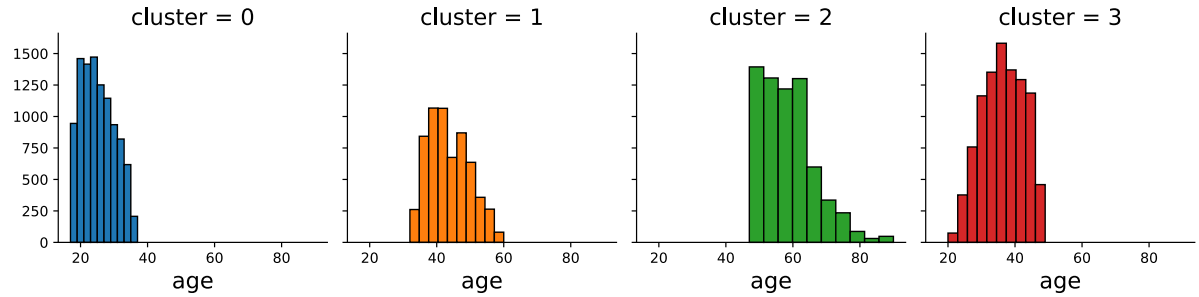
Fonte: O autor (2022).

A partir da execução do algoritmo, adquire-se o agrupamento por similaridade de características das pessoas do dataset Adult.

3.3.6 Visualização do Agrupamento

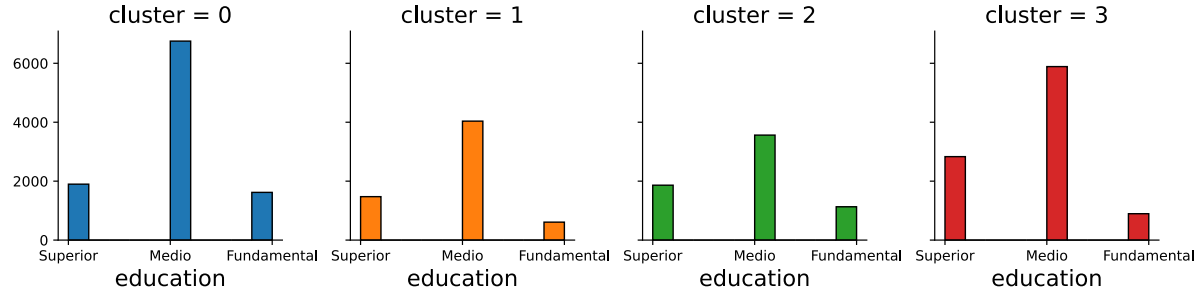
Os seguintes gráficos exibem a distribuição das pessoas, por suas características, em cada grupo (cluster) gerado, com um total de 4 grupos (0, 1, 2 e 3). No eixo X está disposto o valor para os atributos dos usuários e no eixo Y apresenta-se a quantidade de pessoas com o mesmo valor para o atributo.

Figura 8 - Visualização dos grupos isolando o atributo idade



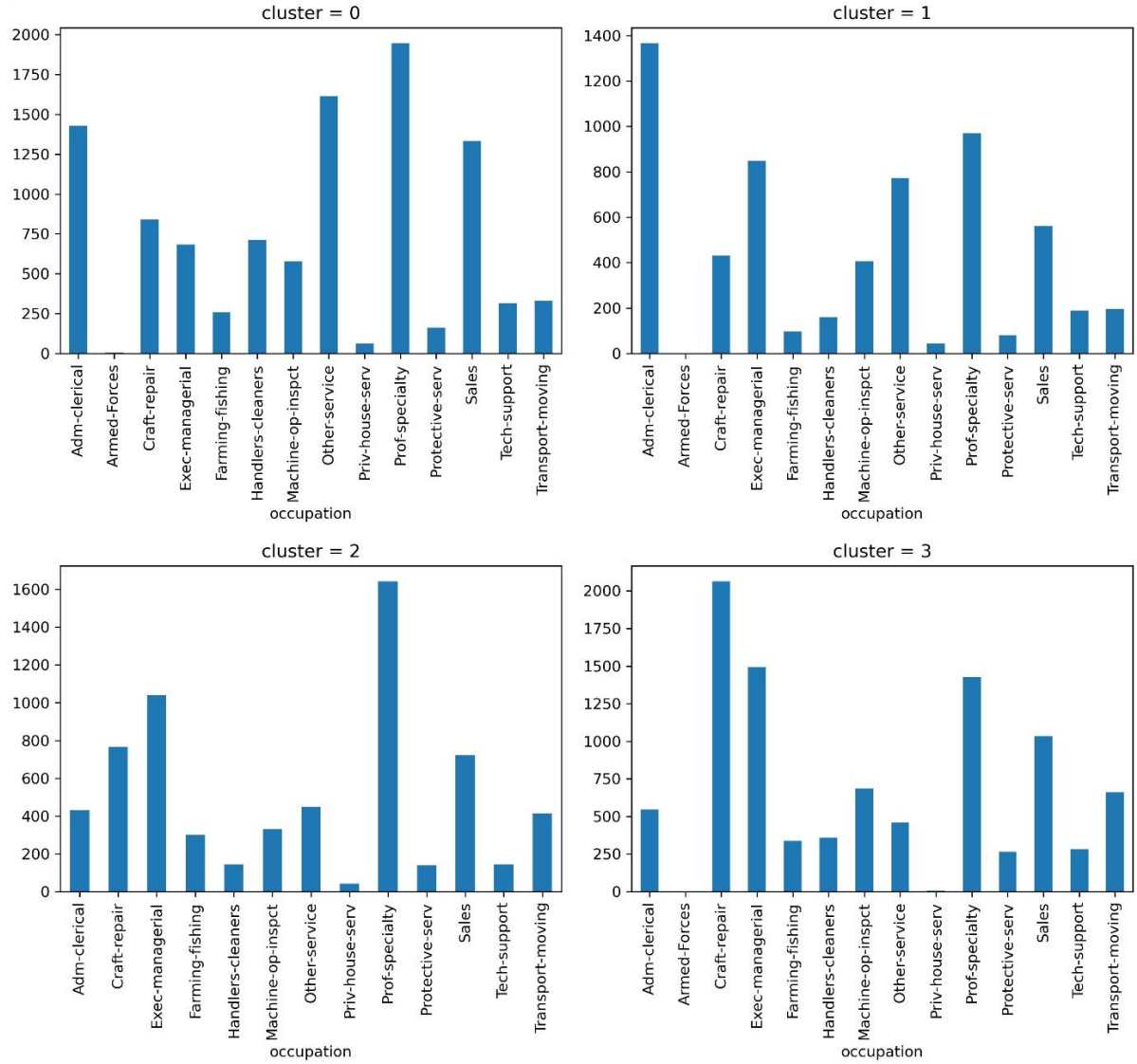
Fonte: O autor (2022).

Figura 9 - Visualização dos grupos isolando o atributo educação



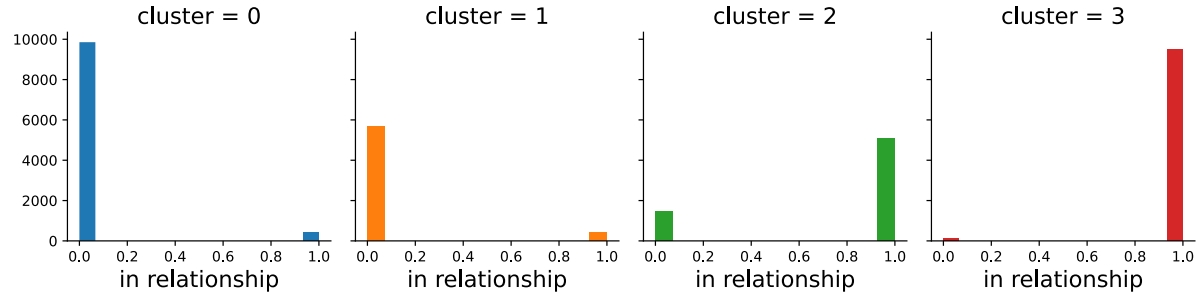
Fonte: O autor (2022).

Figura 10 - Visualização dos grupos isolando o atributo profissão



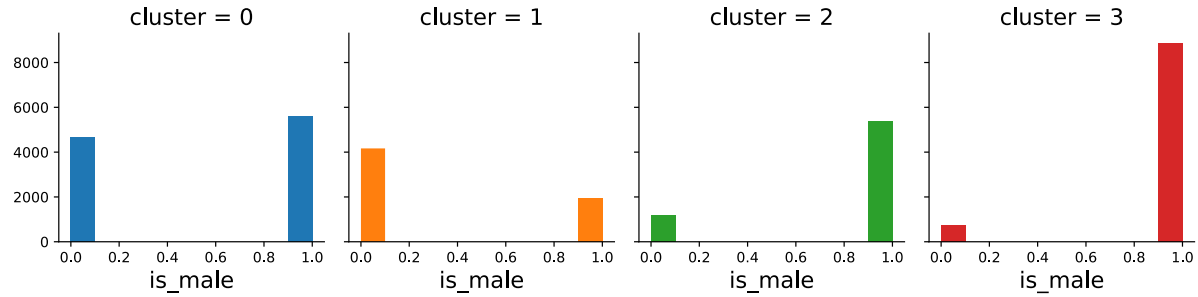
Fonte: O autor (2022).

Figura 11 - Visualização dos grupos isolando o atributo relacionamento



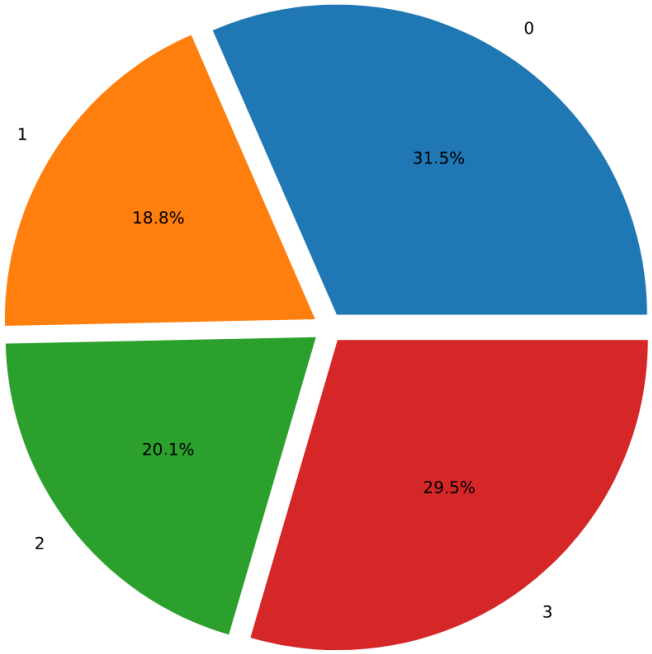
Fonte: O autor (2022).

Figura 12 - Visualização dos grupos isolando o atributo sexo



Fonte: O autor (2022).

Figura 13 – Alocação total de pessoas em cada grupo



Fonte: O autor (2022).

Pode-se perceber algumas características pessoais em cada grupo, majoritariamente:

- Grupo 0: Jovens de 20 a 30 anos, solteiros, de ambos os sexos. Também é o grupo mais genérico, com a maior quantidade de pessoas e que trabalham em várias áreas;
- Grupo 1: Mulheres de 35 a 55 anos, que possuem como ocupação as categorias “adm-clerical”, “prof-specialty” e “exec-managerial”, não estão em um relacionamento;

- Grupo 2: Homens, de 45 a 80 anos, que estão em um relacionamento, como ocupação tem a observação “prof-specialty”;
- Grupo 3: É o grupo com maior parcela de pessoas com ensino superior, formado majoritariamente por homens que estão em um relacionamento, tem como ocupação as categorias “craft-repair”, “exec-managerial” e “prof-specialty”.

3.4 Desenvolvimento da Central de Avaliações de Produtos

Para desenvolvimento do protótipo da central de avaliações, é necessário que haja o gerenciamento de usuários, produtos e avaliações.

3.4.1 API RESTful

REST[9] é uma arquitetura de software que impõe condições sobre como uma API[10] deve funcionar. O REST foi criado inicialmente como uma diretriz para gerenciar a comunicação em uma rede complexa como a internet. Pode-se usar a arquitetura REST para oferecer suporte a comunicação confiável e de alto desempenho em escala (AMAZON WEB SERVICES, INC., 2022).

As APIs REST se comunicam por meio de solicitações HTTP para executar funções de banco de dados padrão, como criar, ler, atualizar e excluir registros (também conhecidos como CRUD) em um recurso. Por exemplo, uma API REST usaria uma solicitação GET para recuperar um registro, uma solicitação POST para criar um, uma solicitação PUT para atualizar um registro e uma solicitação DELETE para excluir um (IBM CLOUD EDUCATION, 2021).

Em concordância com o requerimento de gerenciamento de usuários, produtos e avaliações, adota-se o padrão REST para desenvolvimento da API web.

[9] REST (Representational State Transfer) - Transferência de Estado Representacional
[10] API (Application Programming Interface) - Interface de Programação de Aplicativos

3.4.2 Computação Serverless - AWS Lambda

O AWS Lambda é um serviço de computação sem servidor e orientado a eventos que permite executar código de qualquer tipo de back-end sem provisionar ou gerenciar servidores. A função do Lambda (também chamadas de lambda functions) é o princípio fundamental do Lambda. Cria-se uma função a partir de código, em alguma linguagem suportada, e quando um evento determinado ocorrer, o Lambda invoca a função. O Lambda executa várias instâncias de sua função em paralelo, regidas por simultaneidade e limites de escalabilidade (AMAZON WEB SERVICES, INC., 2022).

Para desenvolvimento das funções Lambda neste projeto, foi escolhida a linguagem Python, pela sua praticidade oferecida e conhecimento do autor. Utiliza-se uma função Lambda para cada operação em um recurso, como exemplos: cadastrar usuário, atualizar produto, editar avaliação.

Figura 14 - API RESTful com funções Lambda

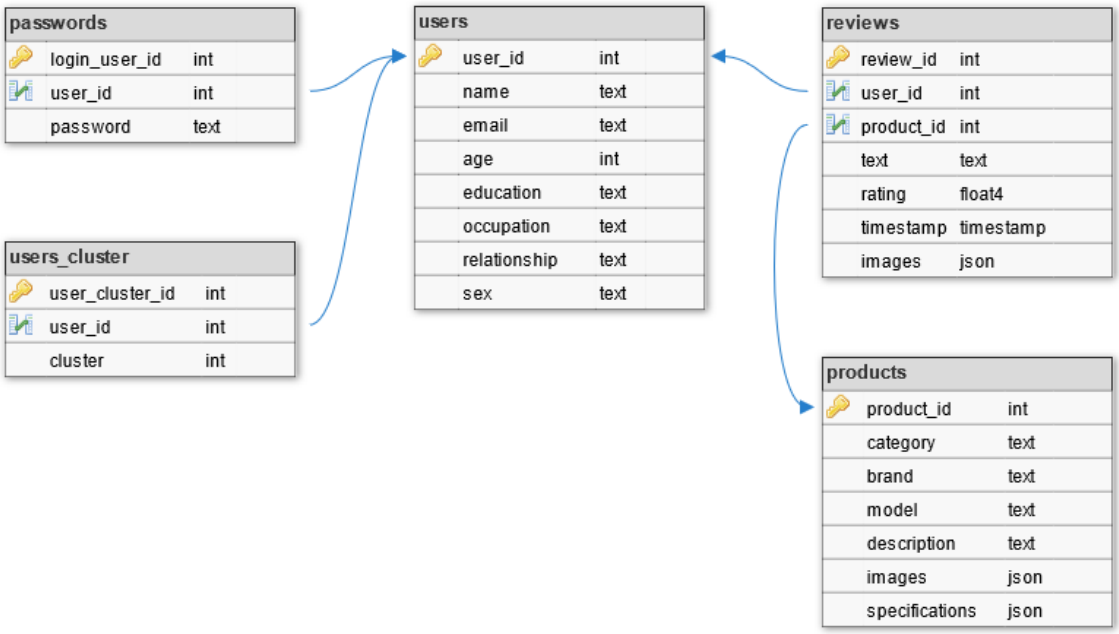
Fonte: O autor (2022).

O Lambda integra-se a outros serviços da AWS para invocar funções com base em eventos específicos. Utiliza-se da integração com o serviço API Gateway para expor uma API RESTful a partir de funções Lambda.

3.4.3 Banco de Dados - PostgreSQL

Para armazenamento dos dados de usuários, produtos e avaliações foi escolhido o banco de dados PostgreSQL. A figura seguir ilustra a modelagem das tabelas do banco de dados.

Figura 15 - Modelagem do Banco de Dados



Fonte: O autor (2022).

3.4.4 Armazenamento de Objetos e Arquivos - Amazon S3

O Amazon Simple Storage Service (Amazon S3) é um serviço de armazenamento de objetos e arquivos.

Neste projeto, o Amazon S3 é utilizado para operações de integração entre o sistema de agrupamento (notebook de aprendizado de máquina) e a base de dados de usuários, isto é, ao finalizar o processamento do algoritmo de agrupamento, um arquivo CSV com as informações dos usuários e seus respectivos grupos é salvo no S3. Posteriormente, é executada uma função Lambda que lê este arquivo e atualiza o grupo de cada usuário no banco de dados.

3.4.5 Interface de Usuário - React

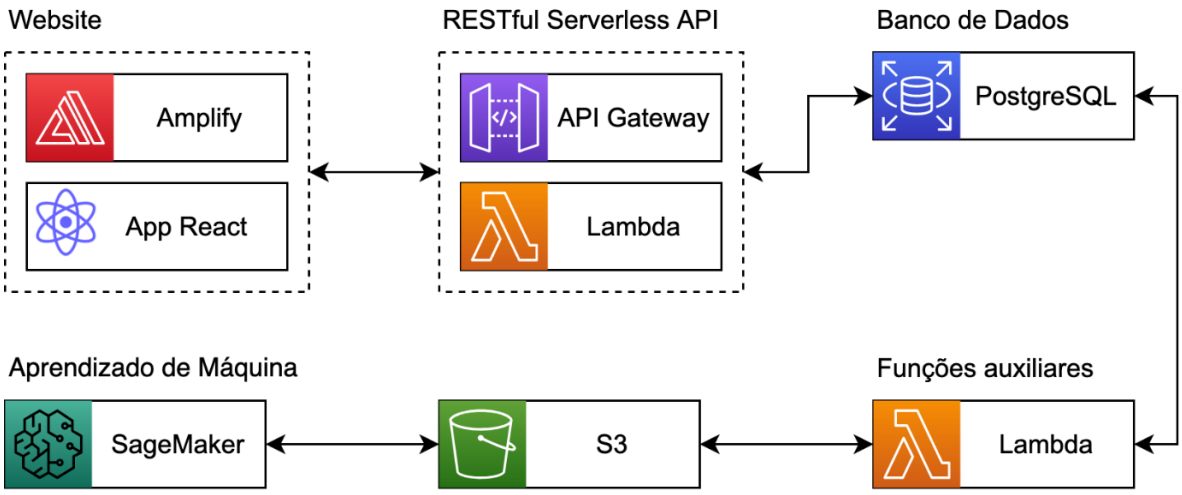
Para construção da interface de usuário (UI) do protótipo da central de avaliações, utiliza-se React. React é uma biblioteca para desenvolvimento front-end, escrita em JavaScript, gratuita e de código aberto, criado pela Facebook.

Também se utiliza da biblioteca de componentes Material UI, de código aberto e mantida pela Google. A biblioteca inclui uma coleção abrangente de componentes (botões, tabelas, listas, ícones etc.) que serão utilizados para construção do protótipo.

3.5 Visão Geral de Arquitetura do Projeto

A arquitetura do projeto é inspirada e faz uso de boas práticas recomendadas pela AWS documentadas em seu framework intitulado "*AWS Well-Architected*"[11].

Figura 16 - Diagrama de Arquitetura de Alto Nível



Fonte: O autor (2022).

3.6 Modelo de Central de Avaliações

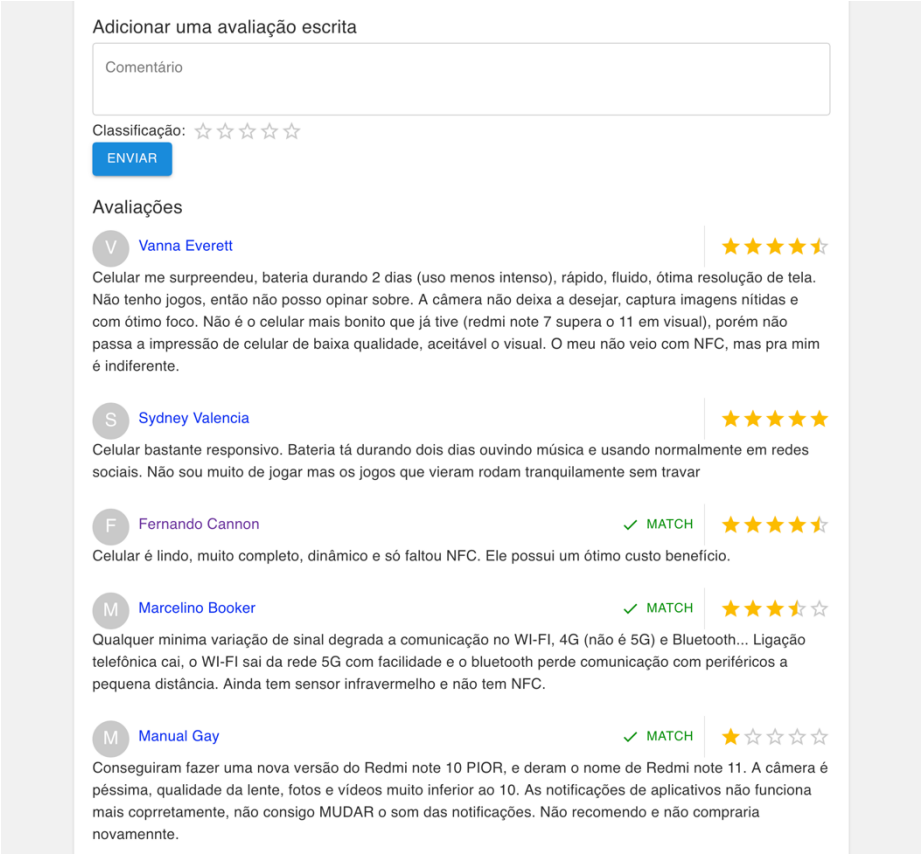
Para que as pessoas possam compartilhar suas opiniões sobre produtos, desenvolveu-se a central de avaliações, um portal exemplificador da aplicabilidade do agrupamento de pessoas similares para recomendação de avaliações.

[11] https://aws.amazon.com/architecture/well-architected

Como usuários deste sistema, foram cadastradas as pessoas do dataset Adult, suplementarmente com a adição de atributos como nome, e-mail e senha para habilitar funções como login e visualização de perfis. Da mesma forma, utilizando da API RESTful, foram cadastrados 3 celulares como produtos e em seguida geradas e cadastradas avaliações fictícias para estes produtos, simulando que os usuários do sistema as escreveram.

As avaliações escritas pelos diversos usuários do sistema ficam disponíveis na página de cada produto no site. As avaliações cujo autor está no mesmo grupo do usuário logado são indicadas pelo símbolo ✔ MATCH, simbolizando então que aquela avaliação é entendida como de maior valor para ele.

Figura 17 – Área de visualização de avaliações na Central de Avaliações de Produtos



Fonte: O autor (2022).

3.7 Metodologia

Para implantação do sistema proposto no presente artigo, aplica-se duas sistemáticas – Produto de Valor Mínimo (*Minimum Value Product*, MVP) e o método Kanban.

3.7.1 MVP

Hyrynsalmi et al. (2018) disserta sobre o conceito de “Produto Mínimo Viável” (MVP), amplamente adotado na indústria de software, bem como na academia. Produtos mínimos viáveis são usados para testar hipóteses sobre o público-alvo, economizar recursos de trabalho de desenvolvimento desnecessário e orientar uma empresa para um modelo de negócios estável.

Desta forma, o desenvolvimento do sistema apresentado é direcionado a funcionalidades essenciais para a principal entrega de valor que se deseja atingir – mostrar avaliações relevantes para os usuários do sistema.

Portanto, juntamente com o professor orientador, foram elaboradas e priorizadas as principais funcionalidades do sistema da seguinte forma:

1. Aprendizado de máquina
    a. Análise de perfis de usuários
    b. Criação de grupos de usuários
2. Back-end
    a. Gerenciamento de usuários, produtos e avaliações
    b. Integração com sistema de aprendizado de máquina
3. Interface de usuário
    a. Visualização de produtos
    b. Visualização de avaliações
    c. Gerenciamento de usuário
        i. Cadastro, login, editar perfil

3.7.2 Kanban

Kanban é um framework popular usado para implementar o desenvolvimento de software ágil. Foi desenvolvido no final da década de 1940 por Taiichi Ohno. O Kanban se concentra na visualização de todo o projeto em quadros para aumentar a transparência do projeto e a colaboração entre os membros da equipe (KISSFLOW INC., 2022).

O método Kanban gira em torno do quadro Kanban. É uma ferramenta para visualizar todo o projeto e acompanhar o seu fluxo. Por meio dessa abordagem gráfica de quadros Kanban, um novo membro ou uma entidade externa pode entender o que está acontecendo agora, tarefas concluídas e tarefas futuras (KISSFLOW INC., 2022). Tipicamente, o quadro Kanban indica as tarefas que estão sendo executadas, as tarefas a fazer e as tarefas concluídas.

Desta forma, utiliza-se o método Kanban para gerenciamento das atividades a executar para a confecção do sistema proposto.

4 CONCLUSÃO

Durante o desenvolvimento deste projeto, foram realizadas diversas pesquisas a respeito das áreas envolvidas, possibilitando um amplo conhecimento delas para que pudessem ser bem aplicadas.

Na área de aprendizado de máquina, pôde-se observar e entender os tipos de aprendizagem, bem como os algoritmos que tratam do problema de agrupamento proposto pelo projeto, assim identificando sua aplicação e funcionamento. Ainda nesse tópico, pôde-se elaborar um modelo de aprendizado de máquina para agrupamento de pessoas a partir de suas características, com o objetivo de integrá-lo a um sistema centralizador de avaliações, assim possibilitando que o objetivo principal de indicar avaliações de pessoas semelhantes ao usuário fosse alcançado.

No tópico de desenvolvimento web, pesquisou-se e adotou-se práticas, arquitetura e linguagens atuais de mercado, como desenvolvimento ágil, ecossistema AWS, TypeScript e Python, resultando numa aplicação integrada, fácil de manter e altamente escalável.

Entende-se que o modelo de agrupamento realizado, baseado no dataset Adult, pode sofrer de falta de características disponíveis para identificar se uma pessoa é semelhante a outra. O ideal seria que, além de mais características, comportamentos e interesses também fossem considerados pela análise - dados que redes sociais e gigantes da internet possuem, como, por exemplo, quais categorias de produto, marcas, estilos de roupa um indivíduo tem mostrado interesse. Assim sendo, o uso do dataset Adult serviu como balizador para implementações futuras que empoderarem de dados mais completos.

Mesmo que a semelhança de características entre indivíduos não implique na mesma preferência e nas mesmas expectativas de compra, este projeto apresenta um conceito para a experiência de compra que não está presente nos maiores e-commerces, que se limitam a mostrar as avaliações de seus produtos ordenadas por data e ainda deixam dúvidas de quem é o avaliador do produto.

Por fim, este projeto propõe uma importante alternativa e meio para que as pessoas encontrem avaliações de produtos de forma mais fácil, e, além disso, apresenta uma forma de melhorar o setor de avaliações de e-commerces, para que o usuário final tenha uma experiência melhor ao realizar compras na internet.

5 REFERÊNCIAS

AHMED, Mohiuddin; SERAJ, Raihan; ISLAM, Syed Mohammed Shamsul. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. **Electronics**, v. 9, n. 8, p. 1295, 2020. Disponível em: https://www.mdpi.com/2079-9292/9/8/1295. Acesso em: 23 jul. 2022.

AMAZON WEB SERVICES, INC. **AWS Lambda**. [*S. l.*], 2022. Disponível em: https://aws.amazon.com/pt/lambda/. Acesso em: 10 abr. 2022.

AMAZON WEB SERVICES, INC. **Introducing Amazon SageMaker**. [*S. l.*], 2017. Disponível em: https://aws.amazon.com/about-aws/whats-new/2017/11/introducing-amazon-sagemaker/. Acesso em: 10 abr. 2022.

AMAZON WEB SERVICES, INC. **What is RESTful API?** [*S. l.*], 2022. Disponível em: https://aws.amazon.com/what-is/restful-api/. Acesso em: 10 abr. 2022.

APRILLIANT, Audhi. **The k-modes as Clustering Algorithm for Categorical Data Type**. [*S. l.*], 2021a. Disponível em: https://medium.com/geekculture/the-k-modes-as-clustering-algorithm-for-categorical-data-type-bcde8f95efd7. Acesso em: 15 jun. 2022.

APRILLIANT, Audhi. **The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical)**. [*S. l.*], 2021b. Disponível em: https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb. Acesso em: 15 jun. 2022.

BHOLOWALIA, Purnima; KUMAR, Arvind. EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. **International Journal of Computer Applications**, v. 105, n. 9, p. 17–24, 2014. Disponível em: https://www.ijcaonline.org/archives/volume105/number9/18405-9674. Acesso em: 18 jun. 2022.

BROWN, Sara. **Machine learning, explained**. MIT Sloan. Disponível em: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained. Acesso em: 20 set. 2022.

CAO, Fuyuan; LIANG, Jiye; BAI, Liang. A new initialization method for categorical data clustering. **Expert Systems with Applications**, [*s. l.*], v. 36, n. 7, p. 10223–10228, 2009. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417409001043. Acesso em: 18 jun. 2022.

GUHA, S.; RASTOGI, R.; SHIM, K. ROCK: a robust clustering algorithm for categorical attributes. *In*: **Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337)**. [s.l.: s.n.], 1999, p. 512–521.

HUANG, Zhexue. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. **Data Mining and Knowledge Discovery**, v. 2, n. 3, p. 283–304, 1998. Disponível em: https://doi.org/10.1023/A:1009769707641. Acesso em: 7 mar. 2022.

HYRYNSALMI, Sami; KLOTINS, Eriks; UNTERKALMSTEINER, Michael; *et al*. What is a Minimum Viable (Video) Game? *In*: AL-SHARHAN, Salah A.; SIMINTIRAS, Antonis C.; DWIVEDI, Yogesh K.; *et al* (Orgs.). **Challenges and Opportunities in the Digital Era**. Cham: Springer International Publishing, 2018, p. 217–231. (Lecture Notes in Computer Science).

IBM CLOUD EDUCATION. **REST APIs**. [*S. l.*], 2021. Disponível em: https://www.ibm.com/cloud/learn/rest-apis. Acesso em: 20 set. 2022.

IBM CLOUD EDUCATION. **What is Machine Learning?**. [*S. l.*], 2020a. Disponível em: https://www.ibm.com/cloud/learn/machine-learning. Acesso em: 20 set. 2022.

IBM CLOUD EDUCATION. **What is unsupervised learning?**. [*S. l.*], 2020b. Disponível em: https://www.ibm.com/cloud/learn/unsupervised-learning. Acesso em: 20 set. 2022.

KISSFLOW INC. **Kanban Methodology: The Simplest Agile Framework**. Disponível em: https://kissflow.com/project/agile/kanban-methodology/. Acesso em: 1 set. 2022.

SINAGA, Kristina P.; YANG, Miin-Shen. Unsupervised K-Means Clustering Algorithm. **IEEE Access**, v. 8, p. 80716–80727, 2020.

UCLA: STATISTICAL CONSULTING GROUP. **What is the difference between categorical, ordinal and interval variables?**. [*S. l.*], [*s. d.*]. Disponível em: https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/. Acesso em: 9 jul. 2022.

VOS, Nelis J. de. kmodes: Python implementations of the k-modes and k-prototypes clustering algorithms for clustering categorical data. Disponível em: https://github.com/nicodv/kmodes. Acesso em: 13 ago. 2022.

XU, Rui; WUNSCH, D. Survey of clustering algorithms. **IEEE Transactions on Neural Networks**, v. 16, n. 3, p. 645–678, 2005.

ZHOU, Shibing; XU, Zhenyuan; LIU, Fei. Method for Determining the Optimal Number of Clusters Based on Agglomerative Hierarchical Clustering. **IEEE Transactions on Neural Networks and Learning Systems**, v. 28, n. 12, p. 3007–3017, 2017.