Aplicação da Clusterização por *K-means* para Criação de Sistema de Recomendação de Produtos baseado em Perfis de Compra

Applying K-means Clustering to Create Product Recommendation System Based on Purchase Profiles

Roniel Venâncio Alencar Santana Graduando em Engenharia de Produção Mecânica. Universidade Federal do Ceará (UFC) – Brasil. roniel_venancio@hotmail.com.

Heráclito Lopes Jaguaribe Pontes Doutor em Engenharia Mecânica. Universidade Federal do Ceará (UFC) – Brasil. hjaguaribe@ufc.br.

RESUMO

O uso de modelos preditivos de machine learning para big data se faz hoje uma das principais tendências a serem exploradas pela ciência de dados. Sua aplicação ao mundo dos negócios na busca por um diferencial competitivo se relaciona diretamente com o Business Intelligence para que assim as empresas passem a tomar decisões mais assertivas. Com isso, o presente artigo propõe-se a aplicar uma técnica de machine learning para a criação de um sistema de recomendação de produtos com base no perfil de compra dos clientes a partir da modelagem em uma empresa distribuidora de produtos. Para tanto foi utilizado o algoritmo de clusterização K-means para realização de agrupamentos dos clientes com base em seu perfil de compra. Por fim, o princípio de funcionamento do sistema de recomendação baseia-se na análise comparativa entre clientes de um mesmo cluster com base em suas distâncias geográficas para assim recomendar aquele item que vende bem em um estabelecimento, mas que não tem o mesmo desempenho em outro. Ao final da aplicação do sistema de recomendação de produtos foram gerados um total de 70 clusters para toda a gama de clientes da empresa foco do estudo. Cada cliente de cada cluster recebeu uma lista contendo 5 produtos recomendados com base na comparação realizada com seus vizinhos próximos de perfil de compra similar.

Palavras-chave: Sistema de recomendação. Ciência de Dados. *Machine learning*. Clusterização. *Business Intelligence*.

ABSTRACT

The use of predictive machine learning models for big data is today one of the main trends to be explored by data science. Its application to the business world for a search of competitive differential is directly related to Business Intelligence so companies can make more assertive decisions. Thus, this paper proposes to apply a machine learning technique to create a product recommendation system based on customers' purchase profile, modeled for a product distribution company. For this purpose, the K-means clustering algorithm was used to group customers based on their purchase profile. Finally, the recommendation system's principle is based on a comparative analysis between customers in the same cluster and based on their geographic distances to recommend that item that sells well in one point of sales but does not perform so well in another. At the end of the application 70 clusters were generated for the entire range of customers of the company focused in the present study. Each customer in each cluster received a list containing 5 recommended products based on the comparison made with their close neighbors of similar buying profile.

Keywords: Recommendation system. Data science. Machine learning. Clustering. Business intelligence

Recebido em 01/02/2020. Aprovado em 06/05/2020. Avaliado pelo sistema *double blind peer review*. Publicado conforme normas da APA.
http://dx.doi.org/10.22279/navus.2020.v10.p01-14.1189

1 INTRODUÇÃO

O correto uso e tratamento de dados se faz hoje mais do que necessário para se gerar informação útil para atender os mais diversos objetivos das organizações. Hoje a ciência de dados em conjunto com a inteligência artificial e suas técnicas são essenciais para que empresas mudem seus modos de tomada de decisão para abordagens cada vez mais inteligentes e assertivas.

Conforme estudo realizado por Fraga *et al.* (2017), hoje já existem empresas que possuem sua cultura organizacional orientada a dados com métodos e práticas correlatas para o desenvolvimento intelectual a respeito do *business intelligence* nos colaboradores. Tal realidade é essencial para a difusão das novas tendências relacionadas à importância deste conhecimento nas organizações.

Ademais, grandes avanços no campo da inovação já podem ser observados em uso nos mais diversos ambientes de desenvolvimento, seja para avanços na tecnologia, contribuições para a sociedade ou impactos ao meio ambiente.

Ferreira *et al.* (2018) realizou um levantamento na literatura científica para a identificação de possíveis interligações e correlações entre diferentes temas ligados à inovação. Os resultados conseguidos demonstraram pouquíssimos títulos na literatura que abordassem em seus trabalhos uma aproximação dos conceitos tão importantes e similares. Tal resultado demonstra a lacuna ainda existente em se tratar os temas de uma maneira integrada para compor cada vez mais uma base sólida de pesquisas.

Métodos e técnicas pouco ainda utilizados por empresas, como por exemplo o *Machine Learning*, *Data Mining*[1] e o *Deep Learning*[2], além de ferramentas de *Business Intelligence*, são importantes para fornecer novas análises e novos *insights* sobre comportamentos existentes em grandes bancos de dados, seja de clientes, financeiros ou de processos. A utilização desse tipo de inovação trará para gestores uma nova forma de tomar decisões, baseando-se sempre em uma gestão por dados como diferencial de mercado.

Diversos autores têm trazido abordagens práticas e inovadoras na forma de usar a gestão por dados (*Data-Driven Decision Making*) na tomada de decisões empresariais (Long, 2018; Ali & Lande, 2019; Cavalcante, Frazzon, Forcellini & Ivanov, 2019; Zhang et al., 2019). Conforme Arunachalam e Kumar (2018), na era do *Big Data*[3], a tomada de decisão baseada em dados prevalece, independentemente do tamanho da empresa. O uso de informações orientadas a dados permitiria aos tomadores de decisão resolver problemas de negócios complexos. Com a disponibilidade de ferramentas e técnicas de análise de código aberto, a tomada de decisões orientada a dados não está longe do alcance das organizações em necessidade.

O *Machine Learning* usa computadores para simular a aprendizagem humana e permite que eles identifiquem e adquiram conhecimento do mundo real, além de melhorarem o desempenho de algumas tarefas com base nesse novo conhecimento (Portugal, Alencar, & Cowan, 2018). Autores como Parashar e Goyal (2016) utilizaram-se de diferentes técnicas de *machine learning* para definir qual a melhor delas na aplicação de seu estudo sobre classificação de locomoções humanas.

Já Cardoso *et al.* (2019) demonstrou um uso prático da técnica de *Machine Learning* ao apresentar um modelo estatístico para avaliar os riscos financeiros existentes em investimentos em projetos de campanha de *crowdfunding*[4]. A adoção de uma ferramenta com tal poder preditivo mostra-se como mais uma alternativa existente criada para auxiliar a gestão estratégica no momento de tomar decisões sobre qual tipo de

[1] Prática de examinar dados que já foram coletados utilizando diversos tipos de algoritmos, normalmente de forma automática, a fim de gerar novas informações e encontrar padrões. Disponível em: https://www.aquare.la/o-que-e-data-mining-mineracao-de-dados/

[2] Tipo de *machine learning* que configura parâmetros básicos sobre os dados e treina o computador para aprender sozinho através do reconhecimento padrões em várias camadas de processamento. Disponível em: https://www.sas.com/pt_br/insights/analytics/deep-learning.html

[3] É o termo que descreve o grande volume de dados, estruturados e não estruturados, presente em empresas. O *big data* pode ser analisado para obter informações que levam a melhores decisões e movimentos estratégicos de negócios. Disponível em: https://www.sas.com/pt_br/insights/big-data/what-is-big-data.html

[4] Financiamento coletivo. Disponível em: https://m.sebrae.com.br/sites/PortalSebrae/artigos/artigosFinancas/entenda-o-que-e-crowdfunding,8a733374edc2f410VgnVCM1000004c00210aRCRD

campanha investir com base nos riscos que ela apresenta, concentrando-se no uso de probabilidades como fonte de informações.

A utilização do *Machine Learning* encontra-se ainda em crescente desenvolvimento nas empresas, com as mais diversificadas contribuições em diferentes áreas de atuação como saúde, transportes, comércio e finanças. Em empresas cujos objetivos sejam alcançar um sucesso de mercado com a venda de produtos num cenário cada vez mais competitivo, o uso de tais conhecimentos se mostra como uma alternativa às constantes incertezas e oscilações existentes.

O exemplo do estudo realizado por Trigueiro *et al.* (2017) demonstra a importância de se conhecer o perfil consumidor de determinado produto para se criar estratégias mais assertivas sobre seu comportamento dentro do mercado. O autor se utilizou de técnicas de análise fatorial para identificar o comportamento de consumo sobre determinados atributos de compra e assim definir diferentes grupos com perfis diferentes de consumo.

Dentre as abordagens existentes para conquistar um mercado consumidor, a adoção de sistemas de recomendações eficientes torna-se essencial para identificar oportunidades e alavancar percentuais de vendas antes não explorados. As principais tarefas de tais sistemas são tipicamente filtrar fluxos de entrada de informação de acordo com as preferências dos usuários ou apontá-los para itens adicionais de interesse no contexto de um determinado objeto (Karimi, Jannach, & Jugovac, 2018).

Para tanto, técnicas existentes no *Machine Learning* já são utilizadas para criação de tais sistemas, baseando-se sempre na observação dos perfis de compra dos clientes e comparando aqueles que sejam similares. Através da realização de agrupamentos menores do banco de dados a partir de atributos escolhidos para a análise é possível se chegar a grupos com características em comum que permitam uma mesma abordagem em termos de recomendações de compra.

Conforme Lopez, Tucker, Salameh, & Tucker (2018) métodos não supervisionados não requerem um conjunto de treinamento que contenha informações *a priori* de rótulos de classe de objetos como entrada. Métodos não supervisionados são capazes de detectar estruturas de *cluster* potencialmente interessantes e novas em um conjunto de dados.

A clusterização é hoje uma das técnicas com amplo potencial de aplicação para ações voltadas para a comercialização de produtos e serviços. De acordo com Nilashi *et al.* (2017), clusterização é definida como um processo de colocar um conjunto de objetos em vários grupos razoáveis de acordo com a similaridade entre eles. São exemplos de aplicações da clusterização: novas ações de marketing, direcionamento de vendas, categorização de produtos, avaliações de crédito de clientes e agrupamentos espacial de clientes espalhados para melhor distribuição de pontos de venda.

O presente artigo se propõe a aplicar a técnica de *Machine learning* de clusterização por *k-means* para criação de um sistema de recomendação de produtos com base no perfil de compra de clientes de uma empresa distribuidora de produtos.

Dessa forma o objetivo da aplicação é tornar o mix de produtos vendidos das lojas com menor variabilidade, mais eficiente, baseando-se no perfil de compra dos clientes de uma mesma região, para lojas pertencentes ao mesmo cluster. Além disso, a metodologia de construção do sistema de recomendações fornece mais uma fonte de embasamento para possíveis novas aplicações com essa mesma finalidade de investigação do perfil de clientes.

As contribuições deste artigo são:

- Desenvolvimento detalhado de um sistema de recomendação de produtos específico para distribuidoras baseado nos principais algoritmos de *Machine learning* presentes na literatura;
- Análise detalhada acerca dos métodos de redução de dimensionalidade, pouco presente em estudos anteriores sobre sistemas de recomendação;
- Fornecimento de uma solução para auxiliar a gestão de vendas com base na identificação de segmentos similares de clientes.

O artigo se divide em 5 sessões: além desta, a sessão 2 cita os trabalhos já realizados na área sobre o tema abordado; na sessão 3 é demonstrada a metodologia utilizada no trabalho, assim como a metodologia para a criação do sistema de recomendações; a sessão 4 mostra o desenvolvimento da metodologia de criação

do sistema caracterizando cada uma de suas etapas descritas; a sessão 5 demonstra a aplicação do sistema de recomendação realizada na empresa modelada. Por fim, a sessão 6 é destinada a discorrer sobre as conclusões retiradas do estudo.

2 REVISÃO BIBLIOGRÁFICA

De forma a se alcançar os diferentes objetivos comerciais de uma organização, diferentes abordagens são requeridas para a construção de sistemas de recomendação eficientes. Ademais, o uso de técnicas variadas de *Machine Learning* como meio principal de abordagem possibilita o uso em diversas aplicações nos mais diferentes segmentos de mercado.

2.1 Aplicações de Sistemas de Recomendação

Chiu e Ko (2017) desenvolveram um sistema inteligente de seleção de músicas de forma a aumentar a performance produtiva de pessoas na realização de atividades, visto estudos anteriores que comprovam o aumento de eficiência no trabalho com base no hábito de ouvir música durante tais exercícios. Foram utilizadas novas tecnologias como sensores de variação cardíaca para interpretar as emoções sentidas e assim ajudar a construir o banco de dados necessário para o estudo. O sistema proposto pode estimular aplicativos inovadores para fábricas e seus resultados significativos gerados por meio da validação experimental indicam que esse sistema gera altos níveis de satisfação, não aumenta a carga de trabalho mental e melhora o desempenho dos usuários.

Fernández-García *et al.* (2019) aplicaram diferentes algoritmos de *Machine Learning* na busca por um melhor modelo que procurasse prever os componentes-base de aplicações desenvolvidas que deveriam ser sugeridos para cada usuário. Como cada cliente possui diferentes necessidades, o sucesso da aplicação passa pelo quanto ela consegue ser útil à resolução dos problemas individuais de cada um dos clientes. O autor ainda cita que prevendo componentes mais de perto alinhados a cada situação, pretende-se melhorar a experiência do usuário em aplicações de *software* e, assim, otimizar as possibilidades de êxito alcançar uma boa posição na crescente competitividade mercado de desenvolvimento de *software*.

Frémal e Lecron (2017) optaram por uma nova abordagem para clusterização de filmes baseando-se em seus gêneros e avaliações recebidas por usuários. Através de uma ponderação realizada sobre essas avaliações foi possível se construir um sistema de recomendações baseado nas previsões de avaliação geradas para cada *cluster* construído. Segundo os autores, os sistemas de recomendação são ferramentas úteis para atividades *on-line*, como sites de comércio eletrônico ou recomendação; com tantos itens e sem um bom sistema de recomendação, os clientes podem perder itens que realmente interessam a eles, levando a um déficit para as empresas.

2.2 Tratamento e processamento de dados

Referindo-se a toda a parte que envolve o processamento e tratamento de dados utilizados para os modelos de classificação utilizados em sistemas de recomendação, importantes passos têm sido utilizados para se atingir uma forma final para os dados que seja mais conveniente de acordo com o problema.

No que tange ao tratamento da base de dados trabalhada, Sangeetha e Prakash (2019) utilizaram a normalização de dados como passo fundamental para manter os valores retirados do pré-processamento de dados na mesma escala de avaliação. Sua aplicação utilizou-se de um modelo bastante conhecido de sistema de recomendações, acrescentando-lhe uma nova abordagem para avaliação de sentimentos em textos de avaliações de produtos. Como a fonte primária de dados coletados são palavras, o processo de normalização foi o agente que, contando com o auxílio de outros algoritmos para análise de textos, traduziu de forma quantitativa e em uma escala padrão o real valor de cada item do banco.

Já Schenatto *et al.* (2017) compararam o uso de diferentes tipos de normalização frequentemente utilizadas para a realização de clusterizações voltadas às chamadas zonas de gestão, termo utilizado na área do agronegócio para se referir a diferentes regiões que devem receber diferente formas de tratamento quanto ao uso de fertilizantes, manejo do solo e irrigação.

Para a redução de dimensionalidade, Majumdar (2018) se destaca por realizar a comparação de três variações para a técnica do *AutoEncoder*: a primeira se dando de uma maneira não supervisionada, a segunda sendo voltada exclusivamente para aplicações em clusterizações e a última, supervisionada, voltada para problemas de classificação de rótulos ou multirrótulos. O estudo fornece relevante embasamento teórico para usos futuros neste campo do conhecimento e utilizações desta importante técnica.

Já Diale, Celik, & Van Der Walt (2019) utilizaram-se de uma nova abordagem para o *AutoEncoder* na busca de conseguir uma representação mais robusta de informações em um espaço de dimensões menores de forma a se construir um sistema de classificações de e-mails com melhor desempenho. Sua consideração baseia-se no processamento das palavras contidas em cada e-mail analisado de modo que se gere um avaliador para cada um deles, para se gerar uma redução de dimensionalidade que agregue as informações do avaliador e assim atribuir maior densidade de informações relevantes mesmo em um espaço reduzido.

2.3 Novas metodologias de sistemas de recomendação

Diversas inovações em metodologias e abordagens para novos sistemas de recomendação têm sido amplamente desenvolvidos para as mais diversas aplicações. O uso de novas técnicas ou a mescla de técnicas já existentes mostraram-se ser capazes de fornecer resultados superiores em termos de eficiência e acurácia para sistemas de recomendação que usam previsões para realizar as sugestões para os clientes.

Wasid e Ali (2018) apresentaram uma nova abordagem para a técnica de clusterização com base na utilização de classificação para multicritérios. Além disso, foi escolhido um método para verificar o nível de similaridade de itens dentro de um mesmo *cluster* de forma a gerar recomendações mais eficientes. Já Nilashi *et al.* (2017) também exploraram essa área propondo um novo método para aumento de acurácia dos sistemas de recomendação baseados em avaliações multicritérios aplicados ao setor de turismo. A forma desenvolvida se mostrou bastante eficaz em experimentos realizados para teste, o que demonstrou o potencial de sua metodologia agregando diferentes técnicas para o desenvolvimento do novo sistema.

Irvan e Terano (2016) propuseram uma abordagem diferenciada para sistemas de recomendação voltados para grupos de pessoas (usuários de serviços *on-line*). A inovação se dá por conta de que os integrantes de um grupo possuem características e gostos distintos, o que compromete a eficiência das sugestões oferecidas devido ao nível elevado de complexidade. Govind, Tene, & Saideep (2018) desenvolveram um sistema que realiza recomendações baseadas em sentimentos identificados na análise de filmes por usuários de famosos serviços de *streaming*. A contribuição da nova abordagem mostra ser mais adequada para gerar recomendações mais intuitivas e direcionadas aos usuários de tais serviços.

A técnica utilizada no presente estudo já foi bastante explorada em trabalhos anteriores para sistemas de recomendação. Autores como Zahra et al. (2015) Kant *et al.* (2018) Wen, Bao, & Ding (2018) e Putriany, Jauhari, & Heroza (2019) aplicaram a clusterização por *k-means* como forma de agrupar os diferentes itens de um conjunto de dados para o desenvolvimento de sistemas de recomendação. A utilização de tal técnica de *Machine learning* é preferível para aplicações em varejo ou atacado por conta de os dados de operações de compra serem capazes de fornecer informações importantes sobre perfis de compra por parte dos clientes com base em seus históricos de movimentação.

3 METODOLOGIA

O presente artigo propõe inicialmente a realizar estudos sobre métodos e técnicas existentes para a criação de um sistema de recomendação de produtos. Baseando-se principalmente em algoritmos de *Machine*

Learning, a metodologia do estudo buscou referências em aplicações anteriormente realizadas nesse campo do conhecimento sobre análise e tratamento de dados para geração de modelos preditivos.

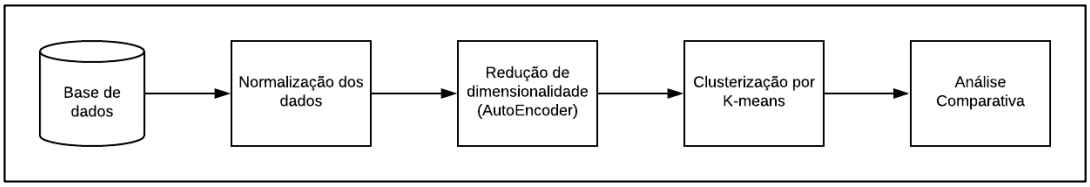
Para tanto, dentre as inúmeras abordagens possíveis para se desenvolver o sistema, a escolhida foi a técnica de clusterização por *K-means*, famoso algoritmo amplamente utilizado para realização de agrupamentos de objetos contidos em um banco de dados.

Para aplicação correta de tal técnica, tendo em vista os objetivos pretendidos, uma série de procedimentos necessários foi realizada.

3.1 Metodologia de desenvolvimento do sistema de recomendações

As principais etapas do desenvolvimento do sistema de recomendação de produtos podem ser observadas de acordo com a figura abaixo:

Figura 1 – Metodologia de criação do sistema de recomendações



Fonte: Elaborado pelos autores (2020).

Primeiramente foi concedido o acesso ao banco de dados a ser utilizado para construção do modelo de clusterização pretendido. O banco de dados é na ordem de milhares de registros de transações realizadas pelos clientes da empresa modelada para aplicação do sistema de recomendações.

Logo após, uma normalização dos dados foi aplicada ao banco de forma que os valores referentes a quantidade de itens comprados por produto fossem considerados a partir de então como um valor entre 0 e 1, representando assim uma proporção de compra comparada a todo o mix adquirido por aquele cliente. Dessa forma fica mais visível observar o perfil de compra dos clientes, observando suas principais preferências.

Em seguida, foi utilizada a técnica de redução de dimensionalidade *AutoEncoder*, uma rede neural que permite diminuir a dimensão de um banco de dados para uma camada de dimensão menor para então reproduzi-la novamente. Tal técnica é utilizada para se trabalhar com uma representação menor e mais relevante dos dados e assim diminuir utilização de recursos e tempo computacional as análises.

Por fim foi aplicado o algoritmo de clusterização para o devido agrupamento dos objetos contidos no banco. A partir de então, o funcionamento do sistema de recomendação se baseia em uma análise comparativa entre clientes que ficaram agrupados num mesmo cliente, observando seus perfis de compras para itens em comum e complementares do mix de compra.

Uma vez construído o sistema de recomendações de produtos, sua aplicação se deu através de uma empresa de distribuição de produtos com atuação em todo o território cearense.

4 DESENVOLVIMENTO

O desenvolvimento da metodologia apresentada se deu através de sua modelagem em uma empresa distribuidora de produtos. Para tanto o estudo foi construído em cima do tratamento dos dados fornecidos pela empresa. Os dados originais contêm informações a respeito de transações realizadas por clientes da empresa e reúne campos referentes aos valores de compra de itens, quantidades, categorias dos itens, descrição e identificação dos clientes.

4.1 Transformação e normalização dos dados

O banco de dados fornecido pela empresa em estudo, em sua forma original, não é o mais adequado para aplicação do modelo de clusterização proposto com base no perfil de compra de produtos pelos clientes existentes. O quadro 1 mostra um exemplo da forma original de um banco de dados.

Quadro 1 – Banco de dados original disponível

Id Cliente	Nome fantasia	Id Produto	Qtd pedida	Valor pedido	Descrição do produto	Categoria do produto
76049000128	XXX	7080	1	18,29	AAA	Alimentos
93081000111	YYY	25247	1	18,29	BBB	Alimentos
12718100011	ZZZ	15603	1	18,29	CCC	Alimentos

Fonte: Elaborado pelos autores (2020).

A primeira transformação necessária então ao banco foi a de torná-lo um *Dataframe* [5]contendo valores que representem a real proporção de compra de um item (produto) por um cliente, sempre com base em todo o mix de compras realizadas por ele. Assim, o banco de dados assume uma nova aparência, em que seus objetos passam a ser os clientes e suas *features* [6]passam a ser cada produto do catálogo da empresa em estudo, como pode ser visto no exemplo do quadro 2.

Quadro 2 – Banco de dados normalizado

Clientes	**Produto A**	**Produto B**	**Produto C**	**...**	**Produto N**
A	0,02	0,13	0	...	0,003
B	0,08	0	0,18	...	0,01
C	0,17	0,001	0,05	...	0,04

Fonte: Elaborado pelos autores (2020).

A normalização dos dados se fez necessária pelo fato de que os diferentes atributos utilizados para as análises possuem diferentes escalas de valores. A normalização para valores proporcionais que fiquem entre 0 e 1 se mostra então uma abordagem adequada por tratar o maior valor daquela escala como sendo 1 e o menor como sendo 0. Esse artifício se aplica na análise exigida por deixar as devidas quantidades de itens comprados em uma mesma escala para comparação frente ao total de itens adquiridos por um mesmo cliente.

Tal processo é de grande importância visto os benefícios de minimizar o risco de inconsistências em análises e trazer facilidade para manuseio dos dados, além de se evitar problemas com os resultados gerados pelo modelo de clusterização utilizado a seguir.

4.2 Redução de dimensionalidade

Em seguida, foi utilizada a técnica de redução de dimensionalidade conhecida por *AutoEncoder* ou Método *Autoassociative* Neural Networks. Essa técnica é uma rede neural treinada de forma não supervisionada para aprender as características mais relevantes das *features* de entrada, reduzi-la até um espaço de menor dimensão e conseguir reconstrui-la na forma mais próxima possível das *features* que entraram.

Além de trabalhar com um conjunto de dados de dimensão reduzida, o que facilita a operação dos modelos empregados e análises realizadas, a aplicação do *AutoEncoder* também permite detectar possíveis

[5] *Dataframe* é semelhante a uma matriz, mas as suas colunas têm nomes e podem conter dados de tipos diferentes. Pode ser visto como uma tabela de uma base de dados, em que cada linha corresponde a um registo (linha) e cada coluna corresponde às propriedades (campos) a serem armazenadas para cada registo da tabela. Disponível em: https://www.dcc.fc.up.pt/~ltorgo/SebentaR/HTML/node16.html

[6] *Features* são as propriedades (campos) a serem armazenadas para cada registo do *dataframe*.

anomalias nos dados, visto que o modelo aprende uma representação da maior parte dos dados, não dando tanta importância para *outliers*, o que é fundamental para se evitar erros nos resultados finais.

Um *AutoEncoder* pode ser dividido em duas partes, conforme a figura 2: *Encoder*, função h(x) que transforma a entrada (*Input layer*) para uma representação menor através de várias camadas (*hidden layer*) e *Decoder*, função r(x) que transforma a representação h(x) em sua reconstrução aproximada da original (*Output layer*) h'(x). Neste cenário foi escolhido utilizar 8 camadas, 4 para o *Encoder* e 4 para o *Decoder*. Além disso o número de neurônios escolhidos foi metade do número de neurônios da camada anterior, para o *Encoder*, e o dobro da camada anterior, para o *Decoder*.

Cada camada do *AutoEncoder* é ajustada conforme a seguinte equação desenvolvida por Liang e Liu (2015):

$$y = W \times X + b \quad (1)$$

Sendo, no estudo proposto, a função de ativação escolhida a seguir:

$$r(c) = \max(0, x) \quad (2)$$

Sendo assim, o modelo matemático final adaptado pode ser descrito através da seguinte forma:

$$Fi = \begin{cases} r(W_1 \times X + b_1), \ i = 1 \\ r(Wi \times F_{i-1} + bi), \ caso\ contrário \end{cases} \quad (3)$$

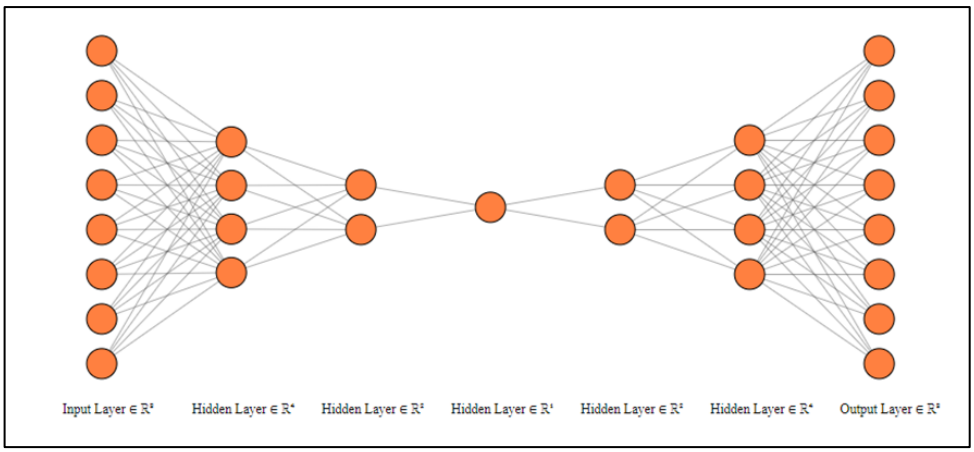
Sendo a camada reduzida igual a:

$$X_{new} = F_4 \quad (4)$$

Por fim, no artigo em questão adotou-se o método do gradiente descendente para minimizar a seguinte função:

$$\frac{1}{n} \times \sum_{j=1}^{n} (X_j - F_{4j})^2 \quad (5)$$

Figura 2 – Imagem representativa do *AutoEncoder*



Fonte: Elaborado pelos autores (2020).

A escolha pela técnica do *AutoEncoder* se deu em vista dos resultados conseguidos com a clusterização estarem mais homogêneos do que com a aplicação de outras técnicas. Com sua utilização,

resultados menos discrepantes para um mesmo *cluste*r passaram a ser obtidos, o que correspondia com os objetivos definidos para o desenvolvimento do sistema de recomendações.

4.3 Clusterização por K-Means

Por fim, a realização dos devidos agrupamentos de clientes para construção do sistema de recomendações ficou por conta do algoritmo de aprendizado não-supervisionado *k-means* esférico. Sua utilização é justificada pelos melhores resultados obtidos com a clusterização em termos de homogeneidade e discrepância dos itens de um *cluster* frente àqueles conseguidos com outras variações do *k-means* ou outros algoritmos, como o KNN[7].

O algoritmo *k-means* foi implementado utilizando como métrica a distância euclidiana e a distância esférica, baseadas na distância cosseno.

Dados dois vetores A e B a distância cosseno pode ser representada pela seguinte expressão matemática:

$$\cos(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \quad (6)$$

enquanto a distância euclidiana pode ser definida como:

$$\sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \quad (7)$$

Vale salientar que o algoritmo *K-means* é aquele que apresenta os melhores resultados quando trabalha com bancos de dados de várias dimensões. Ainda que tenha sido realizada uma redução de dimensões utilizando-se a rede neural *AutoEncoders*, o banco utilizado ainda ficou com múltiplas dimensões a serem levadas em conta com a aplicação dos algoritmos.

Ademais, para definição do número k de *clusters* a serem criados para a aplicação foi utilizado o método de *Elbow*. Tal método indica o número mais adequado de *clusters* para se utilizar no processo baseado na variância que os dados sofrem à medida em que se aumenta o número de *clusters.*

Para a avaliação dos resultados conseguidos com a clusterização realizada, dois outros métodos também foram empregados: Método *Silhouette* e Índice de *Dunn*. O primeiro mede o quanto um item é coeso em seu *cluster* comparado a outros. O valor obtido indica a compatibilidade daquele objeto com seu próprio *cluster*. Já o segundo confere a compactação de cada *cluster* referente a pequenas variações entre seus componentes, assim como diferenças de *cluster* para *cluster*. Ambos os métodos são essenciais para avaliar a eficiência do algoritmo utilizado para os agrupamentos.

A clusterização final realizada se deu pelo mix de produtos comprados pelos clientes. Dessa forma aqueles clientes que possuíssem perfis aproximados de proporção de compra teriam a tendência de serem alocados ao mesmo *cluster*.

4.4 Análise Comparativa

O princípio de funcionamento do sistema de recomendação baseia-se na etapa final de análise dos *clusters* construídos. Uma vez com os agrupamentos realizados, os clientes que foram alocados ao mesmo *cluster* possuem perfis de compra similares e, portanto, serão comparados entre si para a geração de sugestões de novas compras do sistema de recomendação.

[7] K-*Nearest Neighbors* é um dos muitos algoritmos de aprendizagem supervisionada usado no campo de *machine learning*. Ele é um classificador onde o aprendizado é baseado "no quão similar" é um dado item de outro item. Disponível em: https://medium.com/brasil-ai/knn-k-nearest-neighbors-1-e140c82e9c4e

Os clientes que sejam próximos geograficamente e que estejam em um mesmo *cluster* são comparados observando-se a dispersão de seus itens comprados.

Para o cálculo da distância de cada cliente para cada vizinho mais próximo foi utilizado a fórmula da distância Haversine (8):

$$haversin\left(\frac{d}{R}\right) = haversin(\Delta\emptyset) + \cos\left(\emptyset_{1)}\right)\cos(\emptyset_2)\, haversin(\Delta\lambda) \quad (8)$$

Dado o mix de produtos usualmente comprado por aquele cliente alvo, o sistema de recomendações fará sugestões quanto a produtos não adquiridos por ele e que sejam vendidos por seus vizinhos próximos em uma boa proporção.

Da mesma forma, produtos daquele cliente que tenham um percentual de venda baixo, podem também ser recomendados para venda já que a venda do item alcança bons resultados em outros estabelecimentos próximos, o que demonstra uma certa procura na região pelo público consumidor.

5 APLICAÇÃO DO SISTEMA DE RECOMENDAÇÕES

Após o desenvolvimento do sistema de recomendações, a aplicação se deu em uma empresa do ramo de distribuição de produtos dos segmentos de alimentos, higiene e limpeza. A empresa possui atuação em todo território do estado do Ceará, localizado no nordeste brasileiro e possui 6.809 clientes de diferentes portes, o que foi decisivo para realização das análises necessárias quanto à comparação entre eles.

O banco de dados utilizado na aplicação fornece uma base confiável para que a aplicação gere resultados condizentes com a realidade já que é bastante volumoso e conta com campos bastantes significativos para as análises de perfis de compra exigidos pela metodologia, possuindo em torno de 1.800 produtos distintos oferecidos aos clientes da distribuidora em um período aproximado de 15 meses de registro de transações (Junho de 2017 a Setembro de 2018).

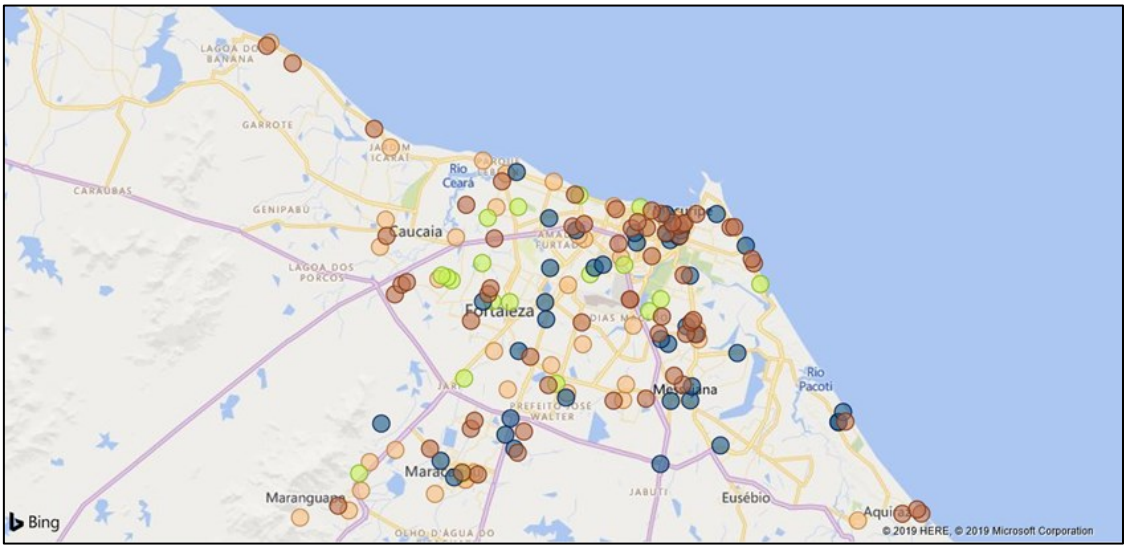
O sistema gerado apresenta bom desempenho quanto a sua velocidade de processamento, apresentando resultados quase que de maneira imediata uma vez realizada a clusterização. A aplicação do sistema se deu através de fornecimento de ferramenta em *Business Intelligence* (BI) para empresa foco do estudo de modo que suas operações pudessem ser realizadas de maneira integrada com seu próprio sistema ERP para um melhor acompanhamento e controle dos dados e resultados.

5.1 Análise e discussão dos resultados

Com base nisso, a criação do sistema de recomendação de produtos para a aplicação proposta conseguiu entregar resultados factíveis dentro do esperado. Uma vez organizados em *clusters,* os clientes puderam ser considerados como parte de um grupo com características de compra semelhantes, o que gera uma nova forma de tratamento para eles em quesitos mercadológicos. Dessa maneira, por exemplo, uma gama de produtos nunca oferecidos a determinado cliente pode passar a fazer parte de seu mix já que ele possui boas vendas em estabelecimentos similares nas proximidades.

A Figura 3 mostra um exemplo de *clusters* de clientes localizados na cidade de Fortaleza/CE com perfis de compra similares. A diferenciação se dá pelas cores.

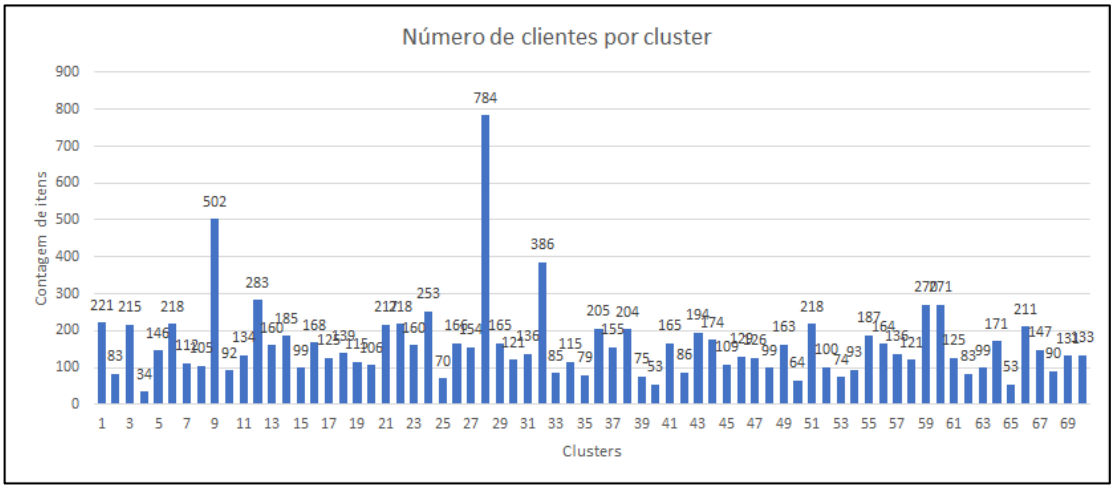
Figura 3 – Exemplo de clusters construídos para o problema



Fonte: Elaborado pelos autores (2020).

Seguindo os procedimentos já descritos, os resultados para o banco de dados em estudo foram os seguintes mostrados na Figura 4.

Figura 4 – Número de clientes por *cluster* gerado



Fonte: Elaborado pelos autores (2020).

A figura 4 representa os 70 *clusters* gerados no modelo para todo o estado do Ceará com a quantidade de clientes presentes em cada um deles. É possível notar-se a ausência de uma homogeneidade no número de clientes, o que demonstra uma divisão não necessariamente igual para particionar o conjunto inicial de dados do banco. Tal resultado já era esperado e conforme o número de clientes mude em cada *cluster*, as formas de abordagem comercial também mudam por parte da gestão de vendas da empresa distribuidora, variando consideravelmente fatores como o preço de determinados produtos para *clusters* que possuem muitos clientes buscando por eles.

O sistema de recomendação fornece seus apontamentos na forma de duas listas de produtos: a primeira se refere aos produtos que o cliente já vende, mas não na proporção ideal com base no público

consumidor; já a segunda mostra os produtos que o cliente ainda não vende em seu estabelecimento, mas que possuem boa proporção de venda em vizinhos próximos.

Através desta simples abordagem o sistema recomenda uma lista contendo 5 produtos para cada cliente de cada *cluster* de forma que assim possam ser mais assertivos em seus pedidos à distribuidora para assim poder garantir uma porcentagem de venda maior de seus produtos. Tal ganho acarretará diversos benefícios, tais como: aumento no faturamento, aumento da porcentagem de vendas de seu mix de produtos ofertados e menores problemas com produtos parados em estoque.

Espera-se que com a adoção prática do sistema desenvolvido seu uso possa vir a ser mais uma alternativa para tomada de decisões gerenciais, mercadológicas e comerciais. O uso cada vez mais de aplicações dentro do campo de estudo da inteligência comercial tende a fornecer ferramentas para que as empresas gradualmente migrem para as inovações que surgem diariamente.

5. CONCLUSÃO

O presente artigo demonstrou uma aplicação voltada para o desenvolvimento de sistemas de recomendação de produtos a partir de uma empresa de atuação no ramo de distribuição de produtos em atacado. Foi demonstrada a utilização de uma técnica de *Machine learning* para o agrupamento e posteriormente análise dos clientes que compõem o objeto de estudo do banco de dados. Com os resultados é possível se afirmar a confiabilidade do sistema quanto ao agrupamento de clientes realizado e direcionar diferentes ações de venda mais assertivas considerando as diferenças existentes entre os estabelecimentos de forma a se observar os resultados na prática.

Para o desenvolvimento do sistema de recomendações de produtos, os procedimentos e análises com *big data* contaram com uma série de dificuldades a respeito da limpeza e correto tratamento do banco de dados fornecido, desde a remoção de erros, valores não formatados no padrão exigido ou preenchimento de dados faltantes. O adequado trabalho com grande volume de dados exige bastante precisão para que seu uso não seja defasado em relação aos objetivos pretendidos para a aplicação.

Da mesma forma, os tratamentos necessários para aplicações com *Big Data* exigem uma capacidade de processamento bastante elevada, o que foi um dos obstáculos enfrentados nas etapas de desenvolvimento do sistema. Para tanto, o processamento dos dados utilizou-se de servidor localizado em nuvem, o que foi o diferencial para uma abordagem mais rápida e sem problemas com perda de informação.

Por fim, o estudo não se propôs a realizar uma comparação entre diferentes técnicas, o que provavelmente forneceria resultados diferentes alcançados e assim uma avaliação de assertividade poderia ser realizada com testes reais na empresa em estudo. Outras formas de clusterização existentes (*KNN, DB-Scan*) e até mesmo outra técnicas de *Machine learning* (Árvore de decisão, Regressão logística) poderiam ser aplicadas ao problema como nova abordagem para seu desenvolvimento. Dessa forma, o estudo encontra sua limitação justamente na oportunidade que estudos futuros têm para contribuir com os avanços alcançados pelo uso de tais técnicas em diferentes ambientes de varejo e logística.

A utilização de técnicas de inteligência comercial, sejam utilizando métodos computacionais, de inteligência artificial ou de algoritmos preditivos, como o *Machine Learning*, tendem a fazer cada vez mais parte da realidade gerencial de empresas que busquem um diferencial competitivo pautado pelo advento da revolução 4.0. A existência de diversas técnicas com bastante potencial explorável depende ainda de uma mentalidade estratégica por parte das organizações que se voltem à gestão por dados e adoção do *business intelligence* para tomada de decisões.

REFERÊNCIAS

Ali, H., & Lande, M. (2019). Data-Driven Decisions in Prototyping and Product Development: A Framework for Uncertainty and Decision-Making. *ASME International Mechanical Engineering Congress and Exposition, Salt Lake City, Vol. 83518, p. V014T14A039.*

Arunachalam, D., & Kumar, N. (2018). Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making. *Expert Systems with Applications, 111, 11-34.*

Cardoso, G. P., Sininbardi, B. G., Sobral, M. S., & de Souza, E. M. (2019). Modelo de avaliação de risco em campanhas kickstarter utilizando machine learning. *Navus-Revista de Gestão e Tecnologia, 9(4), 66-79.*

Cavalcante, I. M., Frazzon, E. M., Forcellini, F. A., & Ivanov, D. (2019). A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing. *International Journal of Information Management, 49, 86-97.*

Chiu, M. C., & Ko, L. W. (2017). Develop a personalized intelligent music selection system based on heart rate variability and machine learning. *Multimedia Tools and Applications, 76(14), 15607-15639.*

Diale, M., Celik, T., & Van Der Walt, C. (2019). Unsupervised feature learning for spam email filtering. *Computers & Electrical Engineering, 74, 89-104.*

Fernández-García, A. J., Iribarne, L., Corral, A., Criado, J., & Wang, J. Z. (2019). A recommender system for component-based applications using machine learning techniques. *Knowledge-Based Systems, 164, 68-84.*

Ferreira, E. de Paula, Junior, M. R. B., Isnard, P. A., de Souza França, R., & de Aguiar Filho, A. S. (2018). Gestão do conhecimento, internet das coisas e inovação: a relação dos temas e a intensidade de pesquisas realizadas. *Navus-Revista de Gestão e Tecnologia, 8(3), 99-112.*

Fraga, B. D., Erpen, J. G., Varvakis, G., & dos Santos, N. (2017). Business Intelligence: métodos e técnicas de gestão do conhecimento e as tendências para avanços do capital intelectual. *Navus-Revista de Gestão e Tecnologia, 7(1), 43-56.*

Frémal, S., & Lecron, F. (2017). Weighting strategies for a recommender system using item clustering based on genres. *Expert Systems with Applications, 77, 105-113.*

Govind, B. S., Tene, R., & Saideep, K. L. (2018). Novel Recommender Systems Using Personalized Sentiment Mining. *IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, 4, 1-5.*

Irvan, M., & Terano, T. (2016). Group Recommendation System for E-Learning Communities: A Multi-agent Approach. *Advances in Social Computing and Digital Education, Singapore, 7, 35-46.*

Kant, S., Mahara, T., Jain, V. K., Jain, D. K., & Sangaiah, A. K. (2018). LeaderRank based k-means clustering initialization method for collaborative filtering. *Computers & Electrical Engineering, 69, 598-609.*

Karimi, M., Jannach, D., & Jugovac, M. (2018). News recommender systems–Survey and roads ahead. *Information Processing & Management, 54(6), 1203-1227.*

Liang, J., & Liu, R. (2015). Stacked denoising autoencoder and dropout together to prevent overfitting in deep neural network. *2015 8th International Congress on Image and Signal Processing (CISP), London, 22, 697-701.*

Long, Q. (2018). Data-driven decision making for supply chain networks with agent-based computational experiment. *Knowledge-Based Systems, 141, 55-66.*

Lopez, C., Tucker, S., Salameh, T., & Tucker, C. (2018). An unsupervised machine learning method for discovering patient clusters based on genetic signatures. *Journal of biomedical informatics, 85, 30-39.*

Majumdar, A. (2018). Graph structured autoencoder. *Neural Networks, 106, 271-280.*

Nilashi, M., Bagherifard, K., Rahmani, M., & Rafe, V. (2017). A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques. *Computers & industrial engineering, 109, 357-368.*

Parashar, A., & Goyal, D. (2016). Clustering Gait Data Using Different Machine Learning Techniques and Finding the Best Technique. *International Conference on Smart Trends for Information Technology and Computer Communications, Singapore, 426-433.*

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications, 97, 205-227.*

Putriany, V., Jauhari, J., & Heroza, R. I. (2019). Item Clustering as An Input for Skin Care Product Recommended System using Content Based Filtering. *Journal of Physics: Conference Series, Palembang, Vol. 1196, No. 1, p. 012004.*

Sangeetha, J., & Prakash, V. S. J. (2019). Improved Feature-Specific Collaborative Filtering Model for the Aspect-Opinion Based Product Recommendation. *Advances in Big Data and Cloud Computing, Singapore, 275-289.*

Schenatto, K., de Souza, E. G., Bazzi, C. L., Gavioli, A., Betzek, N. M., & Beneduzzi, H. M. (2017). Normalization of data for delineating management zones. *Computers and Electronics in Agriculture, 143, 238-248.*

Trigueiro, F. M. C., Neto, D. A. C., de Sousa Santos, T., & Prearo, L. C. (2017). Comportamento de consumo no segmento de veículos automotores nas cidades de Cuiabá e Várzea Grande. *NAVUS-Revista de Gestão e Tecnologia, 7(3), 7-18.*

Wasid, M., & Ali, R. (2018). An improved recommender system based on multi-criteria clustering approach. *Procedia Computer Science, 131, 93-101.*

Wen, T., Bao, J., & Ding, F. (2018). QoS-Aware Web Service Recommendation Model Based on Users and Services Clustering. *Proceedings of the International Conference on Information Technology and Electrical Engineering 2018, Xiamen, 18, 1-6.*

Zahra, S., Ghazanfar, M. A., Khalid, A., Azam, M. A., Naeem, U., & Prugel-Bennett, A. (2015). Novel centroid selection approaches for KMeans-clustering based recommender systems. *Information sciences, 320, 156-189.*

Zhang, Y., Zhang, R., Wang, Y., Guo, H., Zhong, R. Y., Qu, T., & Li, Z. (2019). Big data driven decision-making for batch-based production systems. *Procedia CIRP, 83, 814-818.*