NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Algoritmos de segmentação para classificação de inventários: DBSCAN

Cassandra Marie Custódio

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

ALGORITMOS DE SEGMENTAÇÃO PARA CLASSIFICAÇÃO DE INVENTÁRIOS: DBSCAN

por

Cassandra Marie Custódio

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação, Especialização em Gestão do Conhecimento e Business Intelligence.

Orientador: Mauro Castelli

Novembro 2017

RESUMO

A gestão de inventário constitui um grande desafio para as empresas na medida em que uma má gestão pode ter elevado impacto na rentabilidade da empresa. É importante conseguir manter os níveis de inventário baixos e escoar os produtos o mais rápido possível. Uma técnica frequentemente utilizada para ajudar as empresas a dar resposta a este desafio é a classificação de artigos, pois diferentes artigos tem diferentes níveis de procura o que por sua vez determina o inventário. O método mais utilizado na classificação de artigos é a conhecida classificação ABC. No entanto, esta técnica não permite classificar os artigos de inventário com base em todos os critérios considerados importantes para o negócio o que constitui uma grande limitação. Têm sido apresentadas outras técnicas de classificação mas não parecem existir estudos suficientes que explorem os algoritmos de segmentação como solução para este problema. Assim, este estudo tem como objetivo estudar a classificação de artigos de inventário utilizando técnicas de segmentação, mais precisamente o algoritmo DBSCAN. Este algoritmo apresenta importantes vantagens face aos restantes algoritmos de segmentação e pode constituir uma ferramenta sólida para ajudar as empresas a gerir o seu inventário.

PALAVRAS-CHAVE

Gestão de artigos de inventário; Classificação ABC; Análise de *clusters*; DBSCAN;

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

1. INTRODUÇÃO

Nos dias de hoje, a correta gestão de inventários é um dos problemas mais comuns enfrentados pelas empresas (Partovi, F. Y., & Anandarajan, M., 2002). É difícil encontrar um balanço entre a procura e a oferta, isto é, é extremamente difícil prever com exatidão a quantidade que vai ser vendida, num dado período de tempo, de cada artigo disponível para venda. Isto constitui um problema pois, no que diz respeito ao armazenamento de produtos existem custos acrescidos que as empresas têm de ter em conta. Se a quantidade armazenada de um dado produto for muito superior á procura do mesmo, os custos deste tornam-se mais elevados, tendo impacto direto no lucro. Caso se verifique um nível de armazenamento incapaz de satisfazer a procura na totalidade, o impacto reflete-se não só nas receitas, pois representam vendas que poderiam ter sido feitas, mas também na perda de clientes (Partovi, F. Y., & Anandarajan, M., 2002). Gerir a quantidade produzida ou comprada e determinar quanto de cada produto deve existir em armazém, torna-se cada vez mais difícil nos dias de hoje, devido ao crescimento da variedade de produtos em loja (Partovi, F. Y., & Anandarajan, M., 2002). O armazenamento destes é influenciado pelas suas características, onde a diferença no volume de vendas anuais, previsão de procura, custo unitário do produto ou até mesmo requisitos específicos de armazenamento podem resultar em diferentes estratégias de produção e armazenamento. Isto leva a que na vida real, seja vantajoso para as empresas o agrupamento de produtos com características semelhantes em classes. Facilitando deste modo, a tomada de decisão e implementação de estratégias de compra ou produção específicas para cada classe de produtos, em vez de para cada produto separadamente (Van Kampen, T. J., Akkerman, R., & Pieter van Donk, D., 2012).

Em meados do século XVIII, no âmbito de um estudo feito à economia Italiana, o sociólogo e economista Vilfredo Pareto concluiu que 80% da riqueza do país era detida por 20% do total da população (Guvenir, H. A., & Erel, E., 1998). Esta lógica foi rapidamente verificada em enumeras situações, dando origem à conhecida regra de Pareto ou 80/20. A classificação ABC na gestão de inventário constitui uma prática generalizada nas empresas e tem por base a referida regra de Pareto. Esta classificação consiste no agrupamento de artigos em três grandes grupos: A, B e C. O grupo A refere-se ao conjunto de artigos mais importantes num inventário. Estes representam 15 a 20% dos artigos que são responsáveis por 75 a 80% do total do custo anual das vendas. O grupo B é constituído por artigos importantes, mas não tão importantes como os do grupo A. Estes representam 30 a 40% dos artigos que são responsáveis por 15% do total do custo anual das vendas. O grupo C é constituído por artigos de pouca importância. Estes representam cerca de 40 a 50% dos artigos que são responsáveis por 10 a 15% do total do custo anual das vendas (Chu, C. W., Liang, G. S., & Liao, C. T., 2008). Apesar do seu uso ser frequente, a classificação ABC apresenta algumas limitações. O facto de apenas ter em consideração o impacto do custo de cada artigo no total das vendas leva a que, em alguns casos, o método sobrevalorize artigos de alto custo unitário que não são tão importantes para a atividade da empresa, pois a quantidade vendida não é significativa. Noutros casos, o método pode desvalorizar artigos de baixo custo unitário mas, que no entanto, são artigos de grande importância para o negócio devido ao seu volume de vendas (Flores, Benito E., Olson, L. David e Dorai V.K.,1992).

O facto de esta classificação usar apenas um critério, neste caso o total do custo de vendas, pode levantar outro tipo de problema. O uso de um único critério pode originar problemas e perdas

financeiras como por exemplo, um artigo de classe C com tempos longos de entrega, ou artigos de classe A com uma taxa de obsolência elevada pode causar falhas de produção ou elevados níveis de inventário (Guvenir, H. A., & Erel, E., 1998). Apesar do critério utilizado ser uma referência muito importante para quem gere os armazéns, este não é o único. Existem muitos outros critérios de grande importância que ajudam os gestores no planeamento e controlo da compra ou produção de artigos e o seu armazenamento. Estes dependem de negócio para negocio mas podemos dar como exemplo dos critérios mais utilizados a criticalidade do produto, custo de transporte, *lead time*, custo de armazenamento, custo unitário, quantidade vendida, sazonalidade, obsolescência, margem de lucro, entre outros (Van Kampen, T. J., Akkerman, R., & Pieter van Donk, D., 2012). Em algumas situações, alguns destes critérios têm maior peso na classificação do que propriamente o total do custo de vendas. Posto isto, atualmente o maior desafio na gestão de inventários, é conseguir-se classificar os artigos como A, B ou C usando métodos que sejam capazes de ter em consideração todos os critérios relevantes para melhor dar resposta às necessidades do negócio e que, ainda assim corresponda à realidade do mesmo.

Este trabalho está organizado da seguinte forma: A secção 2 apresenta os objetivos deste estudo bem como a sua importância a nível académico e empresarial. A secção 3 sumariza os estudos feitos nesta área até ao momento. A secção 4 descreve detalhadamente os dados e o algoritmo utilizados neste estudo. A secção 5 apresenta uma análise descritiva dos dados e apresenta a análise e discussão dos resultados obtidos. Por último, são apresentadas as conclusões do estudo e as sugestões para trabalhos futuros na secção 6.

2. OBJECTIVOS E IMPORTÂNCIA DO ESTUDO

Um armazém por muito grande ou muito pequeno que seja tem sempre um custo de aquisição. Existe uma renda que as empresas têm de pagar, e para além disso, o armazenamento de produtos também tem um custo de armazenamento por cada unidade. Ou seja, uma empresa vai querer armazenar o mínimo possível para reduzir ao máximo os custos do mesmo, mas ao mesmo tempo não quer ficar sem produtos disponíveis para venda devido a um baixo nível de *stock*. Para evitar a perda de vendas por falta de *stock* e, por outro lado, reduzir custos de armazenamento, começou a haver uma necessidade de se identificar quais os produtos mais importantes para o negócio, e quais os produtos menos importantes. Isto é, os produtos mais vendidos, terão de existir sempre em *stock* e em boa quantidade, e os produtos menos vendidos em menor quantidade. Esta distinção de importância entre artigos começou a ser feita através da conhecida classificação ABC que classifica como A os produtos mais importantes, C os produtos menos importantes e B os produtos intermédios. Como referido anteriormente, esta não foi suficientemente eficaz para dar resposta à complexidade do problema apresentado, dando origem ao uso e exploração de novas técnicas. Os métodos que têm vindo a ser desenvolvidos para a resolução deste tipo de problemas por vezes enfrentam entraves quando são aplicados na prática. Por um lado, existem métodos que apesar de serem muito eficientes a classificar os artigos, são demasiado complexos para serem implementados nas empresas que privilegiam soluções práticas, simples e de rápida implementação. Por outro lado, existem métodos de fácil implementação mas que devido à sua simplicidade não reúnem condições suficientes para englobar todos os fatores essenciais para resolver este problema de forma robusta e gerar confiança nos resultados. Na revisão de literatura serão apresentado os principais métodos que contribuíram para o avanço e melhorias na resolução deste problema, como por exemplo, AHP, redes neuronais, algoritmos genéticos, programação linear, entre outros.

Este estudo tem como objetivo a aplicação de uma técnica que não parece ter sido ainda suficientemente explorada para resolver este tipo de problema. A segmentação, mais conhecida por Clustering (termo original em inglês), é hoje um método bem conhecido e consiste em agrupar em classes pontos representados no espaço através do cálculo de distâncias entre pontos. Os pontos são coordenadas do tipo (X,Y) onde cada eixo constitui um critério de avaliação (Xu, D., & Tian, Y., 2015). Esta forma de agrupar os pontos pode ser útil na resolução do problema apresentado, pois perante um conjunto de artigos, a segmentação irá agrupa-los consoante valores de critérios semelhantes, e irá distinguir os artigos com características diferentes. O algoritmo permite deste modo encontrar uma classificação de artigos A, B e C consoante os critérios escolhidos. A panóplia de técnicas de segmentação existentes neste momento é enorme, de entre as quais este estudo vai incidir especialmente sobre a exploração do método DBSCAN. DBSCAN, abreviatura para *Density-Based Spatial Clustering of Applications with Noise*, é um método moderno de segmentação que se baseia na densidade dos dados apresentados (Xu, D., & Tian, Y., 2015). A densidade associada a um ponto é obtida contando o número de pontos numa determinada região em torno do mesmo, tendo em consideração uma medida de distância. Os pontos com uma densidade acima de um determinado valor numa determinada região são agrupados no mesmo *cluster* (Birant, D., & Kut, A., 2007). Introduzido por Ester, M., Kriegel, H. P., Sander, J., e Xu, X. em 1996, o DBSCAN foi especialmente desenhado para descobrir grupos com formas arbitrárias. Ou seja, o método tem a capacidade de identificar *clusters* que possuem formas lineares, côncavas e ovais para além das formas esféricas encontradas pelos primeiros métodos de segmentação. Adicionalmente, não é necessário definir *a*

priori o número de *clusters* que irão ser gerados, é eficiente mesmo utilizando um grande volume de dados e é insensível a *outliers*.

Posto isto, este estudo pretende encontrar resposta para a seguinte questão: Podem os algoritmos de segmentação constituir uma alternativa válida para identificar quais os produtos mais importantes no inventário de uma empresa tendo em conta um conjunto de características de negócio? Se sim, as empresas passam a ter à sua disposição um novo conjunto de técnicas para resolver este problema.

Neste sentido este trabalho procura contribuir para o avanço e melhoria das técnicas utilizadas na resolução deste tipo de problemas tanto a nível académico como a nível profissional. Os resultados poderão ser interessantes para as empresas que privilegiem a inovação das técnicas já existentes e implementadas no mercado, pois passarão a ter à sua disposição um conjunto maior de técnicas que podem ser usadas para melhorar os seus processos e diferencia-las da sua concorrência. A constante aplicação dos métodos estudados nas faculdades a problemas reais ajuda a constituir uma base para as empresas poderem inovar e competir. Deste modo, as empresas podem tirar grandes benefícios de uma relação mais próxima com as universidades.

Por fim, o método proposto neste estudo é um método muito recente pelo que pode ainda não ter havido a oportunidade de o explorar em diferentes áreas. Por essa razão este estudo pode também ser do interesse dos investigadores pois permite explorar a possibilidade de expandir a aplicabilidade deste método para outras áreas de estudo, como é o exemplo da classificação de artigos de inventário.

3. REVISÃO DE LITERATURA

A classificação ABC clássica deu origem a inúmeros estudos, onde o principal objetivo dos demais foi a implementação de novas técnicas ou exploração de técnicas já existentes aplicadas à classificação de artigos de inventário para ajudar e facilitar a tomada de decisão. Os principais métodos desenvolvidos e explorados que contribuíram para o avanço e melhorias nesta área são: a matriz de dupla entrada de Flores e Whybark (1986), o método de Analytic Hierarchy Process de Saaty (1980), o uso de técnicas de otimização linear, algoritmos de redes neuronais e os algoritmos genéticos.

Recorrendo a uma matriz de dupla entrada, Flores e Whybark (1986), conseguiram encontrar uma forma de incluir mais do que um critério na classificação de artigos, mais precisamente a inclusão de dois critérios. O método abordado pelos autores consiste em classificar sucessivamente os artigos através da tradicional classificação ABC utilizando diferentes critérios de cada vez. A cada par de critérios é então construída uma matriz conjunta onde os resultados obtidos nas duas classificações são comparados (Partovi, F. Y., & Anandarajan, M.,2002). Apesar deste avanço já ser uma melhoria notória, esta técnica continua a ser uma solução limitada, mais precisamente quando estamos perante a necessidade de incluir mais do que dois critérios na análise (Chu, C. W., Liang, G. S., & Liao, C. T., 2008). Neste estudo, os autores consideraram importante o uso dos seguintes critérios: preço unitário do produto, a quantidade vendida anualmente e o *lead time*.

Flores, B. E., Olson, D. L., & Dorai, V. K., em 1992 e Partovi, F. Y., & Burton, J., em 1993 propuseram a aplicação do método *Analytic Heirarchy Process* (AHP) a esta área de estudo. Introduzido por Saaty em 1980, o método AHP consiste no cálculo sucessivo de matrizes de dupla entrada, onde é feita uma comparação subjetiva da importância de cada critério e de cada artigo. Por comparação subjetiva entende-se uma comparação que depende da interpretação da pessoa que está a analisar o problema. A vantagem deste método na classificação de artigos é a sua capacidade de incorporar tanto critérios qualitativos como quantitativos. Por outro lado, é intuitivo e de fácil implementação. Este método conta também com uma grande desvantagem dada a grande subjetividade envolvida tanto no cálculo dos pesos associados de cada critério, bem como na comparação dos artigos face a cada um dos critérios (Partovi, F. Y., & Anandarajan, M., 2002). Os critérios utilizados em ambos os estudos foram o custo unitário médio, o custo total das vendas por ano, a criticalidade do produto e o *lead time*.

Em 1998, Guvenir, H. A. E Erel, E. adaptaram os algoritmos genéticos ao problema de classificação de produtos de inventário. Os algoritmos genéticos são muito utilizados em problemas de otimização de parâmetros. Neste caso, os parâmetros a otimizar representam os pesos dos respetivos valores de critérios, onde a soma de todos os pesos terá de ser igual a 1. Os pesos ótimos para cada critério são então encontrados utilizando o algoritmo genético. Este método foi aplicado e comparado com o método AHP e avaliado pelos respetivos gestores do inventário. Estes consideraram a classificação feita pelo algoritmo genético mais próxima da realidade do negócio do que a classificação feita pelo método AHP. Para realizar este estudo, os autores consideraram como importantes os seguintes critérios: custo unitário, quantidade vendida por ano, custo total das vendas, *lead time* e o facto de ser um produto substituto ou não.

Em 2002, Partovi, F. Y., e Anandarajan, M., adaptaram as redes neuronais ao problema de classificação de produtos de inventário de uma indústria farmacêutica utilizando dois métodos de aprendizagem diferentes, backpropagation e algoritmo genético. Estes consistem no reajustamento dos pesos dos critérios tendo em conta os valores de entrada dos respetivos critérios para cada artigo. Deste modo, as redes neuronais têm a capacidade de melhorar os resultados automaticamente com base na informação disponível nos dados apresentados. Os resultados obtidos utilizando os dois métodos de aprendizagem foram avaliados pelos gestores de inventário e posteriormente comparados entre si. Deste estudo podemos concluir que as redes neuronais utilizando o algoritmo de genética como método de aprendizagem apresentam melhor precisão na tarefa de classificar os artigos em estudo. Apesar dos bons resultados obtidos neste estudo, a utilização de redes neuronais apresenta algumas desvantagens, como por exemplo suportar um número limitado de critérios a ser utilizado. Caso exista algum critério qualitativo considerado importante para o negócio pode ser difícil de o incluir no modelo apresentado. Por outro lado, este método não consegue nem deve substituir na totalidade a opinião de um profissional. Tendo em conta a experiência e o conhecimento dos gestores de inventário, os critérios considerados importantes para uma boa classificação e utilizados neste estudo são o preço unitário dos respetivos artigos, o seu custo de aquisição, *lead time* e o custo total das vendas anuais.

Em 2005, Canetta*, L., Cheikhrouhou, N., e Glardon, R. propuseram a exploração de uma segmentação em duas fases. A primeira fase consiste na aplicação do método SOM (*Self-Organizing Map*) com o objetivo de reduzir potenciais *outliers* e reduzir a complexidade dos dados. Esta primeira fase foi comparada com outros métodos de segmentação, como é o caso dos algoritmos de segmentação hierárquica utilizando os métodos de agrupamento *Complete*, *Ward* e *Average*, e o método de segmentação por partição *K-means*. Uma segunda fase consiste em analisar o conjunto de dados obtidos pelo remapeamento do método SOM utilizando novamente os já mencionados métodos de segmentação. Comparando os resultados obtidos pelo método SOM com os restantes métodos de segmentação, este apresentou menor precisão nos resultados. No entanto ao utilizar o *K-means* numa segunda fase de análise (SOM e *K-means*) estes algoritmos juntos apresentaram melhores resultados que os iniciais métodos de segmentação. Neste estudo os autores utilizaram como variáveis de estudo os seguintes critérios de negócio: *lead time*, coeficiente de variação do *lead time*, quantidade comprada (mês), coeficiente de variação da quantidade comprada, custo unitário e frequência de utilização do artigo em produtos manufaturados.

Em 2006, Ramanathan, R. sugeriu a implementação de um modelo de classificação simples utilizando técnicas de otimização linear. Este método, posteriormente apelidado de *Ramanathan-model* ou *R-model*, consiste em maximizar a função objetivo ponderada, que permite calcular o desempenho de um determinado artigo através da agregação ponderada dos respetivos critérios sujeitos a algumas restrições. Este método permite calcular separadamente a função objetivo de cada artigo. Isto representa uma vantagem, pois uma vez classificados todos os artigos não é necessário refazer os cálculos caso se queira acrescentar um artigo novo (Ramanathan, R., 2006). No entanto, os pesos de cada função objetivo são determinados pelos respetivos valores de critérios. Se um artigo tiver um valor dominante num dos critérios apresentados, este vai sempre ser classificado como A independentemente dos valores dos restantes critérios. Ora, se o critério dominante for um critério de baixa importância para o negócio, este mesmo artigo sairá mal classificado (Zhou, P., & Fan, L., 2007). Em 2007, Zhou, P. e Fan, L. apresentaram uma versão estendida deste mesmo método. Os autores mantiveram a simplicidade do modelo e propuseram-se a acrescentar, para além

da função objetivo de maximização, uma função objetivo de minimização e suas respetivas restrições. Permitindo deste modo ter em consideração tanto os critérios favoráveis como os menos favoráveis, evitando assim a classificação errada devido a um critério dominante. Ambos os estudos utilizaram como critérios de classificação o preço unitário de cada produto, o custo total de vendas anuais e o *lead time*. Adicionalmente, Ramanathan incluiu o critério de criticalidade do produto no seu estudo.

Em 2007, Wan L, Ng apresentou um modelo simples que se mostrou igualmente capaz de dar resposta ao problema de classificação de artigos de inventário. Este método, posteriormente apelidado de *Ng-model*, consiste primeiramente em transformar os valores dos critérios de cada artigo para uma forma normalizada. Isto é, os valores de critérios estarão representados numa escala de 0 a 1 calculada com base no valor máximo e mínimo de cada critério. Após a normalização dos dados, é calculada a performance de cada artigo, são então ordenados de forma decrescente e finalmente é aplicado o princípio da classificação ABC. Este método tem a vantagem de ser de fácil compreensão, fácil de implementar e não necessita de programas muito avançados nem nenhum tipo de formação específica para executar os cálculos. No entanto, a simplicidade e a facilidade vem sempre com algumas desvantagens associadas. Este método tem a desvantagem de não permitir a utilização de variáveis categóricas. A escala de normalização requer sempre a utilização dos valores extremos de cada critério. Caso ocorra alguma alteração nos dados os extremos poderão sofrer alterações e terão de ser sempre confirmados, pois um valor de extremo inválido poderá levar a uma classificação errada. Para o estudo em questão, o autor utilizou os seguintes critérios: preço unitário, custo total das vendas, e *lead time*.

4. METODOLOGIA

4.1. DADOS

Os dados utilizados para efetuar este estudo foram retirados da base de dados *AdventureWorks* OLTP. A base de dados *AdventureWorks* OLTP é uma base de dados transacional disponibilizada pela Microsoft que contém informação sobre a operação de um fabricante de bicicletas fictício. Nesta base de dados estão incluídas informações sobre produção, vendas, compras e armazenamento de produtos. Adicionalmente, está também disponível informação relativa aos empregados de loja, localização e clientes. Apesar de não conter dados reais, esta base de dados tem vindo a ser melhorada ao longo dos anos com o objetivo de se aproximar o máximo possível de um negócio real. Assim sendo, esta pode ser usada neste estudo pois o que estamos a avaliar é a capacidade do método DBSCAN classificar os artigos de inventário e não o resultado da análise em si.

Como descrito na revisão de literatura, existe um conjunto de critérios utilizados que é comum aos vários estudos apresentados, dos quais serão utilizados neste estudo os seguintes: Custo médio de aquisição, que representa o custo médio que a empresa tem por unidade armazenada; preço médio unitário, que representa o preço médio a que é vendido cada artigo; custo total de vendas, que representa o total de unidades vendidas de cada produto num determinado período de tempo multiplicado pelo custo médio de aquisição. O período de tempo considerado varia de estudo para estudo, podendo-se optar por períodos de tempo diário, semanal, mensal, semestral ou anual, sendo que o mais utilizado é o anual (Van Kampen, T. J., Akkerman, R., & Pieter van Donk, D., 2012). Por este motivo, o período de tempo considerado neste estudo será o anual. O *lead time* é também um critério habitualmente utilizado mas não será incluído neste estudo devido á natureza do negócio. Existem artigos disponíveis para venda que passam por um processo de produção na empresa, como é o caso das bicicletas e alguns dos seus componentes, e outros que são comprados diretamente para revenda, como é o caso dos artigos de vestuário e acessórios desportivos. Deste modo, o significado do tempo de espera até o artigo estar disponível para venda (*lead time*) não é o mesmo, não fazendo sentido utilizar esta informação no estudo.

A base de dados mencionada contêm 121.137 registos de vendas transacionadas durante o período de tempo entre 31-05-2011 e 30-06-2014. Os registos foram filtrados por ano para construir o conjunto de dados para este estudo, sendo que ao analisar os dados identificou-se que o ciclo de comercialização dos produtos inicia aproximadamente a 30 ou 31 de Maio de cada ano e termina a 29 ou 30 de Maio do ano seguinte. Por esta razão optou-se por separar os dados em 3 anos distintos. O primeiro ano contém a informação relativa a 60 produtos comercializados entre 31-05-2011 e 29-05-2012. O segundo ano contém informação relativa a 107 produtos comercializados entre 30-05-2012 a 29-05-2013 e o terceiro ano contém informação relativa a 208 produtos comercializados entre 30-05-2013 a 29-05-2014. Foi filtrada e excluída a informação relativa ao mês de Julho de 2014 para igualar o período de vendas em estudo nos 3 data sets apresentados. De seguida, e com a ajuda de uma folha de Excel, foram calculados os três critérios relevantes para o estudo. O preço médio unitário foi calculado através da média do preço de venda de cada artigo, no respetivo ano, usando os valores disponibilizados nos registos de vendas transacionadas. O custo médio de aquisição foi retirado diretamente da coluna *StandardCost* da tabela *Product* disponibilizada na base de dados. Para calcular o custo total de vendas por artigo, foi calculado a quantidade vendida de cada artigo em cada ano e multiplicado pelos respetivos custos de aquisição. A informação relativa a quantidade

vendida resulta da soma das quantidades vendidas em cada transação. Por fim, os dados foram normalizados utilizando a conhecida fórmula Max-Min para que desta forma não haja discrepâncias nas escalas de valores entre as variáveis que possam influenciar os resultados do algoritmo de segmentação.

4.2. DBSCAN – CONCEITOS

Como mencionado anteriormente, o DBSCAN é um método que constrói *clusters* baseando-se nas regiões de pontos densamente populadas. Isto é, dado um conjunto de pontos, digamos D, num determinado espaço S k-dimensional, o DBSCAN agrupa os pontos que estão densamente próximos num *cluster* e identifica como pontos de ruído os que se encontram em regiões de baixa densidade. Para melhor explicar e compreender o conceito de densidade, pontos densamente próximos e de como são efetivamente formados os *clusters* serão apresentados de seguida alguns conceitos e definições que estão na base da construção deste método.

A densidade associada a um ponto é determinada através da contagem de pontos que se encontram dentro de um raio de distância Eps, designado por vizinhança-Eps do ponto. O conceito de vizinhança é determinado através de uma função de distância entre dois pontos, digamos p e q, de onde resulta a seguinte definição:

Definição 1 (Vizinhança-Eps): A vizinhança-Eps é definida por $N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\}$.

Isto é, para qualquer ponto q pertencente ao conjunto de dados D, q faz parte da vizinhança do ponto p se a distância entre os pontos p e q for menor ou igual ao valor do parâmetro Eps. A partir da noção de vizinhança-Eps de um ponto podemos definir que um ponto p é considerado um **ponto central** se a sua vizinhança-Eps contem um dado número mínimo de pontos. Ou seja, $|N_{Eps}(p)| \geq$ MinPts, designando por MinPts o valor mínimo de pontos.

Definição 2 (ponto diretamente alcançável): Um ponto p é diretamente alcançável através da densidade de um ponto q, com respeito aos parâmetros Eps e MinPts se:

1) $p \in N_{Eps}(q)$
2) $|N_{Eps}(q)| \geq$ MinPts – condição de ponto central

Definição 3 (ponto densamente alcançável): Um ponto p é densamente alcançável a partir de um ponto q se existir uma cadeia de pontos p_1, p_2, …, p_n onde $p_1=q$ e $p_n=p$ de tal modo que p_{i+1} é diretamente alcançável através da densidade do ponto p_i.

Posto isto, podemos definir como sendo um **ponto de fronteira** um ponto que não é um ponto central mas que é um ponto densamente alcançável a partir de um ponto central qualquer. Ou seja, são pontos que não cumprem a condição $|N_{Eps}(q)| \geq$ MinPts mas que ainda assim pertencem à vizinhança de um dado ponto central p pertencente a um *cluster* C. É possível que dois pontos de fronteira que pertencem ao mesmo *cluster* C não sejam densamente alcançáveis a partir um do outro pois a condição de ponto central pode não ser válida. No entanto, deverá existir um ponto central qualquer pertencente a C tal que ambos os pontos de fronteira sejam densamente

alcançáveis a partir desse ponto central. A isto chamamos de pontos que estão conectados por densidade.

Definição 4 (pontos conectados por densidade): Um ponto p é conectado por densidade a um ponto q se existir um ponto central r de tal modo que, p e q são ambos pontos densamente alcançáveis a partir de r, com respeito aos parâmetros Eps e MinPts.

Posto isto, estamos em condições de definir em que consiste efetivamente um *cluster* construído pelo método DBSCAN.

Definição 5 (Cluster): Seja D um conjunto de pontos, um *cluster* C é um subconjunto não vazio de D que respeitando os parâmetros Eps e MinPts satisfazem as seguintes condições:

1) Para quaisquer pontos p e q, se q pertence a C e p é densamente alcançável a partir do ponto q, então o ponto p também pertence ao *cluster* C (condição máxima).

2) Para quaisquer pontos p e q pertencentes a C, p e q são pontos conectados por densidade (condição de conectividade).

Por último e não menos importante, um ponto é considerado ruído ou *outlier* se não satisfazer nenhuma das condições apresentadas anteriormente. Ou seja, **pontos ruído** são pontos que não contêm na sua vizinhança-Eps um número mínimo de pontos (MinPts), não são densamente alcançáveis por nenhum ponto central e consequentemente não serão incluídos em nenhum cluster.

4.3. DBSCAN - ALGORITMO

O algoritmo DBSCAN inicia com um ponto p aleatório pertencente ao conjunto de pontos D e devolve todos os pontos que pertencem à vizinhança-Eps do mesmo. Se o número total de pontos vizinhos for maior ou igual ao valor mínimo MinPts, é formado um novo *cluster*. O ponto p e respetivos pontos vizinhos são associados a este novo *cluster* e a partir daí são então calculados recursivamente todos os pontos densamente conectados a partir de qualquer ponto central pertencentes a este mesmo *cluster*. Caso o número mínimo de pontos vizinhos não seja cumprido, este ponto é marcado como ruído e o algoritmo visita o próximo ponto ainda não visitado. Estes pontos inicialmente classificados como ruído podem ser (ou não) reclassificados mais tarde para pontos de fronteira caso venham a pertencer à vizinhança de algum ponto central. O algoritmo acaba quando não houver mais nenhum ponto por visitar e processar.

Este algoritmo encontra-se disponível em várias linguagens de programação. A linguagem de programação escolhida para a implementação deste algoritmo foi o Python, onde o código se encontra disponível no pacote *sklearn.cluster*. Este contém o DBSCAN com os seguintes métodos: labels_ que nos devolve o número do *cluster* formado para cada observação e atribui o valor -1 a pontos ruído; core_sample_indices_ que nos indica se uma observação é um ponto central (devolve 1) ou não (devolve 0); DBSCAN (eps= x, min_samples= x).fit(Dados) é a função que efetivamente calcula os diferentes *clusters*. Neste estudo esta função será corrida 3 vezes, uma para cada conjunto de dados. Os parâmetros *eps* e *min_samples* (mencionado anteriormente como MinPts) foram definidos com a ajuda de representação gráfica dos dados com os respetivos resultados. Começou-se

por definir eps = 0,5, de seguida diminui-se o valor de eps variando MinPts entre 2 e 6 para que deste modo a distância entre os pontos vizinhos seja o mais reduzido possível e o máximo de pontos mínimos incluídos na vizinhança seja o maior possível. Encontrou-se os valores ótimos para cada conjunto de dados quando os resultados dos *clusters* apresentaram 3 *clusters* distintos. Deste modo, foi-nos possível enquadrar os *clusters* produzidos pelo DBSCAN na classificação ABC. Os valores dos parâmetros utilizados em cada um dos conjuntos de dados foram os seguintes:

- Conjunto de dados 1: eps = 0,2 e MinPts = 3;
- Conjunto de dados 2: eps = 0,15 e MinPts = 3;
- Conjunto de dados 3: eps = 0,3 e MinPts = 3.

5. RESULTADOS

5.1. ANÁLISE DESCRITIVA

Nesta secção irá ser feita uma análise descritiva às variáveis em estudo face aos três anos de atividade da empresa e uma análise aos produtos por categorias.

Tabela 1: Quantidade de produtos por categoria e por ano

Categorias	1º Ano	2º Ano	3º Ano
Acessórios	3	5	23
Bicicletas	30	35	68
Roupa	6	19	28
Componentes	21	48	89
Total	60	107	208

A tabela 1 mostra-nos a evolução da oferta de produtos da empresa ao longo dos três anos. Entende-se por 1º Ano o período de tempo de 31-05-2011 a 29-05-2012, o 2º Ano o período de tempo de 30-05-2012 a 29-05-2013 e o 3º Ano o período de tempo de 30-05-2013 a 29-05-2014, tal como mencionado anteriormente na metodologia. A tabela mostra que a empresa começou o primeiro ano com uma oferta de 60 artigos, dos quais 3 pertencem à categoria de acessórios, 30 pertencem à categoria de bicicletas, 6 pertencem à categoria de roupa e 21 são componentes de bicicleta. No segundo ano a oferta de artigos teve um aumento de 78,33% para 107 artigos. A quantidade de bicicletas vendidas aumentou para 35 modelos diferentes em que apenas 12 tiveram continuidade do ano anterior, e os restantes são modelos novos. Com uma maior diversidade de modelos de bicicletas, vem uma maior diversidade de componentes associado às mesmas, tendo estes aumentado para 48. Aos 3 artigos da categoria acessórios foram adicionados dois novos artigos. Na categoria de roupa, apenas os 4 mais vendidos no primeiro ano continuaram à venda no segundo ano juntamente com 15 outros novos artigos. No terceiro ano a oferta de artigos aumentou em 48,56% para um total de 208 artigos comercializados. Das 68 bicicletas comercializadas no terceiro ano, 24 tiveram continuidade do segundo ano entre as quais continuam incluídas 7 modelos do primeiro ano. Os componentes acompanharam novamente o crescimento de modelos de bicicleta, passando a ser comercializados 89 artigos no terceiro ano. Os acessórios mantiveram os modelos de artigos vendidos no ano anterior no entanto houve uma expansão para outros 18 novos modelos de artigos. A categoria de roupa manteve a maioria das peças comercializadas no segundo ano, investindo apenas em 12 novos artigos.

Tabela 2: Categoria Bicicletas

Variáveis	1º Ano			2º Ano			3º Ano		
	Min	Máx	Média	Min	Máx	Média	Min	Máx	Média
C. Aquisição	486.7	2,171.3	1,212.0	486.7	1,554.9	905.2	294.6	1,554.9	789.8
P. Venda	515.2	3,401.9	1,644.4	583.8	2,054.0	1,144.3	425.3	2,137.3	1,089.9
C. Total Vendas*	85.7	1,441.7	659.3	215.1	1,707.7	788.4	0.5	2,016.9	578.2

* valores em milhares

Podemos observar na tabela 2 que o custo de aquisição e o preço de venda mínimo mantiveram-se praticamente iguais do primeiro para o segundo ano. Isto deve-se aos 12 produtos que se mantiveram à venda de um ano para o outro, mais precisamente as bicicletas de estrada do modelo 650 nas cores vermelho e preto em qualquer tamanho. No terceiro ano os valores diminuíram porque foram introduzidos modelos de bicicletas mais económicos, como por exemplo as bicicletas de montanha modelo 500 em preto e em prateado. Quanto aos valores máximos podemos observar que houve uma descida acentuada no custo e no preço dos produtos mais caros do primeiro para o segundo ano devido à descontinuidade de alguns dos modelos de bicicletas comercializados no primeiro ano, como por exemplo as bicicletas de estrada vermelhas do modelo 150. Do segundo para o terceiro ano o custo de aquisição do produto mais vendido manteve-se pois corresponde a bicicletas de estrada do modelo 250 em preto e em vermelho, produto que aumentou o seu preço médio de venda para 2,137.25 u.m.. No primeiro ano, o produto com menor procura foi a bicicleta de estrada modelo 650 vermelha, tamanho de quadro 58. Apesar de ser o produto com menores vendas este não corresponde ao produto com o custo de aquisição e preço de venda mais baixo, isto porque as bicicletas do mesmo modelo com outras cores e tamanhos de quadro foram vendidas com o mesmo custo de aquisição e preço de venda menor. Verifica-se o mesmo efeito neste produto no segundo ano onde volta a ser o artigo com menor procura. Já no terceiro ano o valor mínimo de procura deve-se a um modelo que vendeu uma única bicicleta, mais precisamente a bicicleta de estrada modelo 650 preta com quadro 48. No primeiro ano, o produto com maior procura foi a bicicleta de estrada modelo 150 em vermelho, que corresponde ao artigo de maior custo de aquisição. No segundo ano o artigo mais vendido foi um dos novos modelos (bicicleta de montanha modelo 200 em preto), mas que neste caso não é o artigo mais caro do ano. Este mesmo produto continuou a ser comercializado no terceiro ano, no qual foi aumentado o seu preço médio de venda mantendo-se na mesma como o produto mais vendido.

Tabela 3: Categoria Acessórios

Variáveis	1º Ano			2º Ano			3º Ano		
	Min	Máx	Média	Min	Máx	Média	Min	Máx	Média
C. Aquisição	13.1	13.1	13.1	8.2	13.1	11.6	0.9	59.5	11.9
P. Venda	20.2	20.2	20.2	12.0	20.2	17.5	2.2	159.0	29.5
C. Total Vendas*	7.4	8.7	8.0	9.3	22.8	16.8	0.0	141.1	22.4

* valores em milhares

Relativamente à categoria Acessórios podemos observar na tabela 3 que no primeiro ano os valores de mínimo, máximo e a média do custo de aquisição e do preço de venda são os mesmos, pois esta categoria apenas teve 3 artigos à venda no primeiro ano, sendo eles os capacetes desportivos em preto, azul e vermelho. A variação do mínimo, máximo e média do custo total de vendas diferencia devido à quantidade vendida de cada um. A diferença entre o mais vendido e o menos vendido é de 1.270,37 u.m. que corresponde sensivelmente a 97 unidades a mais vendidas. No segundo ano, o custo de aquisição e o preço de venda máximo mantiveram-se iguais aos do ano anterior pois referem-se ao mesmo produto. Este mesmo produto quase que triplicou a procura num espaço de um ano. Já o valor mínimo desceu devido à entrada dos dois novos produtos nesta categoria (cadeados de cabo e mini bombas de ar) puxando consequentemente a média do custo de aquisição e do preço de venda ligeiramente para baixo. No terceiro a empresa apostou não só em novos artigos de baixo preço, como também em artigos de valores mais elevados face aos anos anteriores, como é o exemplo das camaras de ar e dos descansos de bicicletas respetivamente.

Tabela 4: Categoria Roupa

Variáveis	1º Ano			2º Ano			3º Ano		
	Min	Máx	Média	Min	Máx	Média	Min	Máx	Média
C. Aquisição	3.4	38.5	21.5	6.9	38.5	25.0	3.4	41.6	25.7
P. Venda	5.6	31.2	18.6	5.6	58.5	34.6	7.2	62.8	39.6
C. Total Vendas*	0.3	38.1	14.0	2.7	90.3	32.7	0.0	159.8	39.1

* valores em milhares

Podemos observar na tabela 4 que no primeiro ano alguns produtos da categoria roupa tiveram um preço de venda mais baixo que o custo de aquisição, mais precisamente 4 artigos dos 6 comercializados nesta categoria. As camisolas de manga comprida e os bonés da marca Logo foram vendidos abaixo do preço de custo levando esta categoria a ter um lucro negativo ou prejuízo. Curiosamente estes produtos continuaram a ser comercializados no ano seguinte. O valor mínimo do custo de aquisição do segundo ano aumentou face ao ano anterior mas continua mais elevado que o valor mínimo do preço de venda, pois corresponde ao valor de um dos 4 artigos com lucro negativo do ano anterior. No terceiro ano estes mesmo 4 produtos continuaram a ser comercializados, mas desta vez com preço de venda superiores ao seu custo de aquisição. Em termos de valores médios, o impacto do lucro negativo no primeiro ano acabou por não ser muito grande pois o custo de aquisição médio rondou os 21,53 u.m. e o preço de venda médio rondou os 18,62 u.m. No segundo ano esta diferença negativa não se verifica e no terceiro ano conseguiu-se aumentar ligeiramente a margem de lucro nos produtos da categoria Roupa face ao segundo ano. No primeiro ano os produtos com menor procura foram precisamente os dois produtos que não tiveram lucro negativo (meias para bicicleta de montanha), tendo estes sido retirados de venda no ano seguinte. O produto com mais procura acabou por ser um dos artigos de lucro negativo, nomeadamente o boné da marca Logo, apesar do preço de venda máximo ter aumentado do primeiro para o segundo ano. O valor mínimo do custo total de vendas no terceiro ano corresponde a uns calções de homem tamanho M que estavam a ser comercializados no ano anterior e que provavelmente devido à necessidade de escoar *stock* continuou a ser vendido no terceiro ano.

Tabela 5: Categoria Componentes

Variáveis	1º Ano			2º Ano			3º Ano		
	Min	Máx	Média	Min	Máx	Média	Min	Máx	Média
C. Aquisição	187.2	868.6	477.2	15.2	868.6	283.2	9.0	868.6	242.0
P. Venda	220.2	1,009.7	581.4	25.3	963.0	349.8	15.0	1,059.3	314.2
C. Total Vendas*	0.2	135.2	55.6	1.4	388.5	93.6	0.1	377.3	63.5

* valores em milhares

Como podemos observar na tabela 5 o valor de mínimo do custo de Aquisição e do Preço de Venda veio sempre a diminuir de ano para ano, isto deve-se à venda de novos tipos de componentes sempre de menor valor. No primeiro ano os componentes disponíveis para venda eram os quadros das bicicletas de vários tamanhos modelos e cores. Já no segundo ano passaram a ser vendidas rodas e auscultadores e no terceiro ano pedais, travões e correntes. O valor de procura mínimo tanto no primeiro como no terceiro ano é muito reduzido pois estes valores correspondem a componentes dos quais só foram vendidas uma ou duas unidades. Já o valor mínimo de procura no segundo ano deve-se à venda de apenas 46 unidades dos auscultadores que por sua vez também representam os valores mínimos de Custo de Aquisição e de Preço de Venda. Apesar destes valores mínimos nos três

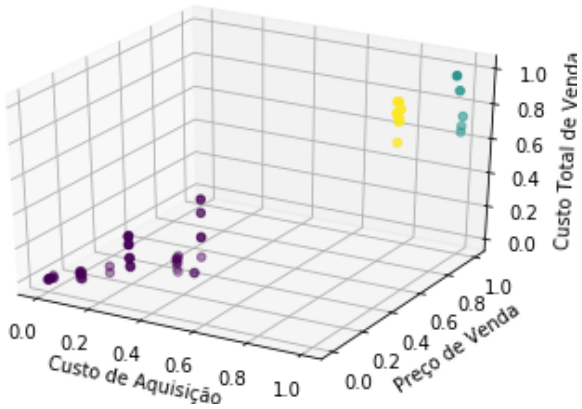
anos, estes não têm grande impacto nos valores médios pois são artigos com poucas unidades vendidas. Os valores máximos no Custo de Aquisição correspondem exatamente aos mesmos artigos nos três anos, mais precisamente quadros de bicicletas de estrada de cor vermelha e tamanhos diferentes. Em nenhum dos casos estes correspondem ao produto com maior procura. Já o preço de venda mais elevado no primeiro ano corresponde a um quadro de bicicleta de montanha prateado, com valores de procura muito próximos dos artigos mais vendidos. No segundo ano este produto deixou de ser vendido, o que fez com que o preço de venda máximo diminuisse dando lugar aos quadros de bicicletas de estrada de cor vermelha que permaneceram nos produtos mais caros no terceiro ano.

5.2. ANÁLISE DE *CLUSTERS*

Nesta secção irão ser apresentados os *clusters* formados pelo algoritmo DBSCAN para cada um dos conjuntos de dados em estudo e as suas respetivas análises.

5.2.1. Ano 1

Figura 1: Clusters Ano 1



Como podemos observar na figura 1, para o primeiro ano, o algoritmo DBSCAN identificou 3 grupos distintos de produtos. O grupo representado a roxo corresponde ao *cluster* número 1 e contém 47 produtos, o grupo representado a amarelo corresponde ao *cluster* número 3 e contém 8 produtos, por fim o grupo representado a azul claro corresponde ao *cluster* número 2 e contem 5 produtos. É também possível observar que nenhum dos produtos apresentados foi marcado como ponto ruído.

Tabela 6: Resultados dos produtos comercializados no Ano 1

	Cluster 1			Cluster 2			Cluster 3		
Variáveis	Min	Max	Média	Min	Max	Média	Min	Max	Média
C. Aquisição	3.4	884.7	435.2	2171.3	2171.3	2171.3	1898.1	1912.2	1905.1
P. Venda	5.6	1049.8	509.6	3198.6	3401.9	3326.5	2616.0	2677.1	2641.3
C. Total Vendas*	0.2	628.1	124.8	0.9	1441.7	1151.7	965.6	1294.5	1178.8

* valores em milhares

Com o auxílio da tabela 6 podemos verificar que os produtos pertencentes ao *cluster* número 1 têm um custo médio de aquisição de 435,17 u.m., são vendidos a um preço médio de 509,56 u.m., o que corresponde a uma margem de lucro média de 74,39 u.m., e têm um custo total de vendas anuais médio de 124.828,13 u.m.. Os produtos pertencentes ao *cluster* número 2 têm um custo médio de aquisição de 2.1771,29 u.m., são vendidos a um preço médio de 3 326,54 u.m., o que corresponde a uma margem de lucro média de 1.155,25 u.m., e têm um custo total de vendas anuais médio de 1.151.654,44 u.m.. O *cluster* número 3 apresenta produtos com um custo médio de aquisição de 1.905,12 u.m., vendidos a um preço médio de 2 641,32 u.m., o que corresponde a uma margem de lucro média de 736,19 u.m., e apresentam um custo total de vendas anuais médio a rondar os 1.178.797,48 u.m..

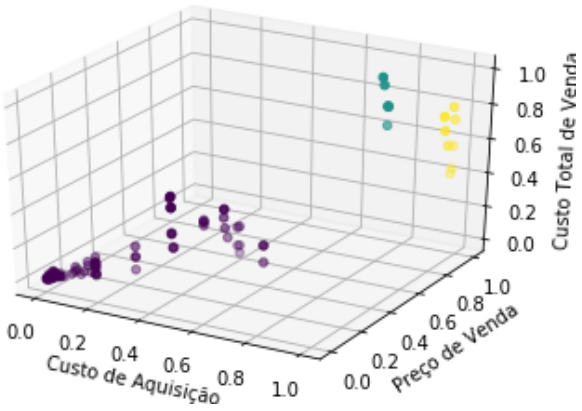
Tendo em conta as características apresentadas e contextualizando os resultados no problema inicial da classificação ABC podemos concluir que o *cluster* número 1 representa os produtos de categoria C pois é o grupo de produtos que apresenta menores valores em todos os critérios. Os produtos deste *cluster* são produtos de baixo custo de aquisição, baixo preço de venda, o que por sua vez se traduz numa margem de lucro baixa quando comparada com os restantes grupos, e o custo total de vendas, ou seja a procura, é bastante menor. O *cluster* número 2 representa os produtos de categoria A, não só por apresentar os valores médios dos critérios de preço de venda e custo total de vendas mais elevados como também por apresentar uma margem de lucro unitária significativamente maior que a margem de lucro unitária obtida nos produtos pertencentes ao *cluster* número 3. Isto acontece porque os artigos do *cluster* 2 têm um custo de aquisição menor que o grupo de artigos do *cluster 3*. Por este mesmo motivo podemos afirmar que o *cluster* número 3 representa os artigos de categoria B, uma vez que numa perspetiva de negócio a margem de lucro é um indicador muito importante para a sua rentabilidade.

Definidas as classes que cada *cluster* representa, estamos em condições de analisar efetivamente que produtos são esses. Os cinco produtos classificados como A são todos bicicletas de estrada vermelhas do modelo 150. Estes, como vimos anteriormente foram os produtos com maior procura e custo de aquisição e preço de venda mais elevado. Isto significa que no primeiro ano de vendas estes foram os produtos considerados mais importantes, o que significa que a empresa tem de os ter em maior quantidade em *stock* e que caso haja uma rotura de stock esta terá um grande impacto na rentabilidade da empresa. Os produtos classificados como C, representam os produtos menos importantes, ou seja, pode haver menor quantidade armazenada e a rotura de *stock* não é tão grave como nos produtos classificados como A. Desta classe fazem parte todos os produtos de roupa, acessórios e componentes e ainda alguns modelos de bicicletas, nomeadamente as bicicletas de estrada modelo 650 em vermelho e preto e as de modelo 450 em vermelho. Já os oito produtos classificados como B são bicicletas de montanha pretas e prateadas do modelo 100. Estas bicicletas

apresentam preço de vendas e procura ligeiramente mais baixos do que o modelo de bicicletas classificado como A e custo de aquisição mais elevado.

5.2.2. Ano 2

Figura 2: Clusters Ano 2



Como podemos observar na figura 2, para o segundo ano, o algoritmo DBSCAN distinguiu 3 grupos distintos de produtos. O grupo representado a roxo corresponde ao *cluster* número 1 e contem 93 produtos. O grupo representado a azul claro corresponde ao *cluster* número 2 e contem 6 produtos. O grupo representado a amarelo corresponde ao *cluster* número 3 e contem 8 produtos. Como também podemos observar, nenhum dos produtos apresentados foi marcado como ponto ruído.

Tabela 7: Resultados dos produtos comercializados no Ano 2

	Cluster 1			Cluster 2			Cluster 3		
Variáveis	Min	Max	Média	Min	Max	Média	Min	Max	Média
C. Aquisição	6.9	868.6	278.8	1252.0	1265.6	1258.8	1518.8	1555.0	1541.4
P. Venda	5.6	963.0	342.7	1662.5	1746.1	1707.0	1798.5	2054.0	1934.9
C. Total Vendas*	1.4	74.4	164.1	1210.7	1707.7	1463.0	757.3	1363.7	1094.4

* valores em milhares

Com o auxílio da tabela 7, conseguimos identificar que os produtos pertencentes ao *cluster* número 1 têm um custo médio de aquisição de 278,75 u.m. e são vendidos a um preço médio de 342,67 u.m., o que corresponde a uma margem de lucro média de 63.91 u.m.. Estes artigos apresentam um custo total de vendas anuais a rondar em média os 164.080,06 u.m.. O *cluster* número 2 apresenta produtos com um custo médio de aquisição de 1.258,80 u.m., vendidos a um preço médio de 1.707,02 u.m., o que corresponde a uma margem de lucro média de 448,21 u.m. e com um custo total de vendas anuais de 1.463.036,10 u.m. em média. Os produtos pertencentes ao *cluster* número 3 têm um custo médio de aquisição de 1.541,39 u.m., são vendidos a um preço

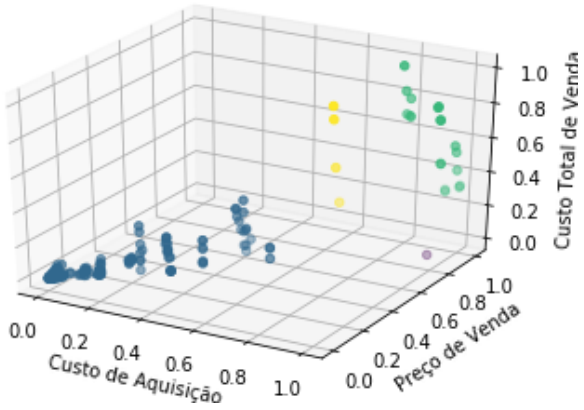
médio de 1.934,91 u.m., o que corresponde a uma margem de lucro média de 393,52 u.m. e e têm o custo total de vendas anuais que ronda em média os 1.094.394, 05 u.m..

Tendo em conta as características apresentadas, podemos afirmar que o *cluster* número 1 representa produtos de categoria C pois é o grupo de produtos com valores menores para os três critérios. O *cluster* número 2 representa produtos de categoria A pois apesar do custo médio de aquisição e o preço médio de venda ser inferior aos dos produtos do *cluster* número 3, estes apresentam uma margem de lucro unitária e um custo total de vendas significativamente superior aos dos produtos do *cluster* número 3. Deste modo, o *cluster* número 3 representa produtos de categoria B.

Definidas as classes que cada *cluster* representa, vamos então analisar que produtos pertencem a cada uma. Os seis produtos classificados como sendo os mais importantes, ou seja classificados como A, são bicicletas de montanha do modelo 200 de cor preta e prateada. Estes foram os produtos que tiveram maior procura na categoria de bicicletas no segundo ano. Como classe B temos representadas oito bicicletas de estrada do modelo 250 nas cores preto e vermelho. Como vimos anteriormente, este modelo de bicicleta é o modelo que tem um maior custo de aquisição, maior preço de venda, mas ainda assim teve uma procura ligeiramente menor que os artigos de classe A. Por último, a classe de artigos C contém todos os produtos das categorias roupa, acessórios e componentes e ainda alguns modelos de bicicletas, como por exemplo as bicicletas de montanha pretas modelo 300, as bicicletas de estrada amarelas modelo 550-W e as bicicletas de estrada pretas e vermelhas no modelo 650. Dentro da categoria de artigos de bicicletas, este último modelo corresponde ao modelo de bicicleta com menor procura no segundo ano de vendas.

5.2.3. Ano 3

Figura 3: Clusters Ano 3



Como podemos observar na figura 3, para o terceiro ano, o algoritmo DBSCAN identificou 3 grupos de produtos distintos e um ponto ruído. O grupo representado a azul corresponde ao *cluster* número 1 e contém 184 produtos. O grupo representado a amarelo corresponde ao *cluster* número 3

e contem 4 produtos e o grupo representado a verde corresponde ao *cluster* número 2 com 19 produtos. O ponto representado a roxo corresponde ao ponto ruído.

Tabela 8: Resultados dos produtos comercializados no Ano 3

	Cluster 1			Cluster 2			Cluster 3		
Variáveis	Min	Max	Média	Min	Max	Média	Min	Max	Média
C. Aquisição	0.9	868.6	234.8	1252.0	1555.0	1430.7	1082.5	1082.5	1082.5
P. Venda	2.2	1059.3	311.5	1932.4	2137.3	2049.2	1419.9	1534.6	1463.6
C. Total Vendas*	0.0	761.2	98.5	588.3	2016.9	1236.3	580.2	1755.8	1239.2

* valores em milhares

Como podemos observar na tabela 8, o *cluster* número 1 apresenta produtos com um custo médio de aquisição de 234,81 u.m.. Estes produtos são vendidos com um preço médio de 311,53 u.m. e uma margem de lucro média de 76,73 u.m.. A procura destes artigos, reflectida no custo total de vendas anuais ronda em média os 98.509,41 u.m.. Os produtos pertencentes ao *cluster* número 2 têm um custo médio de aquisição de 1.430,69 u.m., são vendidos a um preço médio de 2.049,16 u.m. a que corresponde uma margem de lucro média de 618,48 u.m.. O custo total de vendas anuais destes atigos é de 1.236.258,57 u.m. em média. Por fim os produtos pertencentes ao *cluster* número 3 são adquiridos em média por 1.082,51 u.m., vendidos a um preço médio de 1.463,60 u.m., o que gera uma margem de lucro média de 381,09 u.m.. O custo total de vendas anuais ronda em média os 1 239 203,32 u.m..

Perante estas características podemos afirmar que os produtos do *cluster* número 1 representam produtos de categoria C, pois é o grupo de produtos com valores de critérios mais baixos. Apesar do *cluster* número 2 apresentar uma média de custo total de vendas ligeiramente mais baixa que a do *cluster* número 3, a diferença entre os valores de custo de aquisição e os preços de venda é superior, o que leva a uma margem maior de lucro neste *cluster*. Olhando para os valores numa perspetiva de negócio a margem de lucro de um produto é um indicador muito importante e por esse mesmo motivo os produtos do *cluster* número 2 são de categoria A e os do *cluster* número 3 são de categoria B.

Posto isto, podemos afirmar que os artigos mais importantes para a empresa neste terceiro ano de vendas são os produtos de categoria bicicletas, nomeadamente as bicicletas de estrada vermelhas e pretas modelo 250, as bicicletas de montanha pretas e prateadas do modelo 200 e as bicicletas de passeio azuis e amarelas do modelo 1000. Como identificado anteriormente na análise descritiva desta categoria, as bicicletas de estrada modelo 250 foram consideradas as mais caras para o terceiro ano e as bicicletas de montanha modelo 200 foram consideradas as mais vendidas. As bicicletas de passeio do modelo 1000 são um modelo novo comercializado apenas no terceiro ano e encontram-se com valores de custo, preço e procura idênticos aos dos modelos de bicicletas já existentes. A classe B apenas contém quatro artigos, sendo estes bicicletas do modelo 350 de cor amarela. Este é também um modelo novo comercializado no terceiro ano. Por fim, da classe C fazem parte todos os artigos pertencentes às categorias roupa, acessórios, componentes e ainda alguns modelos de bicicletas, nomeadamente as bicicletas de montanha prateada do modelo 400-W e do modelo 500, as bicicletas de estrada modelo 550-W em amarelo, modelo 650 em preto e vermelho e modelo 750 em preto e ainda as bicicletas de passeio modelo 2000 em azul e modelo 3000 em azul e em amarelo.

Por último, o ponto ruído identificado pelo algoritmo corresponde a um tamanho específico da bicicleta de estrada vermelha do modelo 250 que ao contrário das outras bicicletas deste modelo classificadas como A, apresenta apenas 4 unidades vendidas.

5.3. COMPARAÇÃO ANUAL

Nesta secção iremos fazer uma análise da evolução da classificação A, B e C ao longo dos três anos obtida pelo algoritmo DBSCAN. Esta análise é importante para perceber de que forma se comportam os produtos comercializados de ano para ano. Ou seja, se um artigo classificado como A se mantém A em todos os anos ou se perde importância, ou se existem artigos que passaram de C para B ou de B para A, neste caso ganhando importância. Com base na análise individual feita às classificações A, B e C para cada ano, conseguimos identificar algumas situações deste tipo. Comparando os artigos comercializados no primeiro e segundo ano, podemos observar que os produtos classificados como A no primeiro ano deixaram de ser comercializados no ano seguinte. Sendo as cinco bicicletas de estrada, modelo 150 classificadas como os produtos mais importante comercializados no primeiro ano podemos questionar se estas não deveriam ter continuado a ser vendidas no ano seguinte. Outro ponto relevante que se nota na evolução de três anos refere-se aos artigos classificados como classe C. As categorias de produtos de roupa, acessórios e componentes estão sempre classificadas como C em qualquer ano. Isto deve-se não só ao facto destas categorias estarem a ser comparadas com a categoria bicicletas que tem preços de venda significativamente superiores, mas também porque não existem efeitos nas outras variáveis que permitam compensar esta diferença. Mais precisamente nas variáveis custo de aquisição e custo total de vendas. Podemos também observar que os produtos da categoria bicicletas estão distribuidos pelas 3 classes independentemente do ano. As bicicletas de estrada de modelo 650 continuaram a ser comercializadas no segundo e terceiro ano mantendo-se sempre classificadas como sendo produtos de classe C. O mesmo acontece com as bicicletas de estrada modelo 550 do segundo para o terceiro ano. Sendo estes produtos considerados os menos importantes dentro da categoria bicicletas podemos questionar se estes produtos não deveriam ter sido descontinuados nos anos seguintes. Outro ponto que podemos observar é a passagem de artigos de classe B para classe A, nomeadamente 5 artigos classificados como B no segundo ano que passam a ser classificados como A no terceiro ano. Podemos justificar esta subida de importância devido à enorme quantidade de produtos novos que passaram a ser comercializados no terceiro ano que não existiam no segundo. Isto porque, estes novos artigos apresentam valores de preço de venda e margem de lucro consideravelmente mais baixos que os artigos reclassificados como A. Por último, conseguimos observar que alguns produtos logo no seu primeiro ano de venda foram classificados como produtos muito importantes face aos restantes produtos disponíveis para venda.

5.4. COMPARAÇÃO COM A CLASSIFICAÇÃO ABC CLÁSSICA

Nesta secção iremos proceder a uma comparação da classificação obtida pelo algoritmo DBSCAN com a classificação obtida utilizando a classificação ABC clássica no sentido de perceber se os dois métodos conduzem a resultados semelhantes.

Como este estudo utiliza dados fictícios não é possível avaliar a qualidade das classificações feitas com base na opinião e experiência de gestores especializados na área de gestão de inventários, mais precisamente no negócio de montagem e venda de bicicletas aqui em estudo. Por esta razão comparar os nossos resultados com a classificação feita da forma tradicional é de facto a melhor alternativa para validar se o algoritmo aqui utilizado reúne as condições necessárias para fazer uma classificação melhor que a classificação ABC.

Aplicando a classificação ABC clássica para o conjunto de dados relativos ao primeiro ano de vendas obtivemos que a classe A contém 14 artigos nos quais estão incluídos apenas artigos da categoria bicicletas. A classe B contém 16 artigos dos quais fazem parte artigos de categoria bicicletas e categoria componentes. A classe C contém 30 artigos sendo esta classe constituída por todos os artigos das categorias de roupa e acessórios, os restantes artigos da categoria de componentes e ainda 3 bicicletas de estrada de modelo 650. Para o conjunto de dados relativo ao segundo ano de vendas obtivemos que a classe A contem 25 produtos todos da categoria bicicletas. A classe B contem 28 artigos dos quais 10 são os restantes artigos da categoria de bicicletas e 18 correspondem a artigos da categoria componentes. A classe C contem 54 artigos sendo estes todos os artigos das categorias de acessórios e de roupa, e os restantes artigos da categoria componentes. No conjunto de dados relativos ao terceiro ano de vendas a classe A contem 35 artigos sendo 23 de categoria de bicicletas, 4 de categoria acessórios, 5 artigos de categoria de roupa e 3 artigos da categoria de componentes. A classe B contem 53 artigos, pertencendo estes às categorias roupa, acessórios e componentes. A classe C contem 120 artigos sendo estes das quatro categorias.

Comparando os resultados dos dois métodos foram encontradas duas situações que parecem ser significativas. A primeira incide sobre os produtos classificados como B pelo algoritmo DBSCAN que estão classificados como A pela classificação ABC clássica, como por exemplo bicicleta de estrada 350-W em amarelo vendida no terceiro ano. Isto constitui um problema na medida em que a classificação ABC dá maior importância a artigos que segundo o algoritmo DBSCAN são considerados como B. Isto pode fazer com que a empresa mantenha um nível de *stock* mais elevado do que aquele que efetivamente necessita o que pode levar a custos desnecessários pela aquisição de quantidades que na realidade não são necessárias. A segunda situação encontrada incide sobre os produtos classificados como C pelo algoritmo DBSCAN e que estão classificados como A pela classificação ABC clássica, como por exemplo as bicicletas de estrada modelo 550-W em amarelo vendidas no segundo e terceiro ano. Nesta segunda situação o problema é o mesmo mas de forma mais agravada, ou seja, se a fronteira que separa um artigo A de um B pode ser difícil de definir exatamente, o mesmo já não acontece entre um artigo A e um artigo C. Portanto aqui estamos perante uma diferença significativa entre os resultados dos dois métodos que pode ter consequências para a rentabilidade da empresa. De acordo com os vários estudos feitos nesta área, o uso de mais critérios de avaliação conduz a melhores resultados desde que os critérios adicionais sejam relevantes para avaliar as diferentes características do negócio que determinam os níveis de inventário. Uma vez que seguimos a literatura na escolha dos critérios podemos afirmar que o algoritmo DBSCAN devolve melhores resultados do que a classificação ABC clássica, corrigindo até artigos mal classificados.

6. CONCLUSÕES

Os resultados mostram que o algoritmo DBSCAN foi capaz de agrupar artigos em grupos semelhantes comparáveis com uma classificação ABC para gestão de inventário. O algoritmo mostrou inclusive melhores resultados quando comparado com a abordagem clássica da classificação ABC. Em resposta à questão colocada na secção dos objetivos deste estudo podemos confirmar que os algoritmos de segmentação podem ser utilizados para resolver problemas de classificação de inventário com bons resultados. As empresas passam assim a ter à sua disposição uma nova técnica para as ajudar a gerir os seus inventários de forma mais eficiente.

Apesar dos dados serem fictícios, vale a pena referir que os resultados mostram que a atividade da empresa parece estar muito dependente de um pequeno grupo de produtos, identificados pelo algoritmo DBSCAN, visto que em qualquer um dos conjuntos de dados apresentados os *clusters* A têm um número de artigos bastante pequeno face ao número total. Esta dependência não é tão evidente utilizando a classificação ABC tradicional para classificar os artigos.

No que respeita a investigação futura, seria interessante explorar a classificação dos artigos por categoria, pois poderá fazer sentido num negócio real a identificação de artigos mais importantes ou menos importantes dentro da mesma categoria. Adicionalmente, aplicar esta metodologia num contexto real poderia ser interessante para perceber se o desempenho do algoritmo DBSCAN seria o mesmo.

7. BIBLIOGRAFIA

Chu, C. W., Liang, G. S., & Liao, C. T. (2008). Controlling inventory by combining ABC analysis and fuzzy classification. Computers & Industrial Engineering, 55(4), 841-851.

Partovi, F. Y., & Anandarajan, M. (2002). Classifying inventory using an artificial neural network approach. *Computers & Industrial Engineering*, *41*(4), 389-404.

Guvenir, H. A., & Erel, E. (1998). Multicriteria inventory classification using a genetic algorithm. European journal of operational research, 105(1), 29-37.

Flores, B. E., Olson, D. L., & Dorai, V. K. (1992). Management of multicriteria inventory classification. Mathematical and Computer modelling, 16(12), 71-82.

Partovi, F. Y., & Burton, J. (1993). Using the analytic hierarchy process for ABC analysis. International Journal of Operations & Production Management, 13(9), 29-44.

Canetta*, L., Cheikhrouhou, N., & Glardon, R. (2005). Applying two-stage SOM-based clustering approaches to industrial data analysis. Production Planning & Control, 16(8), 774-784.

Ramanathan, R. (2006). ABC inventory classification with multiple-criteria using weighted linear optimization. Computers & Operations Research, 33(3), 695-700.

Zhou, P., & Fan, L. (2007). A note on multi-criteria ABC inventory classification using weighted linear optimization. European journal of operational research, 182(3), 1488-1491.

Van Kampen, T. J., Akkerman, R., & Pieter van Donk, D. (2012). SKU classification: a literature review and conceptual framework. International Journal of Operations & Production Management, 32(7), 850-876.

Ng, W. L. (2007). A simple classifier for multiple criteria ABC analysis. European Journal of Operational Research, 177(1), 344-353.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), 165-193.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data & Knowledge Engineering, 60(1), 208-221.

8. ANEXOS

8.1. CÓDIGO *PYTHON* UTILIZADO

```
import numpy as np
import csv
import matplotlib.pyplot as plt
from sklearn.cluster import DBSCAN

# Import CSV file
path = 'C:/Users/cmc/Desktop/New folder/Dados/DataSet3_norm.csv'
file = open(path, newline= '')

reader = csv.reader(file)
header = next(reader) #The first line is the header
dataset3 = [row for row in reader] #Next line is the data

# Data preparation
np_dataset3 = np.array(dataset3)
np_dataset3 = np_dataset3.astype(np.double)

CustoAquisicao = np_dataset3[:,0]
PrecoVenda = np_dataset3[:,1]
CustoTotalVendas = np dataset3[:,2]

# Run DBSCAN algorithm
model = DBSCAN (eps=0.3, min_samples=3, algorithm='auto').fit(np_dataset3)
model.labels_ #displays the cluster number for each data entry

# Data Vizualization
fig = plt.figure()
ax = fig.add_subplot(111, projection="3d")

ax.scatter(CustoAquisicao, PrecoVenda, CustoTotalVendas, c=model.labels_,
marker='o')
ax.set_xlabel('Custo de Aquisição')
ax.set_ylabel('Preço de Venda')
ax.set_zlabel('Custo Total de Venda')
plt.show()
```