ISCTE IUL
Instituto Universitário de Lisboa

AI SE ELA "CAE"
Segmentação de empresas alternativa à Classificação Portuguesa de Atividades Económicas

Dina Isabel Ramos Dinis Fernandes

Dissertação submetida como requisito parcial para obtenção do grau de Mestre em Gestão de Empresas

Orientadora:
Prof.ª Doutora Margarida G. M. S. Cardoso, ISCTE Business School, Departamento de Métodos Quantitativos para Gestão e Economia

Coorientador:
Prof. Doutor Luís Miguel da Silva Laureano, ISCTE Business School, Departamento de Finanças

Outubro 2017

ISCTE IUL
Instituto Universitário de Lisboa

AI SE ELA "CAE"
Segmentação de empresas alternativa à Classificação Portuguesa de Atividades Económicas

Dina Isabel Ramos Dinis Fernandes

Dissertação submetida como requisito parcial para obtenção do grau de Mestre em Gestão de Empresas

Orientadora:
Prof.ª Doutora Margarida G. M. S. Cardoso, ISCTE Business School, Departamento de Métodos Quantitativos para Gestão e Economia

Coorientador:
Prof. Doutor Luís Miguel da Silva Laureano, ISCTE Business School, Departamento de Finanças

Outubro 2017

AGRADECIMENTOS

Um sincero agradecimento a todos aqueles que me permitiram superar com êxito esta etapa da minha vida, nomeadamente:

- À minha orientadora Professora Doutora Margarida Cardoso, pelo conhecimento transmitido e contributos realizados;

- Ao meu coorientador Professor Doutor Luís Laureano, pela sua disponibilidade e preciosa ajuda na extração dos dados;

- Ao professor Raul Laureano pela confiança que sempre depositou em mim;

- Ao professor Nuno Santos, pelas longas conversas sobre estatística;

- Ao Marco, Miguel e Sandra, meu grupo de trabalho no Mestrado Executivo em Análise de Dados aplicada à Gestão, pelas longas horas de trabalho, infinitas discussões e gargalhadas;

- Aos meus colegas do Risco, sempre em Risco, sempre a "pisar o Risco": Andreia, Bruno e Eduardo;

- À minha família;

- Aos meus amigos;

Por último, e mais importantes:

- Ao amor da minha vida, João;

- Ao meu amor maior, Vicente (filho lindo da mãe!).

"Organizing data into sensible groupings is one of the most fundamental modes of understanding and learning."

Anil K. Jain

RESUMO

Todos os dias o tecido empresarial português é avaliado, analisado e descrito nos mais diversos meios de comunicação social.

A maioria destas análises utiliza como agrupamento clássico das empresas a Classificação Portuguesa de Atividades Económicas (CAE), isto é, os estudos, notícias e documentos que são apresentados referem-se, salvo raras exceções, a um setor específico da economia: restauração, turismo, indústria farmacêutica…

Mas será que nos dias que correm, com a existência de um elevado número de empresas a exercer mais do que um tipo de atividade, é correto estudar as empresas assente nesta classificação? E, deste modo, agrupar e analisar empresas com a mesma CAE que apresentam características completamente diferentes? Basta pensarmos na comparação entre um hipermercado e uma mercearia!

Este trabalho pretende encontrar uma forma alternativa de segmentar as empresas, que não pela CAE.

A partir de uma amostra da base de dados *Amadeus,* que inclui vários indicadores de empresas portuguesas, foi construído um modelo de segmentação baseado em indicadores de rentabilidade relativos ao ano de 2014, que permitiu dividir as empresas em três segmentos distintos: empresas padrão, empresas em risco e empresas em destaque.

Palavras-chave: Classificação Portuguesa de Atividades Económicas, Análise de Clusters, Segmentação, Rentabilidade

Classificação JEL; C38; L00

ABSTRACT

The Portuguese business structure is evaluated, analyzed and described on a daily basis in a wide variety of media.

Most of these analyses use the Portuguese Classification of Economic Activities (CAE) as a classic grouping of companies, i. e., the studies, news and documents that are presented refer, with rare exceptions, to a specific sector of the economy: catering, tourism, pharmaceutical industry…

Nowadays with a large number of companies carrying out more than one type of activity, is it correct to study companies based on this classification? And, in the same way, to group and analyze companies with the same CAE representing entirely different characteristics? Suffice to think of the comparison between a hypermarket and a grocery!

This dissertation aims at finding an alternative way of segmenting companies, other than CAE.

Based on a sample of the Amadeus database, which includes several indicators of Portuguese companies, a segmentation model was built based on profitability indicators for the year 2014, which allowed companies to be divided into three distinct segments: standard companies, companies at risk and prominent companies.

Keywords: Portuguese Classification of Economic Activities, Cluster Analysis, Segmentation, Profitability

Classificação JEL: C38, L00

LISTA DE ABREVIATURAS E SIGLAS

AIC - *Akaike Information Criterion*

AT – Autoridade Tributária e Aduaneira

BIC - *Schwarz´s Bayesian Information Criterion*

CAE – Classificação Portuguesa de Atividades Económicas

CITA - Classificação Internacional Tipo de Todos os Ramos de Atividade Económica

CRISP-DM - *Cross Industry Standard Process for Data Mining*

DM – *Data Mining*

IBM - *International Business Machines Corporation*

KDD – *Knowledge Discovery in Databases*

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

1 - INTRODUÇÃO

Todos os dias é publicada uma grande diversidade de notícias acerca do tecido empresarial português. Realizada uma curta pesquisa na Internet podem ser encontradas as seguintes manchetes de jornais:

- ✓ *"Mediação imobiliária, alojamento local, construção e hotéis na mira do Fisco"* (Sapo, 2017*);*
- ✓ *"Em 2016, Fisco prioriza restaurantes, hotéis e imóveis"* (Eco, 2017*);*
- ✓ *"Fisco quer apertar o controlo ao comércio online e obras"* (Jornal Económico, 2017*);*
- ✓ *"Fisco inspeciona mais de seis mil cabeleireiros e instaura 400 autos de notícia"*(Dinheiro Vivo, 2017*);*
- ✓ *"Ação Menu. 500 inspetores do Fisco apertam restaurantes"* (Dinheiro Vivo, 2017*),*
- ✓ *"Turismo bate recorde de receitas e dormidas em 2016"* (Diário de Notícias, 2017*),*
- ✓ *"Receitas da restauração sobem 2,2% para 3,7 mil milhões de euros em 2016"* (Di*ário de Notícias, 2017),*
- ✓ *"Bolsa nacional começa semana a recuperar apoiada pela energia"* (Jornal de Negócios, 2017*).*

Estes títulos apresentam todos um ponto em comum: a classificação de empresas por setor de atividade. Para tal é utilizada a Classificação Portuguesa de Atividades Económicas (CAE). A CAE torna-se assim imprescindível. Sem esta classificação, seria impossível avaliar resultados, selecionar empresas para inspeção ou comparar desempenhos em cada setor.

Mas será que nos dias que correm, com a existência de um elevado número de empresas a exercer mais do que um tipo de atividade, é correto estudar as empresas divididas desta forma? Por outro lado, faz sentido agrupar e analisar empresas com a mesma CAE, que apresentam características completamente diferentes? Basta pensar que um hipermercado e uma mercearia têm a mesma CAE. Não se estará a perder informação com este tipo de classificação?

O objetivo do presente estudo, será encontrar uma forma alternativa de agrupar as empresas, não utilizando a CAE. Esta nova tipificação das empresas poderá permitir outro tipo de análises. Por exemplo, no caso da AT, os títulos das notícias deixam transparecer que as empresas alvo de análise são selecionadas por CAE, isto é, são escolhidas aquelas que apresentam maiores desvios relativamente ao grupo de empresas do seu setor de atividade. Se o setor de atividade destas empresas não for tido em consideração, estas empresas irão ser analisadas e selecionadas de forma diferente, pois podem apresentar resultados *standard* comparativamente a empresas com iguais valores de indicadores.

Neste trabalho, a partir de uma amostra da base de dados *Amadeus*[1], que inclui diversos indicadores de empresas portuguesas do ano de 2014, foi construído um modelo de segmentação baseado em indicadores de rentabilidade, que permitiu dividir as empresas em três segmentos distintos: empresas padrão, empresas em risco e empresas com bom desempenho.

[1] base de dados de informações financeiras comparáveis para empresas públicas e privadas localizadas na Europa da responsabilidade da empresa Bureau Van Dijk (https://amadeus.bvdinfo.com/version-2017623/home.serv?product=amadeusneo)

2 - CONTEXTUALIZAÇÃO

2.1. HISTÓRIA DA CLASSIFICAÇÃO PORTUGUESA DE ATIVIDADES ECONÓMICAS

A primeira publicação da Classificação Portuguesa de Atividades (CAE) data de 1953 e tem como origem a Classificação Internacional Tipo de Todos os Ramos de Atividade Económica (CITA), realizada em 1949 pelos Serviços de Estatística das Nações Unidas.

A CITA foi já alvo de algumas revisões, com o objetivo de melhor se adaptar à realidade económica mundial. Após a elaboração de diversos trabalhos estatísticos, Portugal concluiu que a simples tradução da CITA não se adaptava à realidade portuguesa. Assim, em 1964 surgiu a primeira CAE "portuguesa". Desde então, a sua atualização tem sido constante. Neste momento encontra-se vigente a CAE – Revisão 3, elaborada em 2007 e harmonizada com as últimas classificações das Nações Unidas e da União Europeia.

O organismo responsável pela publicação da CAE[2] é o Instituto Nacional de Estatística[3].

2.2. OBJETIVOS DA CLASSIFICAÇÃO PORTUGUESA DE ATIVIDADES ECONÓMICAS

A CAE visa identificar e classificar todas as atividades económicas existentes, a fim de serem enquadradas todas as empresas. Desta forma, depois de realizada a classificação, é possível agrupar ou classificar as empresas (bens e serviços com, e sem fins lucrativos) por atividade; descrever e comparar a nível nacional e mundial a estrutura empresarial; assim como organizar os dados estatísticos económico-sociais. Os objetivos da CAE são, na sua maioria, estatísticos.

2.3. COMO FUNCIONA?

Cada empresa existente em Portugal é classificada numa atividade económica se exercer exclusivamente uma atividade. No caso de exercer várias atividades em simultâneo terá

[2] http://www.sicae.pt/Consulta.aspx
[3] http://smi.ine.pt/Categoria

que identificar a sua atividade principal, sendo-lhe atribuída mais do que uma CAE: a CAE principal e a(s) CAE secundária(s).

É importante referir que, embora uma empresa exerça mais do que uma atividade, os seus resultados serão "imputados" à sua atividade principal. Por exemplo, no *site* http://www.sicae.pt/Consulta.aspx , a consulta por "*Bertrand*" mostra que a empresa com o Número de Identificação Fiscal 508289335, com a denominação BERTRAND & IRMÃO – COMPRA E RESTAURO DE IMÓVEIS LDA, tem como CAE principal o código 68100 - Compra e venda de bens imobiliários, e como Classificações secundárias, os códigos 41200 - Construção de edifícios (residenciais e não residenciais), 68311 - Atividades de mediação imobiliária e 70220 - Outras atividades de consultoria para os negócios e a gestão. Neste caso, os resultados obtidos nas CAE´s secundárias não irão ter reflexo, pois esta empresa será sempre analisada pela sua CAE principal - Compra e venda de bens imobiliários.

2.4. A ATUAL CLASSIFICAÇÃO PORTUGUESA DE ATIVIDADES ECONÓMICAS (CAE)

Na atual Classificação Portuguesa de Atividades Económicas (CAE) existem três setores básicos da atividade económica, nos quais as empresas desenvolvem a sua atividade:

- ✓ Setor primário: agricultura, silvicultura e pescas
- ✓ Setor secundário: indústria
- ✓ Setor terciário: comércio e serviços.

Em cada um dos setores existem, por sua vez, várias atividades específicas, as quais se encontram referenciadas na CAE.

Na prática, cada atividade é representada por um código de cinco dígitos, o qual identifica univocamente a atividade desenvolvida por uma empresa.

Desta forma, a CAE apresenta a seguinte estrutura:

- ✓ Secções (primeiro nível), definidas por um código alfabético;
- ✓ Divisões (segundo nível), definidas por um código de dois dígitos;

- ✓ Grupos (terceiro nível), definidas por um código de três dígitos;
- ✓ Classes (quarto nível), definidas por um código de quatro dígitos;
- ✓ Subclasses (quinto nível), definidas por um código de cinco dígitos.

A CAE define as seguintes secções fundamentais:

A - Agricultura, produção animal, caça e silvicultura;

B - Pesca;

C - Indústrias Extrativas;

D - Indústrias Transformadoras;

E - Produção e distribuição de eletricidade, gás e água;

F - Construção;

G - Comércio por grosso e a retalho, reparação de veículos automóveis, motociclos e de bens de uso pessoal e doméstico;

H - Alojamento e restauração (restaurantes e similares);

I - Transportes, armazenagem e comunicações;

J - Atividades financeiras;

K – Atividades imobiliárias, alugueres e serviços;

L - Administração pública, defesa e segurança;

M - Educação;

N - Saúde e ação social;

O - Outras atividades de serviços coletivos;

P - Famílias com empregados domésticos;

Q - Organismos internacionais e outras instituições extraterritoriais.

As divisões iniciam-se na divisão 01 – "Agricultura, produção animal, caça e atividade dos serviços relacionados" e acabam na divisão 99 – "Organismos internacionais e outras instituições extraterritoriais".

Os grupos, por sua vez, começam no grupo 011 – "Agricultura" e terminam no grupo 990 – "Organismos internacionais e outras instituições extraterritoriais".

Exemplo:

Figura 1 – Exemplo de uma Secção da CAE

Secção	Divisão	Grupo	Classe	Subclasse	Designação
L					Actividades imobiliárias.
	68				Actividades imobiliárias.
		681	6810	68100	Compra e venda de bens imobiliários.
		682	6820	68200	Arrendamento de bens imobiliários.
		683			Actividades imobiliárias por conta de outrem.
			6831		Mediação e avaliação imobiliária.
				68311	Actividades de mediação imobiliária.
				68312	Actividades de angariação imobiliária.
				68313	Actividades de avaliação imobiliária.
			6832		Administração de imóveis por conta de outrem; administração de condomínios.
				68321	Administração de imóveis por conta de outrem.
				68322	Administração de condomínios.

Fonte: https://pt.scribd.com/doc/26315287/Nova-Tabela-CAE

3 – REVISÃO DE LITERATURA

3.1 – DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

Kock *et al*. (1997) distinguem três conceitos básicos estreitamente relacionados: dados, informação e conhecimento.

Os dados são portadores de conhecimento e informação, isto é, tanto o conhecimento como a informação podem ser comunicados através de dados, que só se tornam informação ou conhecimento quando são interpretados pelo leitor.

A informação, ao contrário do conhecimento, é descritiva. Informar é descrever o que já aconteceu ou o que está a acontecer.

Já o conhecimento tem uma componente fortemente preditiva, pois com base em informação (passada e presente) fornece previsões daquilo que ainda irá acontecer no futuro.

A pesquisa em grandes volumes de dados, com o objetivo de extrair conhecimento continua a ser uma tarefa árdua. Frawley *et al*. (1992) designam esta tarefa como Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD).

Fayyad *et al*. (1996) afirma que o KDD é uma tentativa de solucionar um problema que a era digital trouxe a todos nós: excesso de dados.

O processo KDD é interativo e envolve várias etapas com muitas decisões:

1. Conhecimento e identificação do objetivo do negócio;
2. Selecionar um conjunto de dados a ser analisado;
3. Limpeza dos dados, tratamento de dados;
4. Redução e representação dos dados;
5. Associar o objetivo de negócio ao método de *Data Mining* a utilizar (Classificação, Regressão, Agrupamento);

6. Escolher o modelo a ser utilizado;
7. Pesquisa de padrões nos dados; e,
8. Interpretação dos resultados obtidos e repetição de todo o processo caso se considere pertinente.

3.2 – *DATA MINING*

Fayyad *et al*. (1996) define *Data Mining* como sendo uma etapa do KDD que consiste em descobrir padrões ou modelos através da análise dos dados.

Friedman (1998) por seu lado, considera que *Data Mining* é um conceito vagamente definido, pois a sua definição depende em grande parte da visão e do *background* do investigador.

Já Hand (1998), define *Data Mining* como sendo uma disciplina que envolve várias áreas: Estatística, Tecnologias de Bases de Dados, Reconhecimento de Padrões, *Machine Learning*. Reconhece também a necessidade imperiosa de ouvir os contributos dos Estatísticos nesta disciplina.

Carrier *et al*., (2003) diz-nos que o *Data Mining* assume a construção de um modelo a partir dos dados.

Ngai *et al*. (2009) enuncia várias formas de modelação de dados que podem ser utilizadas no *Data Mining*:

- ✓ Associação;
- ✓ Classificação;
- ✓ Agrupamento;
- ✓ Previsão;
- ✓ Regressão;
- ✓ Descoberta de sequências; e,
- ✓ Visualização.

As tarefas de classificação, previsão e regressão são exemplos de *Data Mining* supervisionado. As tarefas de associação e agrupamento são exemplos de *Data Mining* não-supervisionado.

Berry *et al.* (2004) descreve estas duas metodologias:

- ✓ no *Data Mining* supervisionado existe sempre uma variável alvo – algo a ser classificado, estimado ou previsto, por exemplo, um procedimento de classificação inicia-se com um conjunto de classes e exemplos de registos cuja classificação correta já é conhecida;
- ✓ no *Data Mining* não-supervisionado, não existe uma variável alvo. Esta tarefa é descritiva por natureza e serve para encontrar padrões que não estão vinculados a uma só variável.

A técnica mais comum no *Data Mining* não-supervisionado é o *Clustering*, que tem como objetivo encontrar grupos sem qualquer informação sobre quais as variáveis devem ser consideradas mais importantes.

Data Mining não-supervisionado também é designado por Aprendizagem não-supervisionada e *Data Mining* supervisionado por Aprendizagem supervisionada.

3.3 – ANÁLISE DE AGRUPAMENTO

Nas tarefas de classificação pode estabelecer-se a distinção entre Análise Discriminante (aprendizagem supervisionada que utiliza um conjunto de classes como variável alvo) e a Análise de Agrupamento (*Clustering*) que não utiliza qualquer informação anterior sobre a constituição das classes (aprendizagem não supervisionada) (Jain *et al*., 2010). A Análise de Agrupamento é então, uma classificação não supervisionada, de padrões (observações, conjuntos de dados, vetores de características) em grupos (*clusters*) (Jain *et al*., 1999). Esta classificação, recorre a um conjunto de métodos e algoritmos que agrupa objetos de acordo com características intrínsecas medidas ou de semelhança (Jain *et al*., 2010). Mais recentemente, Ferrari *et al.* (2015) afirma que devido à sua natureza

não-supervisionada, a procura de uma solução de Agrupamento com uma boa qualidade pode tornar-se um processo complexo. Assim, muitos anos volvidos sobre os primeiros desenvolvimentos em Agrupamento, selecionar o melhor algoritmo de Agrupamento, capaz de produzir uma boa solução, pode ainda ser um processo lento e caro. O objetivo será encontrar uma solução em que os grupos obtidos sejam coerentes na sua composição, e difiram de modo expressivo entre eles.

A Análise de Agrupamento é amplamente utilizada nos mais diversos contextos, o que demonstra a sua grande utilidade numa fase de análise exploratória dos dados.

Agrawal *et al.* (1998) enuncia as principais condições nos algoritmos de agrupamento em *Data Mining*:

- ✓ Capacidade de trabalhar com conjuntos de dados de dimensões elevadas;
- ✓ Escalabilidade – o algoritmo terá de ser aplicável à totalidade dos dados, independentemente da dimensão da amostra;
- ✓ Deverá garantir a interpretação e compreensão dos resultados finais; e,
- ✓ Insensibilidade à ordem dos registos de entrada.

Jain *et al*. (1988) identificam os seguintes passos na Análise de Agrupamento:

1. Extração e seleção dos dados;
2. Definição de uma medida de proximidade;
3. Agrupamento;
4. Abstração de dados; e,
5. Avaliação dos resultados.

Reis (2001) sugere também cinco etapas, a saber:

1. Seleção dos dados a serem agrupados;
2. Definição do conjunto de variáveis a partir das quais se irão agrupar os dados;
3. Definição de uma medida de semelhança ou dissemelhança;
4. Escolha do algoritmo de agrupamento; e,
5. Validação dos resultados obtidos.

A seleção das variáveis a serem consideradas deve atender ao conhecimento prévio do negócio, isto é, deverá levar em consideração estudos anteriores a fim de identificar a informação relevante. Simultaneamente, os dados devem ter um grande poder discriminatório, devendo procurar a diversidade dentro das variáveis existentes. Se as variáveis estiverem em diferentes escalas de medida, poderá ser necessário proceder a uma estandardização de modo a anular o efeito de mais peso das variáveis que apresentam maiores valores e maior dispersão. Relativamente às medidas de (dis)semelhança, dependem do tipo de medida; devendo adequar-se às variáveis base de agrupamento a distância Euclidiana para dados quantitativos, ou dissemelhança de *Jaccard* para dados qualitativos (por exemplo).

Uma estrutura de grupos obtida pode ser (Cardoso, 2001):

- ✓ uma partição, em que cada entidade é afeta a um só grupo e a reunião dos grupos perfaz o conjunto das entidades;
- ✓ uma estrutura difusa, em que cada entidade tem um grau de pertença a cada grupo (valor entre 0 e 1) e, para cada entidade, a soma desses graus de pertença perfaz 1;
- ✓ uma estrutura sobreposta, em que uma entidade pode pertencer simultaneamente a vários grupos.

Os algoritmos de Agrupamento podem ser tipificados quanto ao modo de processamento das observações: *Batch* – a solução é atualizada para um conjunto de observações ou Incremental – a solução é atualizada para cada nova observação. Podem ser ou não baseados em modelos, nomeadamente Modelos de Segmentos Latentes – (Wedel *et al*., 2000). Podem ser também divididos em (Han *et al*., 2011):

- ✓ Métodos hierárquicos;
- ✓ Métodos não hierárquicos ou de otimização, partição;
- ✓ Métodos com base na densidade;
- ✓ Métodos baseados em grelha.

Os métodos hierárquicos decompõem hierarquicamente um conjunto de objetos. Esta decomposição pode suceder de duas formas distintas: aglomerativa ou divisiva. A técnica

aglomerativa, na fase inicial conta com tantos grupos como objetos. Nas iterações seguintes, o algoritmo vai juntando grupos entre si até formar um único grupo (o nível máximo da hierarquia), ou até ser atingida a condição de paragem definida no início do procedimento.
Para a junção dos grupos pode ser escolhido um de três tipos de métodos:

- ✓ Métodos de ligação (*single linkage, complete linkage, average linkage, median linkage*);
- ✓ Métodos de centróide;
- ✓ Método de *Ward.*

A técnica divisiva, na fase inicial conta com um só grupo (constituído por todos os objetos). Nas iterações seguintes, o algoritmo vai dividindo os grupos, em grupos mais pequenos até existir em cada grupo um único objeto ou até ser atingida a condição de paragem definida.

Os métodos hierárquicos apresentam algumas limitações:

- ✓ atendendo a que uma vez realizada uma iteração (divisão ou junção), nunca mais é desfeita, não é possível retroceder no processo eventualmente promovendo melhores soluções;
- ✓ não é definido à partida o número de grupos, colocando-se muitas vezes a questão de quantos grupos serão adequados para obter a solução final;
- ✓ são adequados apenas para bases de dados de dimensões reduzidas;
- ✓ são sensíveis aos dados que contenham muito ruído.

Os métodos não hierárquicos ou de otimização agrupam os objetos num conjunto de segmentos cujo número é definido à partida. Estes métodos constroem K grupos e afetam os objetos a um dos K grupos procurando uma partição de modo a que os objetos do mesmo grupo sejam o mais próximos possível e os objetos de grupos diferentes estejam o mais afastados possível. Para obter uma melhor estrutura de grupos (grupos coesos e bem separados e facilmente interpretáveis), realiza-se o procedimento várias vezes com diferentes valores de K e comparam-se soluções.

Os métodos não hierárquicos mais conhecidos são o *k-means* e o *k-medoid.* Uma vez que não utilizam a matriz de similaridade, estes métodos tornam-se mais rápidos que os métodos hierárquicos.

Os métodos com base na densidade, são métodos que conseguem criar grupos de regiões densas, separadas por dados dispersos.

Os métodos baseados em grelha trabalham com uma estrutura em grelha que divide num número finito de campos o espaço dos objetos.

Podem ainda considerar-se métodos mistos, recorrendo a algoritmos combinados, considerando, por exemplo, métodos hierárquicos e não hierárquicos – v. capítulo 3.3 - Análise de Agrupamento.

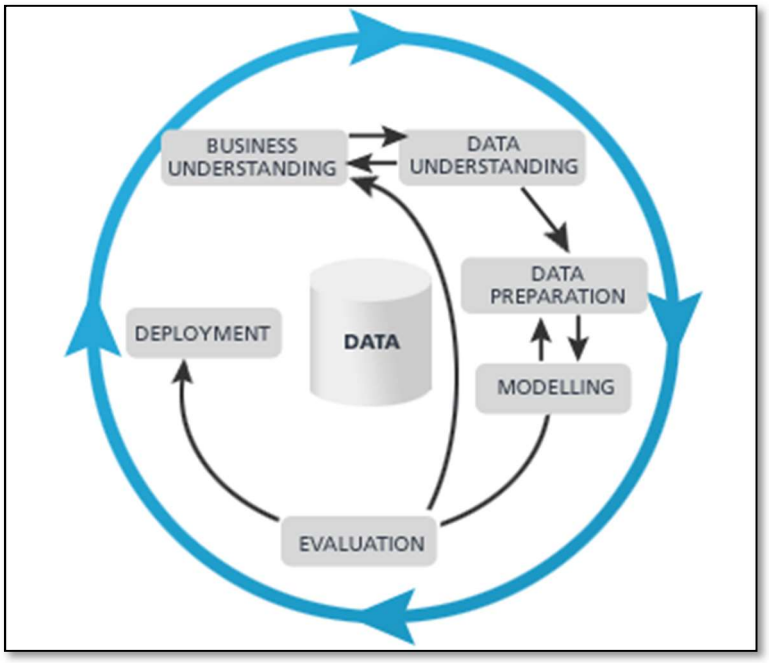
4 – METODOLOGIA

4.1 – CRISP-DM

Na realização deste trabalho, foram utilizados os *softwares* IBM SPSS *Modeler* 18.1, e IBM SPSS *Statistic* 24.

O IBM SPSS *Modeler* utiliza a metodologia CRISP-DM *(Cross Industry Standard Process for Data Mining).* O CRISP-DM engloba seis fases, as quais se encontram enunciadas e esquematizadas na figura seguinte:

Figura 2 – Fases do CRISP - DM



Fonte:https://www.function1.com/2012/03/analyzing-the-word-analytics-in-an-ever-increasing-analytical-world

Relativamente ao trabalho em questão, as seis fases do CRISP-DM traduziram-se nas etapas seguintes:

1. Compreensão do problema:
   Foram identificados os objetivos deste estudo e elaborado um plano de trabalho com as várias possibilidades existentes de exploração dos dados;

2. Compreensão dos dados:

   Nesta fase, foram recolhidos, estudados, analisados e selecionados os dados que melhor descrevem a realidade das empresas portuguesas. Os dados foram extraídos da *Amadeus*, (https://amadeus.bvdinfo.com/version-2017623/home.serv?product=amadeusneo), uma base de dados de informações financeiras comparáveis para empresas públicas e privadas localizadas na Europa da responsabilidade da empresa Bureau Van Dijk.

   De seguida, foi testada a sua qualidade (incoerências e *missings*). Foram selecionados os dados do ano de 2014, pois foi o ano que se mostrou mais consolidado.

3. Preparação dos dados:

   Esta foi a fase mais exigente e demorada deste trabalho. Os dados extraídos da *Amadeus* geraram 527 ficheiros em formato EXCEL, os quais tiveram de ser compilados e organizados em formato SAV. Além desta tarefa, foram eliminadas variáveis, assim como registos; não se mostrando ser necessária a construção de novas variáveis.

4. Modelação:

   Se na primeira fase foi decidido utilizar a Análise de Agrupamento, só aqui se tornou necessária a escolha concreta da técnica a aplicar. Efetuada a pesquisa bibliográfica e a comparação dos resultados obtidos, optou-se por utilizar o algoritmo *Two-Step* e a medida de distância Log-Verosimilhança.

5. Avaliação:

   Após a obtenção do modelo foi avaliada a sua qualidade e observado se foram atingidos os objetivos inicialmente propostos. Foram identificados alguns problemas e, dependendo dos resultados alcançados, foi decidido após várias tentativas, a revisão do modelo proposto.

6. Implementação:

   O objetivo nesta fase é implementar a execução prática do modelo, assim como realizar a sua monitorização. Esta tarefa encontra-se de momento fora do âmbito desta dissertação.

É de salientar que neste percurso, foi necessário reanalisar várias vezes o problema inicial, tendo em vista uma melhoria e aperfeiçoamento do modelo obtido.

4.2 – AMOSTRA E VARIÁVEIS

Os dados utilizados no presente trabalho foram extraídos do sítio *Amadeus*. Para obter a amostra inicial foram definidos dois critérios de pesquisa: "Todas as empresas ativas" e "localização em Portugal" (*All active companies and Region/Country/region in country: Portugal).* Desta forma, foi extraída uma amostra de 378.728 empresas, relativa ao ano de 2014. Para estas 378.728 empresas foram selecionadas as seguintes variáveis:

Tabela 1 – Variáveis extraídas

Variável	Número de observações	Tipo de variável
0 CAE	378 728	Qualitativa
0 CAESECCAO	378 728	Qualitativa
0 CATEGORIA	378 728	Qualitativa
0 CIDADE	378 728	Qualitativa
0 NIF	378 728	Qualitativa
0 NOME	378 728	Qualitativa
0 NUMERO FILIAIS	378 728	Quantitativa
0 NUMEROEMPREGADOS	251 543	Quantitativa
0 NUMEROEMPRESASNOGRUPO	378 707	Qualitativa
0 NUMEROACIONISTAS	378 728	Quantitativa
0 NUMEROSUBSIDIARIOS	378 728	Quantitativa
0 NUTS1	378 726	Qualitativa
0 NUTS2	378 726	Qualitativa
0 NUTS3	378 726	Qualitativa
0 REGIAO	354 353	Qualitativa
ATIVO CIRCULACAO LIQUIDA TH	307 259	Quantitativa
ATIVO PERMANENTE TH	307 259	Quantitativa
ATIVO TOTAL TH	307 259	Quantitativa
ATIVOS CORRENTES TH	307 259	Quantitativa
ATIVOS FIXOS INTANGIVEIS TH	254 586	Quantitativa
ATIVOS FIXOS TANGIVEIS TH	256 322	Quantitativa
CAPITAL EQUIVALENTE TH	297 403	Quantitativa
CAPITAL TH	307 339	Quantitativa
CREDORES TH	295 726	Quantitativa
DESPESAS FINANCEIRAS TH	146 798	Quantitativa
DEVEDORES TH	306 604	Quantitativa
DIVIDA LONGO PRAZO TH	193 718	Quantitativa
E15 RECEITA OPERACIONAL POR EMP TH	239 884	Quantitativa
E25 CUSTO EMP VS RECEITA OPERACIONAL %	225 563	Quantitativa
E35 CUSTO MEDIO EMPREGADO TH	238 000	Quantitativa
E45 FUNDO ACCIONISTA EMPREGADO TH	189 917	Quantitativa
E5 LUCRO POR EMP TH	248 634	Quantitativa

Variável	Número de observações	Tipo de variável
E55 WORKING CAPITAL POR EMP TH	244 228	Quantitativa
E65 ATIVO TOTAL POR EMP TH	250 266	Quantitativa
EMPRESTIMOS TH	291 902	Quantitativa
FINANCEIRAS P TH	296 930	Quantitativa
FLUXO CAIXA	225 674	Quantitativa
CAPITAL PROPRIO TH	307 467	Quantitativa
O12 TABELA INTERESSE X	70 473	Quantitativa
O19 RETORNO VN STOCK X	152 567	Quantitativa
O26 MANUTENCAO STOCK NUMERO DIAS	254 353	Quantitativa
O33 CREDITOS NUMERO DIAS	256 542	Quantitativa
O5 ATIVOS LIQUIDOS VOLUME NEGOCIOS X	231 065	Quantitativa
OPERACIONAL PL EBIT TH	296 900	Quantitativa
OUTRAS RESPONSABILIDADES CORRENTES TH	291 961	Quantitativa
OUTRAS RESPONSABILIDADES NAO CORRENTES TH	193 718	Quantitativa
OUTROS ATIVOS CORRENTES	302 554	Quantitativa
OUTROS ATIVOS PERMANENTES TH	255 748	Quantitativa
OUTROS FUNDOS SHAREHOLDERS TH	307 339	Quantitativa
PASSIVO CIRCULANTE TH	307 422	Quantitativa
PL ANTES IMPOSTO	296 963	Quantitativa
PL DEPOIS IMPOSTO	296 907	Quantitativa
PL POR PERÍODO RESULTADO LIQUIDO TH	296 963	Quantitativa
PROVISOES TH	75 420	Quantitativa
RECEITA FINANCEIRA TH	18 428	Quantitativa
RECEITA OPERACIONAL VN TH	267 448	Quantitativa
RESPONSABILIDADES NAO CORRENTES	307 344	Quantitativa
S15 TAXA LIQUIDEZ X	287 197	Quantitativa
S25 CAPITAL PROPRIO TAXA LIQUIDEZ X	191 880	Quantitativa
S35 RACIO SOLVABILIDADE BASEADO ATIVO %	278 249	Quantitativa
S45 RACIO SOLVABILIDADE BASEADO RESPONSABILIDADES %	129 666	Quantitativa
S5 RACIO CORRENTE X	288 927	Quantitativa
S55 ENGRENAGEM %	208 430	Quantitativa
STOCK TH	302 543	Quantitativa
TRIBUTACAO TH	202 351	Quantitativa
V12 ROCE USING PL ANTES IMPOSTO	75 082	Quantitativa
V19 ROA USING PL ANTES IMPOSTO	282 641	Quantitativa
V26 ROE USING RESULTADO LIQUIDO %	222 100	Quantitativa
V33 ROCE USING RESULTADO LIQUIDO %	75 091	Quantitativa
V40 ROA USING RESULTADO LIQUIDO %	282 791	Quantitativa
V47 MARGEM LUCRO	248 755	Quantitativa
V5 ROE USING PL ANTES IMPOSTO	222 017	Quantitativa
V61 EBITDA MARGEM %	205 257	Quantitativa
V68 EBIT MARGEM %	249 212	Quantitativa
V75 FLUXO CAIXA VS RECEITA OPERACIONAL %	205 120	Quantitativa
VENDAS TH	262 870	Quantitativa
WORKING CAPITAL TH	291 803	Quantitativa

4.3 – ALGORITMO *TWO-STEP*

Atendendo à dimensão da base de dados considerada, optou-se por uma Análise de Agrupamento recorrendo ao Algoritmo *Two-Step* (Chiu *et al*., 2001). Neste método há a possibilidade de definir automaticamente o número de grupos, uma questão difícil no âmbito da Análise de Agrupamento.

No *Two-Step* pressupõe-se que as variáveis utilizadas são independentes, que as variáveis contínuas seguem distribuição Normal e que as variáveis categoriais seguem distribuição Multinomial. Só assim, pode ser especificada a função de verosimilhança. No entanto, no caso destas condições não se encontrarem cumpridas, o algoritmo mostra-se robusto, sendo capaz de produzir bons resultados.

Tal como o nome indica, o *Two-Step* envolve duas etapas:

1. na etapa 1, a fim de ser reduzida a quantidade de informação a tratar, os dados são sumarizados através da construção de uma árvore de objetos simbólicos. Os nós da árvore são subconjuntos de observações semelhantes entre si (é de referir que a ordem das observações influencia o resultado obtido nesta primeira fase).

   Os dados são resumidos através do cálculo de estatísticas sumário (*Cluster Feature Entry*), as quais dependem do número de observações incluídas no *subcluster*, das médias e variâncias de cada variável contínua e da frequência associada a cada categoria de cada variável categórica. A dimensão da árvore é controlada limitando o número máximo de níveis da árvore e o número máximo de folhas por nó. A árvore é construída iterativamente: cada objeto é dirigido para o nó folha que se mostra mais próximo (de acordo com a medida de distância utilizada). O processo termina quando todos os objetos tiverem entrado na árvore;

2. na etapa 2 é utilizado um método hierárquico a partir dos objetos criados na etapa 1.

Se as variáveis forem todas quantitativas pode ser utilizada a Distância Euclidiana.

O número de grupos a formar pode atender aos critérios de Teoria da Informação que consideram o ajustamento do modelo (via função de verosimilhança) e uma penalização da sua complexidade (medida pelo número de parâmetros distribucionais e número de grupos):

- ✓ *Schartz Bayesian Criterion* (BIC);
- ✓ *Akaike Information Criterion* (AIC)

Depois de encontrado o agrupamento, existe a necessidade de avaliar o seu resultado. Para tal, poderá ser utilizado o Índice de Coesão-separação, Índice Silhueta (Kaufman *et al*., 1990), o qual reflete a ligação interna das observações pertencentes a um mesmo grupo e o afastamento entre grupos. O Índice Silhueta varia entre -1 e 1. Um valor alto deste índice indica que o objeto está ajustado ao seu grupo e desajustado nos outros grupos. Na prática, um valor de Índice Silhueta maior que 0,5 indica uma que foi realizada uma partição razoável, assim como, um Índice Silhueta menor que 0,2 indica que os dados não exibem uma estrutura de grupo. Depois de avaliado o modelo, é necessário proceder à caracterização dos grupos formados. Para tal, poderá ser útil o cálculo da estatística *V de Cramer* (Siegel, 1988), que mede a associação entre os segmentos e cada uma das variáveis nominais. Quando *V de Cramer* for zero, quer dizer que não existe associação e quando for um, a associação entre os segmentos e a variável nominal em questão é perfeita. De modo semelhante, a estatística *Eta* mede a associação entre os segmentos e cada uma das variáveis contínuas. O significado dos valores de *Eta* é idêntico aos da estatística *V de Cramer*.

5 – RESULTADOS

Para realizar a Análise de Agrupamento, foi utilizado o *software* IBM SPSS *Statistic* 24.

As variáveis de "*input*" que serviram para discriminar os grupos são as constantes da tabela 2:

Tabela 2 – Variáveis de *input*

Variável	Descrição da variável
V5_ROE_USING_PL_BEFORE_TAX	RENDIMENTO DO CAPITAL PRÓPRIO ANTES IMPOSTO
V19_ROA_USING_PL_BEFORE_TAX	RENDIMENTO DO ATIVO DA EMPRESA ANTES IMPOSTO
V26_ROE_USING_RESULTADO_LIQUIDO	RENDIMENTO DO CAPITAL PRÓPRIO
V40_ROA_USING_RESULTADO_LIQUIDO	RENDIMENTO DO ATIVO DA EMPRESA
V47_MARGEM_LUCRO	MARGEM LUCRO DA EMPRESA
V61_EBITDA_MARGEM	RESULTADO GERADO PELA EMPRESA – NÃO INCLUI AMORTIZAÇÕES
V68_EBIT_MARGEM	RESULTADO GERADO PELA EMPRESA JÁ COM CUSTO ANUAL DOS ATIVOS IMOBILIZADOS UTILIZADOS
V75_FLUXO_CAIXA_VS_RECEITA_OPERACIONAL	FLUXO CAIXA VS RECEITA OPERACIONAL

As fórmulas de cálculo das variáveis de "*input*" encontram-se na tabela 3:

Tabela 3– Fórmulas de cálculo das variáveis de *input*

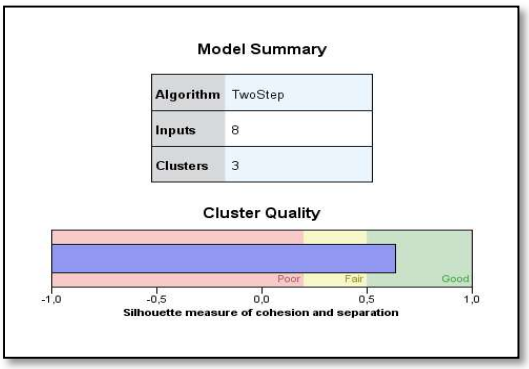
Variável	Fórmula de cálculo
V5_ROE_USING_PL_BEFORE_TAX	(resultado antes imposto / capital próprio) * 100
V19_ROA_USING_PL_BEFORE_TAX	(resultado antes imposto / total ativo) * 100
V26_ROE_USING_RESULTADO_LIQUIDO__	(resultado líquido / capital próprio) * 100
V40_ROA_USING_RESULTADO_LIQUIDO__	(resultado líquido / total ativo) * 100
V47_MARGEM_LUCRO	(resultado antes imposto / rendimentos operacionais) *
V61_EBITDA_MARGEM__	(EBITDA / rendimentos operacionais) * 100
V68_EBIT_MARGEM__	(EBIT / rendimentos operacionais) * 100
V75_FLUXO_CAIXA_VS_RECEITA_OPERACIONAL__	(fluxo caixa / rendimentos operacionais) * 100

Figura 3 – Opções no *software* SPSS *Statistics*

A medida de distância selecionada foi a Log-Verosimilhança e o Critério da teoria de informação foi o *Schwarz´s Bayesian Information Criterion* (BIC).

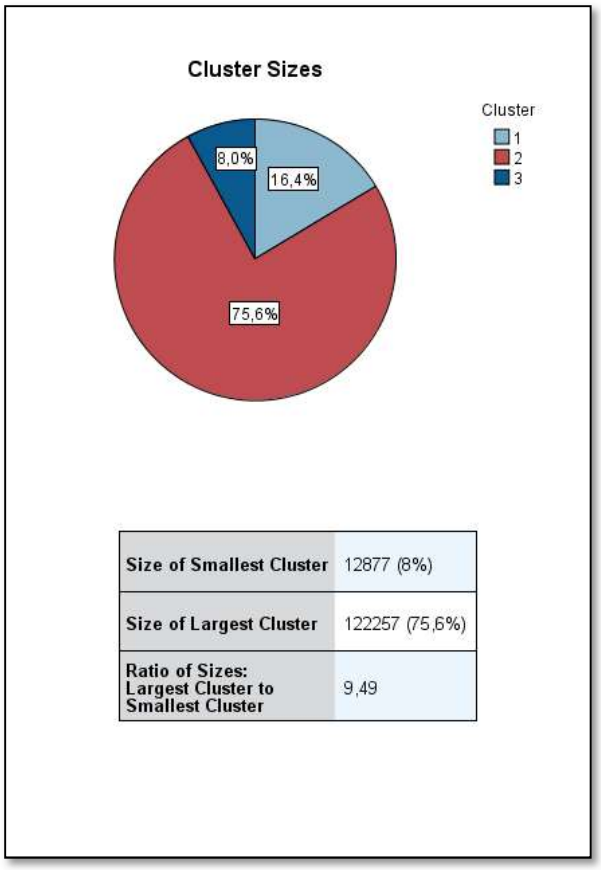
Os resultados obtidos foram os seguintes:

Figura 4 – Resultado do modelo no *software* SPSS *Statistics*



O valor da medida Silhueta obtido foi 0,6, o que se traduz numa boa qualidade do modelo.

Figura 5 – Dimensão dos *clusters* no *software* SPSS *Statistics*



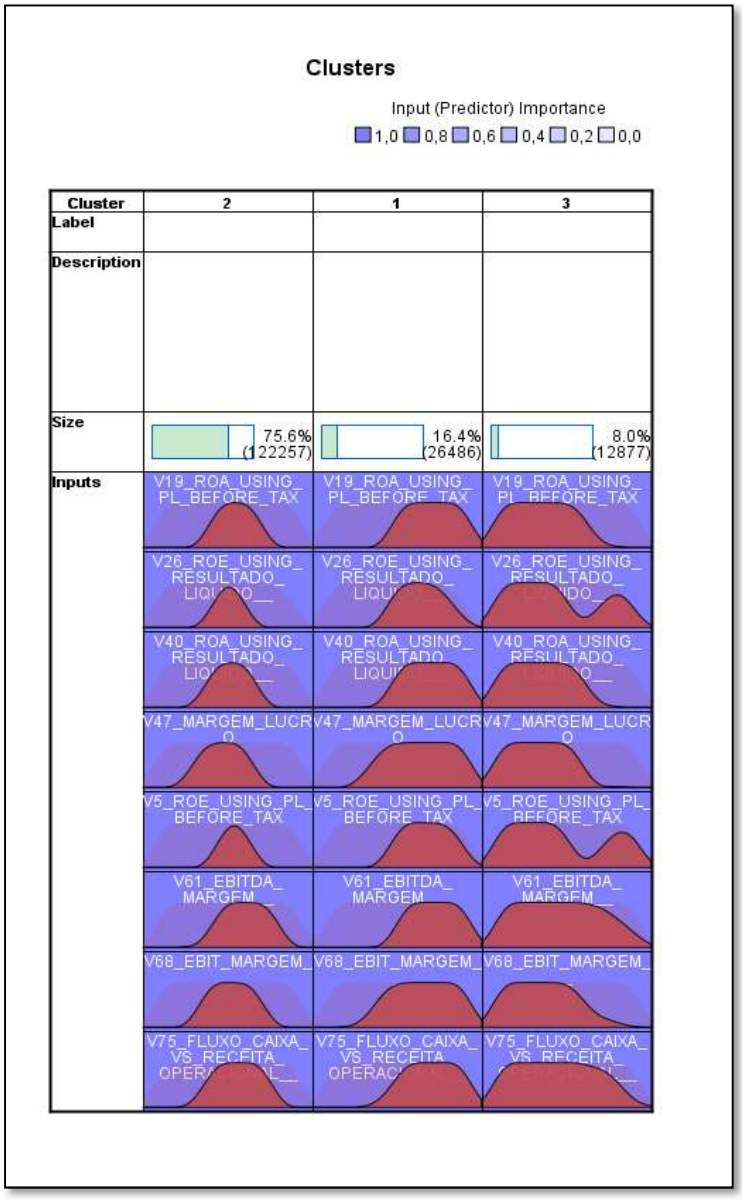
O modelo originou três *clusters*, com as dimensões e percentagens seguintes:

Tabela 4 – Dimensão dos *clusters*

Distribuição das empresas nos *clusters*			
		Número de empresas	Percentagem
Cluster	1	26 486	16,4%
	2	122 257	75,6%
	3	12 877	8,0%
	Total	161 620	100,0%

Das 161.620 empresas, aproximadamente 76% ficaram afetas ao *cluster* 2; 16% ao *cluster* 1 e 8% ao *cluster* 3.

Figura 6 – Distribuição das variáveis de *input* por *cluster*



As oito variáveis de *input* mostraram-se todas com elevada importância no modelo.

Tabela 5 – Medida de associação *Eta*

Medida *Eta* entre cada variável e os *cluster´s*	
Variável	*Eta*
V68_EBIT_MARGEM * Cluster´s	0,795
V47_MARGEM_LUCRO * Cluster´s	0,794
V61_EBITDA_MARGEM * Cluster´s	0,765
V75_FLUXO_CAIXA_VS_RECEITA_OPERACIONAL * Cluster´s	0,763
V40_ROA_USING_RESULTADO_LIQUIDO * Cluster´s	0,699
V19_ROA_USING_PL_BEFORE_TAX * Cluster´s	0,694
V26_ROE_USING_RESULTADO_LIQUIDO * Cluster´s	0,553
V5_ROE_USING_PL_BEFORE_TAX * Cluster´s	0,537

Os indicadores mais correlacionados com os grupos são os que representam os valores mais elevados da medida *Eta*: V68_EBIT_MARGEM e V47_MARGEM_LUCRO.

Relativamente ao indicador V68_EBIT_MARGEM (tabela 6):

- ✓ o grupo 1, apresenta de média 34% e desvio padrão 19%;
- ✓ o grupo 2, tem uma média 5% e é o que apresenta menor variabilidade, pois o desvio padrão é de 7%;
- ✓ o grupo 3, tem média de -29% e desvio padrão 23%, sendo o grupo que apresenta maior variabilidade neste indicador.

Relativamente ao indicador V47_MARGEM_LUCRO (tabela 6):

- ✓ o grupo 1, apresenta de média 33% e desvio padrão 19%;
- ✓ o grupo 2, tem uma média 5% e é o que apresenta menor variabilidade, pois o desvio padrão é de 7%; o grupo 3, tem média de -29% e desvio padrão 23%, sendo o grupo que apresenta maior variabilidade neste indicador.

Tabela 6 – Estatísticas descritivas das variáveis de *input*

Estatísticas descritivas das variáveis de *input* por *cluster*					
	Número de empresas	Mínimo	Máximo	Média	Desvio padrão
Cluster 1					
V5_ROE_USING_PL_BEFORE_TAX	26 486	-159,39	974,50	68,07	84,52
V19_ROA_USING_PL_BEFORE_TAX	26 486	-8,25	100,00	27,78	20,78
V26_ROE_USING_RESULTADO_LIQUIDO__	26 486	-223,40	943,09	56,09	72,46
V40_ROA_USING_RESULTADO_LIQUIDO__	26 486	-29,47	99,69	22,85	17,72
V47_MARGEM_LUCRO	26 486	-49,54	99,88	33,29	19,19
V61_EBITDA_MARGEM__	26 486	-13,69	100,00	45,13	21,58
V68_EBIT_MARGEM__	26 486	-59,98	99,88	34,87	19,18
V75_FLUXO_CAIXA_VS_RECEITA_OPERACIONAL__	26 486	-32,90	100,00	37,53	18,90
Cluster 2					
V5_ROE_USING_PL_BEFORE_TAX	122 257	-156,94	344,11	16,49	30,63
V19_ROA_USING_PL_BEFORE_TAX	122 257	-25,94	49,71	5,15	7,75
V26_ROE_USING_RESULTADO_LIQUIDO__	122 257	-225,39	264,60	11,34	25,90
V40_ROA_USING_RESULTADO_LIQUIDO__	122 257	-42,68	39,29	3,69	6,59
V47_MARGEM_LUCRO	122 257	-54,38	42,47	3,88	6,95
V61_EBITDA_MARGEM__	122 257	-39,97	73,69	9,67	9,15
V68_EBIT_MARGEM__	122 257	-44,96	56,08	4,81	6,95
V75_FLUXO_CAIXA_VS_RECEITA_OPERACIONAL__	122 257	-42,27	61,53	7,50	8,05
Cluster 3					
V5_ROE_USING_PL_BEFORE_TAX	12 877	-996,82	999,84	-126,50	200,10
V19_ROA_USING_PL_BEFORE_TAX	12 877	-99,79	56,17	-19,58	17,24
V26_ROE_USING_RESULTADO_LIQUIDO__	12 877	-996,82	999,84	-132,76	196,94
V40_ROA_USING_RESULTADO_LIQUIDO__	12 877	-99,87	59,76	-20,10	17,49
V47_MARGEM_LUCRO	12 877	-99,93	67,27	-31,31	22,95
V61_EBITDA_MARGEM__	12 877	-99,09	100,00	-18,00	22,37
V68_EBIT_MARGEM__	12 877	-99,90	81,80	-29,20	22,51
V75_FLUXO_CAIXA_VS_RECEITA_OPERACIONAL__	12 877	-99,99	100,00	-20,84	22,21

Analisando a tabela 6, verifica-se que a variável V5_ROE_USING_PL_BEFORE_TAX apresenta média de 68,07 no *cluster* 1, média de 16,49 no *cluster* 2 e média -126,50 no *cluster* 3.

Esta tendência repete-se relativamente às outras variáveis.

De uma forma geral, pode concluir-se o seguinte:

- ✓ o grupo 2 representa as empresas com resultados "normais", pois os valores dos oito indicadores de rentabilidade são valores intermédios que correspondem ao perfil da amostra global;
- ✓ o grupo 1 é constituído pelas empresas com rentabilidade acima da média, pois os oito indicadores de rentabilidade apresentam valores superiores ao do grupo 2;

- ✓ o grupo 3 engloba as empresas com rentabilidade negativa, pois os oito indicadores de rentabilidade apresentam valores inferiores aos do grupo 2

Cruzando esta informação com as variáveis qualitativas de cada uma das empresas, obtêm-se outro tipo de conclusões:

Tabela 7 – Número de empresas por *cluster* e zona de Portugal

Distribuição do número de empresas por cluster e zona de Portugal				
Zonas	*Cluster*			Total
	1	2	3	
Norte	8 663	45 383	4 293	58 339
Algarve	1 337	4 825	661	6 823
Região Centro	4 592	29 112	2 717	36 421
Área Metropolitana de Lisboa	9 141	31 628	3 711	44 480
Alentejo	1 836	7 492	861	10 189
Região Áutonoma da Madeira	338	1 704	265	2 307
Região Áutonoma dos Açores	579	2 113	369	3 061
Total	26 486	122 257	12 877	161 620

Tabela 8 – Percentagem das empresas por *cluster* e zona de Portugal

Percentagem de empresas por cluster e zona de Portugal				
Zonas	*Cluster*			Total
	1	2	3	
Norte	15%	78%	7%	100%
Algarve	20%	71%	10%	100%
Região Centro	13%	80%	7%	100%
Área Metropolitana de Lisboa	21%	71%	8%	100%
Alentejo	18%	74%	8%	100%
Região Áutonoma da Madeira	15%	74%	11%	100%
Região Áutonoma dos Açores	19%	69%	12%	100%
Total	16%	76%	8%	100%

Nas tabelas 7 e 8, é possível verificar a homogeneidade dos *clusters* relativamente às diferentes zonas geográficas portuguesas. Destaca-se pela negativa a Região Autónoma da Madeira com 12% das empresas no grupo das Empresas em Risco, e pela positiva as regiões do Algarve e Área Metropolitana de Lisboa com percentagens de 20% e 21% de empresas no grupo Empresas em destaque.

Tabela 9 – Número de empresas por *cluster* e secção da CAE

Distribuição do número de empresas por cluster e secção da CAE				
Secções da CAE da base de dados *Amadeus*	*Cluster*			Total
	1	2	3	
A. Agriculture, forestry and fishing	1 616	4 403	543	6 562
B. Mining and quarrying	41	314	37	392
C. Manufacturing	1 916	18 465	1 521	21 902
D. Electricity, gas, steam and air conditioning supply	269	95	11	375
E. Water supply; sewerage, waste management and remediation activities	74	431	33	538
F. Construction	1 854	12 329	1 361	15 544
G. Wholesale and retail trade; repair of motor vehicles and motorcycles	2 855	38 209	3 379	44 443
H. Transportation and storage	1 353	7 362	749	9 464
I. Accommodation and food service activities	1 022	6 601	1 536	9 159
J. Information and communication	989	2 786	319	4 094
K. Financial and insurance activities	825	1 440	112	2 377
L. Real estate activities	2 577	2 916	615	6 108
M. Professional, scientific and technical activities	4 026	12 000	1 092	17 118
N. Administrative and support service activities	1 113	3 866	378	5 357
O. Public administration and defence; compulsory social security	4	6	2	12
P. Education	279	1 385	217	1 881
Q. Human health and social work activities	4 946	6 990	506	12 442
R. Arts, entertainment and recreation	342	1 035	203	1 580
S. Other service activities	385	1 624	263	2 272
Total	26 486	122 257	12 877	161 620
As secções da CAE da presente tabela referem-se às que estão definidas na base de dados *Amadeus*, que mostram ser diferentes das secções da CAE portuguesa				

Tabela 10 – Percentagem das empresas por *cluster* e secção da CAE

Percentagem de empresas por cluster e secção da CAE				
Secções da CAE da base de dados *Amadeus*	*Cluster*			Total
	1	2	3	
A. Agriculture, forestry and fishing	25%	67%	8%	100%
B. Mining and quarrying	10%	80%	9%	100%
C. Manufacturing	9%	84%	7%	100%
D. Electricity, gas, steam and air conditioning supply	72%	25%	3%	100%
E. Water supply; sewerage, waste management and remediation activities	14%	80%	6%	100%
F. Construction	12%	79%	9%	100%
G. Wholesale and retail trade; repair of motor vehicles and motorcycles	6%	86%	8%	100%
H. Transportation and storage	14%	78%	8%	100%
I. Accommodation and food service activities	11%	72%	17%	100%
J. Information and communication	24%	68%	8%	100%
K. Financial and insurance activities	35%	61%	5%	100%
L. Real estate activities	42%	48%	10%	100%
M. Professional, scientific and technical activities	24%	70%	6%	100%
N. Administrative and support service activities	21%	72%	7%	100%
O. Public administration and defence; compulsory social security	33%	50%	17%	100%
P. Education	15%	74%	12%	100%
Q. Human health and social work activities	40%	56%	4%	100%
R. Arts, entertainment and recreation	22%	66%	13%	100%
S. Other service activities	17%	71%	12%	100%
Total	16%	76%	8%	100%
As secções da CAE da presente tabela referem-se às que estão definidas na base de dados *Amadeus*, que mostram ser diferentes das secções da CAE portuguesa				

Na tabela 10, pode ser analisada a distribuição percentual das empresas em cada um dos grupos segundo a secção da CAE a que pertencem. Conclui-se que as secções D, K e Q apresentam uma percentagem de empresas acima da média no grupo Empresas em destaque, e por outro lado as secções I e O apresentam uma percentagem acima da média de empresas no grupo Empresas em Risco. Estes valores podem resultar do tipo de atividade muito específica dos setores K e O, uma vez que este tipo de atividade é supervisionada e fortemente regulamentada.

Tabela 11 – Número de empresas por *cluster* e dimensão

Percentagem de empresas por cluster e dimensão				
Dimensão	*Cluster*			Total
	1	2	3	
Grande	450	3 372	119	3 941
Média	2 730	26 601	1 354	30 685
Pequena	23 226	91 878	11 387	126 491
Muito grande	80	406	17	503
Total	26 486	122 257	12 877	161 620

Tabela 12 – Percentagem das empresas por *cluster* e dimensão

Percentagem de empresas por cluster e dimensão				
Dimensão	*Cluster*			Total
	1	2	3	
Grande	11%	86%	3%	100%
Média	9%	87%	4%	100%
Pequena	18%	73%	9%	100%
Muito grande	16%	81%	3%	100%
Total	16%	76%	8%	100%

Nas tabelas 11 e 12, pode ser observado que as empresas mais pequenas são aquelas que apresentam uma maior percentagem no grupo de Empresas em Risco.

6 - DISCUSSÃO

Da análise do modelo explanado anteriormente, resultam três segmentos de empresas no tecido empresarial português:

Grupo 1 – Empresas portuguesas com rentabilidade acima da média, o qual poderá ser denominado por "Empresas em destaque"

Grupo 2 – Empresas portuguesas com rentabilidade "normal", o qual será denominado "Empresas padrão"

Grupo 3 – Empresas portuguesas com rentabilidade negativa, o qual poderá ser denominado por "Empresas de risco".

Este modelo permitiu desta forma, identificar as empresas que apresentam resultados fora da rotina tradicional portuguesa. Poderá servir para estudar os casos de sucesso, assim como evidenciar as empresas que estão a ultrapassar uma fase mais delicada.

Este estudo permitiu demonstrar que independentemente da CAE, existem outras formas de segmentar as empresas portuguesas.

Utilizando e cruzando a segmentação obtida com outras variáveis qualitativas (Zona de Portugal, Secção da CAE e Dimensão da empresa), é possível também realizar outro tipo de estudos.

Fica assim demonstrado que os oito indicadores de rentabilidade escolhidos e a técnica de segmentação foram adequados ao objetivo em questão. Possivelmente, existirá outra combinação de variáveis que permitam obter outro tipo de resultados e conclusões. Este poderá ser um desafio futuro: segmentar as empresas portuguesas por outro tipo de informação/variáveis.

Outro desafio futuro passa por alargar o presente estudo a empresas localizadas fora de Portugal.

Poderá também ser criado, com base nos grupos obtidos, um sistema de classificação das empresas, tendo como objetivo classificar ao final de um ano de existência, por exemplo, cada uma das novas empresas a operar em Portugal.

Outra ideia será estudar as empresas ao longo do tempo, durante vários anos de atividade, a fim de identificar as principais alterações de comportamento.

Por último, é esperado que os presentes resultados sejam encorajadores para "deixar cair" a "CAE" e complementar os estudos e análises do tecido empresarial português com classificações alternativas ao uso da CAE.

7 – REFERÊNCIAS BIBLIOGRÁFICAS

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. 1998. Automatic subspace clustering of high dimensional data for data mining applications. ***ACM***, 27: 94-105.

Berry, M. J., & Linoff, G.. 2004. ***Data mining techniques: for marketing, sales, and customer support***. Indianapolis: John Wiley & Sons, Inc.

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. 2001. A robust and scalable clustering algorithm for mixed type attributes in large database environment. ***ACM, Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining:*** 263-268.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. 1996. From data mining to knowledge discovery in databases. ***AI magazine***, *17*: 37.

Ferrari, D. G., & De Castro, L. N.. 2015. Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods. ***Information Sciences***, *301*: 181-194.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. 1992. Knowledge discovery in databases: An overview. ***AI magazine***, *13*(3): 57.

Friedman, J. H. 1998. Data Mining and Statistics: What's the connection?. ***Computing Science and Statistics***, *29*(1): 3-9.

Giraud-Carrier, C., & Povel, O. 2003. Characterising data mining software. ***Intelligent Data Analysis*****,** *7*(3): 181-192.

Han, J., Pei, J., & Kamber, M.. 2011. ***Data mining: concepts and techniques.*** Waltham: Elsevier.

Hand, D. J.. 1998. Data mining: statistics and more?. ***The American Statistician***, *52*(2): 112-118.
Jain, A. K. 2010. Data clustering: 50 years beyond K-means. ***Pattern recognition letters***, *31*(8): 651-666.

Jain, A. K., & Dubes, R. C. 1988. ***Algorithms for clustering data***. Englewwod Cliffs: Prentice-Hall, Inc..

Jain, A. K., Murty, M. N., & Flynn, P. J. 1999. Data clustering: a review. ***ACM computing surveys (CSUR)***, *31*(3): 264-323.

Kaufman, L., & Rousseeuw, P. J. 2009. ***Finding groups in data: an introduction to cluster analysis***. NY: John Wiley & Sons.

Kock, N. F., McQueen, R. J., & Corner, J. L. 1997. The nature of data, information and knowledge exchanges in business processes: implications for process improvement and organizational learning. ***The Learning Organization***, *4*(2): 70-80.

Cardoso, M. G. M. S. 2001. Modelos de Segmentos Latentes: Aplicações em Marketing. In M. A. M. Ferreira, R. Menezes & M. G. M. S. Cardoso (Eds.), ***Temas em Métodos em Métodos Quantitativos 2***: 206-230. Lisboa: Ed. Sílabo

Ngai, E. W., Xiu, L., & Chau, D. C. 2009. Application of data mining techniques in customer relationship management: A literature review and classification. ***Expert systems with applications*****,** *36*(2): 2592-2602.

Siegel, S., & Castellan, N. J. 1988. **Nonparametric statistics for the behavioral sciences.** New York: McGraw-Hill

Reis, E. 2001. ***Estatística multivariada aplicada*****.** Lisboa: Edições Sílabo

Wedel, M., & Kamakura, W. A. 2000. ***Market segmentation: Conceptual and methodological foundations***. Netherlands: Kluwer Academic Publishers.

Consultas *online*:

"Mediação imobiliária, alojamento local, construção e hotéis na mira do Fisco". **Sapo**, https://casa.sapo.pt/Noticias/Mediacao-imobiliaria,-alojamento-local,-construcao-e-hoteis-na-mira-do-Fisco/?ID=24228, consulta em 20170727.

"Em 2016, Fisco prioriza restaurantes, hotéis e imóveis". **Eco**, https://eco.pt/2017/03/10/em-2016-fisco-prioriza-restaurantes-hoteis-e-imoveis/, consulta em 20170727.

"Fisco quer apertar o controlo ao comércio online e obras". **Jornal Económico**, http://www.jornaleconomico.sapo.pt/noticias/fisco-quer-apertar-o-controlo-ao-comercio-online-e-obras-132248, consulta em 20170727.

"Fisco inspeciona mais de seis mil cabeleireiros e instaura 400 autos de notícia". **Jornal Económico**, https://www.dinheirovivo.pt/economia/fisco-inspeciona-mais-de-6-mil-cabeleireiros-e-instaura-400-autos-de-noticia/, consulta em 20170727.

"Ação Menu. 500 inspetores do Fisco apertam restaurantes". **Dinheiro Vivo**, https://www.dinheirovivo.pt/economia/acao-menu-de-caca-ao-iva/, consulta em 20170727.

"Turismo bate recorde de receitas e dormidas em 2016". **Diário de Notícias,** https://www.dn.pt/volta-ao-mundo/interior/turismo-bate-recorde-de-receitas-e-dormidas-em-2016-8540296.html, consulta em 20170727

"Receitas da restauração sobem 2,2% para 3,7 mil milhões de euros em 2016". **Diário de Notícias**, http://www.dnoticias.pt/pais/receitas-da-restauracao-sobem-2-2-para-3-7-mil-milhoes-de-euros-em-2016-LB1506159, consulta em 20170727.

"Bolsa nacional começa semana a recuperar apoiada pela energia". **Jornal de Negócios,** http://www.jornaldenegocios.pt/mercados/bolsa/detalhe/bolsa-nacional-comeca-semana-a-recuperar-apoiada-pela-energia, consulta em 20170727.