LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO

DECISÃO ECONÓMICA E EMPRESARIAL

TRABALHO FINAL DE MESTRADO

RELATÓRIO DE ESTÁGIO

ANÁLISE DE MERCADO:CLUSTERING

ERIKSON MANUEL GERALDO VIEIRA DE MADUREIRA

OUTUBRO-2016

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO EM
DECISÃO ECONÓMICA E EMPRESARIAL

TRABALHO FINAL DE MESTRADO
RELATÓRIO DE ESTÁGIO

ANÁLISE DE MERCADO:CLUSTERING

ERIKSON MANUEL GERALDO VIEIRA DE MADUREIRA

ORIENTAÇÃO:

PROFESSOR DR. JOSÉ PEDRO GAIVÃO
CO-ORIENTADORA ENGENHEIRA CLÁUDIA NABAIS

OUTUBRO-2016

Agradecimentos

Quero deixar um agradecimento inicial ao mestrado de Métodos Quantitativos para a Decisão Económica e Empresarial do Instituto Superior de Economia e Gestão, aos professores e toda a sua coordenação pelo conhecimento que me transmitiram durante o mestrado que será uma mais-valia para a vida profissional e pessoal futura. Agradecer a Coordenação do Mestrado por tudo indicado atrás, pela sua ajuda para referências bibliográficas e pela sua grande disponibilidade e ajuda para a minha integração no estágio curricular, que sem este não seria possível a realização do TFM.

Agradeço ao meu Orientador, Prof. Doutor José Pedro Gaivão pela sua paciência, motivação e confiança que me deu nesta caminhada desde o início do estágio até a finalização da realização da escrita do TFM, onde foi incansável na sua ajuda constante. Para além do conhecimento que me passou como professor na disciplina de computação, que me despertou um gosto na área de informática e programação, também me transmitiu muito conhecimento que me ajudou no estágio para realização do trabalho pedido e assim a escrita do TFM.

À minha co-orientadora da empresa onde estagiei, Cláudia Nabais, do departamento de gestão da Quidgest, pela paciência e ajuda que me foi dada para ter conhecimento do funcionamento da empresa e para adquirir uma base de dados para a realização do estudo pedido.

À Quidgest, empresa onde estive 4 meses a estagiar, que me integrou muito bem, onde dei os primeiros passos a nível profissional e tive uma experiência muito enriquecedora a nível pessoal. Em especial ao seu presidente, o Doutor João Paulo Carvalho que sempre

me ajudou no trabalho do estágio, passou-me conhecimentos importantes para uma futura vida profissional e me proporcionou esta experiência nas suas instalações.

Quero deixar um especial agradecimento à minha família em especial aos meus pais pelo financiamento do mestrado e pela motivação feita ao longo de todo o mestrado especialmente durante a realização do TFM. Aos meus amigos e à minha namorada pela grande paciência durante todo o mestrado e realização do TFM, pelo apoio prestado, que sem eles não teria a força necessária para a conclusão do TFM.

Resumo

O presente trabalho tem como objetivo descrever as atividades realizadas durante o estágio efetuado na empresa Quidgest. Esta é uma empresa localizada em Lisboa, que atua na área de informática e *software*. O estágio desenvolveu-se entre 19 de Outubro de 2015 a 31 de Janeiro de 2016.

Este trabalho final de mestrado teve como objetivo o desenvolvimento de uma ferramenta/programa de análise de mercado na empresa Quidgest. Este projeto envolveu a implementação computacional de diversos métodos de análise de dados e foi realizado por etapas, resultando num processo complexo. Foi utilizado para a criação do programa, o *software* Excel e como linguagem de programação, o *Visual Basic for Applications* (VBA).

Tendo a empresa a necessidade de estudar as suas diversas vertentes de negócio e deste modo preparar a melhor estratégia a nível interno e externo, optou-se inicialmente por extrair e identificar as informações presentes no banco de dados da empresa. Para isso, foi utilizado um processo conhecido na análise de dados denominado por Extração de Conhecimento em Bases de Dados (ECBD). O maior desafio na implementação deste processo deveu-se há grande acumulação de informação pela empresa, que se foi intensificando a partir de 2013.

O processo de ECBD desenvolve-se em diferentes fases. A extração da informação dos dados é feita na base de dados da empresa denominada por Quigenio. Das fases do processo de ECBD, a que tem maior relevância é a fase *Data mining*, onde é feito um estudo das variáveis caracterizadoras necessárias para a análise em foco. Nesse estudo deu-se relevância à correlação entre as variáveis e a importância que estas representam para a análise pretendida pela empresa.

Foi escolhida a técnica de análise cluster da fase de data mining para que que toda análise possa ser eficiente, eficaz e se possa obter resultados de fácil leitura. Após o desenvolvimento do processo de ECBD, foi decidido que a fase de data mining podia ser implementada e automatizar uma das suas técnicas de modo a facilitar um trabalho futuro de uma análise realizada pela empresa. Para implementar essa fase, utilizaram-se técnicas de análise cluster e foi desenvolvida uma interface gráfica (programa em vba/excel) centrada no utilizador, para que a empresa pudesse usufruir do presente, sem ter de lidar com a complexidade envolvida. A interface gráfica criada e implementada é compatível com o *software* usado pela empresa e pelas capacidades técnicas dos seus recursos. Poderá constituir uma mais-valia para a organização, esperando que possa vir a ser instrumento de trabalho futuro onde a ferramenta poderá ser melhorada e aproveitada.

Para testar o programa criado foi utilizado um caso concreto da empresa. Esse estudo consistiu em determinar quais os atuais clientes que mais contribuíram para a evolução da empresa e as variáveis que caracterizaram melhor esses clientes nos últimos 3 anos de 2013 a 2015. Aplicando o caso referido no programa criado, obtiveram-se resultados e informações que foram analisadas e interpretadas. Com essas análises verificou-se a eficiência do programa e da técnica utilizada, dando assim um contributo para a criação de um conjunto de ferramentas capazes de auxiliar a empresa na definição de futuras estratégias de negócio.

Palavras-chaves: Quidgest; Excel; VBA; ECBD; Quigenio; data mining; análise cluster; interface gráfica; software

Abstract

This work aims to describe the activities performed during the internship accomplished in Quidgest company. This is a company located in Lisbon of IT and software area. The internship were carried out during the period of 19 October 2015 to 31 January 2016.

This final work of master had as objective a development of a market analysis tool in Quidgest Company. This project involved the computational implementation of several methods of data analysis and was done through several steps, creating a complex process. It was used for the criation of the program the Excel software and with a programming language, Visual Basic of Applications (VBA).

Having the company the need to study several aspects of the business in order to prepare a better strategy internally and externally, it was decided to extract and identify the information contained in the company's database. It was used a process known in the data analysis by Knowledge Discovery in Databases (KDD). The biggest challenge in implementing this process was the large accumulation of information in the company, which has been intensifying since 2013.

The process of Knowledge Discovery in Databases developed in different phases. The extraction of data information is made in the company's databases that is Quigenio. About stages of process KDD, which is most relevant is data mining, where the characterizing variables needed for analysis in focus were studied. From the study of these variables, we gave importance the correlation between them and the importance of the characterizing variables had to study elements.

It was decided from the data mining phase to used cluster analysis techniques so that any analysis can be efficient, effective and can get results easy to read. After the development of KDD process, it was decided that the data mining phase could be automated to facilitate future work of an analysis carried out by the company. To automate this phase, we used cluster analysis techniques and a graphical interface was developed in vba / excel user-centered, so that the company could take advantage of this, without having to deal with the complexity involved. The created and implemented graphical interface is compatible with the software used by the company and the technical capabilities of its resources. It could be an asset to the organization, hoping that might be future working tool where the tool can be improved and used.

To test the created program we used a case of the company. This study has gone through see what current costumers that most contributed for the evolution of the company and the variables that best characterize these customers in the last three years from 2013 to 2015. After being made the analysis of the results by cluster analysis technique that created program holds in its programming, were withdrawn conclusions initially if the program worked well and if it was efficient. Subsequently, were analyzed and interpreted these information of analysis from created program.

Applying that case on created program, we obtained results and information that were analyzed and interpreted. With these analyzes it was verified the efficiency of the program and the technique used, thus giving an contribute to the creation of a set of enterprise capable of assisting in the definition of future business strategies tools.

Keywords: Quidgest; Excel; VBA; KDD; Quigenio; Data mining; Cluster analysis; Graphical interface; Software

Índice

Lista de Tabelas

Lista de Figuras

Introdução

Na sociedade atual verifica-se um grande aumento da quantidade de dados armazenados em meios tecnológicos. Alguns estudos mostram que em média a cada 20 meses as empresas no mundo dupliquem o volume de dados acumulados nos seus computadores [3]. Com o aumento do volume de dados, a análise feita pelos especialistas só se tornou viável com o uso de *software* e tecnologias, podendo assim haver uma melhor gestão das bases de dados e ter uma melhor eficiência das análises feitas [1].

O grande desenvolvimento de *software* acontece, pois o uso dos computadores tem aumentado bastante, como também as diversas áreas do conhecimento humano. Com o aumento das bases de dados, houve uma maior procura de soluções que automatizem diversos processos. Essa automatização tem como objetivo melhorar a qualidade de diversas ferramentas e aumentar a produtividade em diversas áreas.

As instituições detendo muita informação gravada nas bases de dados, deparam-se com outro problema que é a ausência de conhecimento concreto desses dados sem nenhuma conclusão relevante. Consequentemente, aparece a necessidade de se analisar as bases de dados para extrair conhecimento, para serem tomadas decisões importantes e verificar a evolução de certas variáveis e áreas da empresa. Essa necessidade foi o objetivo essencial do estágio realizado, onde se exigiu que fosse feita uma análise dos dados, de modo a responder certas questões e a obter informações relevantes sobre o negócio da empresa.

A informação retirada das bases de dados é adquirida através de um processo estruturado que transforma os dados brutos em conhecimento, denominado por extração de conhecimento em bases de dados. O processo de extração de conhecimento em bases de dados (ECBD) passa por diversas fases na procura de informações existente nas bases de

dados. Das fases desse processo vai-se dar mais atenção à fase de data mining, por ser a fase com maior desenvolvimento atualmente e requer estudos mais aprofundados devido a complexidade de algumas das suas técnicas.

Algumas técnicas usadas na fase de data mining são: a associação, classificação, análise cluster[1], previsão, entre outras. Neste TFM a técnica que se vai abordar será a análise cluster. A análise cluster tem vários métodos que têm sido utilizados em muitas aplicações, tais como reconhecimento de padrões, análise de dados, processamento de imagens e pesquisa de mercado [7]. Na Quidgest existe diversas formas de utilizar essa análise cluster mas é feita de forma empírica por experientes trabalhadores.

Foi feito um estudo da técnica análise cluster, referindo alguns dos seus métodos e algoritmos respetivos. Dos métodos e algoritmos apresentados descreveu-se a razão da escolha do método e algoritmo que se achou mais eficiente para o estudo e fez-se uma demonstração de um caso específico, aplicando-se o método e algoritmo escolhido. Escolhendo o método e algoritmo mais apropriado criou-se um programa de análise cluster. O programa foi criado para dar contributo importante e que futuramente pudesse ser vantajoso na análise à base de dados existente. O programa permite ler uma base de dados, onde o utilizador é que coloca as variáveis que necessita e os respetivos elementos de estudo. O programa foi implementado e estruturado em VBA/Excel contento uma folha de inserção dos dados para estudo e cinco folhas de resultados. O programa criado permite que haja uma análise cluster das bases de dados, sem que a empresa necessite de se preocupar com toda a complexidade computacional existente. Por fim, foi realizado um estudo de um caso concreto da empresa. O caso de estudo consistiu na análise dos

[1] Um cluster é um conjunto de elementos agrupados entre si por uma similaridade num grupo homogéneo.

clientes que a empresa detém, para encontrar o peso destes na evolução da empresa nos anos de 2013 a 2015 e respetiva interpretação de resultados.

Capítulo 1: Quidgest e o seu funcionamento

A empresa Quidgest foi criada em 1988. É uma empresa de origem nacional, de consultoria e desenvolvimento de sistemas de informação de gestão que aposta na investigação em engenharia de *software*. Tem empresas constituídas em Portugal, Timor-Leste e Moçambique, tendo a Quidgest investido com grande sucesso na internacionalização das suas atividades [17]. A empresa tem atualmente cerca de 85 colaboradores. As suas principais áreas funcionais são: gestão da Quidgest, investigação & desenvolvimento, a área de marketing & comunicação, gestão patrimonial, gestão de recursos humanos, gestão financeira, gestão bancária, gestão documental e processos de negócio, gestão de sistemas de saúde, projetos especiais e consultoria de negócio nacional e internacional. [19]

1.1 Plataformas

Existem três grandes plataformas de trabalho e funcionamento da empresa de modo a ter um controlo de gestão da empresa. As plataformas de maior importância são o Genio, o BSC e o Quigenio. Numa primeira abordagem tentou-se perceber como se trabalhava com essas plataformas. [19]

1.1.1 Genio

O Genio é uma plataforma de geração automática de código desenvolvida pela Quidgest desde 1991. Tem como objetivo aumentar a produtividade do desenvolvimento de *software*, reduzindo o tempo de realização de código. Permite a produção de um milhão

de caracteres de código por segundo. É com esta plataforma que a empresa trabalha realizando assim os vários programas pedidos pelos seus clientes.

1.1.2 BSC

A ferramenta *Balanced Scorecard* (BSC), tem como objetivo definir a relação causa-efeito entre ações e as medidas de avaliação de desempenho, interligando os diferentes *key performance indicators* (KPI's) definidos de forma a implementar a estratégia empresarial feito com um mapa estratégico. Os KPI's são indicadores de desempenho, fundamentais e determinantes para alcançar os objetivos estratégicos da organização. [16]

1.1.3 Quigenio

A plataforma Quigenio é uma ERP[2]. Nesses sistemas costumam haver alterações constantes. No Quigenio está armazenado as informações sobre a empresa. As principais dificuldades do ERP e consequentemente no Quigenio referem-se à atualização constante dos dados do sistema mas tendo sido melhorado nos últimos tempos.

1.2 Tarefas na empresa no decorrer do estágio

No decorrer do estágio e a pedido da Quidgest, definiram-se 5 objetivos gerais e foi feita a escolha de um objetivo mais específico. O objetivo mais específico seria verificar quais os clientes que nos últimos 3 anos, de 2013 a 2015, contribuíram para a evolução da empresa. Esse objetivo, seria a solução do primeiro ponto dos 5 objetivos gerais.

[2] ERP do inglês *Enterprise Resource Planning,* é um sistema de informação integrado de gestão empresarial. Pode ser vista como uma grande base de dados onde se armazenam informações que interagem e se realimentam entre si. [18]

1.2.1 Objetivos gerais

Os objetivos gerais definidos para o estágio foram:

- Análise de variáveis da plataforma Quigenio.
- Estudo das competências e formação dos técnicos.
- Estudo das tarefas de desenvolvimento, onde é feito um estudo aprofundado das oportunidades de projeto, das estimativas de orçamentação.
- Estudo das tarefas de investigação, do departamento investigação & desenvolvimento (Genio) da empresa, onde existe uma grande dificuldade de contabilizar o valor dos seus projetos, atividades e trabalho para a empresa.
- Estudo dos incidentes onde existe imprevisibilidade e "relação" com o cliente. Requer um estudo do tratamento dos incidentes e razão do aparecimento destes.

1.2.2 Soluções para os objetivos gerais

De modo a cumprir as tarefas gerais definidas para o estágio e posteriormente alcançar o objetivo mais concreto, foi elaborado um estudo prévio. Com o decorrer do estágio, o conhecimento que foi adquirido da empresa, permitiu encontrar algumas soluções para os problemas identificados:

- Para a análise de variáveis a solução encontrada passou por um estudo mais intensivo, onde foi criada uma base de dados fidedigna e segura com certas variáveis. Essa base de dados foi criada usando o *software* VBA/Excel aprendida na disciplina de computação do mestrado. Utilizando a informação existente na base de dados criaram-se algumas estatísticas usando os conhecimentos da disciplina análise de dados. Também foram criadas tabelas de forma estruturada de maneira a construir gráficos para se poder comparar variáveis de diferentes departamentos ou tipos de projetos.

- No estudo das competências e formação dos técnicos foram realizados inquéritos na empresa para aferir as competências e formação dos trabalhadores. Tendo as competências de cada técnico, o nível dessas competências e as competências necessárias para cada projeto, poderá responder-se a questões tais como: quais os melhores técnicos para a realização de cada projeto tendo em conta certas restrições. Este problema pode ser formalizado usando o *staffing problem* ou mais conhecido em investigação operacional como *assignment problem* de modo a tornar eficiente a escolha de afetação dos técnicos aos projetos.
- No estudo das tarefas de desenvolvimento foi feita uma análise de quais as questões que seriam importantes, existindo uma análise de todo o processo desde o início da oportunidade de projeto até a finalização de um projeto.
- No estudo das tarefas de investigação não se conseguiu encontrar uma solução concreta devido à dificuldade de encontrar um valor acrescentado ao trabalho realizado pelo departamento Génio.
- No estudo dos incidentes, a solução passaria por criar uma espécie de central de triagem no aparecimento de um incidente, para melhorar a rapidez de resolução. Essa central de triagem foi implementada na plataforma Quigenio de maneira estruturada, onde cada incidente teria um "ranking" para ser resolvido.

Capítulo 2: Análise de bases de dados

Na obtenção de informações relevantes das bases de dados, utilizou-se um processo eficiente designado por extração de conhecimento em bases e dados.

2.1 Extração de conhecimento em bases de dados

O processo de extração de conhecimento em bases de dados foi denominado por Fayyad[3] em 1989. Esse processo tem diversas fases iterativas que permite obter conhecimentos das bases de dados [4].

O processo de ECBD tem como objetivo extrair informação de certos dados e, através disso, conseguir retirar conclusões relevantes. No processo de ECBD é necessário traçar um objetivo concreto que se quer estudar ou um problema que se quer solucionar. Para que o processo seja eficiente, a informação retirada terá que ser compreensível e útil para quem está a analisar, o que implica um grande tratamento da base de dados em estudo. O processo ECBD passa por várias fases até se conseguir retirar algum conhecimento dos dados. Essas várias fases são atividades interligadas, cada uma com a sua importância no processo. Essas fases são: seleção dos dados, processamento dos dados, transformação de dados, *Data mining* e interpretação dos resultados obtidos.
A figura 1 ilustra essas fases existentes no processo de ECBD:

[3] Usama M. Fayyad nasceu a julho de 1965 na Tunísia tendo atualmente a nacionalidade americana. É analista e pioneiro no estudo de análises de dados e bases de dados.

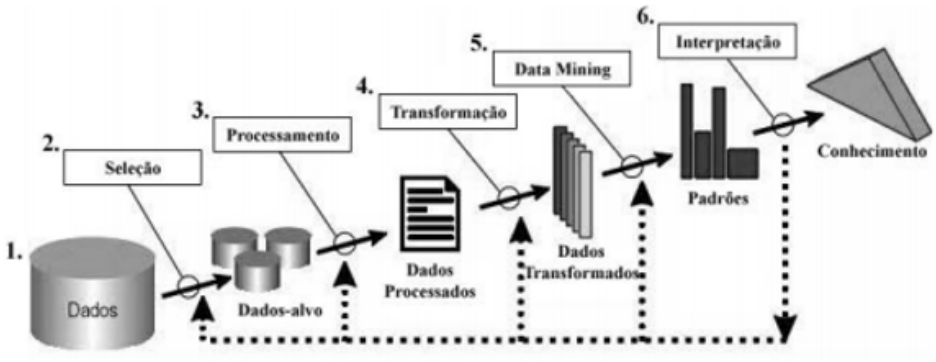


Figura 1: Fases do processo de extração de conhecimento em bases de dados

Ao longo do processo poderá haver interseção entre fases, onde os resultados conseguidos numa fase poderão ser utilizados para aperfeiçoar e complementar os resultados de fases seguintes. Assim mostra que o processo de ECBD é iterativo, procurando aperfeiçoar os resultados a cada iteração. As fases têm diferenças relativamente ao tempo e esforço necessário. Cada uma dessas fases é descrita de seguida.

2.1.1 Seleção dos dados

Esta primeira fase do processo ECBD consiste na extração de dados, dando início ao processo iterativo. A extração de dados é feita com a escolha das variáveis caracterizadoras e elementos mais significativos para se realizar o estudo. Um dos problemas existente nesta fase é a quantidade de variáveis que se escolhe, onde existem opiniões opostas de diversos analistas. Alguns analistas pensam que um aumento do número de variáveis permite obter uma melhor análise e divisão dos clusters. Outros acham que esse aumento do número de variáveis leva a uma análise mais fraca e a uma divisão dos clusters de forma ineficiente. A fase da seleção dos dados é importante, pois permite otimizar o tempo de processamento que existirá nas fases seguintes. Nesta fase procura-se reduzir o número de variáveis, utilizando apenas aquelas com maior ligação aos elementos em estudo.

2.1.1.1 Classificação de variáveis

Em relação a classificação das variáveis que podem ser utilizadas, estas podem ser do tipo qualitativo ou quantitativo. As variáveis qualitativas são as variáveis que não possuem valores quantitativos e são representadas por categorias ou por uma classificação. As variáveis qualitativas estão divididas em nominais ou ordinais. As variáveis nominais são aquelas que não têm ordenação entre as categorias como por exemplo o sexo (masculino, feminino). Nas variáveis ordinais existe uma ordem nas categorias e nos valores viáveis como por exemplo a escolaridade (1ª, 2ª classe).

As variáveis quantitativas são variáveis que assumem valores numéricos. Estas variáveis podem ser divididas em contínuas ou discretas. As variáveis contínuas assumem valores contínuos ou reais no qual os fracionários também pertencem, como por exemplo o peso, altura e idade. As variáveis discretas apenas assumem valores finitos ou infinitos contáveis no qual só pertencem os valores inteiros como por exemplo o número de filhos, o número de empregados ou o número de títulos do Benfica.

2.1.2 Processamento dos dados

A fase de processamento dos dados tem como objetivo garantir a qualidade dos dados escolhidos, para que a informação seja relevante, útil e consiga responder ao problema proposto. Nesta fase faz-se uma eliminação de dados duplicados, tratamento de *outliers*[4], eliminação de valores inexistentes das variáveis, entre outras.

2.1.3 Transformação dos dados

Esta fase poderá ser vantajosa para a fase seguinte que é a fase de data mining porque permitirá ultrapassar algumas limitações existentes quer de espaço de memória quer de

[4] Outliers são valores fora do esperado para uma variável que podem implicar mudanças no estudo e na análise.

tempo de processamento. Nesta fase existem inúmeras técnicas usadas para atingir essa melhoria. Essas técnicas passam por uma redução dos dados, utilizar amostras representativas dos dados, reduzir o número de variáveis caracterizadoras de modo a ficar só com as variáveis mais significativas ou transformar os dados de valores contínuos para discretos.

2.1.4 Data Mining

A fase data mining é uma das fases mais importantes do processo de ECBD, devido à capacidade dessa fase de conseguir retirar informação com relevância da base de dados de acordo com o objetivo pedido. Esta fase é muito importante para encontrar tendências e padrões e relevante para a tomada de decisão em termos de organização e gestão.

2.1.4.1 Conceito:

Data mining é o processo de analisar bases de dados de modo a identificar certos padrões entre variáveis para formar grupos ou subconjuntos desses dados. Na fase data mining utilizam-se diversas ferramentas e técnicas. Com essas técnicas é analisado um conjunto de dados, podendo assim encontrar padrões nesses dados que revelarão informações importantes. Essa informação pode ser apresentada de diversas maneiras tais como: clusters, hipóteses, regras, árvores de decisão, grafos, ou dendrogramas. Para que esta fase de data mining seja eficaz e eficiente requer que as fases de seleção, processamento e transformação dos dados sejam bem conseguidas. Portanto, é essencial a remoção de ruídos, redundâncias e ter as bases de dados organizadas.

2.1.4.2 Tipos de informação obtidos com data mining

Com o uso da data mining, é possível descobrir informações relacionadas com associações, sequências, classificação, análise cluster ou previsões.

- Associações: são padrões frequentes entre conjuntos de dados que detêm uma certa correlação e associação. Por exemplo: num estudo de mercado de supermercados em que na compra de batatas fritas, o cliente compra em 50% das vezes um refrigerante de tipo fanta. Essas variações económicas muitas vezes estudadas detêm informações relevantes que auxiliam na tomada de decisão para atingir melhorias em termos de produtividade, rentabilidade do *stock* e criação de eventuais promoções nos supermercados.
- Sequências: são acontecimentos que ocorrem sequencialmente ao longo do tempo. Exemplo: após a compra de um apartamento, duas semanas depois estima-se que 70% das vezes é comprado uma televisão e um mês depois 30% das vezes é comprado um guarda-roupa. Pode ser útil descobrir padrões sequenciais para campanhas de marketing, instituições de investimentos financeiros, mercado imobiliários entre outros.
- Classificação: é o processo que permite encontrar um conjunto de modelos que descrevem e distinguem classes ou conceitos de elementos. Exemplo: num banco existe o registo dos clientes do histórico de créditos tendo-se classificado como mau, médio e bom. Daí é criado um modelo de classificação de clientes através do histórico de crédito onde a classificação "bom" seria atribuída aos clientes com uma taxa de débito menor que 10%. Assim essa regra poderia ser utilizada para a classificação e reconhecimento desses padrões.
- Análise cluster: consiste em identificar, classificar e agrupar os elementos existentes nos dados que ainda não estão definidos. Exemplo: pode-se aplicar a técnica de análise cluster na base de dados de um supermercado para se identificar

os grupos homogéneos de clientes que o supermercado tem dependendo de diferentes variáveis.

- Previsão: esta técnica é um pouco diferente das anteriores, uma vez que não tem como objetivo identificar e classificar elementos, mas sim estimar e prever valores das variáveis do estudo. Consegue-se essa previsão através da identificação de padrões e tendências nos dados. Exemplo: Permite estimar um valor futuro das vendas de um supermercado analisando tendências e padrões das vendas.

Os tipos de informação que se podem obter com data mining levam a que análise de dados consiga identificar e retirar conclusões sobre padrões, tendências ou outras informações relevantes. Dos tipos de informação possíveis decidiu-se dar mais ênfase à análise cluster devido à necessidade da empresa de encontrar padrões de grupos homogéneos.

2.1.5 Interpretação

A interpretação dos dados é uma fase crítica no processo da ECBD. É uma fase em que é essencial a avaliação dos resultados por parte do analista ou das pessoas que tomam as decisões das instituições, podendo ter que se repetir todo o processo de ECBD ou alterar as técnicas de algumas fases dependendo da qualidade dos resultados ou da informação que se obteve. As técnicas e o processo de ECBD só têm efeito se os resultados tiveram as características de compreensibilidade e importância levando a que se consiga cumprir o objetivo proposto.

Capítulo 3: Análise Cluster

Neste capítulo descrevem-se alguns métodos de análise cluster e respetivos algoritmos. Será analisado qual dos métodos é o mais eficiente para o estudo que se pretende elaborar neste TFM.

A análise cluster é um processo de partição ou aglomeração de uma população com elementos heterogéneos, tendo variáveis caracterizadoras em vários subgrupos mais homogéneos. No processo de análise cluster a partição ou aglomeração é feita de acordo com uma similaridade na qual se faz a distinção entre os elementos. Para uma maior eficiência na análise cluster, os clusters resultantes da análise devem ter uma elevada homogeneidade interna (dentro dos clusters) e uma elevada heterogeneidade externa (entre os clusters) [5].

3.1 Medidas de Similaridade

A medida de similaridade é importante para a definição dos clusters em qualquer dos métodos existentes e técnicas de análise cluster. Existem diversas medidas de similaridade e a sua escolha deve ser feita de acordo com a natureza dos dados e o tamanho destes. A similaridade pode-se subdividir em semelhança ou dissemelhança. Semelhança mede o grau de proximidade entre os elementos. A dissemelhança reflete o grau de diferença ou distância entre dois elementos. Em geral no cálculo da dissemelhança entre dois elementos, utiliza-se uma medida de distância definida no espaço de características. As medidas de similaridade entre elementos são medidas quantitativas.

Dados quaisquer elementos x, y e z, uma medida de similaridade deve seguir estas regras:

1. $d(x, y) \geq 0$, as distâncias são números não negativos.

2. d(x, x) = 0, a distância de um elemento a si próprio é zero.
3. d(x, y) = d(y, x) (simetria), a distância entre dois elementos é independente da ordem.
4. d(x, z) ≤ d(x, y) + d(y, z), conhecida por desigualdade triangular, especifica que a menor distância entre dois pontos é uma reta.

Destacam-se as seguintes medidas de similaridade. [20], [21], [22]

3.1.1 Distância Euclidiana

A distância euclidiana entre dois elementos $x = [x_1, x_2, \ldots, x_p]$ e $y = [y_1, y_2, \ldots, y_p]$, é definida por:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_P - y_P)^2} = \sqrt{\sum_{i=1}^{P}(x_i - y_i)^2}$$

Exemplo: x = (3,4) e y = (1,2)

$$d(x,y) = \sqrt{(3-1)^2 + (4-2)^2} = \sqrt{8} \approx 2{,}83$$

3.1.2 Distância de Manhattan

Considerando a notação anterior a distância de Manhattan é definido por:

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_P - y_P| = \sum_{i=1}^{P} |x_i - y_i|$$

Exemplo: utilizando os pontos do exemplo anterior, temos:

$$d(x,y) = |3-1| + |4-2| = |2| + |2| = 4$$

3.1.3 Distância de Chebychev

A distância de Chebychev é definida por:

d(x,y) = máximo $\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_P - y_P|\}$

Exemplo: Considerando os pontos x = (9,5) e y = (2, 4), a distância Chebychev será

d(x,y) = máximo $(|9 - 2|, |5 - 4|) = |9 - 2| = |7| = 7$

3.1.4 Comparação entre as Medidas de similaridade

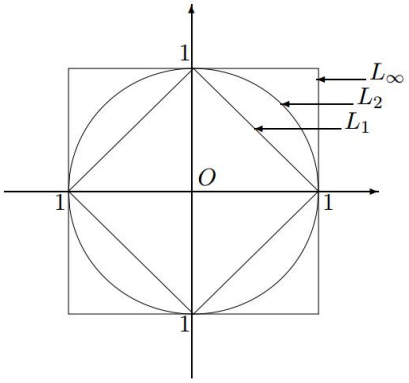


Figura 2: Posição relativa de pontos à distância unitária de um outro ponto O

L1- Manhattan, L2- euclidiana, L_∞- Chebychev

A figura 2 mostra a distância unitária ao ponto O em relação a distância utilizada. [20]
Das medidas de similaridade apresentadas foi escolhida a distância euclidiana para realizar a análise cluster.

A análise cluster pode ser implementada usando 2 tipos de métodos: hierárquicos e não-hierárquicos.

3.2 Métodos Hierárquicos

O método hierárquico origina sucessões de clusters progressivamente mais abrangentes (métodos aglomerativos) ou menos abrangentes (métodos divisivos) que leva a que os clusters sejam formados em etapas sucessivas. No método hierárquico não é necessário

definir-se o número de clusters inicial mas se um elemento entra num cluster não poderá mais o abandonar. Neste método usa-se uma matriz de dados ou de similaridade. [22]

ELEMENTO	X	Y
1	4	3
2	2	7
3	4	7
4	2	3
5	3	5
6	6	1

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{bmatrix} 0 & 4,47 & 4 & 2 & 2,24 & 2,83 \\ 4,47 & 0 & 2 & 4 & 2,24 & 7,21 \\ 4 & 2 & 0 & 4,47 & 2,24 & 6,32 \\ 2 & 4 & 4,47 & 0 & 2,24 & 4,47 \\ 2,24 & 2,24 & 2,24 & 2,24 & 0 & 5 \\ 2,83 & 7,21 & 6,32 & 4,47 & 5 & 0 \end{bmatrix}$$

Figura 3: Exemplo de uma matriz de similaridade

Os diferentes clusters, nos métodos hierárquicos, são normalmente representados por um diagrama bidimensional denominado por dendrograma. O dendrograma é uma representação gráfica do processo de criação dos clusters. Cada ramo representa um elemento e a raiz representa a aglomeração de elementos. O dendrograma permite identificar os clusters agrupados ao longo de todo o processo (tracejado vertical) e observar o incremento nos valores da distância entre os clusters (tracejado horizontal).

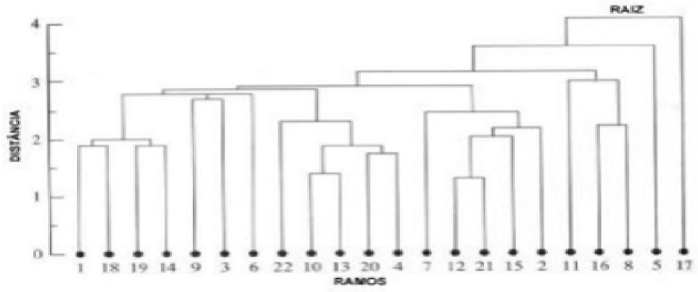


Figura 4: Exemplo de um Dendrograma

Da representação do dendrograma consegue-se retirar informações acerca da estrutura de dados fazendo um corte horizontal no dendrograma.

3.2.1 Métodos Aglomerativos

Os métodos aglomerativos são os métodos hierárquicos mais utilizados. Este método começa com cada elemento no seu próprio cluster e a cada iteração vai-se aglomerando esses elementos, finalizando o método quando existir uma aglomeração de todos os elementos num só cluster ou quando o utilizador impor uma paragem. Essa aglomeração progressiva é feita através de uma medida de similaridade e um algoritmo deste método. Os métodos aglomerativos são:

1) Métodos de ligação: onde existe a ligação simples (*single linkage*) conhecido como método do vizinho mais próximo ou distância menor, ligação completa (*complete linkage*) que é o método do vizinho mais afastado ou distância maior, ligação média (*average linkage*) que passa por uma ligação através da distância média e ligação pela mediana (*median linkage).*
2) Método de centróide
3) Métodos de soma de erros quadráticos ou variância (método de *Ward*).

Os métodos aglomerativos possuem uma desvantagem bastante significativa, devido ao facto de serem inexequíveis para grandes bases de dados pois têm alta complexidade computacional. O método aglomerativo tem uma complexidade de tempo da ordem de O (n^2logn) e a complexidade de espaço da ordem de O (n^2), onde n é o número de elementos [7]. De modo genérico, os métodos aglomerativos utilizam um algoritmo padrão, conforme descrito na figura 5. A diferença entre os métodos ocorre no terceiro passo pois pode não ser a menor distância mas outro tipo de distância, onde a função distância é definida de acordo com cada método.

Entrada: Uma base de dados com N elementos.
Saída: Um conjunto de grupos.
1. Iniciar com N grupos, contendo um elemento em cada grupo e uma matriz de similaridade $D_{N \times N}$;
2. Repetir;
3. Localizar a menor distância d_{UV} (maior similaridade);
4. Atualizar a matriz D, retirando os elementos U e V;
5. Atualizar a matriz D, adicionando as novas distâncias do grupo (U,V);
6. Até N-1, quando todos elementos estarão em um único grupo.

Figura 5: Algoritmo de métodos aglomerativos

3.2.2 Métodos Divisivos

O método divisivo começa com a existência de um único cluster e a cada iteração vai-se dividindo em clusters mais pequenos até existir uma paragem ou até que o número de clusters seja igual ao número de elementos existentes. O método divisivo de forma geral funciona de maneira oposta aos métodos aglomerativos. Na figura 6 estão representadas as diferentes iterações e a comparação entre o método aglomerativo e o método divisivo:

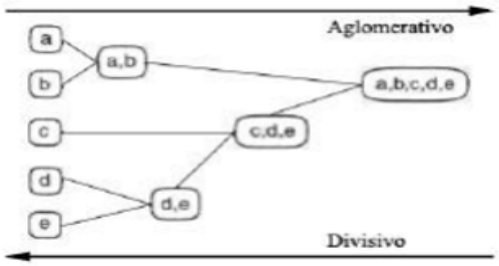


Figura 6: Método aglomerativo e divisivo

Analisando a figura 6 consegue-se verificar que o método divisivo exige mais iterações do que o método aglomerativo. O método divisivo é fundamentado pela lógica de que todas as partições dos elementos é feita em dois clusters até que o número de clusters seja igual ao número de elementos ou que exista uma paragem pelo analista. Essa lógica leva a uma complexidade de tempo da ordem de O (2^n). Este número cresce exponencialmente à medida que n aumenta, sendo n o número de elementos, tornando o método inexequível numa implementação computacional para grandes bases de dados.

3.3 Escolha do número de clusters

Dos diversos problemas da análise cluster, um dos que possui vários estudos é a escolha do número de clusters. Nos métodos hierárquicos essa escolha só é feita no fim do método e é preciso analisar a representação dos resultados num dendrograma, utilizando a técnica de corte. Nos métodos não-hierárquicos a escolha do número de clusters é feita no início do estudo. Essa escolha no início provoca muitas dificuldades ao analista, pois à partida não conhece o padrão dos dados que detém nem nenhuma informação pertinente.

3.4 Métodos Não-Hierárquicos

Nos métodos não-hierárquicos é necessário escolher o número de clusters no início, mantendo assim esse número fixo, originando partições através dessa informação. Neste método agrupa-se os n elementos em k clusters inicialmente e a cada iteração é feita uma melhoria e mudança dos elementos entre clusters. O processo é finalizado quando cada elemento estiver no cluster mais apropriado em relação a técnica usada. Deve-se optar pelo valor k que melhor represente uma divisão de clusters e onde se consiga fazer uma melhor análise ou representação gráfica [2]. Dos métodos não-hierárquicos, os que se vão abordar serão o método k-medoid e o método k-means.

3.4.1 k-medoid

No método k-medoid começa-se por se escolher k elementos aleatórios para os k clusters fixados inicialmente. Usa-se o valor desses elementos como um valor de referência, denominado de medóide. Esse medóide será o elemento central do cluster. Tendo esse medóide em cada cluster vai-se juntando cada elemento ao cluster, escolhendo o mais próximo ao medóide. Ao longo desse processo iterativo, vai-se mudando um dos medóides por um elemento que não seja medóide de modo a analisar e verificar se o cluster melhora de qualidade e homogeneidade.

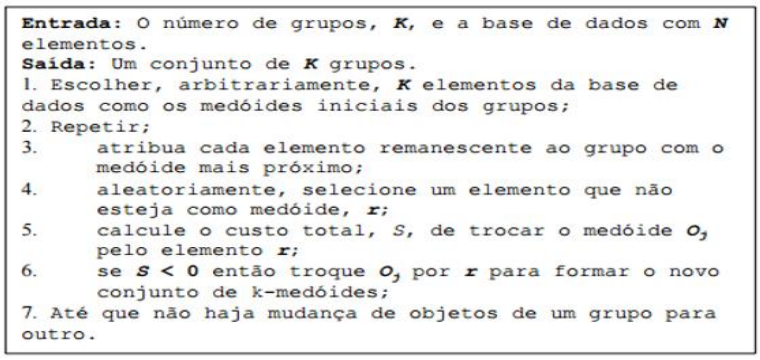


Figura 7: Algoritmo k-medoid

Neste método as características que mais se destacam são: os resultados mantém-se inalterados independentemente da ordem dos elementos na base de dados, os clusters formados têm a propensão em criar grupos esféricos, o tempo de processamento é elevado para grandes bases de dados e o método é robusto na presença de ruídos uma vez que o medóide é menos influenciado pelos ruídos do que a média.

Devido ao grande tempo de processamento para grandes bases de dados, desenvolveram-se formas de otimizar esse tempo e o próprio método. Uma maneira de otimizar o método k-medoid para grandes bases de dados é obter uma amostra representativa dos dados em estudo e escolher os medóides dessa amostra [8].

3.4.2 Método k-means

Um dos métodos não-hierárquicos mais usados é o método k-means, que analisa a proximidade entre clusters usando uma medida de similaridade entre os centros dos clusters. O método k-means tem como critério de entrada, os N elementos e as suas variáveis caracterizadoras e os k clusters escolhidos inicialmente que se quer agrupar.

Veja-se o algoritmo na figura 8:

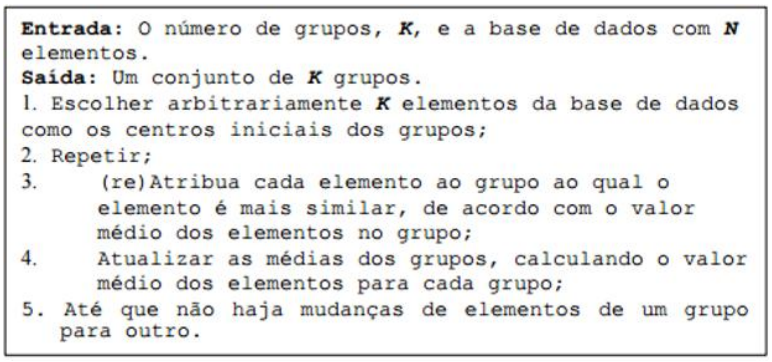


Figura 8: Algoritmo k-means

Tendo a informação do número de clusters, atribui-se k elementos aleatórios a cada cluster de modo que inicialmente esses elementos sejam considerados os centróides de cada cluster. Tendo k elementos como centróides, associa-se cada elemento ao cluster, escolhendo aquele que é mais similar ao centróide. De seguida, vai se atualizando o centróide ao longo das iterações e reagrupam-se os elementos de cada cluster que é mais similar. Repete-se todo o processo até não haver mais mudanças. O método k-means tem uma complexidade de tempo de ordem de O(nkl), do qual n é o número de elementos, k é o número de clusters e l o número de iterações do algoritmo [7].

Exemplo: considerando os elementos e variáveis da tabela I, assumindo que k=2, atribui-se aleatoriamente os centróides aos 2 clusters sendo os primeiros k elementos.

Elemento	Coordenada X	Coordenada Y
A	4	3
B	2	7
C	4	7
D	2	3

Tabela I: Elementos e as suas variáveis

Inicialmente o centróide C1 do primeiro cluster é o elemento A com (4,3) e o centróide C2 do segundo cluster é o elemento B com (2,7). Calculando a distância euclidiana de cada elemento ao centróide obtém-se os resultados contidos na seguinte tabela.

Distância		**Valor**
d(A, C1)	$\sqrt{(4-4)^2+(3-3)^2}$	0
d(A, C2)	$\sqrt{(4-2)^2+(3-7)^2}$	4,47
d(B, C1)	$\sqrt{(2-4)^2+(7-3)^2}$	4,47
d(B, C2)	$\sqrt{(2-2)^2+(7-7)^2}$	0
d(C, C1)	$\sqrt{(4-4)^2+(7-3)^2}$	4
d(C, C2)	$\sqrt{(4-2)^2+(7-7)^2}$	2
d(D, C1)	$\sqrt{(2-4)^2+(3-3)^2}$	2
d(D, C2)	$\sqrt{(2-2)^2+(3-7)^2}$	4

Tabela II: Distâncias euclidianas entre os elementos iniciais

Nesta tabela II, verifica-se que o elemento C está mais próximo do cluster 2 e o elemento D está mais próximo do cluster 1 originando assim dois novos clusters. O primeiro cluster ficará com os elementos {A,D} e o segundo cluster ficará com os elementos {B,C}. Calculando os centróides de cada cluster, temos:

Cluster	**X**	**Y**
(A,D)	$\frac{(4+2)}{2}=3$	$\frac{(3+3)}{2}=3$
(B,C)	$\frac{(2+4)}{2}=3$	$\frac{(7+7)}{2}=7$

Tabela III: Cálculo dos centróides de cada cluster

Calculando a distância euclidiana de cada elemento ao centróide de cada cluster de modo a verificar a qual cluster é mais similar esse elemento.

Distância		**Valor**
d(A, C1)	$\sqrt{(4-3)^2+(3-3)^2}$	1
d(A, C2)	$\sqrt{(4-3)^2+(3-7)^2}$	4,12
d(B, C1)	$\sqrt{(2-3)^2+(7-3)^2}$	4,12
d(B, C2)	$\sqrt{(2-3)^2+(7-7)^2}$	1
d(C, C1)	$\sqrt{(4-3)^2+(7-3)^2}$	4,12
d(C, C2)	$\sqrt{(4-3)^2+(7-7)^2}$	1
d(D, C1)	$\sqrt{(2-3)^2+(3-3)^2}$	1
d(D, C2)	$\sqrt{(2-3)^2+(3-7)^2}$	4,12

Tabela IV: Distâncias euclidianas dos elementos aos centróides de cada cluster

Na tabela IV observa-se que cada elemento está corretamente associado ao cluster com o centróide mais próximo. Como não se verifica nenhuma mudança dos elementos, dá-se por finalizado o processo do algoritmo k-means com os clusters {A,D} e {B,C}.

Neste método as características que mais se destacam são: a existência de alguma sensibilidade a ruídos caso exista um elemento com um valor extremamente alto, a tendência a formar clusters esféricos, o número de clusters é o mesmo durante todo o processo e este método não é apropriado para descobrir clusters com formas não convexas ou de tamanhos muito distintos.

Dos métodos não-hierárquicos e dos métodos hierárquicos apresentados, foi decidido utilizar no trabalho o método não-hierárquico e o algoritmo k-means, devido às seguintes características: menor complexidade computacional, fácil implementação em grandes bases de dados e por ter o conhecimento do algoritmo k-means, uma vez que que foi um dos algoritmos abordado no mestrado.

Capítulo 4: Interface gráfica/ Programa de Análise Cluster

Depois de se fazer um estudo aprofundado das técnicas de data mining, escolheu-se a técnica que melhor se adequava ao estudo e aos tipos de dados da empresa. Essa técnica escolhida foi a análise cluster e aí decidiu-se a aplica-la com um algoritmo do método não-hierárquico que é o *k-means*. Essa aplicação foi feita através da criação de uma interface gráfica (programa) de modo a que a empresa não precisasse de conhecer os detalhes da implementação do algoritmo e onde pudesse obter resultados de fácil leitura. Esse programa foi criado em VBA Excel. Em seguida irá ser explicada como foi feita a programação e como está estruturada o programa de análise cluster.

4.1 Estrutura do programa

O programa está estruturado em seis folhas de Excel. Na primeira folha é feita a leitura dos dados que se quer estudar e inicia-se a análise através do botão que foi criado tendo como nome "Análise Cluster Atualização" (Figura B 3). Após a execução, os resultados irão aparecer nas cinco folhas de resultados existentes. A programação em VBA pode ser consultada no Anexo B.

4.2 Folha de resultados 1

Na folha de resultados 1, será apresentado quais os elementos obtidos para cada cluster. Foi feita a programação da folha de resultados 1 de modo a facilitar a sua apresentação dos resultados, tendo-se gravado numa matriz o elemento pertencente a cada cluster como se verifica no Anexo C.

4.3 Folha de resultados 2

Na folha de resultados 2 são apresentadas algumas estatísticas e a caracterização da qualidade da análise cluster. As caracterizações de qualidade da análise que estão

apresentadas são: a média de distâncias dentro dos clusters e a distância euclidiana entre os centros de cada cluster. Também são apresentados os seguintes resultados: o número de elementos presentes na análise, o número de variáveis caracterizadoras, o número de clusters escolhidos para o estudo, o número de iterações que são encontradas através do número de vezes que se utiliza a etapa de organização dos elementos aos seus clusters e o número de elementos que contém cada cluster. (Anexo D)

4.4 Folha de resultados 3

Na folha de resultados 3 denominada por ANOVA estão representadas tabelas Anova quer a nível dos clusters quer a nível das variáveis caracterizadoras. Nesta folha de resultados está também representada a média do valor de cada variável caracterizadora em cada cluster retratado pelos seus elementos e um gráfico de barras a representar esses valores.

4.4.1 Anova

Anova é a análise de variância em português, é conhecida como *analysis of variance*. A Anova foi desenvolvida por *Ronald Fisher* e é usada na análise de vários estudos comparativos e com associação. A significância desse estudo é definida através de uma relação de dois desvios. A relação existente é independente de certas transformações que se possa fazer aos elementos do estudo. Para testar a significância estatística do estudo terá que se fazer um teste F. [12], [13], [14], [15]

4.4.2 Teste F

O teste F é um teste estatístico que detém como estatística de teste uma distribuição conhecida como F-Snedecor sob a hipótese nula. É normalmente utilizado para se fazer comparações estatísticas entre modelos de conjuntos de dados e para se poder verificar

qual o modelo que se enquadra melhor na população do qual se retirou os dados. Este teste desempenha um papel importante na Anova como já foi referido para testar os diferentes clusters como por exemplo. [25]

Na Anova a fórmula do teste F é:

$$F= \frac{valor\ da\ variação\ entre\ os\ grupos}{valor\ da\ variação\ dentro\ dos\ grupos} = \frac{MS\ Between\ Group}{MS\ Within\ group}$$

4.4.3 Teste de hipóteses:

No teste F da análise a hipótese nula (H_0) é as médias populacionais serem iguais. A hipótese alternativa (H_1) é as médias populacionais serem diferentes, ou seja, pelo menos uma das médias é diferente das demais. Essa análise de variância compara as médias de vários grupos ao mesmo tempo para verificar se elas possuem médias iguais ou não.

4.4.4 Tabelas Anova

Tabelas exemplo de como foi programado e está representado na folha de resultados:

Summary	Contagem	Soma	Média	Variância
Cluster 1 ou variável 1				
Cluster 2 ou variável 2				
…				

Tabela V: 1º tipo de Anova

Source of Variation	SS	DF	MS	F	P-value	F critic
Between Groups						
Within Groups						
Total						

Tabela VI: 2º tipo de Anova

Estas duas tabelas anteriores representam exemplos da Anova que aparecerem na folha de resultados. [12], [13], [14], [15]

As variáveis existentes nas tabelas são:

- A contagem sendo um número fixo, que consiste no número K de clusters escolhido ou o número das variáveis caracterizadoras inseridas na base de dados.
- A soma que é o somatório de todas as distâncias dos elementos ou de cada cluster ou de cada variável caracterizadora.
- Existindo K clusters, i=1…, K, cada um com n_i elementos e n= $n_{1+\ldots}n_k$. X_{ij} com j= 1 … n_i e i=1…, K .
- A Média do cluster = $\overline{X}i = \frac{\sum_j xij}{n_i}$ que será a relação entre a soma e o número de variáveis do cluster i. A média total é $\overline{X} = \frac{\sum_{ij} xij}{n}$
- A Variância do cluster: $Si^2 = \frac{\sum_j (Xij - \overline{Xi})^2}{n_i - 1}$, que representa a razão da diferença da distância de cada elemento para o valor médio ao quadrado e o número de elementos menos um. Variância total: $S^2 = \frac{\sum_{ij} (Xij - \overline{X})^2}{n-1}$
- SS= *sum of square* = soma dos quadrados. SS_{Total}= SS_B + SS_W. SS_B é o SS relacionado com *between groups* e SS_W com *within groups*. $SS_B = \sum_i n_i (\overline{Xi} - \overline{X})^2$. $SS_W = \sum_i \sum_j (Xij - \overline{Xi})^2$ Sendo o somatório das diferenças entre o valor de cada elemento e a média do valor de cada cluster ou variável caracterizadora ao quadrado.
- DF= *degrees of freedom* = graus de liberdade. Graus de liberdade é o número de valores no cálculo final de uma estatística que são livres para variar. O DF varia dependendo de certos valores e variáveis. O DF sobre o estudo dos clusters: DF

between groups = (K-1) No qual K = número de clusters escolhido para o estudo. DF *within groups* = (n_1-1)+…+ (n_K-1) .

- MS = *mean sum of square* = Média da soma do quadrado. $MS_B = \frac{SS_B}{DF_B}$ É o MS relacionado com *Between Groups* e $MS_W = \frac{SS_W}{DF_W}$ É o MS relacionado com *Within Groups*.

O Teste F: $F = \frac{MS_B}{MS_W}$. Neste teste requer uma comparação entre o valor do teste e um valor que prova a significância estatística desse teste. Essa comparação pode ser feita através do p-value do teste em comparação com um α sendo o nível de significância normalmente fixada em 0,05 ou 0,1. A comparação também pode ser feita com a comparação entre o valor do F observado no teste e o F critic.

P-value: Este p-value é a probabilidade de se obter um resultado igual ou maior do que o observado no teste. Para a hipótese H_0 ser rejeitada, o p-value observado terá que ser menor ou igual que α = 0,05 ou 0,1 em casos mais extremos. Na programação utilizou-se uma função da distribuição F com constantes para esse cálculo que é o valor observado em F e os dois tipos de graus de liberdade quer do *between groups* quer do *within groups*.

F Critic: O F critic é outra forma de testar a significância estatística do teste. Quando o valor observado do cálculo de F for maior que o F critic então aí rejeita-se a hipótese H_0. Na programação feita calcula-se o F critic sendo a função inversa da distribuição F com constantes de α = 0,05 e os graus de liberdade respetivos. Toda a programação feita para o cálculo dessas variáveis encontra-se no Anexo E.

4.5 Folha de resultados 4

A folha de resultados 4 contém na primeira linha as variáveis que irão ser representadas em cada coluna. A primeira coluna tem os elementos a serem estudados e na segunda coluna o "id" de cada elemento utilizado na programação. Na terceira coluna será indicado o número do cluster que cada elemento pertence. Nas restantes colunas aparecerão o cluster escolhido no estudo e a distância de cada elemento ao centro desse cluster, onde foi programado para que a distância menor de cada elemento esteja sombreada a amarelo. Aparecerá o nome de cada variável caracterizadora usada na análise e o valor usado respetivamente. (Anexo F)

4.6 Folha de resultados 5

Na folha de resultados 5 estão representadas as variáveis caracterizadoras em duas tabelas distintas. Numa tabela estão os elementos com maior valor em cada variável caracterizadora e o seu valor respetivo e noutra tabela está os elementos com menor valor de cada variável caracterizadora. Para representar melhor essa divisão os que têm maior valor estarão sombreados a verde e os que têm menor valor estarão a vermelho. (Anexo G)

Capítulo 5: Caso de Estudo

Criado o programa de análise cluster decidiu-se aplicá-lo num caso de estudo concreto para se obter respostas e informações. Para se conseguir retirar informações relevantes sobre o caso de estudo e utilizar-se esse programa criado de forma eficiente, tinha-se que passar por todo o processo de ECBD já desenvolvido e explicado. O caso de estudo passou por retirar informações sobre os clientes da empresa que nos anos de 2013 a 2015 inclusive mais contribuíram ou não para a evolução da empresa.

5.1 Seleção dos dados

No início deste caso de estudo começou-se por escolher as variáveis caracterizadoras dos clientes que poderiam ser relevantes para o estudo. Essas variáveis e os seus valores foram retiradas da plataforma Quigenio através da lógica e conhecimento do funcionamento da empresa. Essas variáveis caracterizadoras retiradas são: valor de faturação, horas soma da empresa usadas para a faturação, incidentes totais, horas não faturadas, valor em projetos, número de projetos, horas em projetos e a média de avaliação dos projetos.

5.2 Processamento dos dados

Nesta fase começou-se por melhorar a qualidade dos dados que se tinha retirado em bruto. Depois de se ter uma base de dados, decidiu-se eliminar os dados duplicados, eliminar os clientes que não tinham nenhum valor nessas características mostrando que nada contribuíram para empresa nos últimos anos.

5.3 Transformação dos dados

Nesta fase de transformação de dados, tentou-se reduzir o número de variáveis caracterizadoras e melhorar a velocidade de processamento do programa. As variáveis que se retiraram do estudo foram: número de projetos, horas em projetos e a média de

avaliação dos projetos. Dos dados selecionados, depois processados e transformados ficaram-se com 137 clientes em estudo e com 5 variáveis caracterizadoras. As variáveis caracterizadoras dos clientes relativamente aos anos de 2013 a 2015 utilizadas são: o valor de faturação da empresa em relação ao cliente, as horas utilizadas com o cliente, os incidentes totais existentes com o cliente, as horas não faturadas com o cliente e o valor ganho em projetos da empresa pelo cliente. Ficando com cinco variáveis caracterizadoras, fez-se um teste de correlação das variáveis para verificar e manter-se as mais relevantes e quais as variáveis que tinham menos correlação entre elas. A Figura H1 demonstra esse teste de correlação feito.

Feito o estudo da correlação, foram retiradas as variáveis que tivessem correlações superiores a 0,5. Como se verifica na Figura H 1, é assinalado a amarelo os valores das variáveis que são superiores a 0,5. Das variáveis em estudo verifica-se que as horas soma 2013-2015, os incidentes totais e as horas não faturadas são as mais correlacionadas com restantes variáveis. No estudo ficou-se com as variáveis valor de faturação 2013-2015 que seria uma variável obrigatória para o estudo, o valor em projetos 2013-2015 e com 137 clientes. Com a base de dados decidiu-se aplicar o programa e verificar os resultados.

5.4 Discussão dos resultados

Classificaram-se os clientes da empresa como "maus", "razoáveis" e "bons" relativamente ao investimento e ao peso que têm na faturação da Quidgest. Assim escolheram-se inicialmente 3 clusters para agrupar os clientes. Após a execução do programa, pode-se ler os resultados na folha 1 (Figura H 7). Constata-se que no cluster 1 existem 104 clientes, no cluster 2 existem 28 clientes e no cluster 3 existem 5 clientes.

Na folha de resultados2 (Figura H 8) verificam-se: que ocorreram doze iterações durante a utilização do algoritmo k-means. Esse número de iterações representa o número de vezes que se precisou para agrupar e modificar os elementos pelos clusters. Outras informações como a média de distâncias dentro dos clusters mostram que o cluster 2 é o cluster que tem a menor média de distância dentro dos clusters, o cluster 1 é o que tem maior média e o cluster 3 é o intermédio. Dos resultados das distâncias dentro dos clusters, uma das interpretações que se pode retirar é que o cluster 2 é o cluster que tem maior homogeneidade interna. Relativamente às distâncias entre os clusters verifica-se que o cluster 1 e o cluster 3 têm o maior afastamento. Uma possível interpretação é que os clusters 1 e 3 são os mais heterogéneos do estudo.

Na folha de resultados 3 consegue-se verificar as informações mais relevantes para este estudo de caso. Consegue-se aqui dividir os clusters e caracteriza-los. O cluster 1 representa os clientes "maus" que foram os que menos investiram e contribuíram para a evolução da empresa, o cluster 2 representa os clientes "razoáveis" e o cluster 3 representa os clientes "bons" que representaram uma grande vantagem para a empresa como se verifica na figura 9.

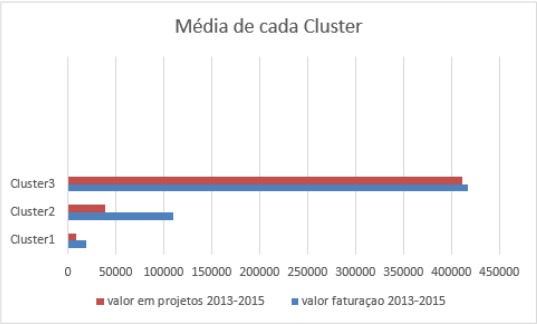


Figura 9: Gráfico da média de valores de cada cluster

Na Figura H 9 está representado a folha de resultados 3. Nessa figura verifica-se que a

média geral do valor do cluster 1 é mais pequena e a do cluster 3 a maior. Na variância de cada cluster verifica-se o oposto, onde o cluster 3 detém uma menor variância e o cluster 1 detém uma maior variância. Esses resultados representam que o cluster 3 é mais convexo e existe menor variabilidade entre os clientes do cluster 3 do que no cluster 1. Foi feito o estudo do teste F que permite comparar as médias destes 3 diferentes clusters, para verificar se eles possuem médias iguais ou não, de modo a apurar se existe divisão desses clusters. Para verificar a significância dos clusters formados como referido o valor de F observado tem que ser maior que o F critic ou se o valor p-value for menor que 0,05. Verifica-se nesta folha que o F observado é de 106,954 maior que o F critic que é de 9,55 ou o p-value que é de 0,00163 menor que 0,05 o valor fixado. Conclui-se que existe significância estatística nesta análise cluster feita.

Na folha de resultados 4 estão apresentados os clientes em estudo e a sua distância euclidiana a cada cluster. Existe certa especificidade nesse resultado que é o sombreado a amarelo da distância menor de cada cliente, que representa o cluster que o cliente pertence. (Figuras H 10, H 11, H 12 e H13)

Na folha de resultados 5 visualiza-se os clientes com maior e menor valor em cada variável. Na variável faturação verifica-se que o cliente 43 é o maior investidor com cerca de 633 mil e o cliente 137 o menor investidor com 140. Na variável valor em projetos o cliente 108 investiu mais em projetos com pouco mais de 1,22 milhões e por outro lado o cliente 2 não investiu nada em projetos como se verifica na figura 10.

Clientes com maior resultados nessas variáveis		
	Cliente	
valor faturaçao 2013-2015	Cliente 43	633156
valor em projetos 2013-2015	Cliente 108	1226668
Clientes com menor resultados nessas variáveis		
valor faturaçao 2013-2015	Cliente 137	140
valor em projetos 2013-2015	Cliente 2	0

Figura 10: Folha de resultados 5

Conclusão

Este TFM surgiu no decorrer de um estágio feito na Quidgest motivado pelo aumento da informação existente na base de dados da empresa denominada por Quigenio. Tendo a necessidade de se extrair conhecimento da base de dados, usou-se um processo conhecido por extração de conhecimento em base de dados que é constituído por diversas fases, nomeadamente a fase designada por data mining. Um dos objetivos do estágio passava por classificar os clientes da Quidgest que nos anos 2013 a 2015 inclusive contribuíram para a evolução da empresa. Assim escolheu-se uma técnica de data mining designada por análise cluster.

Com o objetivo de automatizar a data mining, foi criada uma interface gráfica usando a linguagem de programação *VBA/Excel*. A programação feita exigiu um grande estudo e perceção de todas as noções envolvidas e aprendidas quer a nível académico como a nível empresarial. Após uma reflexão acerca dos diversos métodos da análise cluster, decidiu-se implementar em VBA o método k-means usando a distância euclidiana como medida de similaridade.

Na aplicação do método ao caso de estudo consideraram-se 137 clientes com 2 variáveis caracterizadoras. Decidiu-se agrupar os clientes usando 3 clusters. Conclui-se que o cluster com 104 elementos representa os clientes com menor peso na evolução da empresa. O cluster com 28 elementos representa os clientes com um impacto médio e o cluster com 5 elementos representa os clientes que foram cruciais no desenvolvimento da empresa. Com os resultados desta análise, conseguiu-se garantir uma homogeneidade interna e uma heterogeneidade externa, pois a distância entre os elementos de cada cluster

é menor que a distância entre os clusters. Em suporte destas observações, realizou-se um teste F e obteve-se um p-value de 0,00163.

O programa feito em VBA tem algumas limitações. Uma das limitações mais difíceis de resolver consiste na natureza dos dados introduzidos. Estes devem ser tratados a priori para não produzir resultados inesperados na análise cluster.

Para o futuro, o programa construído em VBA poderá ser melhorado e adaptado na empresa. Sendo a empresa Quidgest uma empresa que trabalha no desenvolvimento de software, poderá integrar parte do código desenvolvido neste TFM nos seus programas, pois como foi referido ao longo do trabalho, a maioria das empresas têm grandes bases de dados mas não dispõem de um método para a extração de conhecimento.

Neste trabalho conseguiu-se aplicar conhecimentos aprendidos no mestrado num ambiente empresarial. Foi feita uma iniciação no mundo profissional, ganharam-se novas valências e construi-se uma ferramenta de análise cluster. Espero que tenha dado um contributo positivo à empresa que poderá permitir uma melhoria nos processos de análise de dados e consequentemente na tomada de decisão.

Referências Bibliográficas

[1] Aurélio, Marco & Vellasco, Marley & Lopes, Carlos Henrique (1999). *Descoberta de conhecimento e data mining,* Pontifícia Universidade Católica, Laboratório de Inteligência Computacional Aplicada.

[2] Bussab, Wilton de Oliveira & Miazaki, Édina Shizue & Andrade, Dalton Francisco de (1990). *Introdução à análise de agrupamentos*, São Paulo: Associação Brasileira de Estatística.

[3] Diniz, Carlos Alberto R. & Louzada Neto, Francisco (2000). *Data mining: uma introdução*, São Paulo: Associação Brasileira de Estatística.

[4] Fayyad, Usama M. et al. (1996). *Advances in knowledge discovery and data mining*, Massachusetts: MIT Press.

[5] Hair-Jr, Joseph F. et al (2005). *Análise Multivariada de Dados*,5 ed, Bookman, p.381-419

[6] Jackson, Joyce (2002). *Data mining: a conceptual overview*, Communications of the Association for Information Systems, v. 8, p. 267-296

[7] Jain, A. K. & Murty, M. N. & Flynn, P. J. (Sept. 1999). *Data clustering: a review*. ACM Computing Surveys, New York, v. 31, n. 3, p. 265-323.

[8] Kaufman, Leonard & Rousseeuw, Peter J. (1990). *Finding groups in data: an introduction to cluster analysis*, New York: Wiley.

[9] Sarstedt, Marko & Mooi, Erik (2014). A concise Guide to Market Research. *The Process, Data, and Methods Using IBM SPSS Statistics*, Second Edition: Springer.

[10] Sharma, S. (1996). *Applied Multivariate Techniques*, John Wiley & Sons.

[11] Análise de cluster: Métodos hierárquicos e de particionamento, Tese de Mestrado da Universidade Presbiteriana Mackenzie de São Paulo. Disponível em: http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF

[12] Wikipedia.org, Anova. Disponível em: http://sweet.ua.pt/gladys/ME/Acetatos/Aula10-ANOVA-1xPage.pdf [Acesso em 18/12/2015]

[13] Wikipedia.org, Anova. Disponível em: https://pt.wikipedia.org/wiki/An%C3%A1lise_de_vari%C3%A2ncia [Acesso em 18/12/2015]

[14] Wikipedia.org, Anova. Disponível em: https://en.wikipedia.org/wiki/Analysis_of_variance [Acesso em 18/12/2015]

[15] Wikipedia.org, Anova. Disponível em: https://en.wikipedia.org/wiki/One-way_analysis_of_variance [Acesso em 18/12/2015]

[16] Wikipedia.org, Balanced scorecard. Disponível em: https://pt.wikipedia.org/wiki/Balanced_scorecard [Acesso em 02/11/2015]

[17] Wikipedia.org, História da empresa Quidgest. Disponível em: https://pt.wikipedia.org/wiki/Quidgest [Acesso em 02/11/2015]

[18] Wikipedia.org, Sistema integrado de gestão empresarial. Disponível em: https://pt.wikipedia.org/wiki/Sistema_integrado_de_gest%C3%A3o_empresarial [Acesso em 05/11/2015]

[19] Site da empresa Quidgest.Disponível em: http://www.quidgest.pt/ [Acesso em 02/11/2015]

[20] Slides de análise cluster do IST. Disponível em: https://fenix.tecnico.ulisboa.pt/downloadFile/3779579704252/SlidesACluster.pdf [Acesso em 14/12/2015]

[21] Técnicas de Análise Cluster. Disponível em: http://www.fsma.edu.br/si/edicao4/FSMA_SI_2009_2_Tutorial.pdf [Acesso em 14/12/2015]

[22] Técnicas de Análise Cluster. Disponível em: http://www.modcs.org/wp-content/uploads/2012/10/apresentacao_clustering.pdf [Acesso em 16/12/2015]

[23] Tese de Mestrado do Instituto Superior de Engenharia do Porto. Disponível em: http://www.dei.isep.ipp.pt/~paf/proj/Julho2003/Clustering.pdf

[24] Tese de mestrado da Universidade da Madeira do Departamento de Matemática e Engenharias. Disponível em: http://repositorio.uma.pt/bitstream/10400.13/224/1/GuidaCaldeiraMestrado.pdf

[25] Wikipedia.org, Teste F. Disponível em: https://en.wikipedia.org/wiki/F-test [Acesso em 18/12/2015]

Anexos

Anexo A: Programação feita no Início do programa em VBA Excel

```
Private Sub Workbook_Open()

analise.Show

End Sub
```

Figura A 1: Programação do início do programa

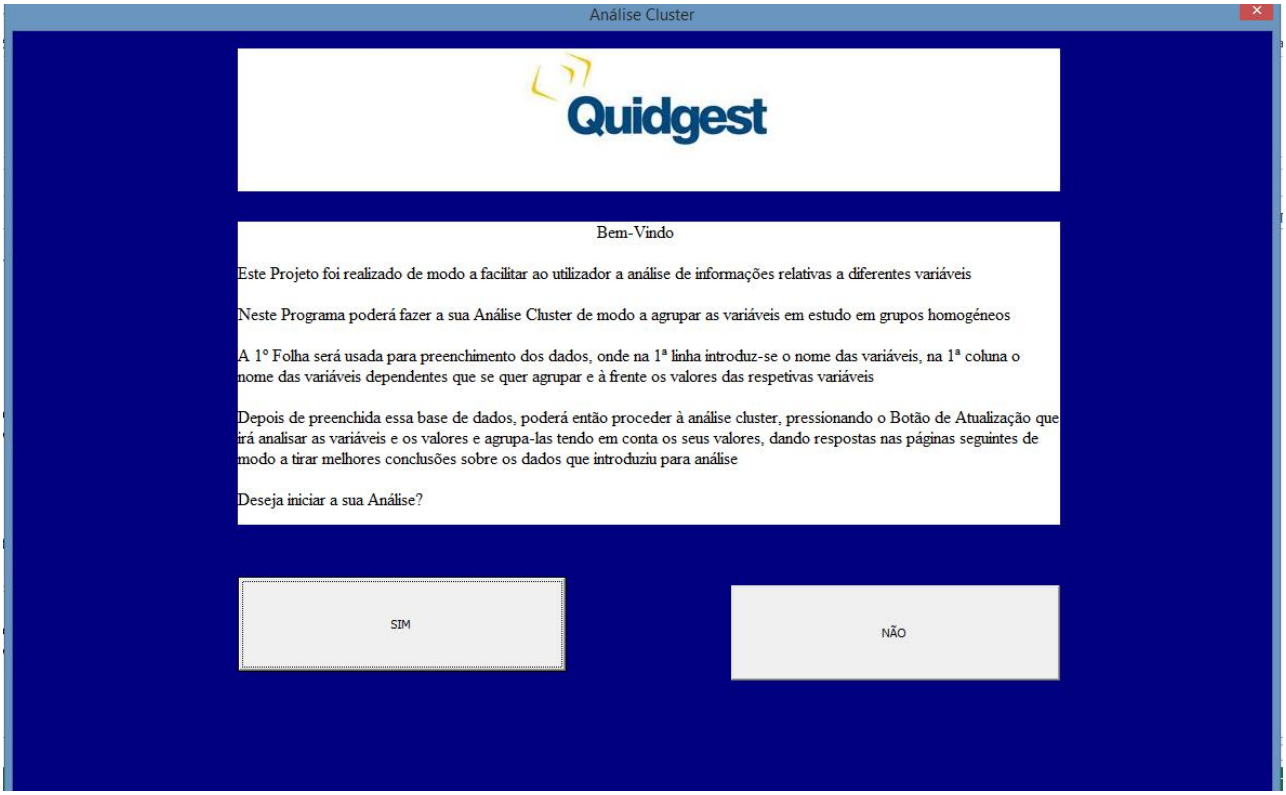


Figura A 2: Mensagem inicial do programa

Anexo B: Folha principal de inserção dos dados:

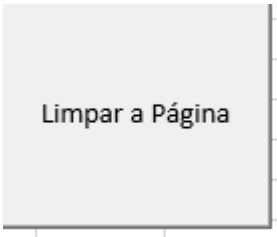


Figura B 1: Botão para limpar a página de inserção dos dados

```
Sub limpar()
Dim op As Long

op = MsgBox("Tem a certeza de querer apagar os dados?" & vbCrLf & "Irá perder os dados", vbOKCancel)

If op = vbOK Then
Sheet1.Select
Cells.Clear
Sheet2.Select
Cells.Clear
Sheet3.Select
Cells.Clear
Sheet4.Select
Cells.Clear
Sheet5.Select
Cells.Clear
Sheet6.Select
Cells.Clear
Sheet1.Select
Else
MsgBox "Dados mantidos"
End If


End Sub
```

Figura B 2: Programação do botão de limpar a página

Figura B 3: Botão que realiza a análise cluster

```
Option Explicit

Sub algoritmo_kmeans()

Dim k, n, i, j, l, a, d, g, h, aux2 As Integer
Dim t(), v(), between, within, mediatotal, MSb, MSw, Fcrit, pvalue As Double
Dim b(), dist(), grupo(), tamanho(), dmin(), igual(), contar(), meddist(), entredist() As Double
Dim entrecent(), aux3, aux4, menor, sum(), average(), variance() As Double
Dim resultado(), c(), variaveis(), aux, res, res2, letra1, letra2 As String

Sheet1.Select

k = InputBox("quantos grupos de cluster?")


'contagem de nº de clientes/variaveis para cluster

n = 0

Do While Cells(n + 2, 1) <> ""
n = n + 1
Loop

'contagem do numero de variáveis
l = 0
Do While Cells(1, l + 2) <> ""
l = l + 1
Loop


ReDim t(1 To n)
```

Figura B 4: Variáveis utilizadas na programação

```
'vetores com o nome de cada cliente e o titulo

aux = Cells(1, 1)

For i = 1 To n
    t(i) = Cells(i + 1, 1)
Next


'ou entao criar uma matriz com os valores das variáveis
ReDim v(1 To n, 1 To l)

For i = 1 To l
    For j = 1 To n
        v(j, i) = Cells(j + 1, i + 1)
    Next
Next


'vetor dos centroides
ReDim c(1 To k)

For i = 1 To k
    c(i) = t(i)
Next
```

Figura B 5: Início da programação da análise cluster

```
'vetor da media de cada grupo k
ReDim b(1 To l, 1 To k) 'linha é o numero de variáveis e coluna é o número de clusters

For d = 1 To l
    For g = 1 To k
        For j = 1 To n
            If c(g) = t(j) Then
            b(d, g) = b(d, g) + v(j, d)
            End If
        Next
    Next
Next


ReDim igual(1 To n)
h = 0

For i = 1 To n
    igual(i) = 0
Next

aux2 = 0

Do While h / n <> 1

aux2 = aux2 + 1 'contagem do nº de iterações do algoritmo
```

Figura B 6: Início do algoritmo k-means

```
'Distância Euclediana
ReDim dist(1 To n, 1 To k)
ReDim grupo(1 To n)

'ter num vetor todas as distancias do ponto da variavel até a media do centroide
'calculo das distâncias

For a = 1 To k
    For i = 1 To n
    dist(i, a) = 0
        For j = 1 To l
            dist(i, a) = dist(i, a) + (v(i, j) - b(j, a)) ^ 2
        Next j
        dist(i, a) = Sqr(dist(i, a))
    Next i
Next a



'saber para cada n qual é o grupo de cada e onde é o cluster do qual tem distancia menor
ReDim dmin(1 To n)

For i = 1 To n
grupo(i) = 1
dmin(i) = dist(i, 1)
    For a = 1 To k - 1
        If dist(i, a + 1) < dmin(i) Then
        dmin(i) = dist(i, a + 1)
        grupo(i) = a + 1
        End If
    Next a
Next i
```

Figura B 7: Medida de similaridade do algoritmo k-means

```
'calcular o tamanho de cada grupo cluster
ReDim tamanho(1 To k)

For i = 1 To k
    For j = 1 To n
        If grupo(j) = i Then
        tamanho(i) = tamanho(i) + 1
        End If
    Next j
Next i
```

Figura B 8: Tamanho de cada cluster

```
'recalcular a média de cada grupo cluster para voltar a calcular as distancias entre as variaveis para cada grupo
ReDim meddist(1 To k)

For a = 1 To k
    For j = 1 To l
        b(j, a) = 0
        For i = 1 To n
            If grupo(i) = a Then
            b(j, a) = b(j, a) + v(i, j)
            meddist(a) = meddist(a) + dist(i, a) 'variavel que contabiliza a media das distancias dos membros de cada cluster
            End If
        Next i
        b(j, a) = b(j, a) / tamanho(a)
    Next j
    meddist(a) = meddist(a) / l
Next a
```

Figura B 9: Cálculo dos centróides de cada cluster

```
'ver se mudou alguma coisa em relação a iteração anterior

h = 0

For i = 1 To n
    If grupo(i) = igual(i) Then
    h = h + 1
    End If
Next

    If h <> n Then
        For i = 1 To n
        igual(i) = grupo(i)
        Next
    End If

Loop
```

Figura B 10: Paragem do algoritmo k-means

Anexo C: Folha 2
Folha de Resultados 1 e a sua programação

```
' TER ATENÇÃO A ESTE CALCULO EM BAIXO

'distância entre clusters
ReDim entredist(1 To k, 1 To k)
ReDim entrecent(1 To k)

For j = 1 To k
    For i = 1 To l
        entrecent(j) = entrecent(j) + b(i, j)
    Next i
    entrecent(j) = entrecent(j) / l
Next j

For i = 1 To k
    For j = 1 To k
    entredist(i, j) = entredist(i, j) + (entrecent(i) - entrecent(j)) ^ 2
    entredist(i, j) = Sqr(entredist(i, j))
    Next
Next
```

Figura C 1: Distância entre os clusters

```
'variavel que grava o nome das variaveis
ReDim variaveis(1 To l)

For i = 1 To l
    variaveis(i) = Cells(1, i + 1)
Next

'mostrar os resultados obtidos do cluster na folha de resultados
ReDim resultado(1 To n, 1 To k)
ReDim contar(1 To k)
```

Figura C 2: Resultados a aparecer na folha de resultados1

```
'FOLHA RESULTADOS1
'meter a folha toda com o formato normal (apagar o que tem anteriormente)
Sheet2.Select

'apagar tudo da pagina
Cells.Clear


    'meter fundo em braco
Cells.Select
    With Selection.Interior
        .Pattern = xlSolid
        .PatternColorIndex = xlAutomatic
        .ThemeColor = xlThemeColorDark1
        .TintAndShade = 0
        .PatternTintAndShade = 0
    End With
```

Figura C 3: Estrutura da folha de resultados1

```
'introduzir o nome dos grupos e das variáveis

Cells(1, 2) = "ANÁLISE CLUSTER(K-MEANS)"
For i = 1 To n
    For j = 1 To k
    Cells(4, j) = "Cluster" & j
    Next j
Next i

'ter numa "matriz" n x k os nomes e o cluster de cada
For i = 1 To n
    For j = 1 To k
        If j = grupo(i) Then
        resultado(i, j) = t(i)
        End If
    Next j
Next i

'para aparecer no excel o resultado do algoritmo
For j = 1 To k
a = 0
    For i = 1 To n
        If j = grupo(i) Then
        Cells(5 + a, j) = t(i)
        a = a + 1
        End If
    Next i
Next j
```

Figura C 4: Elementos de cada cluster

Anexo D: Folha 3

Folha de Resultados 2

```
'FOLHA RESULTADOS2
'Folha nº 3 aparecer os resultados 2ª parte
Sheet3.Select

'apagar tudo da folha
Cells.Clear


    'meter fundo em braco
Cells.Select
    With Selection.Interior
        .Pattern = xlSolid
        .PatternColorIndex = xlAutomatic
        .ThemeColor = xlThemeColorDark1
        .TintAndShade = 0
        .PatternTintAndShade = 0
    End With
```

Figura D 1: Estrutura da folha de resultados2

```
'colocar na sheet3 alguns resultados


'nº de variaveis

Cells(1, 1) = "INPUT DATA"
Cells(2, 1) = "Variável Dependente"
Cells(1, 2) = "Número de Variáveis"
Cells(2, 2) = n
Cells(5, 1) = "Número de Clusters"
Cells(5, 2) = k
Cells(6, 1) = "#Iterações"
Cells(6, 2) = aux2
Cells(8, 1) = "Variáveis"
Cells(9, 1) = "Nº Variáveis Selecionadas"
Cells(9, 2) = l
Cells(13, 1) = "Clusters"
Cells(13, 2) = "Nº de Observações"
Cells(13, 3) = "Média de Distâncias dentro dos Clusters"
```

Figura D 2: Resultados a aparecer na folha de resultados2

```
Cells(16 + k, 1) = "Distância entre os centros dos clusters"
For i = 1 To k
    For j = 1 To k
    Cells(16 + k, j + 1) = "Cluster" & j
    Cells(16 + k + j, 1) = "Cluster" & j
    Cells(16 + k + i, j + 1) = entredist(i, j)
    Next
Next
```

Figura D 3: Distância entre os centros dos clusters

```
'media das distâncias todas


'resultado da media de distancias de cada cluster
For i = 1 To k
    Cells(13 + i, 1) = "Cluster" & i 'introduzir nome de cluster na tabela da media das distancias de cada cluster
    Cells(13 + i, 2) = tamanho(i) 'n de observações
    Cells(13 + i, 3) = meddist(i) 'media da distancia
    For j = 1 To l
        Cells(10, j) = variaveis(j) 'colocar o nome das variaveis na tabela de resultados
    Next
Next

Cells.Select
    With Selection
        .HorizontalAlignment = xlCenter
        .VerticalAlignment = xlBottom
        .WrapText = False
        .Orientation = 0
        .AddIndent = False
        .IndentLevel = 0
        .ShrinkToFit = False
        .ReadingOrder = xlContext
        .MergeCells = False
    End With

Range("A1").Select
```

Figura D 4: Média de distâncias de cada cluster

Anexo E: Folha 4
Folha de resultados 3 (Anova)

```
'Folha nº4 a aparecer
'Resultados da ANOVA E DOS VALORES DA MÉDIA DE CADA GRUPO

Sheet4.Select
Cells.Clear
Cells.Delete


    'meter fundo em braco
Cells.Select
    With Selection.Interior
        .Pattern = xlSolid
        .PatternColorIndex = xlAutomatic
        .ThemeColor = xlThemeColorDark1
        .TintAndShade = 0
        .PatternTintAndShade = 0
    End With


'introdução das respostas em relação as medias das variáveis de cada grupo
Cells(2, 2) = "Média de cada Cluster"

For i = 1 To k
    Cells(3, i + 2) = "Cluster" & i
    For j = 1 To l
        Cells(3 + j, 2) = variaveis(j)
        Cells(3 + j, i + 2) = b(j, i)
    Next
Next
```

Figura E 1:Estrutura da folha de resultados3

```
'ANOVA IMPLEMENTAÇÃO

'Estudo de Grupos de Cluster

ReDim sum(1 To k): ReDim average(1 To k): ReDim variance(1 To k)

Cells(5 + l, 2) = " Anova: Single Factor"
Cells(6 + l, 2) = "Summary"
Cells(6 + l, 3) = "Contagem"
Cells(6 + l, 4) = "Soma"
Cells(6 + l, 5) = "Média"
Cells(6 + l, 6) = "Variância"

'calculo da soma e da média
For i = 1 To k
    Cells(i + 6 + l, 2) = "Cluster" & i
    Cells(i + 6 + l, 3) = l
    For j = 1 To l
        sum(i) = sum(i) + b(j, i)
    Next
    average(i) = sum(i) / l
        Cells(i + 6 + l, 4) = sum(i)
        Cells(i + 6 + l, 5) = average(i)
Next
```

Figura E 2: Soma e média da Anova

```
'verificar variancia
For i = 1 To k
    mediatotal = mediatotal + average(i) 'para calcular between
    For j = 1 To l
    variance(i) = variance(i) + ((b(j, i) - average(i)) ^ 2)

    Next
    variance(i) = variance(i) / (l - 1)
    Cells(i + 6 + l, 6) = variance(i)
Next
mediatotal = mediatotal / k 'soma da media total para calculo do between

Cells(8 + l + k, 2) = "ANOVA": Cells(9 + l + k, 2) = "Source of Variation": Cells(9 + l + k, 3) = "SS": Cells(9 + l + k, 4) = "df"
Cells(9 + l + k, 5) = "MS": Cells(9 + l + k, 6) = "F": Cells(9 + l + k, 7) = "P-value": Cells(9 + l + k, 8) = "F crit"
Cells(10 + l + k, 2) = "Between Groups": Cells(11 + l + k, 2) = "within Groups": Cells(12 + l + k, 2) = "Total"
```

Figura E 3: Variáveis da Anova

```
'Calcular between
Cells(10 + l + k, 4) = k - 1 'df

For i = 1 To k
    between = between + l * ((average(i) - mediatotal) ^ 2)
Next
Cells(10 + l + k, 3) = between 'SS between

MSb = between / (k - 1)
Cells(10 + l + k, 5) = MSb
```

Figura E 4: Between groups

```
'Calcular o Within
For i = 1 To k
    For j = 1 To l
    within = within + ((b(j, i) - average(i)) ^ 2)
    Next
Next
Cells(11 + l + k, 3) = within: Cells(11 + l + k, 4) = k * (l - 1): Cells(12 + l + k, 3) = between + within
Cells(12 + l + k, 4) = (k - 1) + (k * (l - 1))
MSw = within / (k * (l - 1)): Cells(11 + l + k, 5) = MSw: Cells(10 + l + k, 6) = MSb / MSw

pvalue = WorksheetFunction.FDist(MSb / MSw, k - 1, k * (l - 1))
Fcrit = WorksheetFunction.FInv(0.05, k - 1, k * (l - 1))
Cells(10 + l + k, 7) = pvalue
Cells(10 + l + k, 8) = Fcrit
```

Figura E 5: Within groups

```
'Linhas
'Estudo das variáveis
Cells(14 + l + k, 2) = "Summary": Cells(14 + l + k, 3) = "Contagem": Cells(14 + l + k, 4) = "Soma": Cells(14 + l + k, 5) = "Média"
Cells(14 + l + k, 6) = "Variância"

ReDim sum(1 To l): ReDim average(1 To l)

For i = 1 To l
    Cells(14 + l + k + i, 2) = variaveis(i)
    Cells(14 + l + k + i, 3) = k
    For j = 1 To k
        sum(i) = sum(i) + b(i, j)
    Next
    average(i) = sum(i) / k
        Cells(14 + l + k + i, 4) = sum(i)
        Cells(14 + l + k + i, 5) = average(i)
Next
```

Figura E 6: Estudos das variáveis

```
'verificar variancia linhas
ReDim variance(1 To l)

mediatotal = 0
For i = 1 To l
    mediatotal = mediatotal + average(i) 'para calcular between
    For j = 1 To k
    variance(i) = variance(i) + ((b(i, j) - average(i)) ^ 2)

    Next
    variance(i) = variance(i) / (k - 1)
    Cells(14 + l + k + i, 6) = variance(i)
Next
```

Figura E 7: Variância 2ª tabela Anova

```
Cells(16 + 2 * l + k, 2) = "Source of Variation": Cells(16 + 2 * l + k, 3) = "SS": Cells(16 + 2 * l + k, 4) = "df"
Cells(16 + 2 * l + k, 5) = "MS": Cells(16 + 2 * l + k, 6) = "F": Cells(16 + 2 * l + k, 7) = "P-value": Cells(16 + 2 * l + k, 8) = "F crit"
Cells(17 + 2 * l + k, 2) = "Between Groups": Cells(18 + 2 * l + k, 2) = "within Groups": Cells(19 + 2 * l + k, 2) = "Total"
```

Figura E 8: Variáveis da 2ª tabela Anova

```
'Calcular between linhas

Cells(17 + 2 * l + k, 4) = l - 1 'df
mediatotal = mediatotal / l 'soma da media total para calculo do between
between = 0
For i = 1 To l
    between = between + k * ((average(i) - mediatotal) ^ 2)
Next
Cells(17 + 2 * l + k, 3) = between 'SS between

MSb = between / (l - 1)
Cells(17 + 2 * l + k, 5) = MSb
```

Figura E 9: Cálculo Between da 2ª tabela de Anova

```
'Calcular o Within Linhas
within = 0
For i = 1 To l
    For j = 1 To k
    within = within + ((b(i, j) - average(i)) ^ 2)
    Next
Next
Cells(18 + 2 * l + k, 3) = within: Cells(18 + 2 * l + k, 4) = l * (k - 1): Cells(19 + 2 * l + k, 3) = between + within
Cells(19 + 2 * l + k, 4) = (l - 1) + (l * (k - 1))
MSw = within / (l * (k - 1)): Cells(18 + 2 * l + k, 5) = MSw: Cells(17 + 2 * l + k, 6) = MSb / MSw 'F

pvalue = WorksheetFunction.FDist(MSb / MSw, l - 1, l * (k - 1))
Fcrit = WorksheetFunction.FInv(0.05, l - 1, l * (k - 1))
Cells(17 + 2 * l + k, 7) = pvalue
Cells(17 + 2 * l + k, 8) = Fcrit
```

Figura E 10: Cálculo Within da 2ª tabela de Anova

```
Cells.Select
    With Selection
        .HorizontalAlignment = xlCenter
        .VerticalAlignment = xlBottom
        .WrapText = False
        .Orientation = 0
        .AddIndent = False
        .IndentLevel = 0
        .ShrinkToFit = False
        .ReadingOrder = xlContext
        .MergeCells = False
    End With
```

Figura E 11: Formatação da folha de resultados

```
'Gráfico de Barras
'tentativa de grafico

Range("M7").Select


Range("B3:I" & k + 3).Select
    ActiveSheet.Shapes.AddChart2(216, xlBarClustered).Select
    ActiveChart.SetSourceData Source:=Range("ANOVA!$B$3:$I$" & l + 3)

    ActiveChart.ChartTitle.Text = "Média de cada Cluster"
    Selection.Format.TextFrame2.TextRange.Characters.Text = "Média de cada Cluster"
```

Figura E 12: Gráfico de barras

Anexo F: Folha 5: Folha de resultados 4

```
'Folha nº 5 aparecer os resultados 4ª parte
Sheet5.Select

'meter a folha toda com o formato normal (apagar o que tem anteriormente)
 Range("F1000").Select
    Selection.Copy
    Cells.Select
    Selection.PasteSpecial Paste:=xlPasteFormats, Operation:=xlNone, _
        SkipBlanks:=False, Transpose:=False
    Application.CutCopyMode = False
    Range("A1").Select

Cells.Select
    Selection.ClearContents
```

Figura F 1: Formatação da folha de resultados4

```
'resultado da distancia de cada variavel ao ponto médio
Cells(3, 1) = aux
Cells(3, 2) = "Nº de ID"
Cells(3, 3) = "Cluster i"

For i = 1 To l
    Cells(3, 3 + k + i) = variaveis(i)
    For j = 1 To n
    Cells(3 + j, 3 + k + i) = v(j, i)
    Next
Next

For i = 1 To n
    Cells(i + 3, 1) = t(i)
    Cells(i + 3, 2) = i
    Cells(i + 3, 3) = grupo(i)
    For j = 1 To k
    Cells(3, j + 3) = "cluster" & j
    Cells(i + 3, j + 3) = dist(i, j)
    Next
Next
```

Figura F 2: Resultado da distância de cada variável ao centróide do cluster

```
 'meter fundo em braco
 Cells.Select
     With Selection.Interior
         .Pattern = xlSolid
         .PatternColorIndex = xlAutomatic
         .ThemeColor = xlThemeColorDark1
         .TintAndShade = 0
         .PatternTintAndShade = 0
     End With
```

Figura F 3: Aperfeiçoamento da apresentação dos resultados

```
'meter a amarelo as distancias de onde este tem o seu cluster

For i = 1 To n
     For j = 1 To k
          If grupo(i) = j Then
          Cells(i + 3, j + 3).Select

          With Selection.Interior
          .Pattern = xlSolid
          .PatternColorIndex = xlAutomatic
          .Color = 65535
          .TintAndShade = 0
          .PatternTintAndShade = 0
          End With
          End If
     Next
Next

Cells.Select
     With Selection
          .HorizontalAlignment = xlLeft
          .VerticalAlignment = xlBottom
          .WrapText = False
          .Orientation = 0
          .AddIndent = False
          .IndentLevel = 0
          .ShrinkToFit = False
          .ReadingOrder = xlContext
          .MergeCells = False
     End With

Range("A1").Select
```

Figura F 4: Apresentação das distâncias de cada elemento a cada cluster

Anexo G: Folha 6

Folha de resultados 5

```
'Folha de resultados 5
Sheet6.Select

Range("F1000").Select
    Selection.Copy
    Cells.Select
    Selection.PasteSpecial Paste:=xlPasteFormats, Operation:=xlNone, _
        SkipBlanks:=False, Transpose:=False
    Application.CutCopyMode = False
    Range("A1").Select

Cells.Select
    Selection.ClearContents

Cells.Select
    With Selection.Interior
        .Pattern = xlSolid
        .PatternColorIndex = xlAutomatic
        .ThemeColor = xlThemeColorDark1
        .TintAndShade = 0
        .PatternTintAndShade = 0
    End With
```

Figura G 1: Estrutura da folha de resultados5

```
Cells(1, 1) = "Clientes com maior resultados nessas variáveis"
Cells(2, 2) = aux


For i = 1 To l
    aux4 = v(1, i)
    Cells(i + 2, 1) = variaveis(i)
    For j = 1 To n
        If v(j, i) > aux4 Then
        aux4 = v(j, i)
        res = t(j)
        End If
    Next
    Cells(i + 2, 2) = res
    Cells(i + 2, 2).Select
    With Selection.Font
        .Color = -11489280
        .TintAndShade = 0
    End With
    Cells(i + 2, 3) = aux4
Next
```

Figura G 2: Elementos com maiores resultados

```
'menor resultados nas variáveis

Cells(l + 5, 1) = "Clientes com menor resultados nessas variáveis"

For i = 1 To l
    menor = v(1, i)
    Cells(l + 5 + i, 1) = variaveis(i)
    For j = 1 To n
        If v(j, i) < menor Then
        menor = v(j, i)
        res2 = t(j)
        End If
    Next
    Cells(l + 5 + i, 2) = res2
    Cells(l + 5 + i, 2).Select
    With Selection.Font
        .Color = -16776961
        .TintAndShade = 0
    End With
    Cells(l + 5 + i, 3) = menor
Next

Range("A1").Select


MsgBox "Atualizado", vbInformation


End Sub
```

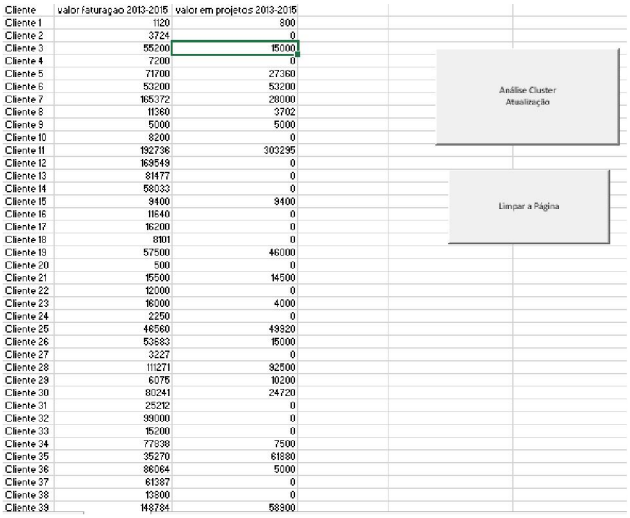
Figura G 3: Elementos com menores resultados

Anexo H

Anexos do caso de estudo

	valor faturaçao 2013-2015	*Horas soma 2013-2015*	*incidentes totais*	*horas não faturadas*	*valor em projetos 2013-2015*
valor faturaçao 2013-2015	1				
Horas soma 2013-2015	0,454590221	1			
incidentes totais	0,384984803	0,504356455	1		
horas não faturadas	0,241246555	0,512564136	0,56118289	1	
valor em projetos 2013-2015	0,653455378	0,221311087	0,090244546	0,080412626	1

Figura H 1: Estudo de correlação das variáveis caracterizadoras



Cliente	valor faturaçao 2013-2015	valor em projetos 2013-2015
Cliente 1	1120	800
Cliente 2	3724	0
Cliente 3	55200	15000
Cliente 4	7200	0
Cliente 5	71700	27360
Cliente 6	53200	53200
Cliente 7	165372	28000
Cliente 8	11360	3702
Cliente 9	5000	5000
Cliente 10	8200	0
Cliente 11	192736	303295
Cliente 12	169549	0
Cliente 13	81477	0
Cliente 14	58033	0
Cliente 15	9400	9400
Cliente 16	11640	0
Cliente 17	16200	0
Cliente 18	8101	0
Cliente 19	57500	46000
Cliente 20	500	0
Cliente 21	15500	14500
Cliente 22	12000	0
Cliente 23	16000	4000
Cliente 24	2250	0
Cliente 25	46560	43820
Cliente 26	53683	15000
Cliente 27	3227	0
Cliente 28	111271	92500
Cliente 29	6075	10200
Cliente 30	88241	24720
Cliente 31	25212	0
Cliente 32	99000	0
Cliente 33	15200	0
Cliente 34	77838	7500
Cliente 35	35270	61880
Cliente 36	86064	5000
Cliente 37	61387	0
Cliente 38	13800	0
Cliente 39	148784	58800

Figura H 2: Página de inserção dos dados parte 1

Cliente 40	121550	0
Cliente 41	20550	0
Cliente 42	10680	3000
Cliente 43	633156	0
Cliente 44	15000	5000
Cliente 45	4028	0
Cliente 46	29320	49440
Cliente 47	4900	0
Cliente 48	3000	0
Cliente 49	27002	4200
Cliente 50	35956	22200,19922
Cliente 51	60316	15360
Cliente 52	180	0
Cliente 53	66800	65320
Cliente 54	5100	5100
Cliente 55	74860	40460
Cliente 56	30175	0
Cliente 57	55958	9900
Cliente 58	7500	0
Cliente 59	102300	85000
Cliente 60	65844	4746
Cliente 61	55265	68000
Cliente 62	5000	0
Cliente 63	65055	63580
Cliente 64	1300	0
Cliente 65	75680	0
Cliente 66	43200	37500
Cliente 67	80389	0
Cliente 68	2620	0
Cliente 69	198613	0
Cliente 70	2640	0
Cliente 71	18000	18000
Cliente 72	23992	17371,19922
Cliente 73	300	0
Cliente 74	2200	0
Cliente 75	13488	0
Cliente 76	9680	0
Cliente 77	17274	4800
Cliente 78	30358	0
Cliente 79	4946	110295,7969

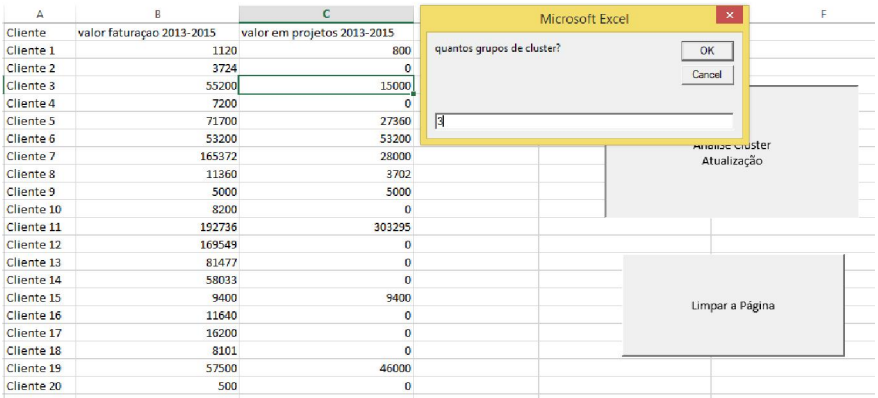
Figura H 3: Página de inserção dos dados parte 2

Cliente 80	12000	16000
Cliente 81	85464	33859,60156
Cliente 82	282427	257480,4688
Cliente 83	9960	3000
Cliente 84	12805	19780
Cliente 85	37800	22800
Cliente 86	31224	16920
Cliente 87	100189	0
Cliente 88	42400	42400
Cliente 89	9850	10000
Cliente 90	3300	0
Cliente 91	25870	0
Cliente 92	180	0
Cliente 93	12096	0
Cliente 94	5200	0
Cliente 95	8280	12600
Cliente 96	49816	0
Cliente 97	7543	0
Cliente 98	1536	0
Cliente 99	125000	2952
Cliente 100	132132	33408
Cliente 101	15954	0
Cliente 102	3572	0
Cliente 103	201975	85641,95313
Cliente 104	3537	0
Cliente 105	108035	179388
Cliente 106	9500	77001
Cliente 107	59625	59625
Cliente 108	613334	1226668,375
Cliente 109	49300	30600
Cliente 110	36225	31225
Cliente 111	33600	19800
Cliente 112	55041	0
Cliente 113	3800	0
Cliente 114	2028	0
Cliente 115	14505	12000
Cliente 116	286965	86250
Cliente 117	15120	0
Cliente 118	8226	0
Cliente 119	15120	0

Figura H 4: Página de inserção dos dados parte 3

Cliente 120	850	0
Cliente 121	51560	24000
Cliente 122	980	0
Cliente 123	43704	0
Cliente 124	6000	0
Cliente 125	52800	0
Cliente 126	6120	0
Cliente 127	25000	0
Cliente 128	14980	0
Cliente 129	22437	22471
Cliente 130	54072	0
Cliente 131	363771	267924
Cliente 132	10400	10000
Cliente 133	17847	0
Cliente 134	5000	0
Cliente 135	5000	0
Cliente 136	2500	0
Cliente 137	140	0

Figura H 5: Página de inserção dos dados parte 4



A	B	C
Cliente	valor faturaçao 2013-2015	valor em projetos 2013-2015
Cliente 1	1120	800
Cliente 2	3724	0
Cliente 3	55200	15000
Cliente 4	7200	0
Cliente 5	71700	27360
Cliente 6	53200	53200
Cliente 7	165372	28000
Cliente 8	11360	3702
Cliente 9	5000	5000
Cliente 10	8200	0
Cliente 11	192736	303295
Cliente 12	169549	0
Cliente 13	81477	0
Cliente 14	58033	0
Cliente 15	9400	9400
Cliente 16	11640	0
Cliente 17	16200	0
Cliente 18	8101	0
Cliente 19	57500	46000
Cliente 20	500	0

Figura H 6: Escolha do número de clusters

ANÁLISE CLUSTER(K-MEANS)

Cluster1	Cluster2	Cluster3
Cliente 1	Cliente 5	Cliente 11
Cliente 2	Cliente 7	Cliente 43
Cliente 3	Cliente 12	Cliente 82
Cliente 4	Cliente 13	Cliente 108
Cliente 6	Cliente 19	Cliente 131
Cliente 8	Cliente 28	
Cliente 9	Cliente 30	
Cliente 10	Cliente 32	
Cliente 14	Cliente 34	
Cliente 15	Cliente 36	
Cliente 16	Cliente 39	
Cliente 17	Cliente 40	
Cliente 18	Cliente 53	
Cliente 20	Cliente 55	
Cliente 21	Cliente 59	
Cliente 22	Cliente 61	
Cliente 23	Cliente 63	
Cliente 24	Cliente 65	
Cliente 25	Cliente 67	
Cliente 26	Cliente 69	
Cliente 27	Cliente 81	
Cliente 29	Cliente 87	
Cliente 31	Cliente 89	
Cliente 33	Cliente 100	
Cliente 35	Cliente 103	
Cliente 37	Cliente 105	
Cliente 38	Cliente 107	
Cliente 41	Cliente 116	
Cliente 42		
Cliente 44		

Figura H 7: Folha de resultados 1

INPUT DATA	Número de Variáveis
Variável Dependente	137
Número de Clusters	3
#Iterações	12
Variáveis	
Nº Variáveis Selecionadas	2
valor faturaçao 2013-2015	valor em projetos 2013-2015

Clusters	Nº de Observações	Média de Distâncias dentro dos Clusters
Cluster1	104	2142466,889
Cluster2	28	1619430,268
Cluster3	5	1909194,743

Distância entre os centros dos clusters	Cluster1	Cluster2	Cluster3
Cluster1	0	60862,51515	400240,4575
Cluster2	60862,51515	0	339377,9423
Cluster3	400240,4575	339377,9423	0

Figura H 8: Folha de resultados 2

Média de cada Cluster

	Cluster1	Cluster2	Cluster3
valor faturaçao 2013-2015	18826,2	110346	417085
valor em projetos 2013-2015	8847,23	39052,3	411070

Anova: Single Factor

Summary	Contagem	Soma	Média	Variância
Cluster1	2	27673,5	13836,7	5E+07
Cluster2	2	149398	74699,2	2,5E+09
Cluster3	2	828154	414077	1,8E+07

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1,9E+11	2	9,3E+10	106,954	0,00163	9,55209
within Groups	2,6E+09	3	8,7E+08			
Total	1,9E+11	5				

Summary	Contagem	Soma	Média	Variância
valor faturaçao 2013-2015	3	546257	182086	4,4E+10
valor em projetos 2013-2015	3	458969	152990	5E+10

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1,3E+09	1	1,3E+09	0,02711	0,87721	7,70865
within Groups	1,9E+11	4	4,7E+10			
Total	1,9E+11	5				

Dados | resultados1 | resultados2 | **ANOVA** | resultados 4 | resultados5

Figura H 9: Folha de resultados 3 (Anova)

Cliente	Nº de ID	Cluster i	cluster1	cluster2	cluster3	valor faturaçao 2013-2015	valor em projetos 2013-2015
Cliente 1	1,00	1,00	19 449,12	115 730,71	584 249,80	1 120,00	800,00
Cliente 2	2,00	1,00	17 502,87	113 548,98	582 962,56	3 724,00	0,00
Cliente 3	3,00	1,00	36 890,49	60 163,23	536 499,50	55 200,00	15 000,00
Cliente 4	4,00	1,00	14 609,67	110 291,51	580 503,01	7 200,00	0,00
Cliente 5	5,00	2,00	56 021,06	40 376,20	516 259,33	71 700,00	27 360,00
Cliente 6	6,00	1,00	56 113,50	58 871,41	510 375,13	53 200,00	53 200,00
Cliente 7	7,00	2,00	147 792,06	56 124,81	458 368,44	165 372,00	28 000,00
Cliente 8	8,00	1,00	9 067,41	105 109,03	574 944,30	11 360,00	3 702,00
Cliente 9	9,00	1,00	14 351,50	110 713,04	578 538,14	5 000,00	5 000,00
Cliente 10	10,00	1,00	13 827,15	109 356,87	579 797,35	8 200,00	0,00
Cliente 11	11,00	3,00	341 970,90	276 789,24	248 893,03	192 736,00	303 295,00
Cliente 12	12,00	2,00	150 982,22	70 922,89	479 845,98	169 549,00	0,00
Cliente 13	13,00	2,00	63 272,38	48 564,51	530 670,13	81 477,00	0,00
Cliente 14	14,00	1,00	40 192,60	65 282,09	545 798,85	58 033,00	0,00
Cliente 15	15,00	1,00	9 442,41	105 211,17	572 315,77	9 400,00	9 400,00
Cliente 16	16,00	1,00	11 398,04	106 150,80	577 376,55	11 640,00	0,00
Cliente 17	17,00	1,00	9 228,79	101 924,41	574 183,61	16 200,00	0,00
Cliente 18	18,00	1,00	13 903,38	109 449,34	579 867,17	8 101,00	0,00
Cliente 19	19,00	2,00	53 628,25	53 300,93	512 422,70	57 500,00	46 000,00
Cliente 20	20,00	1,00	20 350,03	116 581,58	585 253,01	500,00	0,00
Cliente 21	21,00	1,00	6 558,77	97 972,51	564 391,51	15 500,00	14 500,00
Cliente 22	22,00	1,00	11 174,56	105 816,13	577 123,80	12 000,00	0,00
Cliente 23	23,00	1,00	5 610,99	100 647,23	571 467,10	16 000,00	4 000,00
Cliente 24	24,00	1,00	18 789,48	114 934,18	584 008,65	2 250,00	0,00
Cliente 25	25,00	1,00	49 559,41	64 705,36	517 414,38	46 560,00	49 920,00
Cliente 26	26,00	1,00	35 395,64	61 556,72	537 523,93	53 683,00	15 000,00
Cliente 27	27,00	1,00	17 933,47	114 015,79	583 315,07	3 227,00	0,00
Cliente 28	28,00	2,00	124 674,87	53 455,70	441 597,84	111 271,00	92 500,00
Cliente 29	29,00	1,00	12 822,78	108 189,34	574 130,18	6 075,00	10 200,00
Cliente 30	30,00	2,00	63 432,80	33 342,72	512 571,69	80 241,00	24 720,00
Cliente 31	31,00	1,00	10 911,08	93 663,82	567 928,24	25 212,00	0,00
Cliente 32	32,00	2,00	80 660,45	40 667,17	519 765,46	99 000,00	0,00
Cliente 33	33,00	1,00	9 561,54	102 848,81	574 882,23	15 200,00	0,00
Cliente 34	34,00	2,00	59 027,16	45 302,65	527 216,07	77 838,00	7 500,00
Cliente 35	35,00	1,00	55 523,62	78 469,37	517 412,69	35 270,00	61 880,00
Cliente 36	36,00	2,00	67 347,75	41 823,24	523 896,23	86 064,00	5 000,00
Cliente 37	37,00	1,00	43 470,60	62 626,54	543 598,30	61 387,00	0,00
Cliente 38	38,00	1,00	10 175,28	104 145,32	575 861,81	13 800,00	0,00
Cliente 39	39,00	2,00	139 263,43	43 259,65	442 728,73	148 784,00	58 900,00

Figura H 10: Folha de resultados 4 parte 1

Cliente	Nº de ID	Cluster i	cluster1	cluster2	cluster3	valor faturaçao 2013-2015	valor em projetos 2013-2015
Cliente 40	40,00	2,00	103 104,07	40 627,68	506 279,58	121 550,00	0,00
Cliente 41	41,00	1,00	9 013,60	97 920,56	571 155,00	20 550,00	0,00
Cliente 42	42,00	1,00	10 027,51	105 986,39	575 921,55	10 680,00	3 000,00
Cliente 43	43,00	3,00	614 393,48	524 266,34	464 397,41	633 156,00	0,00
Cliente 44	44,00	1,00	5 425,97	101 244,52	571 458,38	15 000,00	5 000,00
Cliente 45	45,00	1,00	17 241,26	113 263,58	582 747,04	4 028,00	0,00
Cliente 46	46,00	1,00	41 927,22	81 689,32	530 224,00	29 320,00	49 440,00
Cliente 47	47,00	1,00	16 498,88	112 445,45	582 129,28	4 900,00	0,00
Cliente 48	48,00	1,00	18 131,27	114 229,09	583 476,15	3 000,00	0,00
Cliente 49	49,00	1,00	9 404,26	90 337,90	563 655,42	27 002,00	4 200,00
Cliente 50	50,00	1,00	21 719,37	76 275,11	544 498,44	35 956,00	22 200,20
Cliente 51	51,00	1,00	41 997,83	55 356,52	532 794,56	60 316,00	15 360,00
Cliente 52	52,00	1,00	20 638,68	116 883,14	585 480,83	180,00	0,00
Cliente 53	53,00	2,00	74 098,97	50 855,30	492 181,07	66 800,00	65 320,00
Cliente 54	54,00	1,00	14 228,52	110 587,15	578 396,72	5 100,00	5 100,00
Cliente 55	55,00	2,00	64 336,24	35 514,09	504 449,47	74 860,00	40 460,00
Cliente 56	56,00	1,00	14 389,87	89 176,79	564 515,18	30 175,00	0,00
Cliente 57	57,00	1,00	37 146,70	61 708,43	539 768,09	55 958,00	9 900,00
Cliente 58	58,00	1,00	14 372,08	110 010,99	580 291,22	7 500,00	0,00
Cliente 59	59,00	2,00	112 991,66	46 646,88	453 222,72	102 300,00	85 000,00
Cliente 60	60,00	1,00	47 196,31	56 190,45	537 093,05	65 844,00	4 746,00
Cliente 61	61,00	2,00	69 475,42	62 224,64	498 608,36	55 265,00	68 000,00
Cliente 62	62,00	1,00	16 414,56	112 351,68	582 058,48	5 000,00	0,00
Cliente 63	63,00	2,00	71 643,39	51 506,30	494 645,31	65 055,00	63 580,00
Cliente 64	64,00	1,00	19 632,68	115 828,11	584 683,84	1 300,00	0,00
Cliente 65	65,00	2,00	57 538,04	52 219,02	534 355,15	75 680,00	0,00
Cliente 66	66,00	1,00	37 617,31	67 164,12	528 530,10	43 200,00	37 500,00
Cliente 67	67,00	2,00	62 195,25	49 219,05	531 358,87	80 389,00	0,00
Cliente 68	68,00	1,00	18 463,89	114 586,26	583 745,89	2 620,00	0,00
Cliente 69	69,00	2,00	180 004,33	96 520,02	465 519,19	198 613,00	0,00
Cliente 70	70,00	1,00	18 446,34	114 567,46	583 731,69	2 640,00	0,00
Cliente 71	71,00	1,00	9 189,98	94 715,45	560 153,87	18 000,00	18 000,00
Cliente 72	72,00	1,00	9 967,11	89 034,34	556 345,54	23 992,00	17 371,20
Cliente 73	73,00	1,00	20 530,33	116 770,05	585 395,39	300,00	0,00
Cliente 74	74,00	1,00	18 833,61	114 981,21	584 044,17	2 200,00	0,00
Cliente 75	75,00	1,00	10 332,96	104 434,62	576 080,35	13 488,00	0,00
Cliente 76	76,00	1,00	12 725,05	107 975,75	578 754,58	9 680,00	0,00
Cliente 77	77,00	1,00	4 334,68	99 174,85	570 003,19	17 274,00	4 800,00
Cliente 78	78,00	1,00	14 534,63	89 012,31	564 389,77	30 358,00	0,00

Figura H 11: Folha de resultados 4 parte 2

Cliente	Nº de ID	Cluster i	cluster1	cluster2	cluster3	valor faturaçao 2013-2015	valor em projetos 2013-2015
Cliente 79	79,00	1,00	102 393,71	127 219,62	510 218,83	4 946,00	110 295,80
Cliente 80	80,00	1,00	9 887,33	101 011,78	565 838,90	12 000,00	16 000,00
Cliente 81	81,00	2,00	72 114,41	24 440,18	501 594,33	86 464,00	33 859,60
Cliente 82	82,00	3,00	362 345,02	278 053,83	204 275,50	282 427,00	257 460,47
Cliente 83	83,00	1,00	10 620,73	106 663,74	576 429,85	9 960,00	3 000,00
Cliente 84	84,00	1,00	12 578,91	99 623,09	562 771,21	12 605,00	19 780,00
Cliente 85	85,00	1,00	23 551,73	74 344,37	542 780,08	37 800,00	22 800,00
Cliente 86	86,00	1,00	14 794,41	82 159,35	551 581,76	31 224,00	16 920,00
Cliente 87	87,00	2,00	81 842,38	40 351,59	519 038,67	100 189,00	0,00
Cliente 88	88,00	1,00	41 006,23	68 028,60	525 648,12	42 400,00	42 400,00
Cliente 89	89,00	1,00	9 049,94	104 611,27	571 574,13	9 850,00	10 000,00
Cliente 90	90,00	1,00	17 870,00	113 947,21	583 263,28	3 300,00	0,00
Cliente 91	91,00	1,00	11 185,30	93 247,72	567 612,31	25 670,00	0,00
Cliente 92	92,00	1,00	20 638,68	116 883,14	585 480,83	180,00	0,00
Cliente 93	93,00	1,00	11 116,18	105 726,91	577 056,43	12 096,00	0,00
Cliente 94	94,00	1,00	16 246,46	112 164,17	581 916,90	5 200,00	0,00
Cliente 95	95,00	1,00	11 194,02	105 438,27	570 875,96	8 280,00	12 600,00
Cliente 96	96,00	1,00	32 227,94	72 034,61	551 239,11	49 816,00	0,00
Cliente 97	97,00	1,00	14 338,22	109 970,80	580 260,87	7 543,00	0,00
Cliente 98	98,00	1,00	19 422,29	115 605,96	584 516,04	1 536,00	0,00
Cliente 99	99,00	2,00	106 337,32	38 961,09	501 869,98	125 000,00	2 952,00
Cliente 100	100,00	2,00	115 937,18	22 505,11	473 102,91	132 132,00	33 408,00
Cliente 101	101,00	1,00	9 301,78	102 151,68	574 355,39	15 954,00	0,00
Cliente 102	102,00	1,00	17 634,19	113 691,72	583 070,34	3 572,00	0,00
Cliente 103	103,00	2,00	198 597,34	102 793,17	390 096,60	201 975,00	85 641,95
Cliente 104	104,00	1,00	17 664,48	113 724,59	583 095,17	3 537,00	0,00
Cliente 105	105,00	2,00	192 463,32	140 354,72	386 248,79	108 035,00	179 388,00
Cliente 106	106,00	1,00	68 788,91	107 749,97	526 998,27	9 500,00	77 001,00
Cliente 107	107,00	2,00	65 137,72	54 734,58	501 289,13	59 625,00	59 625,00
Cliente 108	108,00	3,00	1 355 185,53	1 289 739,77	838 877,32	613 334,00	1 226 668,38
Cliente 109	109,00	1,00	37 441,07	61 628,54	529 171,76	49 300,00	30 600,00
Cliente 110	110,00	1,00	28 345,76	74 533,32	537 899,70	36 225,00	31 225,00
Cliente 111	111,00	1,00	18 390,97	79 124,13	547 861,72	33 600,00	19 800,00
Cliente 112	112,00	1,00	37 279,80	67 703,36	547 771,76	55 041,00	0,00
Cliente 113	113,00	1,00	17 437,34	113 477,62	582 908,67	3 800,00	0,00
Cliente 114	114,00	1,00	18 985,62	115 143,00	584 166,36	2 028,00	0,00
Cliente 115	115,00	1,00	5 349,10	99 585,94	566 857,14	14 505,00	12 000,00
Cliente 116	116,00	2,00	279 087,07	182 816,38	349 912,72	286 965,00	86 250,00
Cliente 117	117,00	1,00	9 592,16	102 922,82	574 938,16	15 120,00	0,00

Figura H 12: Folha de resultados 4 parte 3

Cliente	Nº de ID	Cluster i	cluster1	cluster2	cluster3	valor faturaçao 2013-2015	valor em projetos 2013-2015
Cliente 117	117,00	1,00	9 592,16	102 922,82	574 938,16	15 120,00	0,00
Cliente 118	118,00	1,00	15 396,07	111 202,94	581 191,14	6 226,00	0,00
Cliente 119	119,00	1,00	9 592,16	102 922,82	574 938,16	15 120,00	0,00
Cliente 120	120,00	1,00	20 035,42	116 251,86	585 003,93	850,00	0,00
Cliente 121	121,00	1,00	36 070,86	60 682,67	532 382,60	51 560,00	24 000,00
Cliente 122	122,00	1,00	19 918,86	116 129,43	584 911,44	980,00	0,00
Cliente 123	123,00	1,00	26 404,12	77 241,59	555 330,00	43 704,00	0,00
Cliente 124	124,00	1,00	15 581,57	111 414,58	581 350,93	6 000,00	0,00
Cliente 125	125,00	1,00	35 106,85	69 545,99	549 255,50	52 800,00	0,00
Cliente 126	126,00	1,00	15 482,94	111 302,20	581 266,08	6 120,00	0,00
Cliente 127	127,00	1,00	10 788,38	93 856,55	568 074,54	25 000,00	0,00
Cliente 128	128,00	1,00	9 647,12	103 052,37	575 036,05	14 980,00	0,00
Cliente 129	129,00	1,00	14 094,14	89 459,28	553 855,34	22 437,00	22 471,00
Cliente 130	130,00	1,00	36 339,21	68 497,20	548 412,69	54 072,00	0,00
Cliente 131	131,00	3,00	431 401,98	341 476,78	152 751,48	363 771,00	267 924,00
Cliente 132	132,00	1,00	8 504,71	104 083,02	571 182,39	10 400,00	10 000,00
Cliente 133	133,00	1,00	8 901,26	100 405,08	573 034,91	17 847,00	0,00
Cliente 134	134,00	1,00	16 414,56	112 351,68	582 058,48	5 000,00	0,00
Cliente 135	135,00	1,00	16 414,56	112 351,68	582 058,48	5 000,00	0,00
Cliente 136	136,00	1,00	18 569,30	114 699,09	583 831,09	2 500,00	0,00
Cliente 137	137,00	1,00	20 674,82	116 920,85	585 509,31	140,00	0,00

Figura H 13: Folha de resultados 4 parte 4