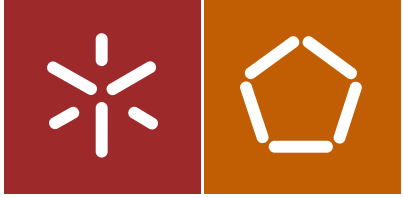
Universidade do Minho
Escola de Engenharia

Rui Flávio Gonçalves da Silva

Data Mining na Caracterização Geo-Espacial e Previsão da Incidência de Pneumonia em Portugal

janeiro de 2016

Data Mining na Caracterização Geo-Espacial e Previsão da Incidência de Pneumonia em Portugal
Rui Flávio Gonçalves da Silva
UMinho | 2016



Universidade do Minho
Escola de Engenharia

Rui Flávio Gonçalves da Silva

Data Mining na Caracterização Geo-Espacial e Previsão da Incidência de Pneumonia em Portugal

Tese de Mestrado
Ciclo de Estudos Integrados Conducentes ao Grau de
Mestre em Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação do
Professora Doutora Maribel Yasmina Santos

janeiro de 2016

DECLARAÇÃO

Nome: Rui Flávio Gonçalves da Silva

Endereço electrónico: a61589@alunos.uminho.pt Telefone: 914631739

Número do Bilhete de Identidade: 13741199

Título dissertação: *Data Mining* na Caracterização Geo-Espacial e Previsão de Pneumonia em Portugal

Orientadora: Professora Doutora Maribel Yasmina Santos

Ano de conclusão: 2016

Designação do Mestrado: Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

Universidade do Minho, 28/01/2016

Assinatura: _____

Agradecimentos

Agradeço à Professora Maribel Yasmina Santos por toda a disponibilidade demonstrada e pelas críticas construtivas. Sem a sua ajuda não teria sido possível. Por cada erro corrigido devo-lhe um sentido obrigado.

Agradeço à minha família, em especial aos meus pais, pela educação que me proporcionaram e pelo apoio quando eu já não acreditava que poderia chegar tão longe.

Agradeço aos meus amigos de curso pelo apoio dado durantes os 5 anos, sem vocês não teria sido a mesma experiência. São amigos que tenho para a vida.

Resumo

O número de portugueses afetados por doenças que atacam o sistema respiratório tem vindo a aumentar de ano para ano. No caso da Pneumonia, esta encontra-se fortemente presente em Portugal, apresentando uma elevada taxa de incidência. A Pneumonia é a terceira principal doença responsável por vítimas mortais em Portugal, colocando Portugal em terceiro lugar nos países com maior número de mortes por Pneumonia na União Europeia. Como referido, esta situação tem vindo a piorar com o passar dos anos, levando à necessidade de encontrar formas de prevenir esta doença. Deste modo, a Fundação Portuguesa do Pulmão tem vindo a promover vários estudos que permitam caracterizar a incidência desta doença na população. Um dos estudos passou pela criação de um sistema de *Business Intelligence* suportado por um *Data Warehouse* que intrega dados com a incidência da doença e dados demográficos recolhidos nos censos de 2011, entre outros.

Após a criação deste *Data Warehouse* e a caraterização geral da doença em Portugal, existiu a necessidade de continuar o estudo através da aplicação de *Data Mining* aos dados. Com a aplicação de *Data Mining* nos dados pretende-se, por um lado, realizar uma caracterização geo-espacial da doença utilizando uma abordagem de *clustering* geo-espacial e, por outro lado prever a incidência futura da mesma tendo em conta a evolução da Pneumonia em Portugal ao longo de uma década. Tendo em conta os dados disponíveis, foram criados três *datasets* com os dados referentes aos casos de pneumonia. O primeiro com os registos que apresentassem coordenadas geográficas para permitir a localização dos casos de doença no espaço. Um segundo *dataset* com todos os registos de indivíduos que pelo menos apresentassem uma outra patologia associada à Pneumonia. E um terceiro com todos os registos. Sobre estes *datasets* foram utilizadas duas técnicas de *Data Mining*, o *clustering* espacial e *time-series forecasting*. Na componente de *clustering* espacial foi utilizada a abordagem F-SNN (*Fast-Shared Nearest Neighbor*), para a criação dos clusters sobre os dois *datasets*, de modo a caracterizar a incidência da Pneumonia em Portugal. Os modelos identificados nesta componente foram avaliados com o auxílio de uma métrica de qualidade. Na componente de *time-series forecasting* foi utilizado apenas o terceiro *dataset* para a previsão de casos de Pneumonia e vítimas mortais. Aqui os modelos foram avaliados atendendo à precisão dos mesmos. Por fim, os modelos de cada componente foram comparados e analisados de forma a caracterizar e prever a doença em Portugal.

Palavras-Chave: *Data Mining* Espacial, Pneumonia, *Clustering* Espacial, F-SNN, *Time-series Forecasting*.

Abstract

The Portuguese number affected by diseases that attack the respiratory system has been increasing from year to year. In the case of pneumonia, it is strongly present in Portugal, with a high incidence rate. Pneumonia is the third leading disease responsible for deaths in Portugal, putting Portugal in third place in countries with the highest number of deaths by pneumonia in the European Union. This situation has worsened over the years, leading to the need to find ways to prevent this disease. Thus, the Portuguese Lung Foundation has promoted various studies to characterize the incidence of this disease in the population. One study went through the creation of a Business Intelligence system supported by a Data Warehouse that integrates data about the disease incidence and demographic data collected in the 2011 census, among others.

After the creation of the Data Warehouse and the general characterization of the disease in Portugal, there was the need to continue the study by applying data mining to the data. With the application of data mining to the data is intended, on one hand, to do a geo-spatial characterization of the disease using a geo-spatial clustering approach, and, secondly, to predict the future incidence of the disease taking into account the evolution of Pneumonia in Portugal over a decade. Given the available data, three datasets were created with data on the cases of pneumonia. The first with the records that include the geographic coordinates, locating the cases of disease in space. A second dataset derived from the first, but with the records of individuals who presented at least another pathology associated with pneumonia. And the third dataset with all the records. On these three datasets Data Mining techniques were used, spatial clustering and time-series forecasting. In the spatial clustering component, the F-SNN (Fast-Shared Nearest Neighbor) approach allowed the creation of clusters on the two datasets to characterize the incidence of pneumonia in Portugal. The identified models were evaluated with the use of a quality metric. In the time-series forecasting, the third dataset was used for predicting cases of pneumonia and fatalities. These models were assessed with an accuracy metric. All the obtained models were compared and analyzed to characterize and predict the disease in Portugal.

Keywords: Data Mining Space, Pneumonia, Spatial Clustering, F-SNN, Time-series Forecasting.

Índice

Índice de Figuras

Índice de Tabelas

Lista de Abreviaturas

Neste documento são utilizadas as seguintes abreviaturas:

- BI – *Business Intelligence*
- BPN - *Backpropagation Network*
- CLIQUE - *Clustering in Quest*
- CRISP-DM - *Cross Industry Standard Process for Data Mining*
- DBSCAN - *Density-based Spatial Clustering of Application with Noise*
- DCBD – Descoberta de Conhecimento em Base de Dados
- DM –*Data Mining*
- DW – *Data Warehouse*
- F-SNN - *Fast-Shared Nearest Neighbor*
- KDD - *Knowledge Discovery in Databases*
- SNN - *Shared Nearest Neighbor*
- STING - *Statistical Information Grid-based*

1. Introdução

Neste capítulo será apresentada a motivação subjacente a esta dissertação e os objetivos que se pretende atingir no final de todo o processo. É também apresentada a estrutura organizacional do documento, onde é realizada a descrição de cada um dos capítulos principais, de forma a compreender os assuntos que são tratados em cada capítulo.

1.1. Motivação

Anualmente mais de um milhão de portugueses adoecem com doenças que atacam o sistema respiratório, tais como gripes, agudizações infeciosas de doenças pulmonares obstrutivas crônicas e tuberculose. Estas infeções podem torna-se graves e pôr em risco a vida dos doentes, sobretudo no caso da Pneumonia. A Pneumonia é a terceira principal causa de morte em Portugal, o que torna Portugal no terceiro país com maior número de mortes na União Europeia. Esta doença afeta principalmente os indivíduos com mais de 65 anos, mas cada vez mais tem vindo a afetar a população mais jovem (Araújo, 2014).

A situação da Pneumonia em Portugal tem vindo a agravar-se com o passar dos anos e cada vez mais é necessário compreender a doença e o seu comportamento de forma a definir novas e melhores técnicas de controle da doença. Deste modo, a Fundação Portuguesa do Pulmão, enquanto organização sem fins-lucrativos, tem vindo a promover diversos estudos que permitam caracterizar a população portuguesa em termos de incidência de infeções respiratórias ou outras patologias que incidam sobre o aparelho respiratório. Depois da implementação de um sistema de BI (*Business Intelligence*) suportado por um DW (*Data Warehouse*) com a incidência de pneumonias em Portugal continental (Leite, 2014). Pretende-se agora compreender os padrões comportamentais da doença, prosseguindo o estudo com a aplicação de técnicas de DM (*Data Mining*) aos dados para realizar a caracterização geo-espacial da doença utilizando uma abordagem de *clustering* espacial e, por outro lado prever a incidência futura da doença tendo em conta a evolução da Pneumonia em Portugal ao longo de uma década.

Os dados guardados no DW contêm a informação da posição geográfica dos eventos, que neste caso são indivíduos a quem foi diagnosticada a doença. A existência desta informação permite-nos descobrir padrões nos eventos e realizar a caracterização espacial dos eventos. De forma a que seja

possível utilizar a informação geográfica, é necessário recorrer a técnicas de DM que permitam manipular e usar esta informação. Assim, é necessário recorrer à utilização do DM espacial, que difere do DM tradicional no facto de utilizar a componente espacial dos eventos. Existem estudos que utilizam dados com características semelhantes e que recorrem ao DM espacial. Por exemplo, o estudo da mortalidade de cancro nos Estados Unidos (Vinnakota & Lam, 2006), no qual são usadas regras de associação suportadas por um sistema informação geográfico, de forma a descobrir padrões espaciais sobre as regiões mais afetadas e que precisam de tenção futura. Outro exemplo, é a utilização do método de *clustering* espacial para a deteção e inferência de doenças, no qual são utilizadas coordenadas geográficas na criação dos clusters, levando à identificação de clusters de várias formas e independentes de fronteiras politicas. (Kulldorff & Nagarwalla, 1995). Na previsão do comportamento de doenças existem vários estudos realizados, alguns deles sobre pneumonia, com recurso à técnica de *time-series forecasting* (Choi & Thacker, 1981).

A questão que se levanta e à qual esta dissertação tem como objetivo responder é: "Quais os modelos que melhor caracterizam geo-espacialmente e preveem a incidência de Pneumonias em Portugal?"

De modo a que seja possível responder a esta questão, foi necessário definir a metodologia a utilizar para orientar todo o processo de desenvolvimento do trabalho. Foram estudadas um conjunto de metodologias utilizadas na área do DM, após várias comparações, a metodologia escolhida foi o CRISP-DM (*Cross Industry Standard Process for Data Mining*). Foram ainda escolhidas duas técnicas de DM para a execução do trabalho, o *clustering* espacial e o *time-series Forecasting*. No *clustering* espacial foi utilizada a abordagem F-SNN, esta foi escolhida por conseguir lidar com um grande volume de dados poupando tempo de processamento (Antunes, Moreira, & Santos, 2014) e pelo facto do seu código ser disponibilizado de forma a permitir adaptações ao problema em estudo. Para a realização das previsões foi utilizado o *Time Series Forecast* do weka devido a permitir de forma simples de realizar várias configurações de modelos e por apresentar diversos resultados, medidas de erros e gráficos para comparação.

1.2. Finalidade e Objetivos

Esta dissertação tem como finalidade identificar os modelos que caracterizam, geo-espacialmente, a incidência de pneumonias em Portugal, assim como identificar os modelos

preditivos que permitam prever a evolução dos casos de Pneumonia e do número de vítimas mortais no país.

Para atingir esta finalidade são definidos como objetivos desta dissertação:

- Revisão de literatura sobre *Knowledge Discovery in Databases* e *Data Mining*
- Compreensão do domínio e dos dados guardados no *Data Warehouse*.
- Identificação e extração dos subconjuntos de dados relevantes.
- Realização da caracterização geo-espacial de Pneumonias em Portugal através da utilização da abordagem de *clustering* F-SNN e identificação dos melhores modelos.
- Realização da previsão do número de casos de Pneumonias e de vítimas mortais, através da utilização da técnica de *time-series forecasting* e identificação dos melhores modelos.
- Avaliação da relevância e acuidade dos modelos identificados.

1.3. Abordagem Metodológica

A metodologia escolhida para a realização deste trabalho sé o CRISP-DM, já que esta metodologia demonstra ser a mais completa (Azevedo & Santos, 2008). É a que fornece mais informação em termos de utilização, explicando o que deve ser efetuado em cada fase de forma a que seja possível controlar e garantir a qualidade do trabalho, o que apresenta ser uma grande vantagem. Esta demonstra ser uma metodologia bem aceite na área de estudo do DM (Piatetsky, 2014).

A metodologia CRISP-DM é descrita como um modelo de processo hierárquico, constituído por quatro níveis de abstração, como representado na Figura 1, integrando fases, tarefas gerais, tarefas especificas e instâncias de processos.

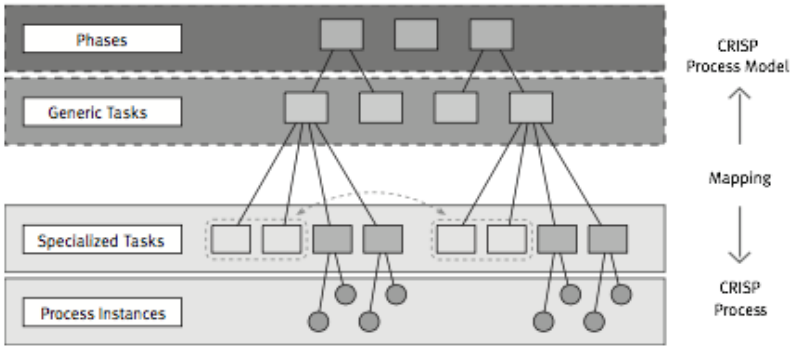


Figura 1 Representação dos quatro níveis de abstração do CRISP-DM (retirado de : (Chapman, Clinton, & Kerber, 2000)

Cada fase presente no topo da estrutura dá origem a uma ou mais tarefas gerais de segundo nível, as quais são capazes de cobrir vários projetos de DM (Chapman et al., 2000). No terceiro nível encontram-se as tarefas específicas correspondentes às tarefas gerais, as quais são mais específicas quanto às ações a realizar. Por exemplo se a tarefa geral for "Limpeza de dados" então a tarefa específica poderá ser "Limpeza de dados de valores nulos". O último nível passa pela realização do registo das ações, decisões e resultados da atividade de *Data Mining* em execução.

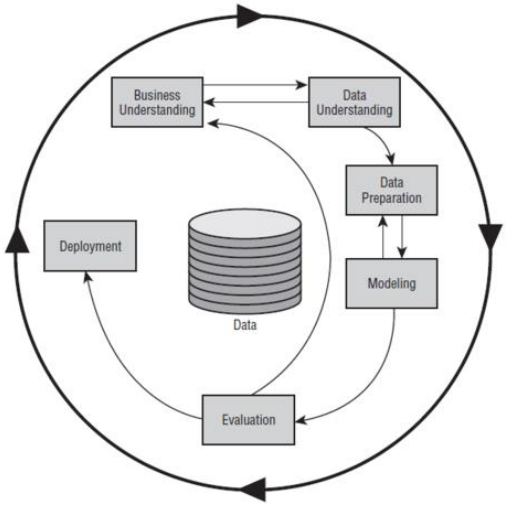


Figura 2 Modelo referencial do CRISP-DM (retirado de : (Chapman et al., 2000)

O CRISP-DM apresenta um modelo referencial cíclico, onde estão representadas todas as fases do projeto, como representado na Figura 2 e com maior detalhe a nível de tarefas na Figura 3. As fases constituintes do CRISP-DM são:

- ***Business understanding*** – Nesta fase o foco é compreender os objetivos e requisitos do ponto de vista do negócio, de modo a formular o problema de *Data Mining* a resolver para alcançar os objetivos.

- *Data understanding* – Esta fase inicia-se com a utilização de todos os dados existentes para o projeto, com o intuito de explorar e compreender os dados, detetar subconjuntos de dados interessantes e identificar problemas de qualidade nos dados

- *Data preparation* – Esta fase contém todas as atividades necessárias para construir uma *dataset* final para alimentar os modelos.

- *Modeling* – Nesta fase é realizada a seleção e aplicação de vários algoritmos e realizados testes para vários parâmetros utilizados pelos algoritmos.

- *Evaluation* – Esta fase tem como objetivo garantir a qualidade dos resultados obtidos nos modelos, sendo necessário reavaliar o processo de criação do modelo, para garantir que este responde aos objetivos traçados no início do projeto.

- *Deployment* – Esta fase tem como objetivo implementar os modelos encontrados, ou seja, implementação destes modelos num cliente de modo que este possa tirar partido deles.

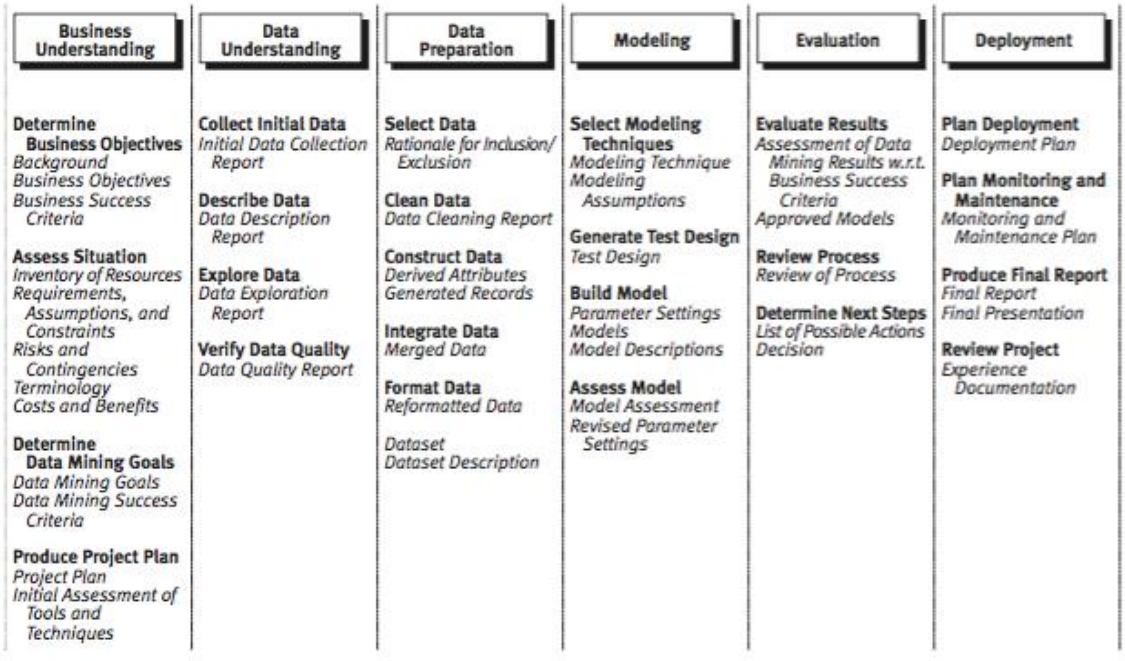


Figura 3 Tarefas Genéricas (negrito) e outputs (itálico) do modelo referencial do CRISP-DM (retirado de : (Chapman et al., 2000)

1.4. Estrutura do Documento

No primeiro capítulo deste documento encontra-se a introdução, onde é descrita a motivação para realização desta dissertação, a finalidade e os objetivos da dissertação, a abordagem metodológica utilizada e a organização da estrutura do documento.

No segundo capítulo é apresentado o enquadramento conceptual. Neste é realizado o estudo sobre a Descoberta de Conhecimento em Base de Dados e *Data Mining*. Na secção de Descoberta de Conhecimento em Base de Dados são estudados o processo de descoberta em base de dados e as suas fases, problemas e desafios. Na secção sobre o tema *Data Mining* são apresentadas as tarefas e técnicas de *Data Mining*. São estudadas com mais detalhe as técnicas de *Data Mining* associadas às árvores de decisão, regras de associação, *clustering* e redes neuronais artificiais.

O terceiro capítulo contém a descrição e seleção dos dois *datasets* utilizados para este trabalho. Neste capítulo é ainda explicada a construção dos modelos de *clustering* geo-espacial recorrendo ao F-SNN. É explicado como foram definidos os parâmetros de entrada do F-SNN, as dimensões utilizadas e seus respetivos pesos. Também neste capítulo são apresentados os resultados obtidos no F-SNN, e é efetuada a avaliação dos modelos e a análise geo-espacial dos modelos.

No quarto capítulo é documentada a aplicação da técnica *time-series forecasting* sobre o segundo *dataset* criado no capítulo três. Neste capitulo é apresentado o processo de transformação dos dados para ficheiros no formato arff. Também é explicada a construção dos modelos e os parâmetros envolvidos na criação dos mesmos. Por fim, são apresentados e analisados os resultados, de modo a escolher os melhores modelos para previsão de Pneumonias e de vítimas mortais.

No quinto capitulo são apresentadas as conclusões sobre o trabalho realizado, assim como as propostas de trabalho futuro e de aspetos que podem ser melhorados neste trabalho.

2. Enquadramento Conceptual

Neste capitulo é apresentado o enquadramento conceptual do trabalho, abarcando os temas de Descoberta de Conhecimento em Base de Dados e *Data Mining*. Na secção 2.1 referente ao DCBD, é apresentado o processo de descoberta em base de dados e as suas fases, desafios e problemas. Na secção 2.2 referente ao *Data Mining*, são exploradas as tarefas e as técnicas de *Data Mining*.

2.1. Descoberta de Conhecimento em Base de Dados

"It has been estimated that the amount of information in the world doubles every 20 months" (Frawley, Piatetsky-Shapiro, & Matheus, 1992).

A presença de bases de dados em diversas áreas, por exemplo no sector da industria da saúde, levou à coleta de dados a um ritmo elevado. Com isto surge a necessidade de tirar maior partido desses dados através de novas teorias computacionais e ferramentas para auxiliar as pessoas a retirar informação útil, ou seja, obter conhecimento, destes volumes de dados. A DCBD (Descoberta de Conhecimento em Base de Dados) ou no inglês KDD (*Knowledge Discovery in Databases*) tornou-se um campo emergente e em desenvolvimento através da necessidade de teorias e ferramentas apropriadas para analise em vastas quantidades de dados (Usama Fayyad, Gregory Piatetsky-shapiro, 1996).

A quantidade de informação no mundo e nas nossas vidas aumenta de dia para dia, com a presença de computadores e outros dispositivos eletrónicos que permitem que seja possível armazenar toda esta informação. Todas as decisões tomadas, escolhas realizadas no supermercado ou hábitos financeiros são registadas por parte das industrias. Mas segundo Written, Frank e Hall *et all* (Witten, Frank, & Hall, 2011) existe um grande buraco entre a quantidade de dados gerados e a compreensão dos mesmos, uma vez que a cada aumento do volume de dados leva à diminuição da sua compreensão.

Segundo (Fayyad, Piatetsky-Shapiro, & Smyth, 1996a), a DCBD continua a evoluir através da intersecção de vários campos de investigação tais como *machine learning*, *pattern recognition*, base de dados, estatística, inteligência artificial, *data visualization* e *high performance computing*. O objetivo em comum de todos os campos anteriormente referidos é a extração de conhecimento a partir de dados no contexto de grandes conjuntos de dados. Neste caso, dados é um conjunto de factos e os padrões traduzem-se numa forma de descrever subconjuntos de dados.

O Processo DCBD é globalmente composto por diversas etapas tais como preparação dos dados, procura de padrões, avaliação do conhecimento e refinamento, através de várias interações. A procura destes padrões não é uma operação trivial e é preciso ter certezas na qualidade dos resultados obtidos, isto é, os padrões obtidos devem poder ser utilizados noutro conjunto de dados e de forma a obter resultados com algum grau de certeza. Os padrões devem ser compreensíveis de forma a que seja possível criar medidas para quantifica-los para que possam ser avaliados. Na DCBD existe uma etapa para a aplicação de algoritmos de descoberta, ou análise de dados, que identificam padrões, a que se dá o nome de *Data Mining* (Fayyad et al., 1996a).

2.1.1. Processo de Descoberta de Conhecimento em Base de Dados

O Processo de DCBD de uma forma generalizada pode ser definido como um processo que utiliza uma fonte de dados, e sobre essa fonte de dados realiza-se a seleção dos dados, pré processamento aos dados, criam-se subconjuntos dos dados e transformações dos dados, de modo a ser possível a aplicação de métodos (algoritmos) de *Data Mining* para a identificação de padrões nos dados.

Segundo (Fayyad et al., 1996a), o Processo de DCBD é um processo interativo e iterativo, que envolve um determinado número de passos com um número elevado de decisões a serem tomadas pelo utilizador. Os passos que constituem o processo de DCBD são a Seleção dos dados, Pré-processamento dos dados, Transformação dos dados, *Data Mining* e Interpretação de Resultados, estes serão detalhados na subsecção 2.1.2.

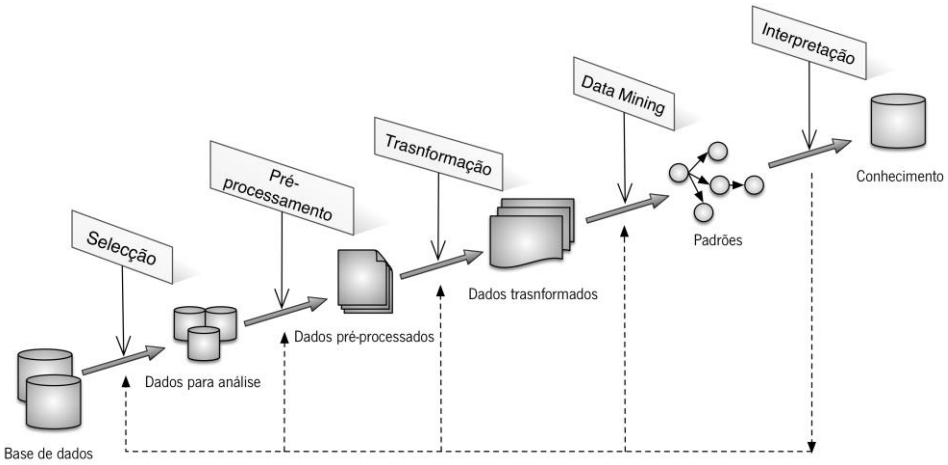


Figura 4 Visão geral dos processos presentes na DCBD (Adaptado de : (Fayyad, Piatetsky-Shapiro, & Smyth, 1996b)

2.1.2. Fases do Processo de Descoberta de Conhecimento

Neste ponto serão brevemente enumeradas e descritas as fases do processo de Descoberta de Conhecimento em Base de dados, representadas na figura 1.

- **Seleção dos dados:** Nesta fase do processo é realizada uma seleção dos dados para responder a um determinado objetivo. Estes proveem de uma ou várias fontes de dados (sistemas transacionais, armazém de dados, etc.). Os dados serão utilizados pelo processo de *Data Mining*. Esta seleção dos dados tem como principal objetivo a eliminação de atributos que não apresentam qualquer interesse para os objetivos do processo.
- **Pré-processamento dos dados:** Nesta fase do processo são realizadas operações gerais sobre os dados selecionados, remoção de ruídos caso seja necessário, recolher informações para modelação, decidir estratégias para lidar com os valores nulos.
- **Transformação dos dados:** Nesta fase do processo o principal foco passa por encontrar recursos úteis para a representação dos dados, dependendo do objetivo da tarefa. A realização desta fase necessita a utilização de métodos de redução de dimensões ou métodos de transformação do número de variáveis consideradas (transformação de valores contínuos em valores discretos, criação de classes de valores), desta forma é possível reduzir o espaço da pesquisa.
- ***Data Mining*:** A fase de *Data Mining* pode se considerada como a mais importante. Esta fase passa pela abordagem de aplicar métodos de *Data Mining* para chegar a um resultado para cada objetivo traçado na primeira fase (como por exemplo, sumarização, classificação, regressão, *clustering*, etc.). A escolha de algoritmos, para serem usados na procura de padrões nos dados, necessita ter em consideração quais os modelos e parâmetros apropriados.
- **Interpretação de Resultados: Nesta** fase é realizada a análise dos resultados obtidos pelos algoritmos de *Data Mining* aplicados na fase anterior do processo. Os modelos obtidos são aplicados a novos conjuntos de dados (conjunto de dados de teste), desta forma os modelos são avaliados pelo seu desempenho em dados desconhecidos. Sendo o processo de DCBD constituído por várias fases em que são tomadas decisões que afetam diretamente os resultados, por vezes levando a resultados indesejáveis ou menos corretos. Como se trata de um processo iterativo e interativo é possível voltar às fases anteriores para aprimorar o processo de forma a melhorar os resultados.

2.1.3. Desafios e Problemas no Processo DCBD

Ao utilizar bases de dados reais estas apresentam problemas nos dados que afetam os resultados e o funcionamento dos algoritmos de *Data Mining*. Na aplicação do processo de DCBD é necessário ter em consideração a disponibilidade de dados suficientes, em geral quanto maior o número de campos existentes mais complexos serão os padrões encontrados, o que leva à necessidade de mais dados para compreender os padrões. Outra consideração importante a ter atenção é a relevância que os atributos apresentam, para identificar aqueles que são relevantes para a tarefa de descoberta, isto é, não importa a quantidade de dados que se utilize, mas sim a utilização de dados que contêm a informação relevante para descoberta. Além disso outra consideração a ter é a existência de ruído nos dados que dificulta a descoberta de padrões nos mesmos. Existe a necessidade conhecimento prévio sobre o domínio em que é conduzido o processo de descoberta, isto é, saber reconhecer quais as relações possíveis, quais os campos importantes, quais os padrões já descobertos, e assim em diante (Fayyad, Piatetsky-Shapiro, & Smyth, 1996c).

Abaixo serão enumerados e resumidos segundo Fayyad, Piatetsky-Shapiro, e Smyth (Fayyad et al., 1996b), alguns dos desafios que são encontrados no processo DCBD com que os utilizadores desde processo se confrontam.

- **Base de dados de grande escala e elevado número de dimensões:** A existência de base de dados com elevada escala, na ordem dos milhões de registos e grande quantidade de campos são comuns. Estes conjuntos de dados permitem criar um espaço de pesquisa de grandes dimensões e pode levar os algoritmos de *data mining* a encontrar padrões que geralmente não são válidos. De forma a ultrapassar este problema é possível a criação de pequenos *datasets* a partir dos dados, redução das dimensões e introdução de conhecimento prévio. Existe a necessidade de uma maior interação entre o utilizador e os algoritmos utilizados.
- **Interação do utilizador:** Sendo o DCBD um processo por definição interativo e iterativo existe o desafio de providenciar um ambiente de produção que seja de alta-performance e resposta rápida que assista o utilizador e lhe permita a utilização de ferramentas e técnicas que o levem a atingir os objetivos.
- ***Overfitting:*** A procura pelos parâmetros adequados através de algoritmos para um modelo em particular usando um conjunto limitado de dados, pode ter como resultado

uma fraca performance do modelo quando aplicado nos dados de teste. Possíveis soluções seriam a aplicação de estratégias de regulação e estratégias estatísticas.

- **Dados em falta:** A falta de dados na base de dados pode levar a erros de operações sobre a base de dados e operações matemáticas perdem precisão nos resultados. A solução para este problema passa pela aplicação de estratégias estatísticas para identificar os valores.
- **Incompreensibilidade dos padrões:** É necessário que os padrões obtidos sejam compreensíveis para o utilizador, isto leva à necessidade das ferramentas apresentarem representações gráficas e outras técnicas de representação para auxiliar a compreensão dos dados e conhecimento.
- **Gerir mudança nos dados e no conhecimento:** As alterações feitas sobre os dados podem resultar na invalidação dos padrões encontrados anteriormente, isto pode resultar devido à modificação, remoção ou acréscimo de novos registos ao longo do tempo. Uma solução para o problema seria a alteração dos métodos para que estes tornem incrementais levando à atualização dos padrões e a utilização destas mudanças para descobrir novos padrões.

2.2. Data Mining

Como já referido anteriormente a componente *Data Mining* está presente no processo de descoberta de conhecimento, sendo uma das componentes mais importante.

Segundo (Witten et al., 2011) o Data Mining é a atividade focada em resolver problemas através da análise de dados armazenados em base de dados. O processo de *Data Mining* procura descobrir padrões nos dados, este processo deve ser automático ou semiautomático, os padrões obtidos devem trazer vantagens significativas para o utilizador.

Segundo (Miller & Han, 2009) citando (Klösgen and Zytkow 1996) e (Han, Kamber, & Pei, 2011), *Data Mining* como a aplicação de funções de baixo-nível com a intenção de revelar informação escondida em base de dados e que o tipo de conhecimento que se pretende adquirir dita quais as funções de *Data Mining* a serem utilizadas.

Segundo (Pujari, 2001), *Data Mining* é uma atividade que procura prever informação oculta em grandes base de dados, utilizando o poder da tecnologia para potencializar a análise sobre informação importante.

2.2.1. Tarefas de Data Mining

As tarefas de *Data Mining* estão divididas em dois focos: previsão e descrição. A previsão tem como objetivo a utilização de variáveis ou campos existentes na base de dados para prever os valores futuros das variáveis de interesse e a descrição tem como objetivo procurar padrões que descrevam os dados que sejam interpretáveis pelo utilizador. Estes dois focos não são completamente isolados um do outro, pois existem modelos de previsão que podem ser descritivos e vice-versa (Fayyad et al., 1996c).

Para atuar nos dois focos anteriormente referidos existe uma variedade de métodos de *Data Mining* que serão mencionados nos seguintes pontos.

- **Associação** – A associação tem como objetivo definir a relação entres campos da base de dados, através da correlação de vários campos, a correlação deve satisfazer o nível de suporte e confiança que o utilizador procura (Fayyad et al., 1996b).

- **Classificação** – A classificação é a função de mapear (classificar) os dados para uma das classes pré-definidas, com este modelo é possível prever as classes para dados não classificados (Fayyad et al., 1996c).

- **Clustering** – O *clustering* ou segmentação tem como função classificar os dados, mas ao contrário da classificação em que as classes estão pré-definidas o *clustering* procura classificar os dados de forma natural, estes são atribuídos a uma classe através da identificação de métricas semelhantes (Witten et al., 2011).

- **Regressão** – A Regressão é a função de mapear os dados com o objetivo de prever o valor real de uma variável e descobrir relacionamentos entre variáveis (Fayyad et al., 1996a).

- **Sequenciação** – A sequenciação é função de analisar e identificar padrões comportamentais nos dados ao longo do tempo, isto é possível através da existência da dimensão tempo (Santos, 2001).

- **Sumarização** – A sumarização permite abstração ou generalização dos dados. Através da sumarização e abstração dos dados é possível obter um conjunto de dados mais pequeno que oferece uma visão geral dos dados com informação agregada (Fu, 1997).

2.2.2. Técnicas de Data Mining

Sendo o *Data Mininig* uma atividade que atua sobre várias áreas de investigação, adotou várias técnicas e conceitos provenientes dessas áreas, como abordagem estatística, *machine learning*, sistemas de base de dados e redes neuronais humanas, são alguns exemplos (Fu, 1997).

2.2.2.1. Árvores de Decisão

Esta técnica apresenta uma abordagem de *"divide-and-conquer"* para resolver problemas de dados com instâncias independentes, permitindo de forma natural a representação a que chamamos árvore de decisão (Witten et al., 2011). Na Figura 5 encontra-se representada uma árvore de decisão simples. Esta permite determinar a quilometragem de um carro através do tipo de transmissão que possui e o peso. Os nós folha estão representados por retângulos na Figura 5, cada um representa uma classe de quilometragem. Através da árvore de decisão apresentada podemos concluir, por exemplo, que um carro de tamanho médio e transmissão automática estará na classe de quilometragem média (Fu, 1997).

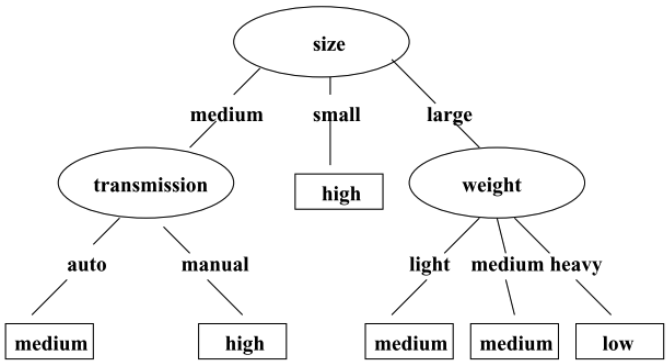


Figura 5 Representação de uma árvore de decisão simples sobre quilometragem de carros (Fu, 1997)

Os nós de decisão presentes na árvore de decisão têm como objetivo testar um determinado atributo, normalmente é comparado com uma constante. Os nós folha permitem aplicar classificações a todas as instâncias que chegam à folha. Para classificar uma instância desconhecida, é necessário descer através da árvore de decisão de acordo com os valores dos

atributos testados nos sucessivos nós, quando se chega a uma folha a instância é classificada de acordo com a classe atribuída aquela folha.

Caso o atributo testado seja um valor nominal, o número de filhos normalmente é o número de valores possíveis para o atributo. Neste caso existe um ramo para cada um dos valores, faz com que não seja possível voltar a testar o mesmo atributo em níveis abaixo na árvore de decisão. Para que seja possível testar pelo menos mais que uma vez o atributo é criado dois subconjuntos de forma a obter dois caminhos, neste caso o atributo poderá ser testado mais que uma vez. Caso o atributo testado seja um valor numérico, normalmente o teste num nó determina se o valor é maior ou menor que a constante pré-determinada, permitindo ter dois caminhos como opção. Em alguns casos pode ser necessário utilizar três caminhos, o teste do nó avalia se o valor é maior igual ou menor que. (Witten et al., 2011).

2.2.2.2. Regras de Associação

Ao contrário das regras de classificação, as regras de associação permitem prever qualquer atributo ou combinação de atributos e não apenas a classe como acontece com as regras de classificação. As regras de associação não são projetadas para serem usadas juntas como um conjunto, como acontece nas regras de classificação. Regras de associação diferentes manifestam diferentes regularidades que fundamentam o conjunto de dados, e que normalmente levam à previsão diferentes. É possível obter um elevado número de regras de associação até mesmo a partir de um pequeno conjunto de dados.

O suporte de uma regra de associação é o número de instâncias que a regra consegue prever corretamente. A confiança é o número de instâncias que são previstas corretamente, expressa em proporção em relação a todas as instâncias a que a regra foi aplicada. Por exemplo, a seguinte regra obtida dos dados representados na Tabela 1 Dados sobre meteorologia (Witten et al., 2011).

IF temperature = cool then humidity = normal

Suporte *= 4*

Confiança *= 100%*

O suporte nesta regra é o número de registo sem que comprem a condição de estar *cool* e a *humidity* se encontra normal, e a confiança é a proporção de registos com o valor *cool* que se

encontram com o valor *humidity* a normal. Deve-se geralmente especificar um valor mínimo para o suporte e para a confiança, de forma a obter as regras com um valor mínimo de *coverage* e *confidence*. Existem situações em que regras de associação se relacionam umas com as outras, de forma a reduzir o número de regras produzidas. Em alguns casos podemos apenas apresentar as regras mais fortes (Witten et al., 2011).

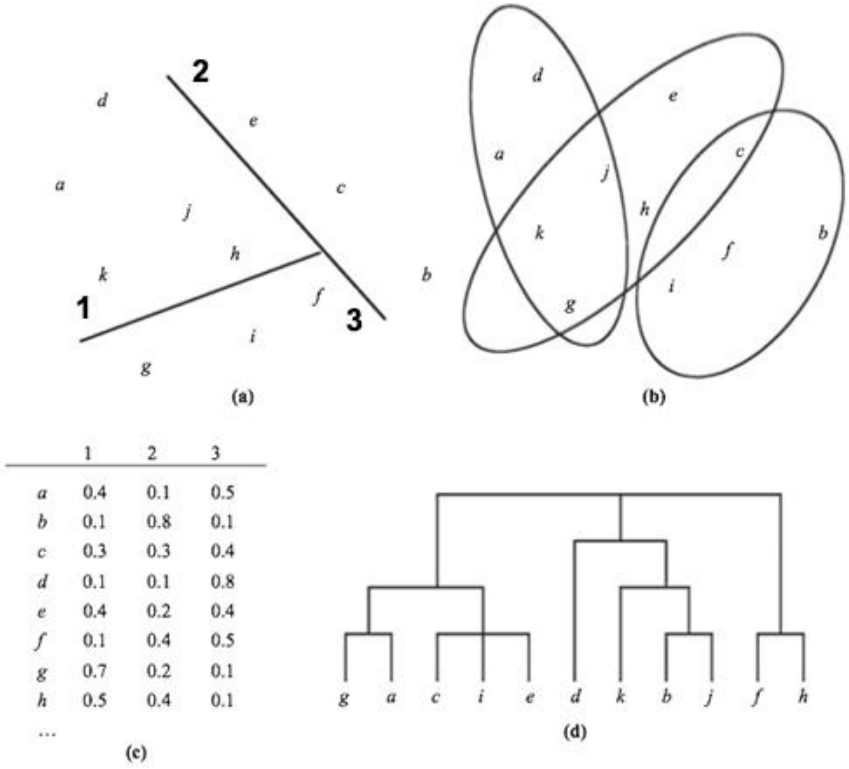
Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Tabela 1 Dados sobre meteorologia (Witten et al., 2011).

2.2.2.3. Clustering

A técnica de *clustering* será a mais abordada em relação às outras técnicas de Data Mining. Esta será a principal técnica utilizada na parte prática do projeto. Esta decisão deve-se à ampla utilização de *clustering* em trabalhos que lidam com dados espaciais, como por exemplo em (Nath, 2006) e em (Jonge, Pelt, & Roos, 2012). Isto não invalida a exploração e aplicação de outras técnicas.

Um *cluster* pode ser interpretado como sendo um classificador. Este tem como finalidade apresentar o mapeamento de instâncias através da atribuição do número do *cluster* a que cada instância está associada, como ilustrado na Figura 6(a). Alguns algoritmos permitem a uma instância pertencer a mais do que um *cluster*, como ilustrado na Figura 6(b). Outros algoritmos associam as instâncias aos clusters através de probabilidades. Neste tipo de algoritmos cada instância apresenta uma probabilidade de pertencer a um *cluster*, como ilustrado na Figura 6(c). Ainda existe outro tipo de algoritmos que produzem uma estrutura hierárquica de clusters. Esta estrutura é criada a partir de um único cluster com todos os registos em análise, posteriormente, vai partindo o cluster inicial num conjunto de sub-clusters, e assim sucessivamente, até que cada registo da origem pertença a um cluster diferente ou até que determinada métrica de qualidade seja atingida, como ilustrado na Figura 6(d) (Witten et al., 2011).



	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1
...			

Figura 6 Representação de vários tipos visualização de *clusters* (adaptado de (Witten et al., 2011)).

O *clustering* é um processo de aprendizagem não supervisionado, utilizado para identificar grupos homogéneos de dados num *data set* (Kotsiantis & Pintelas, 2004). A análise de clusters é o processo de particionar um conjunto de dados em subconjuntos, cada subconjunto é apelidado de *cluster*. Os subconjuntos são constituídos por dados com características semelhantes.

O *Clustering* pode ser também utilizado como uma ferramenta de pré-processamento para outros algoritmos de data mining, como por exemplo um seletor de subconjuntos e classificador. Como um *cluster* representa um conjunto de instâncias que apresentam similaridades e dissimilaridades das outras instâncias presentes em outros clusters, podemos tratar um cluster como uma classe. Neste sentido podemos afirmar que *clustering* é um processo de autoclassificação. Uma vantagem que o *clustering* apresenta é a capacidade de detetar *outliers* . Um *outlier* é um valor que se desvia dos padrões dos restantes valores do cluster e são colocados num *cluster* só para eles.

O *clustering* de dados está em forte desenvolvimento em várias áreas de investigação, como estatística, *machine learning*, *web search*, marketing, base de dados espaciais e muitas outras áreas de aplicação (Han, Kamber, & Pei, 2006).

Para ser possível comparar os vários algoritmos de *clustering* existentes é necessário avaliar uma série de requisitos existentes na análise de *clusters* (Han et al., 2006). Alguns dos requisitos típicos para *clustering* em *data mining* são:

- **Escalabilidade:** A maior parte dos algoritmos de *clustering* comporta-se bem com *data sets* com centenas de registos, mas quando aplicados sobre data sets com milhões de registos o seu comportamento pode-se alterar levando a que o resultado demore em virtude da quantidade de dados, por este motivo é necessário a existência de algoritmos de *clustering* capazes de alta escalabilidade.
- **Habilidade para lidar com diferentes tipos de atributos:** A maior parte dos algoritmos está preparado para lidar com valores numéricos e nominais, mas tem vindo a ser cada vez mais necessário haver algoritmos capazes de lidar com atributos complexos, tais como imagens, gráficos e documentos.
- **Descoberta de clusters com forma arbitrária:** Os algoritmos baseados em medidas de distância tendem a apresentar *clusters* esféricos com tamanhos e densidades parecidas, mas, no entanto, um *cluster* pode ter qualquer tipo de forma, tamanho ou densidade.
- **Necessidade conhecimento do domínio para determinar os parâmetros de entrada:** Na sua maioria, os algoritmos necessitam que o utilizador lhes forneça parâmetros de entrada, como por exemplo o número de clusters. Esta tarefa torna-se complexa com grandes volumes de dados. Os resultados alteram-se com as escolhas realizadas nos parâmetros.
- **Habilidade de lidar com o ruído nos dados:** Os dados recolhidos de atividades do mundo real podem conter *outliers*, valores desconhecidos ou dados errados. Os algoritmos de *clustering* devem ser capazes de lidar com estes problemas nos dados, caso contrário os resultados podem não ser os desejados.
- **Capacidade de lidar com várias dimensões:** A maioria dos algoritmos é capaz de lidar com data sets com poucas dimensões, duas a três dimensões, mas a existência de data sets com um número elevado de dimensões ou atributos, torna-se uma tarefa difícil para identificar os *clusters*.
- **Interpretação e usabilidade:** Os resultados obtidos pelos algoritmos devem ser compreensíveis e fáceis de utilizar.

Métodos de *Clustering*

Segundo (Kotsiantis & Pintelas, 2004) podemos categorizar os algoritmos de *clustering* em métodos de particionamento, métodos hierárquicos, métodos baseados em densidade, métodos baseados em grelha e métodos baseados em modelos, como representado na Figura 7.

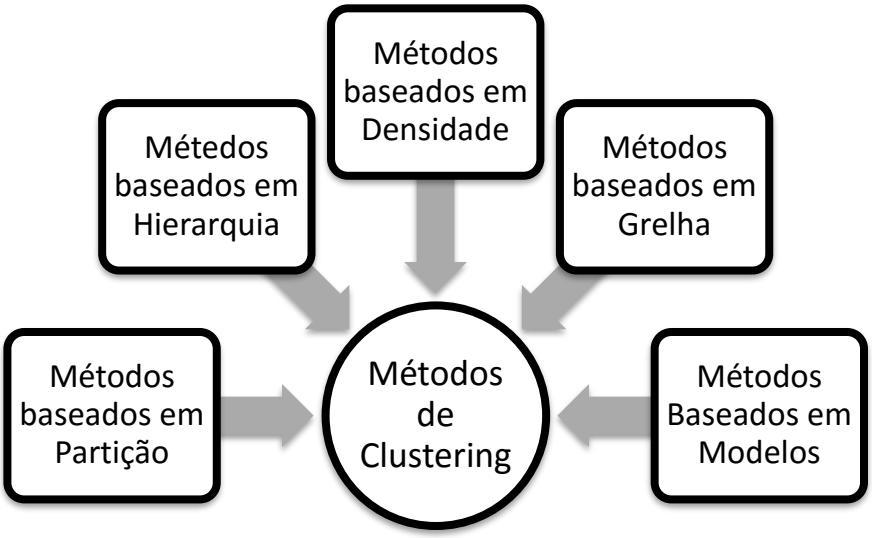


Figura 7 Algoritmos de *Clustering* (adaptado de (Dinis, 2011)**).**

- **Métodos Hierárquicos**

Os métodos hierárquicos constroem os modelos através do particionamento recursivo das instâncias através das abordagens *top-down* ou *bottom-up* (Rokach & Maimon, 2010). É possível dividir os métodos hierárquicos em duas categorias, a categoria com os métodos aglomerativos, que utilizam a abordagem *bottom-up* e a outra categoria com os métodos divisivos que utilizam os a abordagem *top-down*. Ambas as categorias podem ser representadas através de dendrogramas, como por exemplo na Figura 8.

A junção e divisão de *clusters* é realizada de acordo com algumas medidas de similaridade, escolhidas para otimizar um critério, como por exemplo a soma de valores quadrados. Segundo Rokach e Maimon (Rokach & Maimon, 2010) citando Jain, Murty e Flynn (Jain, Murty, & Flynn, 1999) pode-se dividir os métedos hierárquicos de acordo com a medida que se pretende calcular, *single-link clustering* ou *complete-link clustering.*

- ***Single-link clustering:*** este método define a distância entre dois *clusters* através do par de membros mais próximos, de cada *cluster*. Um *single-link clustering* pode ser obtido construindo uma *Minimal Spanning Tree* (MST). Este método tem como vantagens a

sua versatilidade e performance em relação ao *complete-link clustering*. (Rokach & Maimon, 2010).

- ***Complete-link clustering:*** este define a distância entre dois *clusters* através da distância dos membros mais afastados, de cada *cluster* (Rokach & Maimon, 2010). Ao início este método assume cada instância como sendo um *cluster*, constituído apenas por um membro. Com o decorrer da execução do algoritmo os pares mais próximos juntam-se num *cluster*. A cada junção a distância entre *clusters* é atualizada com o comprimento do ramo mais longo que liga os dois *clusters*, como ilustrado na Figura 8. Este método apresenta *clusters* melhor balanceados em relação ao método *single-link clustering* (Guo & Gahegan, 2006).

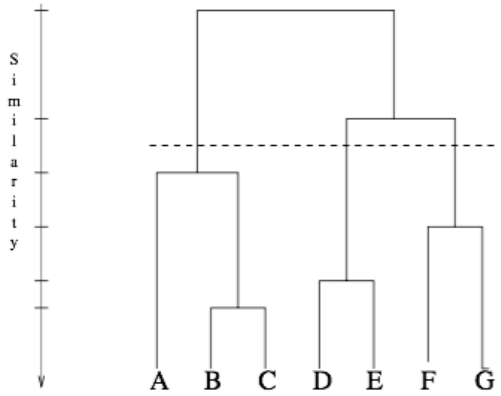


Figura 8 Exemplo de um dendograma obtido através do uso do algoritmo *single*-link (Jain et al., 1999)

- **Métodos baseados em Partição**

Os algoritmos baseados em partição realizam a tarefa de realocar as instâncias, através de transições entre os *clusters*, estes tipos de métodos necessitam tipicamente de um número de *clusters* pré-definido. Para este método obter resultados otimizados necessita de experimentar todas as combinações de partições possíveis (Rokach & Maimon, 2010).

Os algoritmos baseados em partição ao contrário dos algoritmos baseados em hierarquias apresentam apenas uma partição em vez de uma estrutura de *clusters* (Jain et al., 1999).

As técnicas de partição normalmente produzem *clusters* através da otimização dos parâmetros definidos. Na prática os algoritmos são processados várias vezes com diferentes tipos de configurações e o resultado do processamento é o output do *clustering* (Jain et al., 1999).

Devido à característica deste tipo de métodos, a procura exaustiva de todos os agrupamentos dos dados, é computacionalmente proibitiva pois a sua complexidade seria exponencial. Para lidar com este problema são adotadas duas heurísticas ou ainda variações dessas mesmas: o algoritmo K-means e K-Medoids (Duarte, 2008).

K-means

O algoritmo K-means define o centróide de um *cluster* através do valor médio dos pontos do *cluster*. O algoritmo seleciona aleatoriamente um conjunto de registos para formar os *clusters* iniciais (Figura 9(a)), cada *cluster* tem uma média que representa o centro do *cluster.* Os restantes registos são atribuídos aos *clusters* que apresentam maior similaridade, baseada na distância euclidiana entre o registo e a média do *cluster*. Após adicionar um novo registo ao *cluster* é necessário recalcular a média dos *clusters* (Figura 9(b)). O processo de recalcular a média dos *clusters* é repetido até que todos os registos estejam atribuídos a um *cluster*. O processo acaba quando os *clusters* formados durante uma interação é igual à interação anterior (Figura 9(c)) (Han et al., 2011).

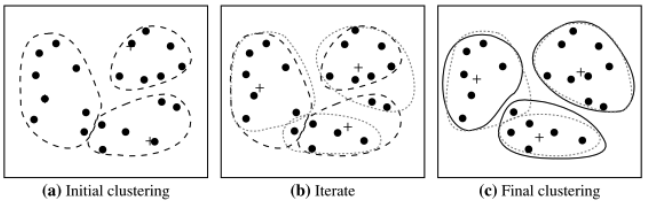


Figura 9 *Clustering* de um conjunto de pontos usando o algoritmo K-means (Han et al., 2011)

O algoritmo K-means apresenta ter bastante sensibilidade em relação aos *outliers*. Como o outlier é um ponto que se encontra bastante longe dos restantes pontos, quando atribuído a um cluster faz com que distorça dramaticamente a média do cluster. Como o K-means utiliza a função squared-error faz com que seja mais gravoso o efeito a quando a existência de outliers (Han et al., 2011).

Para lidar com este problema foi criado uma variação do K-means, de forma a eliminar ou reduzir a sensibilidade aos *outliers*, o K-medoids. Os algoritmos baseados em k-medoids ao contrário do K-means que utiliza o valor médio dos pontos dos *clusters* como referência para calcular a distância entre pontos, usam o ponto mais centralizado no *cluster*. Este usa o mesmo principio que o k-means para gerar os *clusters*, recorrendo a função *squared-error* (Duarte, 2008).

- **Métodos Baseados em Grelha**

Os métodos baseados em grelha colocam os pontos referentes aos objetos numa estrutura em grelha com um numero finito de células, todas as operações de *clustering* ocorrem na estrutura criada. A vantagem deste método é a o tempo de processamento está dependente do tamanho da estrutura em grelha e não na quantidade de dados processados(Rokach & Maimon, 2010). Também é possível integrar este método com outros métodos de *clustering*, tais como métodos baseados em densidade e hierarquia. Alguns dos algoritmos baseados em grelha são o *Statistical Information Grid-based* (STING), *WaveCluster e o Clustering In Quest* (CLIQUE) (Kotsiantis & Pintelas, 2004). O STING procura explorar a informação estatística presentes nas células da grelha. O CLIQUE e o *WaveCluster* são dois algoritmos que que utilizam o método baseado em gralha e em densidade (Han et al., 2011).

STING

Este método divide a área espacial em vários níveis de células retangulares de forma a criar uma estrutura organizada em hierarquias, como na Figura 10. As células dos níveis superiores são compostas a partir das células dos níveis inferiores, o que permite representar os *clusters* de níveis diferentes. O STING apresenta dois problemas, a performance deste está dependente da granularidade do nível mais baixo e os *clusters* resultantes são apena limitados na horizontal ou vertical e nunca na diagonal, o que pode afetar a qualidade dos *clusters* (Kotsiantis & Pintelas, 2004).

O STING utiliza os parâmetros estatísticos das células dos níveis superiores para computar os parâmetros das células dos níveis inferiores. Estes parâmetros incluem os seguintes atributos, *count* (contar), *mean* (média), *stdev* (desvio-padrão), *min* (mínimo), *max* (máximo) e o tipo de distribuição que a célula segue, tais como normal, exponencial ou nenhuma. Quando os dados são carregados para a base de dados, os parâmetros *count, mean, stdev, min, max* dos níveis inferiores são calculados diretamente, a distribuição deve ser carregada caso seja conhecido o seu tipo. O tipo de distribuição das células de nível inferior necessita de ser igual ao do nível superior, de modo que seja possível realizar o teste de hipóteses. Quando o tipo de distribuição dos níveis inferiores difere, deve-se colocar o tipo de distribuição para as células de nível superior como none (nenhum).

O STING para responder às interrogações (query) procede a um conjunto de passos em que são utilizados os parâmetros mencionados anteriormente. Em primeiro é necessário determinar a

camada na estrutura hierárquica onde o processo de resposta á query deverá começar. Tipicamente esta camada contém um pequeno conjunto de células. Para cada célula desta camada é calculado o intervalo de confiança (IC), revelando assim a importância das células em relação à *query*. As células com baixa importância são descartadas nos seguintes passos. O processamento do nível abaixo seguinte apenas examina as células relevantes que restaram. Este processo é repetido até se chegar ao nível mais abaixo da estrutura hierárquica (Figura 10). Caso as condições da *query* sejam encontradas, a região de células relevantes é retornada. Caso contrário, o processo continua até que se satisfaça as condições da query (Han et al., 2011).

Resumo do algoritmo STING (Wang, Yang, & Muntz, 1997) :

(1) Determinar uma camada para começar a responder à query;
(2) Para cada célula da camada criada, são calculados os IC.
(3) Através do IC a célula é rotulada de relevante ou irrelevante.
(4) Se a camada em causa for a ultima camada, salta-se para o passo (6), caso contrário passar para o passo (5).
(5) Descer um nível na estrutura hierárquica e repetir o passo (2) para as células relevantes definidas em (3).
(6) Se as condições da *query* são satisfeitas, passar par ao passo 8, caso contrário, passar para o passo (7).
(7) Retribuir todas as células rotuladas como relevantes e realizar o processamento. Retornar os resultados que vão de encontro com objetivos em questão e passar para o passo (9).
(8) Procurar regiões de células relevantes. Retornar as regiões que vão de encontro com as condições da *query* e passar para o passo 9.
(9) Fim.

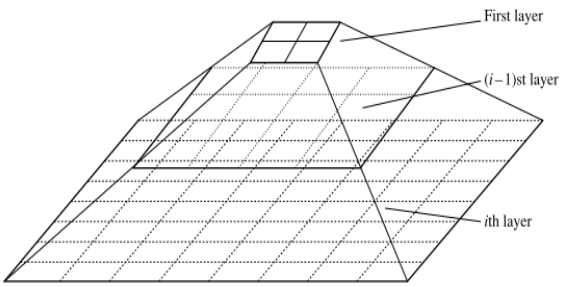


Figura 10 Estrutura Hierárquica do STING, representação de níveis e células (Han et al., 2011)

CLIQUE

É um método baseado em grelha e densidade. CLIQUE particiona cada dimensão em intervalos não sobrepostos, assim, todo o espaço que incorpora os objetos é representado em células, como ilustrado na Figura 11. Para definir quais as células de maior densidade, este utiliza um limiar de densidade. Se o número de objetos mapeado exceder o limiar de densidade definido, é considerado uma célula densa (Han et al., 2011).

O CLIQUE procura descobrir correlações entre os dados em vários subespaços do espaço original. Este pretende identificar quais os subespaços que permitem um melhor agrupamento dos objetos do que no espalho original (Duarte, 2008).

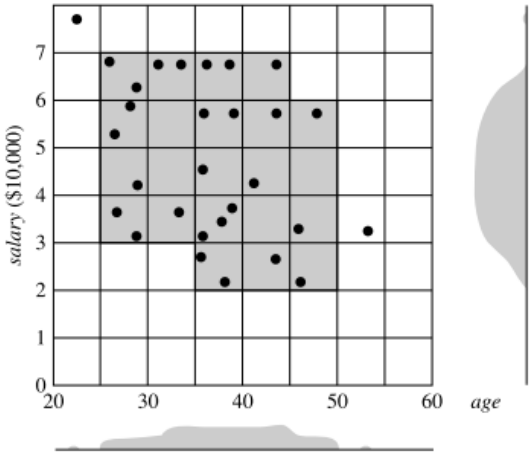


Figura 11 Representação das células de densidade da relação entre a idade e o salário (Han et al., 2011)

A principal estratégia do algoritmo CLIQUE para identificar espaços de pesquisa candidatos baseia-se na propriedade *Apriori* usada em outra técnica de *Data Mininig*, as regras de associação (Han et al., 2011).

Para encontrar grupos nas regiões formadas por células densas é utilizado um algoritmos de pesquisa em profundidade ao conjunto de células densas com o objetivo de agrupar regiões densas que se encontram ligadas (Duarte, 2008).

WaveCluster

O algoritmo *WaveCluster* trabalha com atributos numéricos. Este é reconhecido devido as suas propriedades, fornecer *clusters* de qualidade, capacidade de trabalhar com dados espaciais de grandes dimensões e lidar com *outliers*.

O *WaveCluster* é baseado no funcionamento do processamento de sinal. Este aplica um *wavelet* para filtrar os dados. *Wavelet* é um função com capaz de decompor e descrever ou representar outra função (Torrence & Compo, 1998). As partes de um sinal que apresentam alta frequência correspondem a limites, enquanto os baixos frequência de alta amplitude de um sinal correspondem aos interiores dos clusters. A aplicação de um *wavelet* permite (Berkhin, 2006):

1) Coletar todas as dimensões e atribuir os pontos ás células correspondentes;
2) Aplicação do *wavelet* discreto que acumula as células
3) Procura componentes conectadas (clusters)
4) Atribui os pontos

- **Métodos baseados em Densidade**

Os métodos de partição e hierárquicos foram desenhados para encontrar *clusters* com forma esférica. Estes apresentam problemas quando é necessário encontrar *clusters* com formas arbitrárias. Para encontrar *clusters* com formas arbitrárias, alternativamente, podemos criar *clusters* tendo em conta as regiões densas no espaço, separados por regiões dispersas. Esta é a principal estratégia por de trás dos algoritmos baseados em densidade (Han et al., 2011).

Os algoritmos DBSCAN (*Density-based spatial clustering of applications with noise*) e o OPTICS (*Ordering points to identify the clustering structure*) são os algoritmos mais representativos quando falamos de algoritmos baseados em densidade. O DBSCAN, primeiro introduzido por (Martin Ester, Kriegel, Sander, & Xu, 1996), serve de base para muitos dos algoritmos deste tipo que se encontram disponíveis (J. Moreira, 2013). Por isso será explicado o seu funcionamento. Outro algoritmo que será explorado é o SNN (*Shared Nearest Neighbour*). Este é parecido com o DBSCAN mas apresenta

ligeiras diferenças. Com o SNN é apresentado pela primeira vez a ideia de usar como medida de similaridade o número de vizinhos mais próximos que dois pontos partilham (Jarvis & Patrick, 1973).

DBSCAN

A ideia chave do DBSCAN é que para cada ponto pertencente a um *cluster* o raio da vizinhança deve conter no seu espaço um número mínimo de pontos. A forma da vizinhança é determinada pela escolha da função distância utilizada. O DBSCAN utiliza qualquer função distância. Desta forma permite utilizar funções diferentes para resolver problemas com características diferentes. O DBSCAN necessita como parâmetros de entrada o Eps e o MinPts. O Eps é o valor do raio de um ponto. O MinPts é o número mínimo de pontos que uma vizinhança deve ter para ser considerada um *cluster* (Martin Ester et al., 1996). O processo de *clustering* do DBSCAN baseia-se na classificação dos pontos do *data set* em *core points, border points* e *noise points*, na Figura 12 estão representados os três tipos de pontos (A. Moreira, Santos, & Carneiro, 2005).

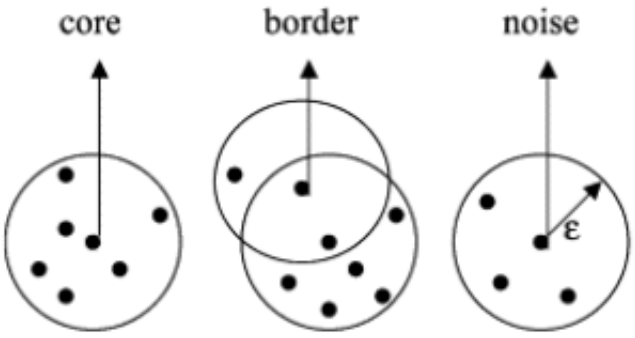


Figura 12 Representação dos tipos de pontos do DBSCAN, Eps = ε e MinPts = 6 (J. Moreira, 2013).

O DBSCAN inicialmente marca todos os pontos do *dataset* como não visitados. Este seleciona aleatoriamente um dos pontos não visitados e procura na vizinhança definida pelo parâmetro Eps, se existe pelo menos o MinPts de pontos. Caso não satisfaça os parâmetros é marcado como *noise point.* Caso contrário, um novo *cluster* é iniciado. Se um ponto fizer parte de um *cluster*, a sua vizinhança também pertencerá a esse mesmo cluster. Assim, todos os pontos encontrados dentro da vizinhança são adicionados ao *cluster* (Figura 13). Este processo continua até que o *cluster* esteja

completamente ligado através da densidade. O DBSCAN passa para outro ponto não visitado para encontrar outros *clusters*. (Han et al., 2011).

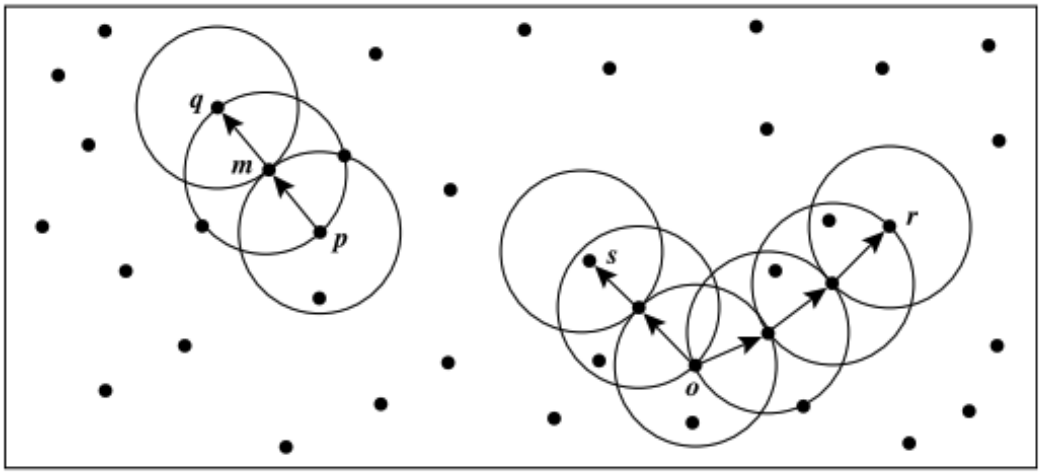


Figura 13 Representação de *clusters* obtidos através da *density-reachability* e *density-connectivity* (J. Moreira, 2013).

SNN

Como já foi anteriormente referido o SNN é o primeiro algoritmo a utilizar como medida de similaridade o número de vizinhos mais próximos entre dois pontos. (Levent Ertöz, Michael Steinbach, 2002) apresenta um SNN melhorado que procura lidar com vários desafios que os algoritmos de *clustering* costumam enfrentar. O SNN segundo estes autores permite encontrar *clusters* na presença de ruído e *outliers* e encontrar *clusters* de várias formas, tamanhos e densidade. Esta abordagem de *clustering* consegue manipular conjunto de dados de grande dimensão onde os conceitos de distância e densidade estão mal definidos. Para superar o problema com conjunto de dados de grande dimensão, é utilizado uma medida de distância baseada no número de vizinhos que dois pontos partilham. Para ultrapassar os problemas com a densidade, foi definido que a densidade de um ponto seria o somatório dos pontos vizinhos semelhantes (Ertöz, Steinbach, & Kumar, 2003).

O SNN necessita como parâmetros de entrada, K, Eps e MinPts. K é o número de vizinhos, Eps define o limiar de densidade e MinPts é o mínimo de densidade para que um ponto possa ser considerado *core point*. O parâmetro mais importante é o K. Esta influência fortemente a

granularidade dos *clusters*. Se K for demasiado pequeno, até um cluster uniforme será dividido em vários *clusters*. O mesmo acontece ao contrário, se K for demasiado grande, o algoritmo irá apenas encontrar alguns *clusters* em conjunto de dados de grande dimensão (Oliveira, Santos, & Pires, 2013).

Os principais passos do algoritmo SNN são (Ertöz et al., 2003):

1) Computar a matriz de similaridade: criação de um gráfico de similaridade com os pontos representados como nodos e nas pontas os pesos da similaridade entre os pontos.

2) Redução da Matriz de similaridade: são apenas mantidos os K com maior número de vizinhos similares.

3) Criação do gráfico SNN: aplicação do limiar de similaridade sobre a matriz de similaridade.

4) Calcular a densidade do SNN em cada ponto: usando o valor Eps, filtra-se os que tenham valor igual ou superior ao Eps definido.

5) Encontrar *core-points*: filtrar os pontos que apresentam uma densidade maior que o MinPts definido.

6) Formar os *clusters*: se dois *core points* estão dentro do mesmo raio, Eps, eles são colocados dentro do mesmo *cluster.*

7) Descartar todos os *noise points*: todos os pontos que não são *core points* e não estão dentro de um raio, Eps, de um *core point* é considerado *noise* e consequentemente é descartado.

8) Atribuir todos os restantes pontos a um *cluster*: *non-noise* e *non-core points* são atribuídos ao core *point* mais próximo.

Para medir a similaridade dos pontos, é necessária uma função de distância, e devido à complexidade computacional, a escolha da função distância é importante. A função distância tem grande influência nos *clusters* obtidos. O exemplo de uma função distância é a distância Euclidiana entre dois pontos.

Avaliação de qualidade do clustering

Para avaliar a qualidade dos *clusters* neste trabalho serão utilizadas medições baseadas na similaridade de pontos sendo cada ponto um registo da base de dados, representado pela sua longitude e latitude. Esta medição necessita de calcular três métricas: medida *Intracluster*, medida *Intercluster* e medida *ModelQuality.* A utilização destas medidas deve-se aos resultados obtidos em (Santos, Moreira, & Carneiro, 2005) e (Galvão, 2014).

A medida *IntraCluster* é utilizada para avaliar a semelhança dentro dos *clusters*. Esta medida é calculada através da soma das similaridades entre todos os pontos de um *cluster* com o ponto médio desse *cluster*, dividido pelo total de pontos do *cluster*. Este cálculo é repetido para todos os clusters. Os resultados de cada *clusters* são agregados para que no final o total seja dividido pelo numero de *clusters* (Galvão, 2014).

A medida *Intercluster* é utilizada para avaliar a semelhança entre *clusters*. Esta medida é calculada através do somatório das similaridades entre todos os pontos de dois *clusters* diferentes. O resultado é dividido pelo número de par de pontos. O cálculo da similaridade é efectuado através da utilização da função distância. Este processo é repetido para cada par de *clusters*. O resultado do processo é somado a cada iteração. No fim o resultado dado pela razão entre o somatório das várias iterações e o número total de pares de *clusters* (Galvão, 2014).

A medida *ModelQuality* é calculada através da diferença entre as medidas *Intracluster* e *Intercluster*. O maior ganho está no equilíbrio entre estas duas medidas. Ou seja quando menor a diferença entre as duas medidas, mais afastados se encontram os pontos (Galvão, 2014).

2.2.2.4. Redes Neuronais Artificiais

As redes neuronais artificiais podem ser caracterizadas como um modelo computacional com particularidades, tais como a capacidade de adaptar-se ou aprender, com o objetivo de generalizar, ou para agrupar ou organizar dados, em que cada operação é executada em processamento paralelo (Hodzic, Konjic, & Miranda, 2006).

Redes neuronais artificiais tentam de certa forma imitar o processo de aprendizagem do cérebro humano, operando como um modelo tipo "*black box*", não necessitando de informação detalhada sobre o sistema (Kalogirou, 2001). Esta informação é usada para executar o processo de treino da rede para poder obter previsões (Lek & Guégan, 1999).

A investigação nesta área levou ao desenvolvimento de vários tipos de redes neuronais, com o objetivo de responder a problemas com natureza diferente: otimização, descoberta de padrões de associação e reconhecimento, tarefas de controlo e previsão (Lek & Guégan, 1999). Estas têm sido aplicadas com sucesso em vários campos da matemática, engenharia, medicina, economia, meteorologia e muitas outras. Por exemplo no campo da meteorologia tem sido utilizado para prever o estado do tempo e no campo da economia na previsão do comportamento dos mercados (Kalogirou, 2001).

A abordagem de redes neuronais artificias mais utilizada é a *backpropagation network* (BPN) ou também chamada de *Multi-layer feed-foward neural network* (Lek & Guégan, 1999).

A rede é constituída normalmente pelos seguintes elementos: camada de entrada (*input layer*), camadas ocultas (*hidden layers*), e camada de saída (*output layer*), representadas na Figura 14. Cada neurónio (nó) está conectado aos neurónios da camada anterior e seguinte através de conexões. Estas conexões têm atribuído um peso que varia durante o processo de treino, estas variações influenciam o resultado final do processo.

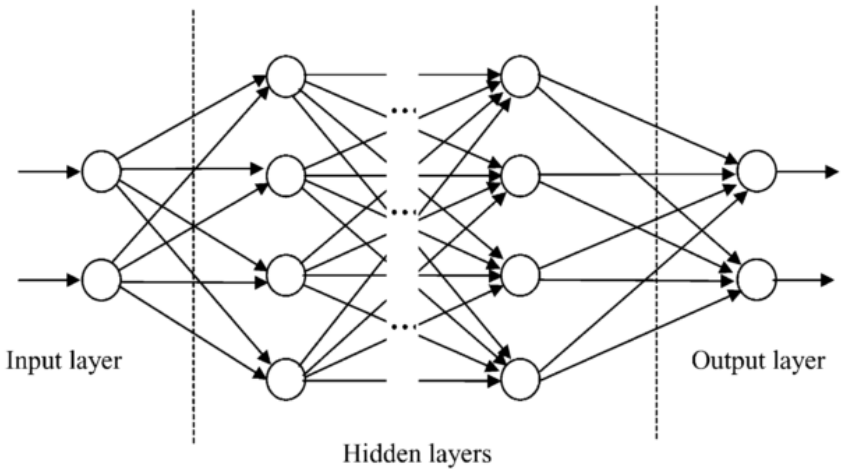


Figura 14 Diagrama de uma BPN (Kalogirou, 2001).

3. Caracterização Geo-Espacial das Pneumonias

Neste capítulo da dissertação será realizada a descrição e seleção dos dados necessários para cumprir os objectivos propostos. Em seguida são definidas as parametrizações e os pesos para a abordagem de *clustering* utilizada, o F-SNN. Após a criação e execução de vários modelos, a qualidade destes é avaliada através da utilização de métricas apropriadas. Por fim, é efetuada a representação no espaço dos resultados referentes a alguns modelos considerados como os mais interessantes.

3.1. Descrição e Seleção dos Dados

Os dados utilizados para a realização deste trabalho encontram-se armazenados num *Data Warehouse* como referido na secção 1.1. Este foi criado anteriormente em uma outra iniciativa promovida pela Fundação Portuguesa do Pulmão. Os dados que integram o *Data Warehouse* correspondem aos registos de ocorrências de pneumonias e aos registos de ocorrências de incêndios em Portugal Continental, correspondentes ao período entre 2002 e 2011. Para complementar o *Data Wahrehouse* foram integrados os resultados dos censos realizados em 2011 (Leite, 2014).

O DW é constituído por quatro tabelas de factos, **Incidencia_Pneumonias**, **Incidencia_Incendios**, **Incidência_Patologias** e **Dados_Estatisticos** e nove tabelas de dimensão, **Incendio**, **Tipo_Incendio**, **Classificacao**, **Causa**, **Local**, **Hospital**, **Tempo_Mes**, **Tempo_Ano** e **Patologias**. Serão só utilizadas as tabelas de factos e as tabelas dimensões relevantes para os objetivos da dissertação, deste modo as tabelas com dados relacionados com os incêndios não serão utilizadas. Para as tabelas selecionadas será feita uma breve descrição da utilidade de cada tabela de factos e os respetivos factos e uma breve descrição de cada tabela de dimensão. Uma descrição mais detalhada sobre as tabelas de factos e dimensões pode ser consultada na dissertação "Business Intelligence no estudo das Pneumonias e da sua incidência em Portugal" (Leite, 2014).

A tabela de factos **Incidencia_Pneumonias** tem como objetivo guardar a informação referente às pneumonias registadas em Portugal Continental desde 2002 a 2011. Esta é constituída pelos factos **n_reingressos** que permite saber o número de vezes que um paciente deu entrada num hospital com sintomas de Pneumonia, o contador de eventos **evento_pneumonia** que nos permite contar o número de casos de pneumonia ocorridos, o campo **dias_internamento** que nos indica o número

de dias que um paciente permaneceu internado devido à doença em estudo e o campo **flag_vitima_mortal** que indica se o paciente faleceu.

A tabela de factos **Incidencia_Patologias** tem como objetivo guardar informação referente a outras patologias que afectam os pacientes portadores de pneumonia. As patologias registadas além da pneumonia são Diabetes Millitus, Doença Cardíaca Crónica, Doença Pulmonar Crônica, Doença Hepática Crónica, Doença Pancreática Crónica e Doença Renal Crónica. Esta é constituída apenas por um facto, o contador de eventos, **evento_patologia**.

A tabela de factos **Dados_Estatisticos** guarda a informação referente aos indicadores populacionais provenientes dos Censos realizados no ano de 2011.

A tabela de dimensão **Local** guarda a informação relativa à localização relativamente ao distrito, concelho e freguesia, para que seja possível identificar os pacientes por regiões.

A tabela de dimensão **Hospital** guarda a informação referente aos hospitais, de forma que seja possível identificar o hospital em que o paciente foi atendido.

A tabela de dimensão **Tempo_Mes** tem como objetivo localizar temporalmente os acontecimentos, através dos campos **dia**, **mes** e **trimeste**.

A tabela de dimensão **Tempo_Ano** permite associar os anos aos acontecimentos e foi criada devido à tabela de factos **Dados_Estastisticos**, que se encontra organizada apenas pelo ano.

A tabela de dimensão **Individuo** guarda a informação relativa aos pacientes, tais como o sexo, a idade e a sua posição geográfica. A posição geográfica nesta tabela é constituída pelas coordenadas "agitadas". A agitação das coordenadas foi necessária devido à grande sobreposição de pontos. Se estes pontos não sofressem esta transformação a sua representação no espaço não seria percetível e não seriam uteis na caracterização geo-espacial da Pneumonia.

A tabela de dimensão **Patologias** tem como objetivo guardar a informação relativa a outras patologias presentes nos indivíduos além da pneumonia.

Para a realização deste trabalho serão criados dois *datasets* com dados extraídos do DW anteriormente referido. O dataset1 criado contém o registo de todos os pacientes com pneumonia, 368 121 registos. Foram removidos os registos que não possuíam dados nos campos referentes às coordenadas geográficas, porque são necessárias as coordenadas dos indivíduos para que seja

possível executar a abordagem de *clustering* utilizada na secção 3.2. Este *dataset* é constituído pelos seguintes campos **longitude**, **latitude**, **idade, mes**, **sexo**, **reingresso**, **internamento**, **vitima_mortal**, **distrito** e **hospital** (Tabela 2). Como referido, os campos **longitude** e **latitude** são necessários para que seja possível localizar os pontos e calcular a distância entre pontos no processo de clustering.

	Tipo	Estatísticas	Intervalo de valores
longitude	real	avg = -8.460 +/- 0.594	[-9.462 ; -6.147]
latitude	real	avg = 39.986 +/- 1.254	[37.000 ; 42.210]
sexo	binominal	mode = M (204065), least = F (164055)	M(204065),F(164055)
reingresso	integer	avg = 0.332 +/- 0.630	[0 ; 19]
internamento	integer	avg = 10.715 +/- 11.563	[9 ; 947]
vitima_mortal	boolean	_	[0 ; 1]
mes	integer	avg = 6.074 +/- 3.699	[1 ; 12]
idade	integer	avg = 65.543 +/- 26.892	[0 ; 111]
distrito	polynominal	mode = Lisboa (24560), least = Barrancos (38)	Lisboa (24560), Loures (8492), Vila Nova de Gaia (8475),...,
hospital	polynominal	mode = Hospital Universidade de Coimbra (13393), least = Hospital Ortopédico Santiago do Outão (1)	Hospital Universidade de Coimbra (13393), Centro Hospitalar Do Médio - Tejo; EPE (12395), Hospital São Teotónio; EPE (11948),..,

Tabela 2 Descrição dos campos do dataset1

Os campos longitude e latitude formam as coordenadas geográficas que permitem localizar no espaço cada paciente. Os campos **sexo**, **idade**, **mes**, **distrito** e **hospital** indicam respetivamente o sexo, a idade de cada paciente, o mês em que foi registado o caso de pneumonia, o distrito onde vive e em que hospital este obteve cuidados hospitalares. Os restantes campos, **reingressos** e **internamento** indicam se o paciente reingressou ao hospital com pneumonia e o tempo em dias que este se encontrou internado, respetivamente.

O dataset2 criado tem como objeto possibilitar o estudo das patologias associadas aos casos de pneumonia. Desta forma foi necessário apenas selecionar os pacientes com patologias. Este processo resultou num *dataset* de menores dimensões com 212 788 registos. Isto porque nem todos os pacientes têm registo de outras patologias. Os campos que constituem o dataset2 são os campos apresentados na Tabela 2 mais os campos da Tabela 3.

	Tipo	Estatísticas	Intervalo de valores
Diabetes_Millitus	boolean	_	[0 ; 1]
Doença_cardiaca_cronica	boolean	_	[0 ; 1]
Doença_pulmonar_cronica	boolean	_	[0 ; 1]
Doença_hepatica_cronica	boolean	_	[0 ; 1]
Doença_pancreatica_cronica	boolean	_	[0 ; 1]
Doença_renal_cronica	boolean	_	[0 ; 1]

Tabela 3 Campos adicionais para o dataset2

Os campos **Diabetes_Millitus**, **Doença_cardiaca_cronica**, **Doença_pulmonar_cromica, Doença_hepatica_cronica**, **Doença_pancreatica_cronica** e **Doença_renal_cronica** indicam, através da utilização de um atributo booleano, se o paciente tem ou não a respetiva patologia.

3.2. Construção dos modelos

Sendo um dos objetivos deste trabalho identificar conjuntos de pacientes com características semelhantes e representá-los geograficamente, foi utilizado um algoritmo de *clustering* baseado em densidade que recorre à similaridade entre prontos para criar os *clusters*. O algoritmo utilizado é o SNN (*Shared Nearest Neighbour*). O algoritmo SNN procura semelhanças diretas que definem um par de pontos em termos da partilha de vizinhos (Antunes et al., 2014). Esta partilha é confirmada através da existência de um vizinho comum (Levent Ertöz, Michael Steinbach, 2002). Devido ao

elevado número de dados foi utilizada a abordagem F-SNN (Antunes et al., 2014) que permite aplicar *clustering* a um grande volume de dados de forma bastante mais rápida em relação ao algoritmo original, o SNN. Na Figura 15 é possível visualizar uma comparação entre o tempo de execução das duas abordagens, destacando-se o desempenho otimizado do F-SNN, quer em contexto de análise de dados 2D ou 3D, acrescentando outra dimensão para além das coordenadas dos pontos.

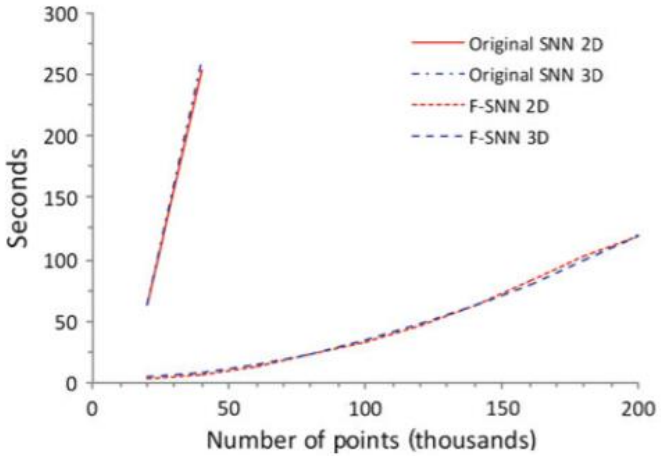


Figura 15 Gráfico de comparação do SNN e F-SNN retirado de (Antunes et al., 2014)

Como é possível verificar pela Figura 15 existe uma grande diferença de desempenho do SNN original para o F-SNN. Como já referido, o F-SNN apresenta ser, em termos de tempo de processamento, mais rápido. Assim, como será necessário processar várias vezes um elevado número de registos para ser possível analisar e comparar resultados dos vários modelos, é necessário optar pela utilização da abordagem que permita efetuar várias corridas do algoritmo em tempo útil, neste caso o F-SNN. Caso se utilize a abordagem tradicional o tempo de processamento seria mais extenso e levaria horas a obter os resultados.

Após a escolha da abordagem referida anteriormente foi necessário, para a construção dos modelos, identificar a parametrização a utilizar no F-SNN. Para calcular estes valores foi utilizada a técnica de cálculo desenvolvida por José Guilherme Moreira (J. Moreira, 2013) para o SNN. Os parâmetros para os quais precisamos atribuir valores são o k (tamanho da lista de vizinhos), Eps (é o raio que limita a área de vizinhança de cada ponto) e MinPts (é o número mínimo de pontos que devem existir na vizinhança defina pelo Eps) (A. Moreira et al., 2005). (J. Moreira, 2013) demonstra que o valor para o parâmetro K é igual a n*0.7%, em que n é o número de registos existentes no *dataset* usado. Para o parâmetro MinPts é igual a K*94% e para o parâmetro Eps é giual a MinPts*18.5%. Estes valores, no caso da implementação usada, são alterados diretamente no código

do F-SNN. O código java da implementação pode ser encontrada através do link *http://ubicomp.algoritmi.uminho.pt/projects/f-snn.* Os valores obtidos para os 368121 e 212787 registos existentes no dataset1 e dataset2 utilizados neste trabalho foram: K= 2576, MinPts=2421, Eps=447 e k= 1489, MinPts=1399, Eps=258, respectivamente.

Havendo uma limitação da capacidade de processamento nas máquinas pessoais no geral, quando é utilizado um conjunto de dados com elevado número de registos, torna-se praticamente impossível conseguir a conclusão do algoritmo, a maior parte das vezes devido à falta de memória. De forma a ultrapassar este problema, José Guilherme Moreira (J. Moreira, 2013) através de execuções do F-SNN sobre um *dataset* de teste no qual já se conhecia o resultado pretendido, concluí que seria possível obter resultados próximos ou iguais dividindo o valor de K por 2 o número de vezes necessárias e recalcular o MinPts e Eps com o novo K. Outra questão que levou ao uso de valores pequenos para k, foi a complexidade quadrática presente no algoritmo do SNN (Antunes et al., 2014). Assim, com a diminuição do valor de K a lista de vizinhos torna-se mais pequena, o que permite executar o *clustering* num período de tempo razoável. Os valores apresentados na terceira coluna da Tabela 4 serviram de referência na experimentação dos modelos para o dataset1 e dataset2 identificados na secção 3.2.

	Parâmetros	Sem redução	Após 5 reduções
Dataset1	K	2576	80
	MinPts	2421	75
	EPS	447	13
Dataset2	K	1489	47
	MinPts	1399	44
	EPS	258	8

Tabela 4 Parâmetros reduzidos para o dataset1 e dataset2

Para além destes parâmetros existe a necessidade de configurar a função distância utilizada pelo algoritmo. Esta alteração é executada diretamente no código. A função foi programada para utilizar três dimensões, longitude e latitude que são utilizadas para calcular a distância geográfica entre dois pontos, e a idade dos indivíduos (equação 1). Foi também alterada para quatro dimensões acrescentado a dimensão mês do registo de ocorrência de pneumonia (equação 2). O uso das dimensões longitude, latitude, idade e mês na função distância permitem ao algoritmo agrupar os indivíduos pela sua posição geográfica, similaridade de idades e os meses em que foi registado o caso de pneumonia.

As funções distância utilizadas são representadas pela Equação 1 e Equação 2, adaptadas de (Oliveira et al., 2013).

Equação 1: $3D\ Distance(p_1, p_2) = W_r * \frac{D_r(x_1,x_2,y_1,y_2)}{Max_r} + W_i \frac{Di(i_1,i_2)}{Max_i}$

Equação 2: $4D\ Distance(p_1, p_2) = W_r * \frac{D_r(x_1,x_2,y_1,y_2)}{Max_r} + W_i \frac{Di(i_1,i_2)}{Max_i} + W_m \frac{Dm(m_1,m_2)}{Max_m}$

As equações 1 e 2 traduzem o cálculo efetuado para determinar a distância entre dois pontos para três e quatro dimensões, respetivamente. p_1 e p_2 correspondem aos pontos em processamento. Os valores (x_1, x_2) e (y_1, y_2) correspondem às coordenadas dos pontos p_1 e p_2, respetivamente. Os valores i_1, i_2, m_1e m_2 correspondem aos valores das dimensões idade e mês utilizados nos pontos p_1 e p_2, em que i e m correspondem às dimensões idade e mês, respetivamente. Os valores W_r, W_i e W_m correspondem aos vários pesos atribuídos às dimensões, desta forma é possível controlar os resultados de acordo com o contexto analítico pretendido. D_r corresponde à função distância que calcula a distância entre os pontos p_1 e p_2, no caso deste trabalho é a distância geográfica. D_i corresponde ao cálculo do valor absoluto da diferença de idades entre i_1 e i_2. O valor D_m corresponde ao cálculo do valor absoluto da diferença de meses entre m_1 e m_2. Os valores Max_r, Max_i, Max_m são utilizados para normalizar as dimensões, estes correspondem à distância máxima, idade máxima e mês máximo, respetivamente. Estas variáveis apresentam uma importância significativa, porque permitem ajustar as escalas das dimensões quando estas apresentam valores de diferentes grandezas, obtendo escalas similares entre as dimensões.

O algoritmo será executado para três e quatro dimensões separadamente. Em primeiro será executado para três dimensões, posição (longitude e latitude) e idade, e em segundo para quatro dimensões, posição, idade e mês. Assim será possível comparar os resultados entre a utilização de três e quatro dimensões. Além da comparação de resultados referida existiu a necessidade de avaliar vários valores para K, utilizando como referência os valores calculados e apresentados na Tabela 4. Desta forma, é possível estudar o comportamento dos resultados com diferentes parametrizações. No entanto, outras parametrizações podem obter resultados igualmente interessantes. A abordagem para parametrizar o algoritmo passou por utilizar os valores de referência calculados na Tabela 4 e através destes valores foram gerados diferentes valores para K e recalculados os parâmetros MinPts e Eps. A forma escolhida para gerar novos valores para K foi através da soma e subtração de 5

unidades aos valores de K e recalculando os valores MinPts e Eps. A opção de utilizar 5 unidades para criar as variações baseou-se na observação do número de *clusters* em execuções do F-SNN realizadas para experimentação, tendo sido verificado que variar o valor em 5 unidades já influenciava bastante o número de *clusters* criados. Como explicado anteriormente quanto maior o valor de K maior seria o tempo de processamento. Assim, a utilização de valores muito superiores ao valor inicialmente calculado levaria a um aumento considerável no tempo de processamento, o que não seria viável. A utilização de vários valores para K tem como objetivo possibilitar o estudo do comportamento dos resultados conforme o aumento ou redução do valor de k. Os parâmetros resultantes podem ser consultados na Tabela 5 e na Tabela 7.

Outra questão a definir são os valores para os pesos das dimensões. Estes pesos influenciam diretamente o resultado do algoritmo, como é possível verificar nas equações 1 e 2. Nesta fase foram utilizadas as parametrizações calculadas na Tabela 4 e aplicando um conjunto possível de combinações de valores para os pesos, podendo estes variar entre 0 e 1. A utilização de vários pesos visa o estudo da influência dos pesos atribuídos a cada dimensão nos resultados obtidos. As configurações utilizadas podem ser consultadas na Tabela 6 e na Tabela 8. Nas configurações de três dimensões, utilizou-se as combinações possíveis variando 0,1 unidades.

Para as configurações de quatro dimensões foram atribuídos pesos às dimensões idade e mês conforme o peso atribuído à posição. Os valores atribuídos às dimensões idade e mês são relativamente próximos. A razão disto acontecer passa pelo interesse de manter o equilíbrio entre estes dois atributos. Nas pré-execuções realizadas, observou-se que quando o valor do peso da posição se mantém fixo e o peso da idade é superior ao peso do mês o número de *clusters* aumenta. Nos casos em que o peso da idade é inferior ao mês o número de *clusters* diminui. Nos casos em que as dimensões idade e mês apresentam um peso superior à dimensão posição, os *clusters* resultantes tendem a ser bastante homogéneos em termos de idade e mês. Este tipo de parametrização origina *clusters* com características semelhantes e muito próximas no espaço criando várias sobreposições de *clusters*, que são apenas possíveis de visualizar recorrendo a gráficos de 3 dimensões.

As configurações utilizadas para as execuções do algoritmo para três dimensões são as apresentadas nas tabelas 5 e 6, enquanto que para quatro dimensões são usadas as parametrizações apresentadas nas tabelas 7 e 8.

Runs	K	MinPts	Eps	W_r-Posição	W_i-Idade
C1	105	98	18	0,5	0,5
C2	100	94	17	0,5	0,5
C3	95	89	16	0,5	0,5
C4	90	84	15	0,5	0,5
C5	85	79	14	0,5	0,5
C6	80	75	13	0,5	0,5
C7	75	70	12	0,5	0,5
C8	70	65	12	0,5	0,5
C9	65	61	11	0,5	0,5
C10	60	56	10	0,5	0,5

Tabela 5 Configurações F-SNN para valores de K diferentes do dataset1 para as três dimensões

Runs	K	MinPts	Eps	W_r-Posição	W_i-Idade
C11	80	75	13	0,9	0,1
C12	80	75	13	0,8	0,2
C13	80	75	13	0,7	0,3
C14	80	75	13	0,6	0,4
C15	80	75	13	0,5	0,5
C16	80	75	13	0,4	0,6
C17	80	75	13	0,3	0,7
C18	80	75	13	0,2	0,8
C19	80	75	13	0,1	0,9

Tabela 6 Configurações F-SNN para K constante e diferentes pesos para as três dimensões do dataset1

Runs	K	MinPts	Eps	W_r-Posição	W_i-Idade	W_m-Mês
C20	105	98	18	0,33	0,33	0,33
C21	100	94	17	0,33	0,33	0,33
C22	95	89	16	0,33	0,33	0,33
C23	90	84	15	0,33	0,33	0,33
C24	85	79	14	0,33	0,33	0,33
C25	80	75	13	0,33	0,33	0,33
C26	75	70	12	0,33	0,33	0,33
C27	70	65	12	0,33	0,33	0,33
C28	65	61	11	0,33	0,33	0,33
C29	60	56	10	0,33	0,33	0,33

Tabela 7 Configurações F-SNN para valores de K diferentes do dataset1 para as quatro dimensões.

Runs	K	MinPts	Eps	W_r-Posição	W_i-Idade	W_m-Mês
C30	80	75	13	0,8	0,1	0,1
C31	80	75	13	0,7	0,2	0,1
C32	80	75	13	0,6	0,2	0,2
C33	80	75	13	0,5	0,3	0,2
C34	80	75	13	0,4	0,3	0,3
C35	80	75	13	0,3	0,4	0,3
C36	80	75	13	0,2	0,4	0,4
C37	80	75	13	0,1	0,4	0,5
C38	80	75	13	0,1	0,5	0,4
C39	80	75	13	0,3	0,3	0,4
C40	80	75	13	0,5	0,2	0,3
C41	80	75	13	0,7	0,1	0,2

Tabela 8 Configurações F-SNN para K constante e diferentes pesos para as quatro dimensões do dataset1

As configurações para o dataset2, presentes na Tabela 9, foram definidas com base nos resultados obtidos nos modelos do dataset1. Os modelos de quatro dimensões apresentaram os resultados mais interessantes. Deste modo optou-se por abordar apenas os modelos com o uso de quatro dimensões. Para este conjunto de modelos, referentes ao dataset2, os valores utilizados para K, MintPts e Eps foram os valores de referência. A utilização dos valores de referência deve-se, tal como os pesos, aos resultados obtidos nos modelos do dataset1, para vários valores diferentes para k. Os *clusters* gerados não apresentavam variações significativas para justificar o mesmo estudo para o dataset2.

Runs	K	MinPts	Eps	W_r-Posição	W_i-Idade	W_m-Mês
C42	47	44	8	0,8	0,1	0,1
C43	47	44	8	0,7	0,2	0,1
C44	47	44	8	0,6	0,2	0,2
C45	47	44	8	0,5	0,3	0,2
C46	47	44	8	0,4	0,3	0,3
C47	47	44	8	0,3	0,4	0,3
C48	47	44	8	0,2	0,4	0,4
C49	47	44	8	0,1	0,4	0,5
C50	47	44	8	0,1	0,5	0,4
C51	47	44	8	0,3	0,3	0,4
C52	47	44	8	0,5	0,2	0,3
C53	47	44	8	0,7	0,1	0,2

Tabela 9 Configurações F-SNN para K constante e diferentes pesos para as dimensões para o dataset2

3.3. Resultados Obtidos

Nesta secção da dissertação são apresentados os resultados da execução dos modelos descritos na secção 3.2. Os modelos serão avaliados em primeiro lugar através da métrica *Model Quality*, em segundo lugar através da exploração estatística e visual dos *clusters* de cada configuração dos modelos.

3.3.1. Avaliação da Qualidade dos Modelos

Após a execução das várias configurações de modelos, apresentados na secção anterior, é necessário identificar aqueles que apresentam ser os melhores modelos. Para tal é necessário recorrer a métricas de qualidade, tais como o *Model Quality*. O cálculo da métrica *Model Quality* foi efetuado através da utilização de uma aplicação em Java, implementada por João Galvão (Galvão, 2014). Esta calcula as medidas *intercluster* e *intracluster* utilizadas para calcular o *Model Quality*. Mas, os valores obtidos para a medida *intercluster* não se enquadravam na ordem de grandeza que seria de esperar para este trabalho. Assim, optou-se por apenas utilizar a medida *intercluster* adaptada para o contexto deste trabalho. Deste modo, recorrendo aos valores obtidos na medida *intracluster* referente a cada modelo, que nos indica a similaridade entre objetos do mesmo cluster, foi criada uma métrica que relaciona esta medida com o número de clusters e os clusters significativos presentes num modelo. Um *cluster* significativo é um *cluster* que apresenta determinadas características, estas características serão explicadas e definidas mais abaixo. A equação que traduz a métrica utilizada encontra-se expressa na equação 3:

Equação 3: $Métrica\ de\ Qualidade = \left| Intracluster - \left(\frac{NºClusters\ Significativos}{NºClusters\ Total} \right) \right|$

Através desta métrica procura-se identificar os modelos que apresentem o maior equilíbrio entre o valor de *intracluster* e o número de *clusters* significativos. Quanto menor o valor obtido na métrica melhor o modelo. É de realçar que os modelos já utilizam parâmetros próximos do ideal e que esta avaliação serve para aprimorar a escolha dos modelos em relação a esses parâmetros.

Como referido anteriormente será explicado o processo para a definição das características para que um *cluster* seja considerado significativo. Como podemos observar na Figura 16, a maior parte dos *clusters* são constituídos por um número reduzido de pontos, alguns deles nem atingem os 100 pontos. Por outro lado, existem alguns *clusters* que se destacam apresentando um elevado agrupamento de pontos e são nestes em que é focada a análise. De modo, a selecionar os *clusters*

para análise foi definido que os *clusters* teriam de conterem pelo menos 1000 pontos agrupados. Este passo é realizado de modo a facilitar a análise dos modelos, pois o número de *clusters* resultantes é demasiado elevado para que sejam possíveis representar todos eles de maneira a que sejam compreensíveis na análise. A utilização desta abordagem pode omitir *clusters* com características interessantes e de certa forma influenciar as conclusões.

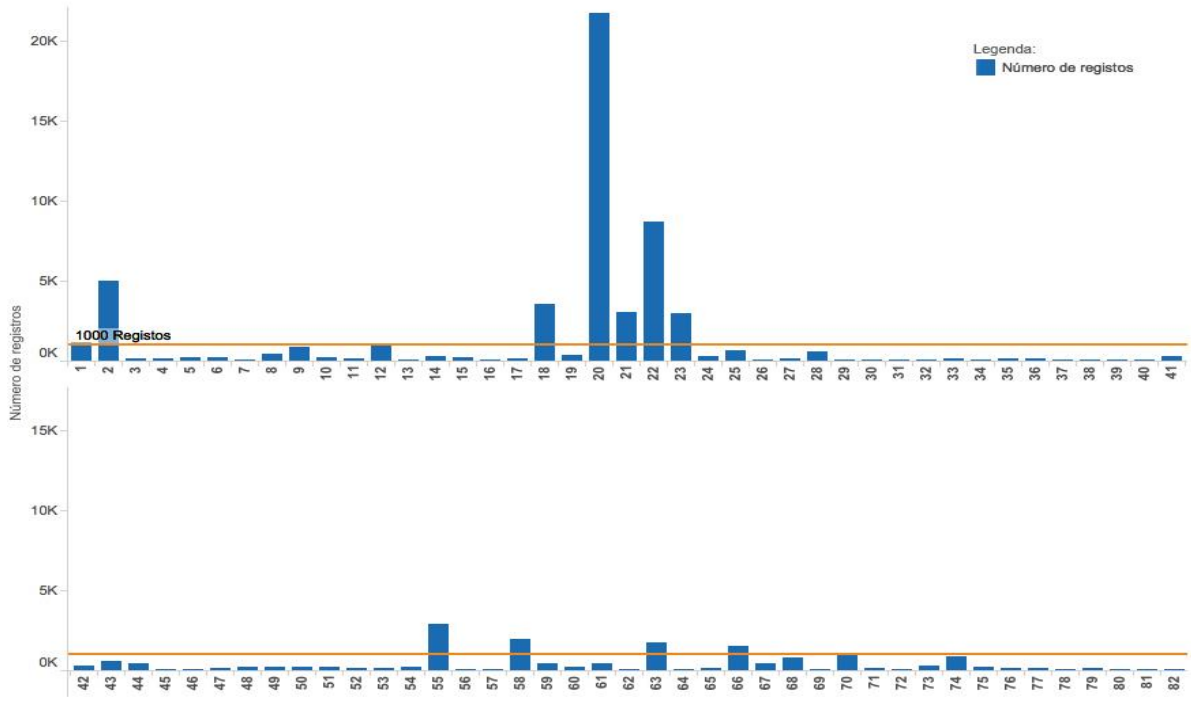


Figura 16 Amostra da distribuição de pontos por *cluster* referente ao modelo C8

Além do número de pontos, os *clusters* significativos são filtrados pelo desvio-padrão do atributo idade. Para definir o valor a utilizar na seleção dos *clusters* através do desvio-padrão das idades, foi utilizada a análise de gráficos, como o representado na Figura 17. Como podemos observar na Figura 17, na parte superior, grande parte dos *clusters* apresentam um desvio-padrão acima dos 10 anos de idade. Assim para abranger um número significativo de *clusters* e que apresentassem a menor dispersão possível de idades, foi definido que o desvio-padrão deveria ser menor ou igual a 10 anos de idade. Quanto ao mês não foi utilizada a mesma abordagem. Como podemos observar na parte inferior da Figura 17, a variação do desvio-padrão dos meses é muito parecida entre os *clusters*. Desta forma, foi optado por não realizar qualquer filtragem pelo desvio-padrão referentes aos meses. Em resumo, os clusters significativos são os clusters que apresentam 1000 ou mais pontos e um desvio padrão para a idade igual ou inferior a 10 anos de idade.

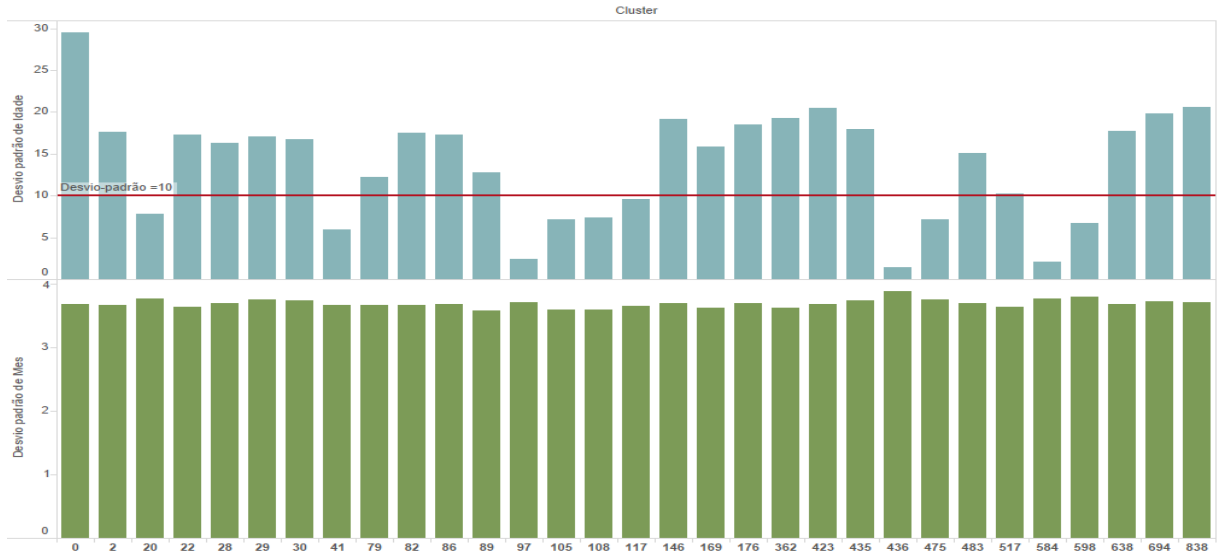


Figura 17 Desvio-padrão das idades e dos meses referentes aos *clusters* selecionados pelo número de registos do modelo C8

Os resultados obtidos para as medidas e para a Métrica de Qualidade utilizada encontram-se apresentados na Tabela 10, Tabela 11, Tabela 12 e Tabela 13. Estas serão analisadas quanto ao comportamento das medidas ao longo dos vários modelos agrupados nas tabelas. Através da Métrica de Qualidade será feita a escolha dos melhores modelos. É de acrescentar que as dimensões latitude e longitude são representadas na tabela como uma dimensão composta pelas duas, a posição.

Runs	K	W_r- Posição	W_i- Idade	NºClusters	Intracluster	NºClusters significativos	Métrica de Qualidade
C1	105	0,5	0,5	495	0,018316	25	0,03219
C2	100	0,5	0,5	547	0,017199	21	0,02119
C3	95	0,5	0,5	567	0,017125	22	0,02168
C4	90	0,5	0,5	596	0,016831	22	0,02008
C5	85	0,5	0,5	620	0,016182	19	0,01446
C6	80	0,5	0,5	703	0,014903	15	0,00643
C7	75	0,5	0,5	762	0,014073	14	0,00430
C8	70	0,5	0,5	844	0,013320	10	0,00147
C9	65	0,5	0,5	1019	0,012105	10	0,00229
C10	60	0,5	0,5	1078	0,011645	9	0,00330

Tabela 10 Resultados da Métrica de Qualidade para diferentes valores de K com três dimensões

Numa avaliação geral sobre os resultados apresentados na Tabela 10 é possível verificar que os valores obtidos para a medida *intracluster* são valores bastante reduzidos. O que apresenta ser um conjunto de distâncias pequenas sendo que o *clustering* é aplicado no território de Portugal Continental. O que permite concluir que no geral os modelos, uns mais que outros, conseguiram

identificar e agrupar pontos que apresentam uma elevada similaridade nos respetivos clusters. Quando o valor de K diminui, o valor para a medida *intracluster* também apresenta uma diminuição, aumentando assim a similaridade entre objetos dentro dos clusters. Se fosse necessário escolher o melhor modelo segundo o *intracluster*, seria o modelo C10. Mas isso não nos garante por si só a qualidade do processo, pois teríamos um conjunto elevado de pequenos *clusters* muito parecidos. Como podemos observar na Tabela 10, o número de *clusters* significativos também diminui com a diminuição do valor de K. O que é indesejável, pois quantos mais clusters significativos, mais rica se torna a análise do modelo.

Desta forma, é necessário procurar o equilíbrio entre o *intracluster* e o número de clusters significativos pelo total de clusters. Através da Métrica de Qualidade criada, o modelo que apresenta o maior equilíbrio entre estas duas medidas é aquele que obteve o menor valor, neste caso foi o modelo C8. Este será representado e analisado com maior detalhe na secção 3.3.2.

Enquanto na Tabela 10 foi analisado o comportamento dos modelos em relação aos diferentes valores de K, Eps e MintPts, na Tabela 11 é analisado o comportamento dos modelos com diferentes pesos nas dimensões.

Runs	K	W_r- Posição	W_i- Idade	nºClusters	Intracluster	NºClusters significativos	Métrica de Qualidade
C11	80	0,9	0,1	1359	0,011060	5	0,00738
C12	80	0,8	0,2	1177	0,012503	11	0,00316
C13	80	0,7	0,3	1018	0,013725	11	0,00292
C14	80	0,6	0,4	849	0,014349	12	0,00021
C15	80	0,5	0,5	703	0,014903	15	0,00643
C16	80	0,4	0,6	570	0,014836	15	0,01148
C17	80	0,3	0,7	485	0,013381	27	0,04229
C18	80	0,2	0,8	501	0,009467	25	0,04043
C19	80	0,1	0,9	876	0,005986	17	0,01342

Tabela 11 Resultados da Métrica de Qualidade para K constante e diferentes pesos para três dimensões

É possível verificar que quando os valores dos pesos atribuídos às dimensões posição e idade se encontram muito díspares, o número de *clusters* tende a aumentar, com maior destaque quando a dimensão posição apresenta um peso superior. Com esta variação encontram-se associados os valores do *intracluster*, já que um maior número de clusters a medida tende a diminuir. É de destacar que quando a dimensão idade apresenta um elevado peso, como nos modelos C18 e C19, o *intracluster* diminui drasticamente, criando clusters de elevada similaridade em relação aos clusters dos restantes modelos da Tabela 11. Quanto ao número de clusters significativos existem dois

modelos que se destacam, os modelos C17 e C18 com 27 e 25 clusters significativos, respetivamente. Apesar de apresentarem um elevado número de clusters significativos, estes devido aos valores do *intraclusters,* apresentam uma grande similaridade entre eles.

Segundo os valores obtidos pela Métrica de Qualidade, o modelo que apresenta o maior equilíbrio entre o *intracluster* e o número de clusters significativos pelo total de clusters, é o modelo C14. Este será representado e analisado com maior detalhe na secção 3.3.2.

Os resultados obtidos na Tabela 12 com diferentes valores para K, MinPts e Eps com peso igual para as três dimensões apresentam um número de *clusters* igual aos apresentados na Tabela 10. Isto deve-se ao facto da igualdade de pesos entre as dimensões. Observando apenas os valores do *intracluster* da Tabela 12, verificamos que os valores no geral são inferiores comparados com os valores apresentados na Tabela 10, o que indica ser um melhoramento na similaridade entre objetos do mesmo cluster, quando utilizado quatro dimensões.

Runs	K	W_r- Posição	W_i- Idade	W_m- Mês	n°Clusters	Intracluster	N°Clusters Significativos	Métrica de Qualidade
C20	105	0,33	0,33	0,33	495	0,01209	25	0,03842
C21	100	0,33	0,33	0,33	547	0,01135	21	0,02704
C22	95	0,33	0,33	0,33	567	0,01130	22	0,02750
C23	90	0,33	0,33	0,33	596	0,01111	22	0,02580
C24	85	0,33	0,33	0,33	620	0,01068	19	0,01997
C25	80	0,33	0,33	0,33	703	0,00984	15	0,01150
C26	75	0,33	0,33	0,33	762	0,00929	14	0,00908
C27	70	0,33	0,33	0,33	844	0,00879	10	0,00306
C28	65	0,33	0,33	0,33	1019	0,00799	10	0,00182
C29	60	0,33	0,33	0,33	1078	0,00769	9	0,00066

Tabela 12 Resultados da Métrica de Qualidade para diferentes valores de K com quatro dimensões

Podemos verificar que existe uma grande diferença nos valores da medida *intracluster* e no número de clusters significativos entre os modelos. Quando o valor de K diminui, os valores das medidas tendem a diminuir e o número de clusters a aumentar. Analisando a Métrica de Qualidade o modelo C29 apresenta o maior equilíbrio entre o *intracluster* e o número de *clusters* significativos pelo total de clusters. Apesar deste apresentar o valor mais baixo para as duas medidas. Desta forma, o modelo C29 será representado e analisado com maior detalhe na secção 3.3.2.

A Tabela 13 apresenta os resultados para K, MinPts e Eps constantes e diferentes pesos para as quatro dimensões. É possível verificar que os modelos com pesos para as dimensões posição e idade quando são maiores que a dimensão mês, apresentam geralmente valores maiores para esta medida. Os valores do *intracluster* registados para os modelos da Tabela 13, apresentam um valor abaixo dos 0,01 nos casos em que o peso dado à dimensão mês é superior ou igual ao da dimensão idade, exceto no modelo 41. Os valores obtidos no número de clusters significativos variam de forma irregular, não aparentando nenhum padrão no comportamento em relação aos pesos das dimensões, *intracluster* e número total de clusters.

Runs	K	W_r- Posição	W_i- Idade	W_m- Mês	nºClusters	Intracluster	Nº Clusters significativos	Métrica de Qualidade
C30	80	0,8	0,1	0,1	1329	0,01029	8	0,00427
C31	80	0,7	0,2	0,1	1134	0,01145	14	0,00090
C32	80	0,7	0,1	0,2	1328	0,00920	7	0,00393
C33	80	0,6	0,2	0,2	1089	0,01060	14	0,00226
C34	80	0,5	0,2	0,3	1041	0,00942	15	0,00499
C35	80	0,5	0,3	0,2	884	0,01109	15	0,00587
C36	80	0,4	0,3	0,3	813	0,01018	10	0,00212
C37	80	0,3	0,3	0,4	703	0,00894	15	0,01240
C38	80	0,3	0,4	0,3	621	0,01012	14	0,01242
C39	80	0,2	0,4	0,4	492	0,00847	19	0,03015
C40	80	0,1	0,4	0,5	501	0,00473	25	0,04517
C41	80	0,1	0,5	0,4	599	0,00498	17	0,02340

Tabela 13 Resultados da Métrica de Qualidade para K constante e diferentes pesos para quatro dimensões

Segundo os valores obtidos pela Métrica de Qualidade, o modelo que apresenta o maior equilíbrio entre o *intracluster* e o número de clusters significativos pelo total de *clusters*, é o modelo C31. Este será representado e analisado com maior detalhe na secção 3.3.2.

Na Tabela 14 encontram-se os resultados das medidas e da Métrica de Qualidade para K, MinPts e Eps constantes e diferentes pesos para as quatro dimensões das configurações executadas sobre o *dataset* 2. Podemos verificar que o comportamento de oscilação dos valores da medida são iguais aos presentes na Tabela 13, mas com valores mais reduzidos. O número de cluster significativos obtido ao logo do modelo é também muito menor.

Runs	K	W_r- Posição	W_i- Idade	W_m- Mês	nºClusters	Intracluster	NºClusters Significativos	Métrica de Qualidade
C42	47	0,8	0,1	0,1	1270	0,00901	4	0,00586
C43	47	0,7	0,2	0,1	1124	0,00964	7	0,00341
C44	47	0,7	0,1	0,2	1244	0,00806	3	0,00565
C45	47	0,6	0,2	0,2	1097	0,00871	8	0,00142
C46	47	0,5	0,2	0,3	1037	0,00786	9	0,00082
C47	47	0,5	0,3	0,2	925	0,00937	4	0,00505
C48	47	0,4	0,3	0,3	884	0,00814	5	0,00249
C49	47	0,3	0,3	0,4	804	0,00703	4	0,00206
C50	47	0,3	0,4	0,3	720	0,00826	6	0,00008
C51	47	0,2	0,4	0,4	619	0,00697	9	0,00757
C52	47	0,1	0,4	0,5	717	0,00424	6	0,00413
C53	47	0,1	0,5	0,4	770	0,00450	4	0,00070

Tabela 14 Resultados da Métrica de Qualidade para K constante e diferentes pesos para quatro dimensões para o *dataset* 2

Segundo os valores obtidos pela Métrica de Qualidade, o modelo que apresenta o maior equilíbrio entre o *intracluster* e o número de clusters significativos pelo total de clusters, é o modelo C50. Este será representado e analisado com maior detalhe na secção 3.3.2.

3.3.2. Análise Geo-espacial dos Modelos

Nesta secção são apresentados os resultados através da representação geo-espacial dos *clusters* obtidos e também são apresentados dados estatísticos, que caracterizam os *clusters*, tais como o desvio-padrão, máximos, mínimos e percentagens.

Para a realização destas tarefas baseadas na análise estatística e representação geo-espacial foi utilizada a ferramenta RStudio[1]. Esta aplicação é uma GUI mais amiga do utilizador em relação à proveniente do R[2]. Através desta ferramenta foi criado um *package* constituído por várias funções que permitem processar estatisticamente os resultados provenientes do F-SNN que são aqui apresentados. O código em R desenvolvido encontra-se no anexo A. Estas funções permitem de forma autónoma calcular várias medidas estatísticas e criar um novo documento com os valores calculados, representar em 3D os pontos dos *clusters* diferenciados através de cores e representar os mesmos pontos no mapa de Portugal utilizando *shapes*, diferenciados por cores. Estes mapas permitem realizar navegação interativa sobre os mesmos. Os mapas gerados através de *shapes* no R serviram para ter uma visualização rápida dos pontos no espaço. Mas, devido a algumas limitações

[1] www.rstudio.com

[2] www.r-project.org

destes para apresentação dos resultados na dissertação dos modelos selecionados, são utilizados os mapas provenientes da ferramenta Tableau[3]. No Tableau serão utilizados os ficheiros criados no R para proceder à representação dos mapas e gráficos relacionados.

Para esta análise são utilizados os modelos que obtiveram os melhores valores na Métrica de Qualidade utilizada na secção 3.3.1. O objetivo agora passa por caracterizar a Pneumonia através da análise deste grupo de modelos selecionados. Os modelos selecionados encontram-se descritos na Tabela 15.

Runs	K	D1-Posição	D2-Idade	D3 –Mês
C8	70	0,5	0,5	-
C14	80	0,6	0,4	-
C29	60	0,33	0,33	0,33
C31	80	0,7	0,2	0,1
C50	47	0,3	0,4	0.3

Tabela 15 Modelos selecionados para análise.

Em primeiro lugar foi realizada a análise sobre as configurações C8, C14, C29 e C31, referentes ao dataset1. Como referido anteriormente, só serão analisados os clusters significativos, que apresentam um número de pontos superior ou igual a 1000 registos e um desvio-padrão de 10 unidades no atributo idade.

O modelo C8 encontram-se representado na Figura 18 e os seus dados estatísticos apresentados na Tabela 16. Através da observação dos resultados apresentados na Figura 18 é possível verificar que os *clusters* se encontram espalhados por todas as regiões de Portugal Continental, especificamente nos distritos de Braga, Vila Real, Bragança, Guarda, Porto, Aveiro, Viseu, Guarda, Coimbra e Lisboa. Existe uma tendência de proximidade de alguns *clusters* às zonas litorais, que são regiões com maior concentração populacional. Estes apresentam estar mais concentrados no espaço em relação aos clusters da zona interior. Os clusters presentes nas zonas interiores encontram-se menos concentrados no espaço, devido à menor concentração populacional. Tal conduz a um maior número de registo de ocorrências de pneumonia nestas regiões.

Em sequência dos resultados apresentados procedeu-se à análise do comportamento das idades ao longo dos *clusters* do modelo C8 e ao comportamento da percentagem de vítimas mortais

[3] www.tableau.com

presentes nos *clusters* e o seu relacionamento com as idades. Perante os resultados apresentados na Tabela 16 é possível observar, através da moda, que as idades mais representativas na maior parte dos *clusters* são elevadas, com mais de 75 anos. Por outro lado, os resultados apresentam em minoria *clusters* com uma média de idades igual ou inferior aos 3 anos. Estes resultados estão relacionados com o número de registos de ocorrências de pneumonias em indivíduos nestas faixas etárias.

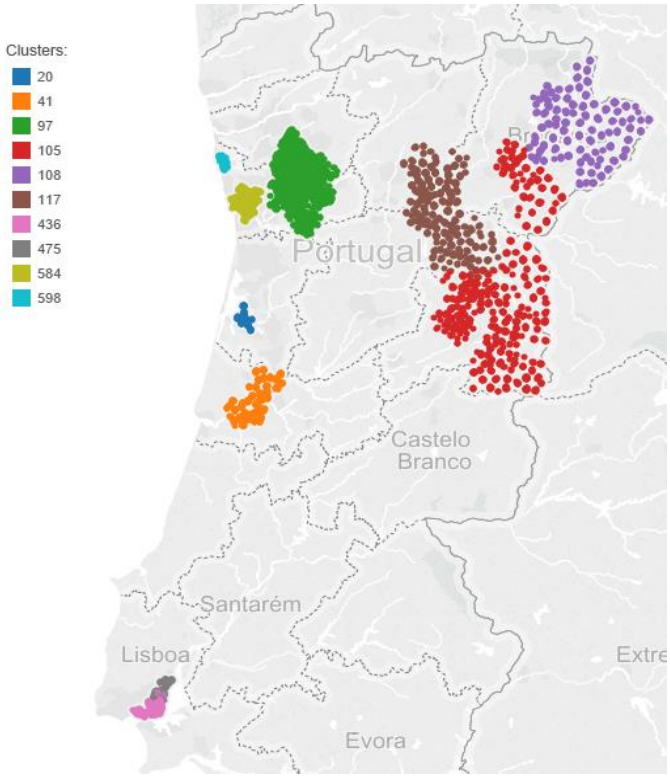


Figura 18 Mapa referente ao modelo C8 (K:70; D1-Posição:0,5; D2-Idade:0,5)

Cluster	Moda idades	Média idades	Desvio-padrão	Min. idade	Max. idade	Número de pontos	Vítimas mortais	% Vítimas mortais
20	81	80	7.8	61	94	1562	351	22,5
41	81	80	5.9	63	96	2112	449	21,3
97	1	3	2.4	0	12	2420	8	0,3
105	84	80	7.2	56	100	1799	394	21,9
108	78	80	7.4	57	100	1376	220	16,0
117	82	80	9.6	43	98	1695	345	20,4
436	0	1	1.4	0	7	1499	7	0,5
475	77	80	7.2	58	90	3340	661	19,8
584	1	3	2.1	0	10	1570	4	0,3
598	82	80	6.7	67	94	1007	243	24,1

Tabela 16 Tabela de dados estatísticos referente aos *clusters* do modelo C8

Através da análise da Tabela 16, referente ao modelo C8, podemos verificar que os *clusters* 97, 436 e 584 são os que apresentam a menor percentagem de vítimas mortais, com percentagens iguais ou inferiores a 0,5%, e com uma moda de idades entre 1 e 3 anos. O *cluster* 97 situa-se na região dos distritos de Braga e Porto, com 2420 indivíduos. O *cluster* 436 situa-se na região do distrito de Lisboa com 1449 pontos. O *cluster* 584 situa-se na região do distrito do Porto com 1570 indivíduos. Por outro lado, o *cluster* 598 é o que apresenta a maior percentagem de vítimas mortais, com 24,1%, e com uma moda de idades de 82 anos e com uma média de 80 anos. O *cluster* 598 situa-se na região do distrito de Porto, com 1007 indivíduos.

Na Figura 19, referente ao modelo C8, é possível verificar que a maioria dos *clusters* do modelo C8 apresentam uma percentagem de vítimas mortais mais elevada para os indivíduos do sexo masculino à exceção do cluster 584.

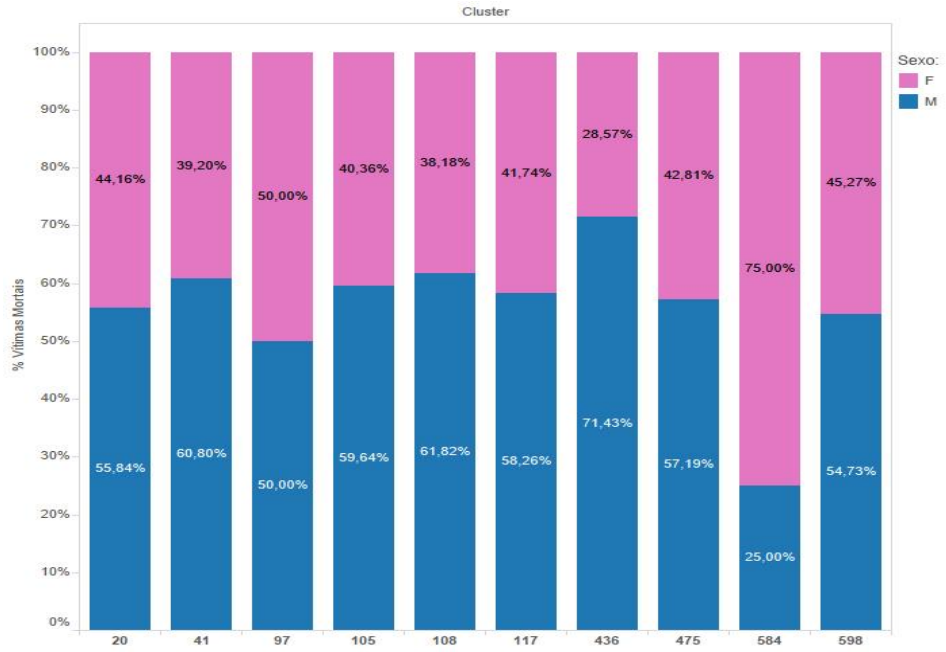


Figura 19 Percentagem de vítimas mortais por sexo e *cluster* do modelo C8.

O cluster 475 apresenta a maior percentagem de vítimas do sexo masculino com 71.43%. Este situa-se na região do distrito de Lisboa. É de salientar que o cluster 584 se destaca devido à elevada percentagem de vítimas mortais em indivíduos do sexo feminino, com 75% de vítimas mortais femininas. O cluster 584 encontra-se localizado na região do distrito de Porto.

O modelo C14 encontra-se representado na Figura 20 e os seus dados estatísticos encontram-se apresentados na Tabela 17. Através da observação dos resultados apresentados na Figura 20 é possível verificar que os *clusters* se encontram espalhados por todas as regiões de Portugal

Continental, especificamente nos distritos de Braga, Viana do Castelo, Bragança, Guarda, Porto, Aveiro, Guarda, Coimbra, Lisboa, Setúbal e Faro. Tal como já referido no modelo anterior, existe uma tendência de proximidade dos *clusters* às zonas litorais e interior norte. Tal conduz a um maior número de registo de ocorrências de pneumonia nestas regiões.

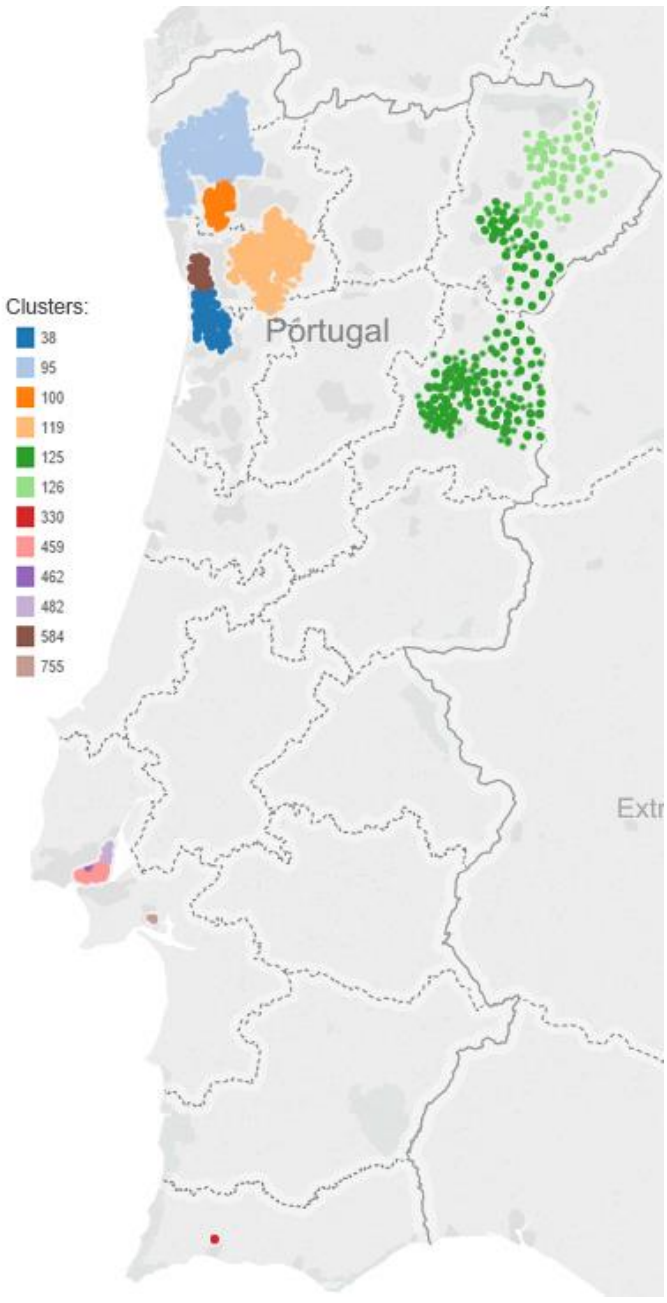


Figura 20 Mapa referente ao modelo C14 (K:80; D1-Posição:0,6; D2-Idade:0,4)

Cluster	Moda idades	Média idades	Desvio-padrão	Min. idade	Max. idade	Número de pontos	Vítimas mortais	% Vítimas mortais
38	1	3	2.5	0	12	1008	5	0,5
95	85	80	6.7	54	97	4962	1072	21,6
100	80	80	5.9	66	97	1110	243	21,9
119	0	3	2.9	0	16	1778	5	0,3
125	84	80	8.8	47	102	1530	326	21,3
126	80	80	7.9	53	100	1025	144	14,0
330	83	80	6.6	68	94	1031	257	24,9
459	0	2	1.4	0	6	1088	4	0,4
462	84	80	7.9	63	95	1576	344	21,8
482	91	80	7.5	65	93	1216	296	24,3
584	1	3	2.1	0	10	1091	2	0,2
755	81	80	9.7	53	93	1557	397	25,5

Tabela 17 Tabela de dados estatísticos referente aos *clusters* do modelo C14

É de destacar no modelo C14, através da análise da Tabela 17, que os *clusters* 38, 119, 459 e 584 são os que apresentam a menor percentagem de vítimas mortais, entre os 0,2% e 0,5%, e com uma média de idades entre os 2 e 3 anos. O *cluster* 38 situa-se na região dos distritos do Porto e Aveiro, com 1008 indivíduos. O *cluster* 459 situa-se na região do distrito de Lisboa com 1088 indivíduos. O *cluster* 119 situa-se na região dos distritos de Porto e Braga, com 1778 indivíduos. O *cluster* 584 situa-se na região do distrito do Porto com 1091 indivíduos. Por outro lado, o *cluster* 755 é o que apresenta a maior percentagem de vítimas mortais, com 25,5%, e com uma média de idades próxima dos 81 anos. O *cluster* 755 situa-se na região do distrito de Setúbal, com 1557 indivíduos.

Na Figura 21, referente ao modelo C14, tal como no modelo C8, existe uma maior incidência de vítimas mortais no sexo masculino. Neste modelo, os *clusters* 38 e 584 destacam-se pela elevada percentagem de vítimas mortais em indivíduos do sexo feminino, com 70% e 100% de vítimas mortais. O *cluster* 38 encontra-se localizado na região do distrito do Porto e Aveiro e o *cluster* 584 na região do Porto. Estes dois clusters encontram-se muito próximos no espaço e apresentam a mesma tendência de vítimas mortais do *cluster* 584 do modelo C8. Apesar de serem clusters com baixa percentagem de vítimas mortais, estas acabam por ser na maioria do sexo feminino. Por outro lado, o *cluster* 459 apresenta a maior percentagem de vítimas mortais em indivíduos do sexo masculino, 75%. O *cluster* 459 encontra-se localizado na região do distrito de Lisboa.

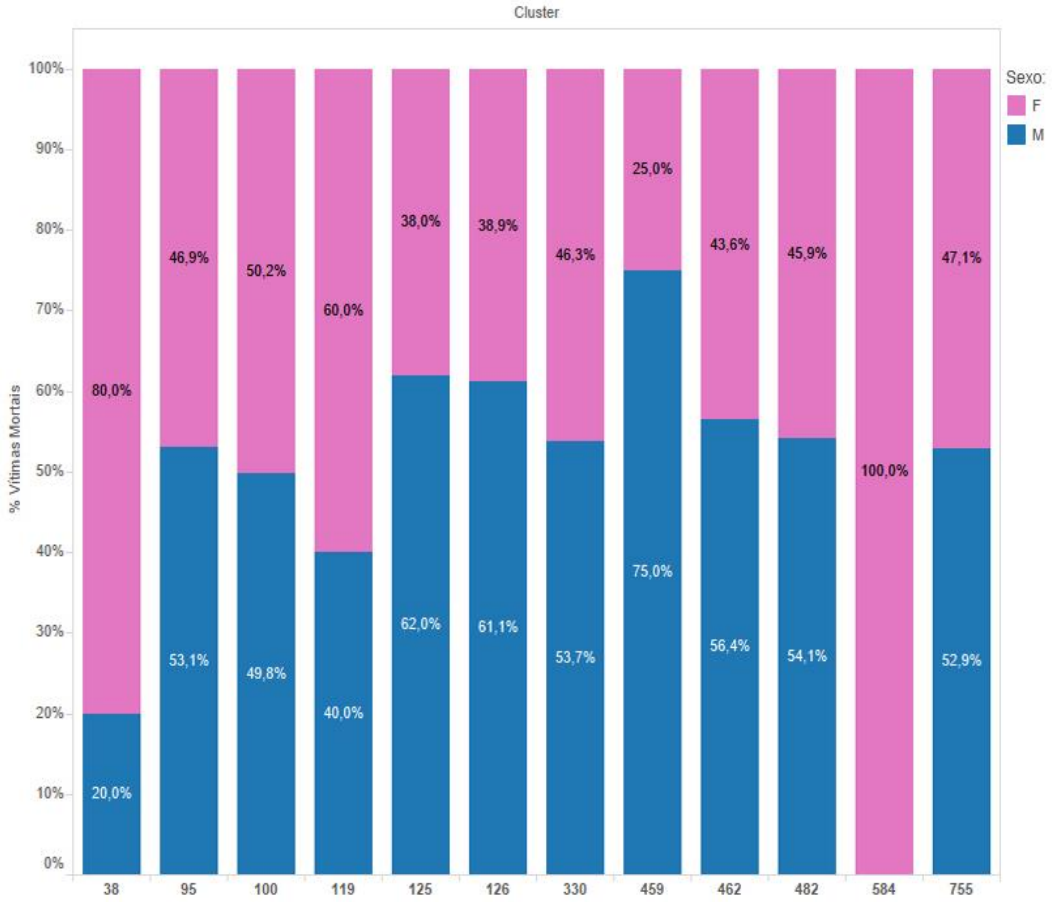


Figura 21 Percentagem de vítimas mortais por sexo e *cluster* do modelo C14.

Se realizarmos o cruzamento entre os dados da Tabela 17 e da Figura 21, verificamos que três dos *clusters* com baixas médias de idade, *cluster* 38, 119 e 584, apresentam a maior percentagem de vítimas mortais do sexo feminino. Verifica-se que nos indivíduos até aos 5 anos de idade, o sexo feminino parece ser mais afetado em termos de mortalidade da doença.

O modelo C29 encontram-se representados na Figura 22 e os respetivos dados estatísticos encontram-se apresentados na Tabela 18 . Através da observação dos resultados apresentados na Figura 22 é possível verificar que os *clusters* se encontram espalhados por Portugal Continental, especificamente nos distritos de Braga, Viana do Castelo, Bragança, Guarda, Porto, Aveiro, Viseu, Vila Real, Bragança, Guarda, Coimbra e Lisboa.

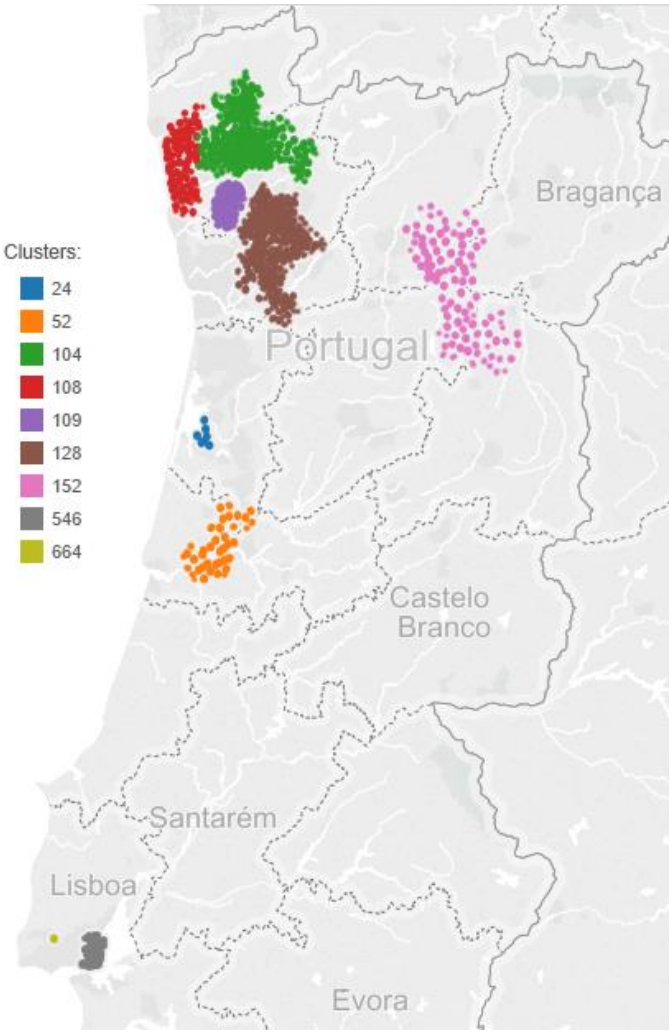


Figura 22 Mapa referente ao modelo C29 (K:60; D1-Posição:0,33; D2-Idade:0,33; D3-Mês:0,33)

Cluster	Moda idades	Média idades	Desvio-padrão	Min. idade	Max. idade	Moda Mês	Número de pontos	Vítimas mortais	% Vítimas mortais
24	82	79	7.9	61	94	1	1370	311	22,7
52	83	82	6.1	66	98	2	1811	389	21,5
104	85	81	6.4	56	96	2	3280	702	21,4
108	82	79	6.7	60	92	2	1530	319	20,8
109	80	81	5.9	66	95	2	1120	250	22,3
128	0	3	2.3	0	11	2	2313	11	0,5
152	82	77	9.6	41	98	1	1397	278	19,9
546	0	1	1.4	0	6	1	1633	7	0,4
664	82	76	9.8	56	91	3	1046	372	35,6

Tabela 18 Tabela de dados estatísticos referente aos *clusters* do modelo C29

Através da análise da Tabela 18, identifica-se que os *clusters* 128 e 546 são os que apresentam a menor percentagem de vítimas mortais, entre os 0,4% e 0,5%, e com uma média de idades entre os 1 e 3 anos. O *cluster* 128 situa-se na região dos distritos do Porto e Aveiro, com 2313 pontos. O *cluster* 546 situa-se na região do distrito de Lisboa com 1633 pontos. Por outro lado, o *cluster* 664 é o que apresenta a maior percentagem de vítimas mortais, com 35,6%, a maior percentagem detetada nos modelos até ao momento, e com uma média de idades nos 76 anos. O *cluster* 664 situa-se na região do distrito de Lisboa, com 1046 pontos. É ainda possível verificar que os meses de Janeiro e Fevereiro são aqueles que apresentam o maior incidência.

Na Figura 23, referente ao modelo C29, é possível observar que existe apenas um *cluster* onde a percentagem de vítimas mortais nos indivíduos do sexo feminino é superior, o *cluster* 128, cuja a região e comportamento já foi detetado nos modelos C8 e C14.

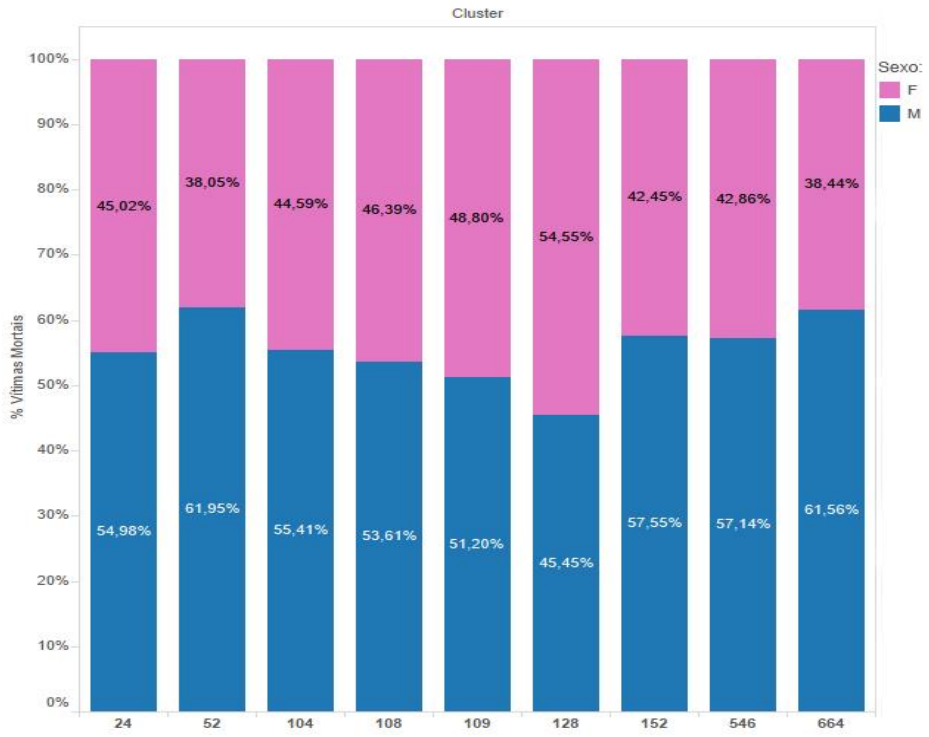


Figura 23 Percentagem de vítimas mortais por sexo e *cluster* do modelo C29.

Neste modelo podemos verificar que existem *clusters* muito parecidos em termos de percentagem de vítimas mortais em indivíduos do sexo masculino. Por exemplo, os *clusters* 152 e 546, e os clusters 52 e 664, com aproximadamente 57% e 62%, respetivamente.

O modelo C31 encontra-se representado na Figura 24 e os seus dados estatísticos na Tabela 19. Através da observação dos resultados apresentados na Figura 24 é possível verificar que os

clusters se encontram associados principalmente às regiões Norte e Sul de Portugal, especificamente nos distritos de Braga, Viana do Castelo, Bragança, Porto, Aveiro, Vila Real e Lisboa.

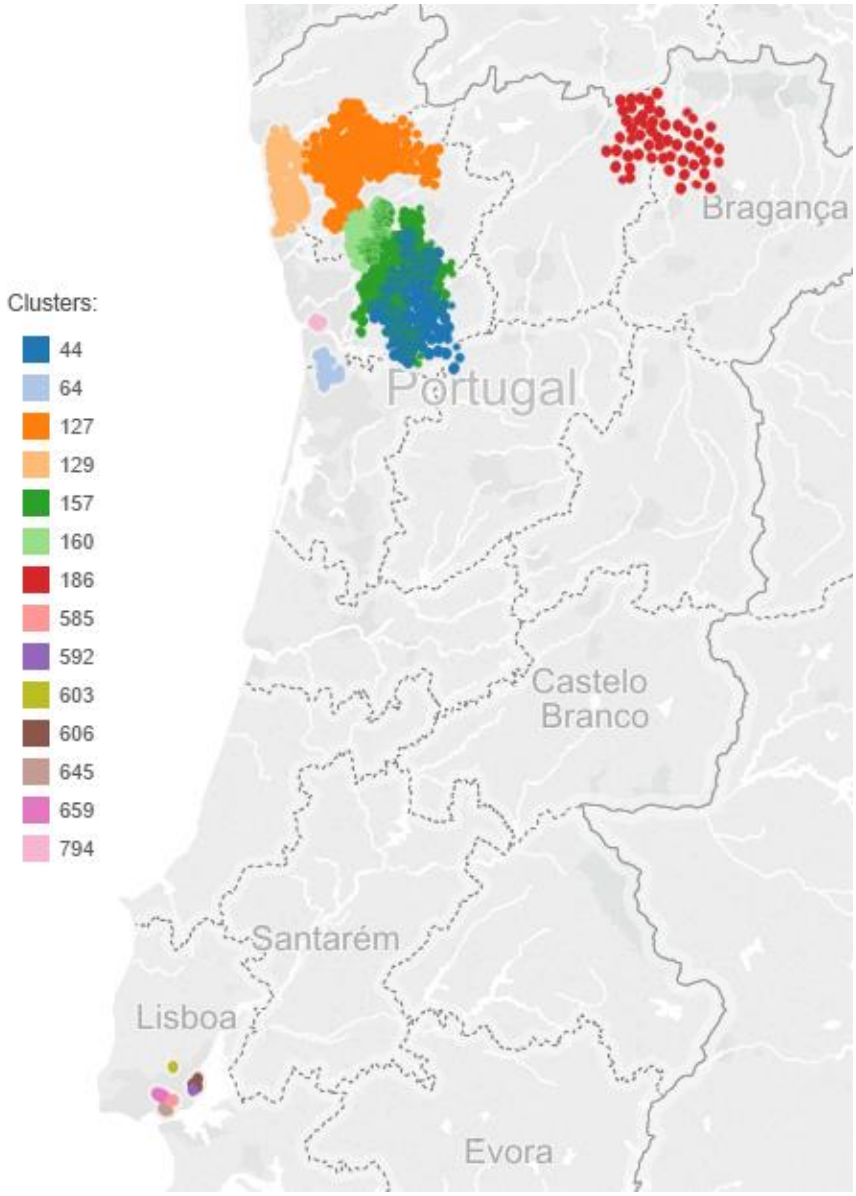


Figura 24 Mapa referente ao modelo C31(K:80; D1-Posição:0,7; D2-Idade:0,2; D3-Mês:0,1;)

Como existe a sobreposição dos clusters 44 e 157, foi necessário recorrer à visualização 3D dos pontos, Figura 25. É possível verificar que estes se situam numa região comum, mas apresentam uma distribuição de idades bastante diferentes, como podemos ver na Tabela 19.

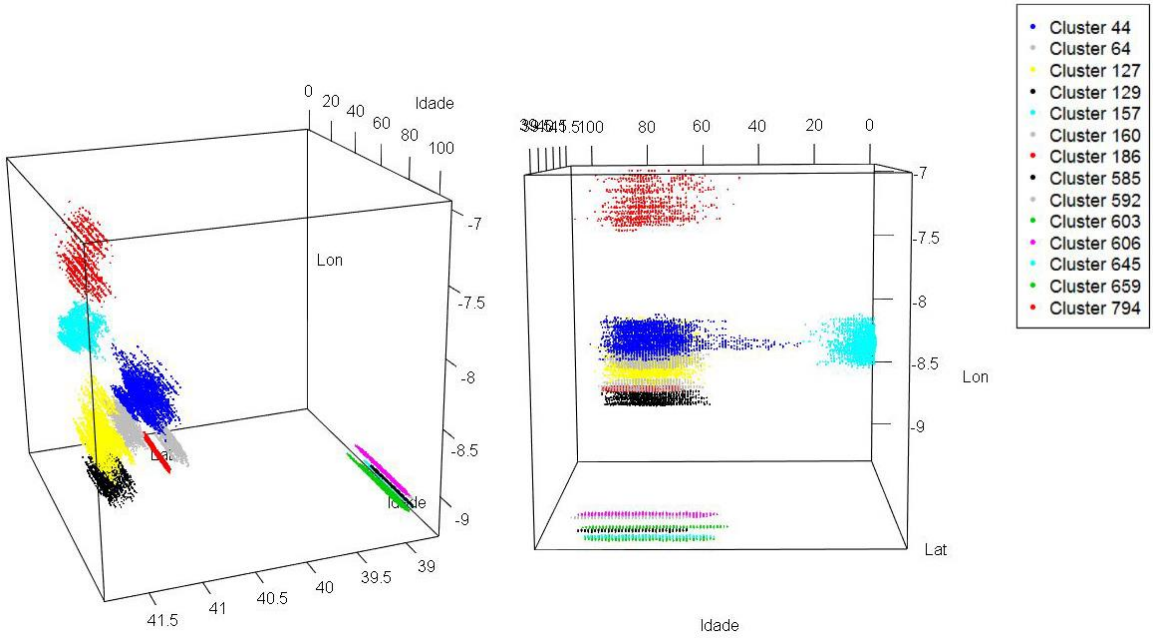


Figura 25 Gráfico 3D dos *clusters* do modelo C31

Cluster	Moda idades	Média idades	Desvio-padrão	Min. idade	Max. idade	Moda Mês	Número de pontos	Vítimas mortais	% Vítimas mortais
44	80	75	9.1	26	99	1	4241	876	20,7
64	81	79	7.4	60	96	1	1082	238	22,0
127	82	81	7.1	57	100	2	3727	770	20,7
129	84	79	7.5	58	97	2	1612	357	22,1
157	0	4	3.8	0	25	2	2506	12	0,5
160	81	79	6.9	58	97	2	2943	578	19,6
186	81	80	8.7	47	106	1	1207	209	17,3
585	84	80	7.9	63	95	1	1488	313	21,0
592	79	79	7.5	62	97	1	1753	401	22,9
603	84	77	9.7	51	94	1	1153	214	18,6
606	80	78	9.2	54	95	1	1355	300	22,1
645	83	78	8.9	55	93	2	1036	218	21,0
659	81	77	8.6	55	93	2	1303	447	34,3
794	81	81	6.8	67	95	1	1656	345	20,8

Tabela 19 Tabela de dados estatísticos referente aos *clusters* do modelo 31

Através da análise da Tabela 19, é possível identificar que o *cluster* 157 é o que apresenta a menor percentagem de vítimas mortais, com 0,5%, e com uma média de idades de 4 anos. O *cluster* 157 situa-se na região dos distritos do Porto e Braga, com 2506 indivíduos. Por outro lado, o *cluster* 659 é o que apresenta a maior percentagem de vítimas mortais, com 34,3%, e com uma média de idades próxima dos 77 anos. O *cluster* 659 situa-se na região do distrito de Lisboa, com 1303

indivíduos. É ainda possível verificar que os meses de Janeiro e Fevereiro são aqueles que apresentam o maior incidência.

Na Figura 26 referente ao modelo C31, é possível observar que globalmente a percentagem vítimas mortais nos indivíduos do sexo masculino é superior aos indivíduos do sexo feminino, como já observado em modelos anteriores. No modelo C31 não existem *clusters* que se destaquem significativamente em termos de percentagens. Os *clusters* 186 e 645, com 62,68% e 64,22% de vítimas mortais em indivíduos do sexo masculino, respetivamente, são os que apresentam as maiores percentagens de vítimas mortais. O *cluster* 186 e 645 encontram-se localizados nas regiões do distrito de Bragança e Lisboa, respetivamente. O *cluster* 129 é o único *cluster* que apresenta uma percentagem de vítimas mortais em indivíduos do sexo feminino acima dos 50%, com 50,70% de vítimas mortais. O *cluster* 129 encontra-se localizado na região dos distritos de Braga e Viana do Castelo.

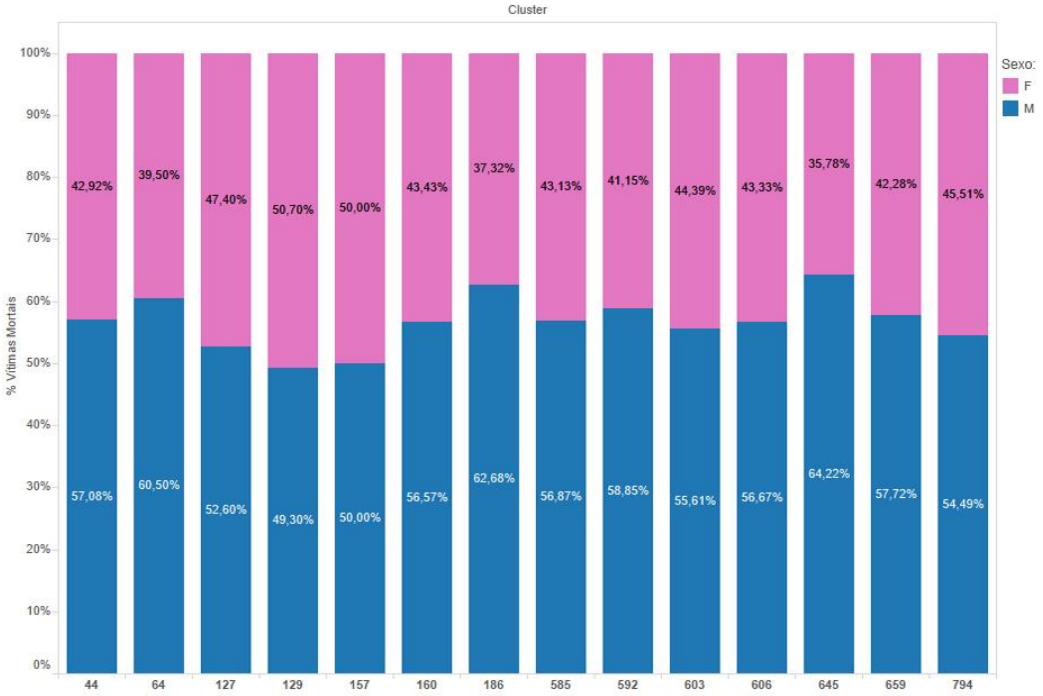


Figura 26 Percentagem de vítimas mortais por sexo e *cluster* do modelo C31.

Os modelos apresentados anteriormente foram escolhidos com base na procura dos modelos que apresentassem um maior equilíbrio entre as medidas *intercluster* e número de clusters significativos pelo total de clusters, que dão origem à Métrica de Qualidade como referido durante a análise dos resultados da secção 3.3.1. Após a representação e análise espacial dos resultados dos

modelos selecionados, foi possível retirar algumas conclusões gerais sobre as características dos modelos.

No caso dos modelos que utilizaram K, MinPts e Eps diferentes, mas com pesos iguais, foi possível verificar que quando o K aumentava as duas medidas também aumentavam, consequentemente levando ao aumento do valor da Métrica de Qualidade. Aliada a este aumento está associada uma redução no número de *clusters* gerados pelos modelos e o aumento de pontos por *clusters*, levando ao aumento de *clusters* significativos. Perante este comportamento, a Métrica de Qualidade utilizada identificou como melhores modelos o modelo C8 e C29, com um valor para K menor em 10 e 20 unidades, respetivamente, que o valor calculado como referência.

No caso dos modelos que utilizam valores para K, MintPts e Eps constantes, mas com pesos diferentes, foi possível verificar que os pesos utilizados influenciam diretamente os valores das medidas e consequentemente a Métrica de Qualidade. No caso dos modelos que utilizavam as dimensões posição e idade, os modelos que apresentavam um maior equilíbrio nos pesos entre as duas dimensões apresentavam os melhores valores para a Métrica de Qualidade. No caso dos modelos que utilizam a dimensão mês, além das duas já referidas, torna-se mais complexo de identificar o comportamento dos resultados perante os pesos. Mas é possível identificar uma variação relacionada com as dimensões idade e mês. Quando o peso da idade é superior ao do mês o *intracluster* tende a aumentar, o inverso também acontece. Assim, a Métrica de Qualidade identificou C14 e C31 como os melhores modelos.

Em termos de resultados sobre a incidência de pneumonias, foi possível concluir através das análises realizadas sobre os modelos na secção 3.3.2, que os *clusters* significativos encontrados, estão localizados principalmente nos centros urbanos, principalmente na zona Litoral. Os distritos que apresentam uma maior ocupação pelos *clusters* e com maiores percentagens de vítimas mortais foram Braga, Porto, Viseu, Coimbra, Setúbal e Lisboa. Além das regiões afetadas também foram identificados os meses que apresentam o maior número de registos de casos de pneumonia, em que se destacam Janeiro, Fevereiro e Março.

Os indivíduos que apresentam a maior percentagem de mortalidade são aqueles com mais de 70 anos de idade, precisamente os indivíduos que apresentam estar na terceira idade, tal como já havia sido verificado em (Leite, 2014). Pelo contrário, os indivíduos que apresentam uma menor taxa de mortalidade são os que se encontram na faixa dos 0 aos 5 anos de idade. Não foram

encontrados *clusters* significativos para outras idades. Em termos de vítimas mortais, também foi possível verificar que globalmente os indivíduos do sexo masculino são os que maior percentagem de mortalidade apresentam. Foi também identificada uma tendência na maioria dos clusters que apresentavam uma média de idades entre os 0 e 5 anos, estes apresentam uma percentagem de vítimas de indivíduos do sexo feminino dominante.

Após a apresentação dos resultados referentes aos modelos aplicados sobre o dataset1, são agora apresentados os modelos selecionados na secção 3.3.1 referentes ao dataset2. Este *dataset* contém informação sobre outras patologias que se encontram relacionadas com a pneumonia. O melhor modelo identificado pela Métrica de Qualidade foi o C50. Este utiliza um K, MinPts e Eps constantes, com diferentes pesos para as três dimensões. O modelo C50 encontra-se representado na Figura 27 e os respetivos dados estatísticos na Tabela 20.

Como podemos observar no mapa apresentado na Figura 27, que corresponde ao modelo C50, é possível verificar que os *clusters* se encontram distribuídos pelos distritos de Braga, Porto, Aveiro, Coimbra, Leiria, Bragança e Guarda.

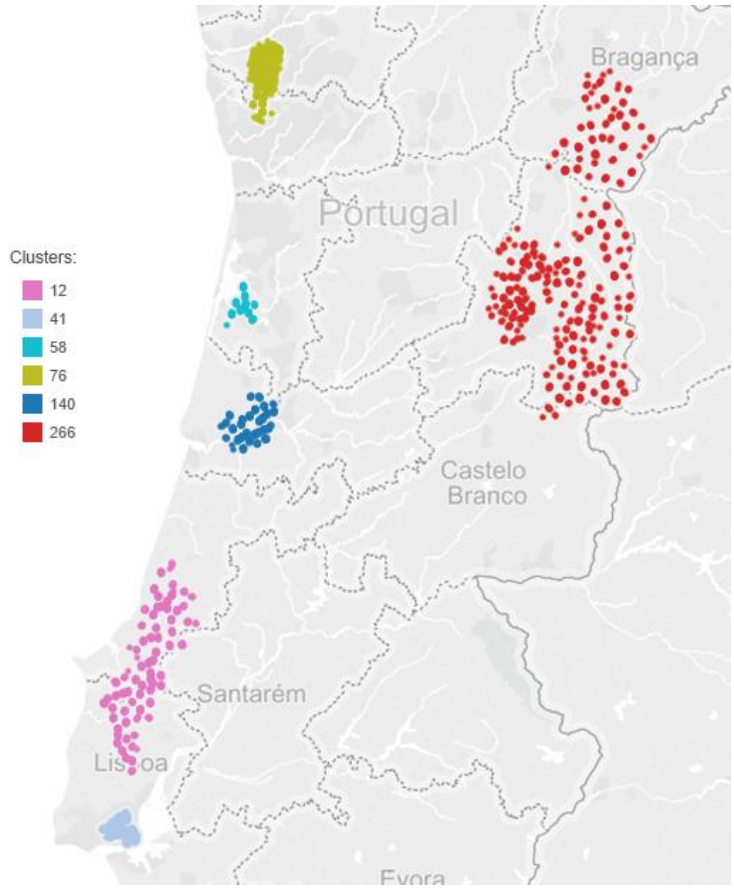


Figura 27 Mapa referente ao modelo C50 (K:47; D1-Posição:0,3; D2-Idade:0,4 D3-Mês: 0.3)

Cluster	Moda idades	Média idades	Desvio-padrão	Min. idade	Max. idade	Moda Mês	Número de pontos	Vítimas mortais	% Vítimas mortais
12	73	77	8,3	46	96	2	2115	534	25,2
41	84	87	3,0	81	96	1	1709	632	37,0
58	72	78	8,0	58	93	1	1544	335	21,7
76	86	80	5,8	67	94	1	1030	227	22,0
140	75	80	6,2	66	96	1	1042	218	20,9
266	84	82	6,4	62	97	2	1123	233	20,7

Tabela 20 Tabela de dados estatísticos referente aos *clusters* do modelo C50

Através da análise da Tabela 20, é possível verificar que a média de idades registada para a maioria dos *clusters* apresentam um valor igual ou superior a 77 anos. Neste modelo não foi encontrado qualquer cluster significativo com idades compreendidas entre os 0 e 5 anos como anteriormente verificado nos modelos do dataset1. É possível identificar que o *cluster* 41 apresenta a maior percentagem de mortalidade, com 37,0%. O *cluster* 41 está localizado na região do distrito de Lisboa, apresenta uma média de idades de aproximadamente 87 anos e é constituído por 1709 pontos. Podemos também verificar, perante os dados da tabela Tabela 20, que o mês que apresenta maior número de casos de pneumonia, na maior parte dos *clusters* apresentados, é o mês de Janeiro seguido do mês de Fevereiro.

Na Figura 28 podemos observar o comportamento das vítimas mortais em relação ao sexo dos indivíduos para o modelo C50. Os *clusters* apresentam uma maior representação de indivíduos do sexo masculino, mas é de destacar que o *cluster* 41, que apresenta o maior número de vítimas mortais, tem como predominância as vítimas do sexo feminino.

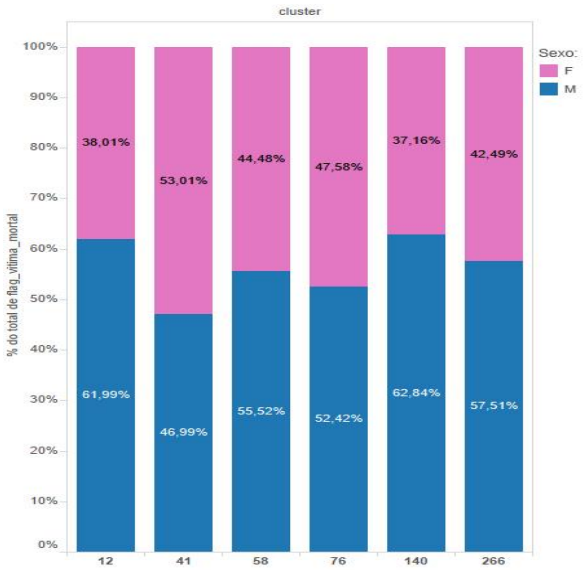
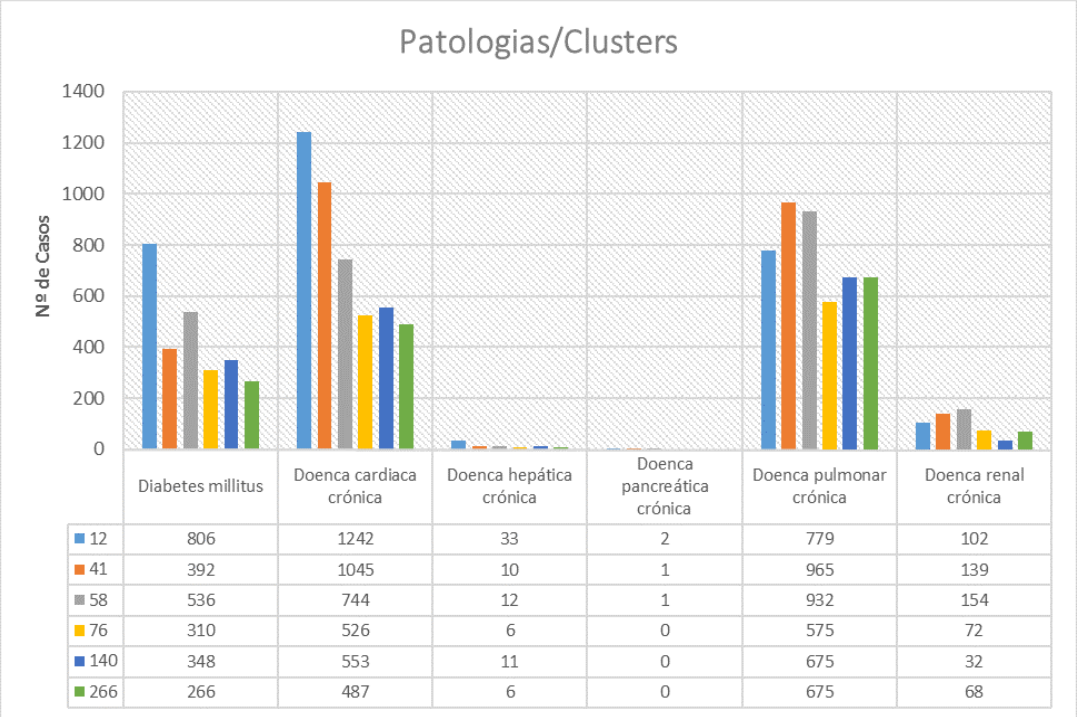


Figura 28 Percentagem de vítimas mortais por sexo e *cluster* de C50

Na Figura 29 encontra-se representado o total de registos de cada patologia presente nos *clusters* do modelo C50. É possível verificar que as patologias que apresentam uma maior presença nos *clusters* são a Diabetes Millitus, a Doença Cardíaca Crónica e Doença Pulmonar Crónica. A Doença Pancreática Crónica e a Doença Renal Crónica apresentam uma baixa presença nos clusters em relação as outras Doenças.



	Diabetes millitus	Doenca cardiaca crónica	Doenca hepática crónica	Doenca pancreática crónica	Doenca pulmonar crónica	Doenca renal crónica
12	806	1242	33	2	779	102
41	392	1045	10	1	965	139
58	536	744	12	1	932	154
76	310	526	6	0	575	72
140	348	553	11	0	675	32
266	266	487	6	0	675	68

Figura 29 Gráfico de Patologias por *cluster* do modelo C50

4. Previsão da Incidência de Pneumonia

Neste capítulo é documentado o processo de utilização de algoritmos de *Time Series Forecasting* com o intuito de encontrar os modelos preditivos para a previsão do número de casos de pneumonia e das vítimas mortais. Na secção 4.1, é apresentada a seleção dos dados e as transformações efetuadas aos dados, de forma a criar os ficheiros com os dados agregados por mês e ano. Com os ficheiros derivados das transformações efetuadas foram utilizados vários algoritmos de previsão no *plugin Forecast* do Weka. Desta forma, foi possível identificar os algoritmos para serem utilizados na criação dos modelos. Foram selecionados quatro algoritmos, um baseado em algoritmo de regressão linear, um baseado em *support vector machines*, um baseado em regras de decisão e por último um baseado em árvores de decisão. Com estes 4 algoritmos foram criados vários modelos. Os modelos encontram-se apresentados na secção 4.2. Na secção 4.3, os modelos foram avaliados através dos resultados obtidos e são comparados entre eles. Por fim os melhores modelos para Portugal Continental serão aplicados a 12 meses e 24 meses e comparados com os valores reais do DW para validação dos modelos.

4.1. Seleção e Transformação dos Dados

Nesta secção é apresentada a necessidade da transformação dos dados e como esta foi executada. Como referido anteriormente será utilizado *o plugin Forecast* do Weka para a criação dos modelos. O Weka é um programa desenhado para auxiliar a aplicação da tecnologia de *machine learning* em dados do mundo real. Este permite a utilização de várias técnicas de *machine learning* através de uma interface unificada (Garner, 1995). Deste modo, para a realização deste objetivo da dissertação, foi necessário instalar um *plugin* de *Forecast* no Weka. O *plugin* adiciona uma nova componente gráfica ao Weka, chamada *Forecast*. Nesta nova componente são inseridas as configurações dos modelos e obtidos os resultados dos mesmos. De forma a que seja possível utilizar os dados no *plugin Forecast* foi necessário realizar algumas transformações nos dados. Para isso foi necessário recorrer a dois processos de ETL criados no *Talend Open Studio for Data Integration* (TOS DI) para duas situações diferentes, que serão explicadas abaixo.

Para esta fase do trabalho foi criado um novo *dataset*, o dataset3, com todos os registos disponíveis, referentes aos casos de pneumonia existentes no DW. Ao contrário dos *datasets* anteriormente utilizados os registos sem coordenadas geográficas e sem outras patologias

associadas são utilizados. O dataset3 é constituído pelos seguintes campos: **sexo**, **idade**, **mes**, **ano,** **n_reingresos**, **dias_internamento** e **vitima_mortais**.

Nestes processos, como já referido, os dados são carregados e agregados por mês e ano. O campo **vitimas_mortais** é somado, de forma a obter o total de vítimas em cada mês e ano. Os campos **sexo** e **idade** foram abordados de forma diferente devido às suas características. O campo sexo é separado e somado em dois novos campos, o **masculino** e o **feminino**. No caso da idade, foi utilizada a média de idades das instâncias agregadas. Os campos **mes** e **ano** apresentam o respetivo mês e ano das instâncias agregadas. Os dados apresentam um intervalo temporal desde 2002 a 2011. É adicionado um campo **casos_pneumonia** que conta o número de eventos de pneumonia para que seja possível utilizar este campo de modo a prever o número de casos de pneumonia. Para cada um dos ficheiros de dados agregados utilizados referentes às regiões/distritos analisadas foram obtidas 120 instâncias agregadas. Além do campo **casos_pneumonia,** foi adicionado outro campo, o **Data**. Este campo é do tipo *date* e permite ao *plugin Forecast* obter o *time stamp* das instâncias agregadas. A criação das instâncias é necessária para que o plugin *Forecast* consiga interpretar os dados e realizar as previsões.

Na Figura 30 encontra-se uma amostra da estrutura de dados presente nos ficheiros arff, neste caso do ficheiro referente a todos os dados de Portugal Continental. No início do ficheiro encontram-se listados os atributos e o seu tipo, seguidos dos registos referentes as instâncias agregadas.

```
1  @RELATION PT
2
3  @ATTRIBUTE idade numeric
4  @ATTRIBUTE mes numeric
5  @ATTRIBUTE ano numeric
6  @ATTRIBUTE masculino numeric
7  @ATTRIBUTE feminino numeric
8  @ATTRIBUTE n_reingressos numeric
9  @ATTRIBUTE dias_internamento numeric
10 @ATTRIBUTE vitimas_mortais numeric
11 @ATTRIBUTE casos_pneumonia numeric
12 @ATTRIBUTE data date yyyy-MM-dd
13
14 @DATA
15 67,02,2011,4800,4446,1708,48433,804,4613,2011-2-01
16 65,02,2009,4047,3651,1480,42887,773,4123,2009-2-01
17 66,02,2010,4346,3794,1535,44460,842,4206,2010-2-01
18 59,02,2003,2705,2462,965,30978,469,3051,2003-2-01
19 58,02,2004,3029,2596,975,33541,499,3242,2004-2-01
```

Figura 30 Amostra do ficheiro Arff dos dados de Portugal

Como referido, foram criados dois processos. O processo ETL 1, representado na Figura 31, tem como objetivo agregar os dados por mês e ano referente a todos os dados sobre Portugal Continental presentes no dataset3. O processo ETL 2 representado na Figura 32 tem como objetivo selecionar os dados dos distritos de Braga, Porto e Lisboa presentes no dataset3. Os dados referentes a Braga e ao Porto serão analisados em conjunto. O motivo da junção dos dois distritos será explicado na secção 4.2. Estes dois distritos apresentam em conjunto uma região de elevada incidência de pneumonias, como verificado na secção 3.3.2. O distrito de Lisboa apresenta uma elevada incidência de pneumonias, como também verificado na secção 3.3.2. Existem outras regiões identificadas na secção 3.3.2 que poderiam ser analisadas, mas a escolha recaiu sobre as duas regiões litorais mais representadas na secção 3.3.2.

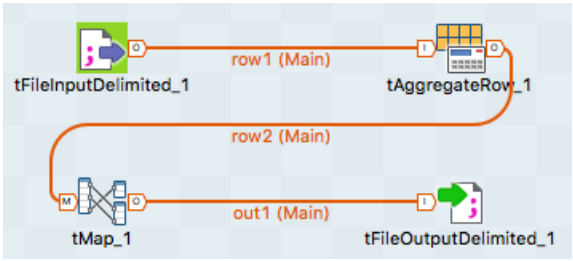


Figura 31 Processo ETL 1

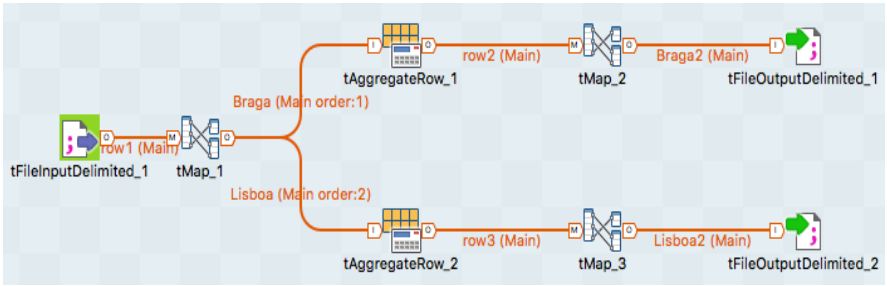


Figura 32 Processo ETL 2

4.2. Construção dos Modelos

Para a construção dos modelos foi necessário escolher os algoritmos e os parâmetros a utilizar para a realização das previsões. A escolha destes algoritmos passou pela observação das percentagens de erro obtidas em alguns testes prévios. Desta forma, foram selecionados quatro algoritmos para identificar e avaliar vários modelos. Foram selecionados 4 algoritmos, o M5P, Linear Regression, SMOreg e o RandomForest. O M5P é um algoritmo que combina árvores de decisão com a utilização de regressão linear nos nós. O Linear Regression, como o próprio nome indica, utiliza a regressão linear para a previsão. O SMOreg implementa um *support vector machine* para

regressão. O RandomForest é um algoritmo de classificação, que cria várias árvores de decisão de forma a usá-las em conjunto para classificar um objeto.

Todos os algoritmos referidos foram testados em condições iguais, como a percentagem de treino e teste, e o número períodos previstos. Como já referido, para criar os modelos, além de selecionar os algoritmos a utilizar é necessário definir os parâmetros. Os parâmetros a definir são o número de previsões que são efetuadas, a periocidade das previsões, a percentagem de treino e os campos *overlayed*. Para o número de previsões foi definido o valor 6 que irá corresponder a 6 meses previstos, pois a periocidade utilizada é o Mês. A utilização de 6 previsões permite obter a média de erros entre as 6 previsões de forma a obter um valor para comparação entre os modelos. Também foi testado o uso de 12 previsões, mas as variações de erro de previsão para previsão mantinham-se no mesmo padrão das 6 previsões. Por isso, de forma a facilitar o processo, foi optado por usar 6 previsões. Para dividir os dados para treino e teste foi utilizado o valor por defeito do *plugin Forecast* do Weka, os 30% para teste e 70% para treino. Ou seja, os dados de treino correspondem ao intervalo temporal entre 2002 e 2008 e os dados de teste ao intervalo temporal de 2009 a 2011.

O *overlayed* refere-se à utilização de outros campos disponíveis no *dataset* para auxiliar a previsão de um determinado campo. Para que seja possível a sua utilização é necessário conhecer os valores destes campos em *overlay*, para o espaço temporal em que se quer efetuar a previsão do campo pretendido. Ou seja, é preciso conhecer os valores futuros dos campos auxiliares para que seja possível utilizá-los para previsões futuras. Não existindo esta informação e para que seja possível estudar modelos com recurso a campos em *overlay* para auxiliar as previsões, neste caso serão geradas previsões sobre os dados de treino apenas, em que são conhecidos os valores dos campos auxiliares. Outro método possivelmente viável para realizar o estudo dos modelos com auxilio de campos em *overlay*, passaria por gerar a previsão dos campos auxiliares em primeiro lugar e em segundo utilizar esses valores para auxiliar o campo que realmente queremos prever. Por exemplo, no caso de queremos prever o número de casos de pneumonia para o ano de 2012, não seria possível utilizar os restantes campos para auxiliar a previsão, pois estes dados não são conhecidos. Mas poderíamos realizar a previsão desses campos e utilizar esses valores obtidos para melhorar a previsão do número de casos de pneumonia. Esta abordagem depende da qualidade das previsões dos campos auxiliares.

Como referido, foram realizadas várias execuções para experimentar quais os algoritmos a selecionar e também foi explorada a utilização de campos para auxiliar a previsão do número de

casos de Pneumonia e de vítimas mortais em Portugal. Através das pré-execuções realizadas averiguou-se que a utilização dos campos em *overlay* presentes no ficheiro de dados beneficiam as previsões na maioria dos modelos. Deste modo, estes campos foram utilizados em alguns dos modelos criados. Os campos utilizados em *overlay* são **idade**, **masculino**, **feminino**, **n_reingressos**, **dias_internamento**, **vitima_mortais** e **casos_pneumonia**. Os modelos com *overlay* estão identificados com o "-O" no nome do modelo.

Como referido na secção 4.1, foram criados dois processos de transformação dos dados, em que no processo 1 os dados de *output* são referentes a Portugal Continental e no processo 2 os dados de output são referentes a Braga-Porto e Lisboa. Braga e Porto durante as execuções de teste dos algoritmos e parâmetros foram explorados em separado e os resultados obtidos eram insatisfatórios. Estes apresentavam previsões de valores negativos e valores de percentagem de erro acima dos 100%. Como já tinham sido realizados testes para os dados de Portugal Continental e obtidos resultados muito diferentes, verificou-se que o motivo para estes resultados estava relacionado com a quantidade de registos nos dois distritos. De modo a melhorar os resultados, os dois distritos foram integrados num só ficheiro. Foram realizados novos testes e os resultados obtidos melhoraram consideravelmente.

As configurações dos modelos que foram criadas e executadas estão apresentadas na Tabela 21, na Tabela 22, na Tabela 23, na Tabela 24, na Tabela 25 e na Tabela 26. As tabelas 21, 23 e 24 contêm as configurações dos modelos referentes à previsão de casos de Pneumonia. As tabelas 22, 24 e 26 contêm as configurações dos modelos referentes à previsão de vítimas mortais.

Portugal- Pneumonias					
Mo.	Algoritmo	Nº Previsões	Periocidade	%teste	Overlayed
F1	M5P	6	Mês	30%	Não
F2	Linear Regression				Não
F3	SMOreg				Não
F4	RandomForest				Não
F1-O	M5P				Sim
F2-O	Linear Regression				Sim
F3-O	SMOreg				Sim
F4-O	RandomForest				Sim

Tabela 21 Configurações dos Modelos de previsão dos casos de Pneumonia em Portugal

Portugal - Vítimas Mortais					
Mo.	Algoritmo	Nº Previsões	Periocidade	%teste	Overlayed
F5	M5P	6	Mês	30%	Não
F6	Linear Regression				Não
F7	SMOreg				Não
F8	RandomForest				Não
F5-O	M5P				Sim
F6-O	Linear Regression				Sim
F7-O	SMOreg				Sim
F8-O	RandomForest				Sim

Tabela 22 Configurações dos Modelos de previsão de vítimas mortais em Portugal

Braga e Porto-Pneumonia					
Mo.	Algoritmo	Nº revisões	Periocidade	%teste	Overlayed
F9	M5P	6	Mês	30%	Não
F10	Linear Regression				Não
F11	SMOreg				Não
F12	RandomForest				Não
F9-O	M5P				Sim
F10-O	Linear Regression				Sim
F11-O	SMOreg				Sim
F12-O	RandomForest				Sim

Tabela 23 Configurações dos Modelos de previsão de casos dos cas Pneumonia em Braga e Porto

Braga e Porto-Vítimas mortais					
Mo.	Algoritmo	Nº Previsões	Periocidade	%teste	Overlayed
F13	M5P	6	Mês	30%	Não
F14	Linear Regression				Não
F15	SMOreg				Não
F16	RandomForest				Não
F13-O	M5P				Sim
F14-O	Linear Regression				Sim
F15-O	SMOreg				Sim
F16-O	RandomForest				Sim

Tabela 24 Configurações dos Modelos de previsão de vítimas mortais– Braga e Porto

Lisboa-Pneumonia					
Mo.	Algoritmo	Nº Previsões	Periocidade	%teste	Overlayed
F17	M5P	6	Mês	30%	Não
F18	Linear Regression				Não
F19	SMOreg				Não
F20	RandomForest				Não
F17-O	M5P				Sim
F18-O	Linear Regression				Sim
F19-O	SMOreg				Sim
F20-O	RandomForest				Sim

Tabela 25 Configurações dos Modelos de previsão dos casos de Pneumonias em Lisboa

Lisboa-Vítimas Mortais					
Mo.	Algoritmo	Nº Previsões	Periocidade	%teste	Overlayed
F21	M5P	6	Mês	30%	Não
F22	Linear Regression				Não
F23	SMOreg				Não
F24	RandomForest				Não
F21-O	M5P				Sim
F22-O	Linear Regression				Sim
F23-O	SMOreg				Sim
F24-O	RandomForest				Sim

Tabela 26 Configurações dos Modelos de previsão de vítimas mortais em Lisboa

4.3. Resultados

Nesta secção serão apresentados os resultados referentes aos modelos descritos na Tabela 21, na Tabela 22, na Tabela 23, na Tabela 24, na Tabela 25 e na Tabela 26 da secção 4.2. Os modelos serão avaliados e comparados de forma a identificar os melhores modelos para cada um dos três casos em estudo, Portugal Continental, Braga-Porto e Lisboa. Os melhores modelos de cada caso serão representados com o gráfico de previsões de forma a estudar o comportamento dos valores previstos. Por fim, o modelo considerado como o melhor na previsão de vítimas mortais em Portugal será aplicado para uma previsão a 12 meses e os resultados obtidos serão comparados com os valores registados no contexto real.

Para avaliar os modelos serão utilizadas as medidas obtidas através do *evaluation* do *Forecast* do Weka. As métricas utilizadas são o *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), e o *Mean Absolute Percentage Error* (MAPE). As medidas MAE e RMSE são duas medidas baseadas no erro absoluto e erros quadráticos respetivamente, sendo as mais utilizadas na avaliação e comparação de modelos de *forecast* num determinado *dataset*. O MAPE é uma medida baseada na percentagem de erros que tem como objetivo permitir a avaliação e comparação da performance dos modelos de *forecast* entre *datasets* diferentes (Hyndman & Athanasopoulus, 2013). Quanto mais próximo de zero for o valor das medidas, melhor é o modelo.

As análises dos modelos serão efetuadas principalmente sobre a medida MAE e MAPE, uma ou outra medida será usada dependendo dos modelos em análise. As seguintes tabelas, Tabela 27, Tabela 29, Tabela 31, Tabela 33, Tabela 35 e a Tabela 37, apresentam os valores médios para cada medida. Estes valores são calculados pela média dos seis meses previstos. Os valores obtidos para cada mês em todos os modelos encontram-se no anexo B.

Na Tabela 27, encontram-se apresentados os resultados dos modelos que têm como objetivo prever o número de casos de pneumonia em Portugal. Os modelos que utilizam atributos em *overlay* apresentam melhores resultados nas medidas de erro como podemos observar na tabela. Esta diminuição é bastante visível nos valores do MAE.

Portugal - Pneumonias					
Modelo	Métrica	Média dos 6 Steps	Modelo	Métrica	Média dos 6 Steps
F1	MAE	513,8	F1-O	MAE	112,9
	RMSE	663,5		RMSE	149,7
	MAPE	14,9		MAPE	3,0
F2	MAE	658,8	F2-O	MAE	298,0
	RMSE	846,0		RMSE	343,6
	MAPE	20,0		MAPE	9,4
F3	MAE	458,2	F3-O	MAE	85,2
	RMSE	590,0		RMSE	113,9
	MAPE	13,9		MAPE	2,4
F4	MAE	319,2	F4-O	MAE	212,4
	RMSE	432,4		RMSE	289,7
	MAPE	8,9		MAPE	6,0

Tabela 27 Resultados dos Modelos de Previsão de Pneumonia em Portugal

Observando a Tabela 27, é possível verificar que o modelo F3-O é o melhor modelo, apresentado o valor de erro mais baixo comparado com os restantes modelos. F3-O utiliza o algoritmo SMOreg. Os resultados obtidos em termos de o modelo de F3-O podem ser consultados na Figura 33 e na Tabela 28.

```
casos_pneumonia:
SMOreg

weights (not support vectors):
 -       0.0493 * (normalized) idade
 -       0.0013 * (normalized) mes
 -       0.0586 * (normalized) ano
 +       0.4182 * (normalized) masculino
 +       0.5873 * (normalized) feminino
 +       0.0128 * (normalized) Month=jan
 -       0.0033 * (normalized) Month=feb
 -       0.0025 * (normalized) Month=mar
 -       0.0105 * (normalized) Month=apr
 +       0.0024 * (normalized) Month=may
 +       0.0013 * (normalized) Month=jun
 -       0.0017 * (normalized) Month=jul
 +       0.0049 * (normalized) Month=aug
 -       0.0074 * (normalized) Month=sep
 +       0.001  * (normalized) Month=oct
 -       0.001  * (normalized) Month=nov
 +       0.0042 * (normalized) Month=dec
 +       0.0069 * (normalized) Quarter=Q1
 -       0.0068 * (normalized) Quarter=Q2
 -       0.0042 * (normalized) Quarter=Q3
 -       0.0166 * (normalized) Lag_casos_pneumonia-2
 +       0.0044 * (normalized) Lag_casos_pneumonia-3
 -       0.0104 * (normalized) Lag_casos_pneumonia-4
 -       0.0217 * (normalized) Lag_casos_pneumonia-5
 -       0.0115 * (normalized) Lag_casos_pneumonia-6
 -       0.0116 * (normalized) Lag_casos_pneumonia-7
 -       0.0231 * (normalized) Lag_casos_pneumonia-8
 +       0.0016 * (normalized) Lag_casos_pneumonia-9
 -       0.0385 * (normalized) Lag_casos_pneumonia-10
 +       0.0361 * (normalized) Lag_casos_pneumonia-11
 -       0.0588 * (normalized) Lag_casos_pneumonia-12
 +       0.0713 * (normalized) data-remapped^2
 -       0.0644 * (normalized) data-remapped^3
 +       0.0364 * (normalized) data-remapped*Lag_casos_pneumonia-1
 -       0.0105 * (normalized) data-remapped*Lag_casos_pneumonia-2
 +       0.0067 * (normalized) data-remapped*Lag_casos_pneumonia-3
 +       0.0063 * (normalized) data-remapped*Lag_casos_pneumonia-4
 +       0.017  * (normalized) data-remapped*Lag_casos_pneumonia-5
 -       0.0236 * (normalized) data-remapped*Lag_casos_pneumonia-6
 +       0.0782 * (normalized) data-remapped*Lag_casos_pneumonia-7
 -       0.028  * (normalized) data-remapped*Lag_casos_pneumonia-8
 +       0.0235 * (normalized) data-remapped*Lag_casos_pneumonia-9
 -       0.0008 * (normalized) data-remapped*Lag_casos_pneumonia-10
 -       0.023  * (normalized) data-remapped*Lag_casos_pneumonia-11
 +       0.017  * (normalized) data-remapped*Lag_casos_pneumonia-12
 0.0855
```

Figura 33 Fórmula do Modelo F3-O para a previsão do número de casos de Pneumonia em Portugal

Como referido na secção 4.2, os resultados das previsões dos modelos que utilizam campos *overlay* apresentam a previsão de 6 meses no ano a seguir ao ultimo ano dos dados de treino, de modo a utilizar os valores dos campos utilizados em *overlay* dos dados de teste. O atributo que apresenta o maior peso no modelo é o ano, Figura 33.

	Casos de Pneumonia	
Data	Valor Real	Valor Previsto
01/01/2009	5581	5392
01/02/2009	4163	4123
01/03/2009	4239	4251
01/04/2009	3631	3675
01/05/2009	3425	3528
01/06/2009	2671	2803

Tabela 28 Tabela de comparação entre os valores reais e previstos para os casos de Pneumonia em Portugal do modelo F3-O

Na Tabela 29 encontram-se os resultados referentes aos modelos para Portugal Continental na previsão de vítimas mortais. O modelo F5, que não recorre à utilização de atributos em *overlay* para auxiliar a previsão do número de vítimas mortais, apresenta ser o modelo com o maior erro. Este modelo, tal como todos os modelos, apresenta melhorias significativas nos resultados quando utilizado atributos em *overlay*.

Portugal - Vítimas Mortais					
Modelo	Métrica	Média dos 6 Steps	Modelo	Métrica	Média dos 6 Steps
F5	MAE	224,8	F5-O	MAE	29,9
	RMSE	287,4		RMSE	37,1
	MAPE	32,8		MAPE	4,4
F6	MAE	145,0	F6-O	MAE	37,9
	RMSE	188,0		RMSE	49,8
	MAPE	21,0		MAPE	5,5
F7	MAE	72,7	F7-O	MAE	58,4
	RMSE	99,3		RMSE	71,4
	MAPE	10,8		MAPE	8,5
F8	MAE	59,7	F8-O	MAE	46,6
	RMSE	78,9		RMSE	56,9
	MAPE	9,1		MAPE	7,1

Tabela 29 Resultados dos Modelos de Previsão Vítimas Mortais em Portugal

É possível observar que o modelo para a previsão de vítimas mortais F5-O apresenta o menor valor para a medida MAE, deste modo é considerado o melhor modelo para previsão de vítimas mortais. F5-O utiliza o algoritmo M5P. O modelo e os resultados de F5-O encontram-se apresentados na Figura 34 e na Tabela 30. Os atributos com maior peso no modelo são o atributo **mês** e **ano**, nestes casos os dados referentes aos meses de Dezembro, Janeiro e Fevereiro que são meses em que existe uma grande incidência de pneumonias.

```
vitimas_mortais:
M5 pruned model tree:
(using smoothed linear models)
LM1 (84/17.217%)

LM num: 1
vitimas_mortais =
        15.143 * idade
        - 42.2197 * ano
        + 0.0038 * dias_internamento
        + 0.1494 * casos_pneumonia
        + 24.3061 * Month=feb,dec,jan
        + 4.1627 * data-remapped
        - 0.0021 * data-remapped*Lag_vitimas_mortais-4
        + 0.0012 * data-remapped*Lag_vitimas_mortais-5
        + 83503.9196

Number of Rules : 1
```

Figura 34 Fórmula do Modelo F5-O para a previsão de vítimas mortais em Portugal

	Vítimas Mortais	
Data	Valor Real	Valor Previsto
01/01/2009	1101	1048
01/02/2009	784	770
01/03/2009	833	755
01/04/2009	715	625
01/05/2009	699	644
01/06/2009	580	584

Tabela 30 Tabela de comparação entre os valores reais e previstos para o número de vítimas mortais em Portugal do modelo F5-O

No gráfico da Figura 35 encontra-se representada a comparação do uso de atributos em *overlay* para os algoritmos utilizados nos modelos na previsão de vítimas mortais, no qual podemos confirmar que o uso de atributos em *overlay* para determinados modelos permite reduzir o erro para as três medidas. Isto acontece devido aos atributos em *overlay* estarem relacionados temporalmente com os acontecimentos previstos. Por exemplo, o campo **casos_pneumonia** que apresenta o total de casos pneumonia relaciona-se com o número de vítimas mortais, pois quando aumenta o número de casos de pneumonia também existe a tendência do número de vítimas mortais aumentarem. A relação entre os dois acontecimentos permite obter maior precisão nos resultados.

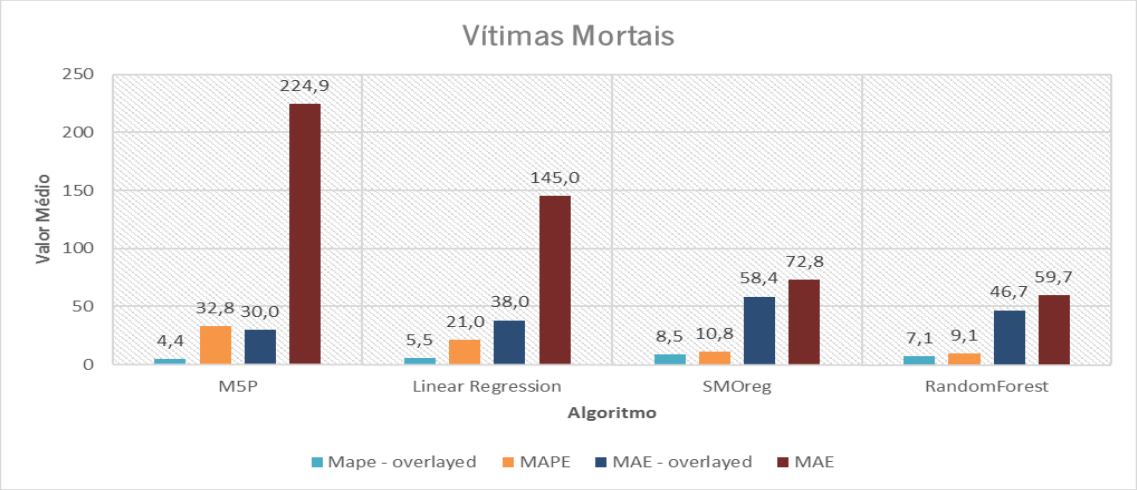


Figura 35 Gráfico de comparação das medidas MAE e MAPE referente à previsão de vítimas mortais em Portugal

Como anteriormente referido, os modelos com os melhores resultados serão explorados através do gráfico de comparação da previsão gerada e os dados de teste. Neste caso, serão representados os modelos F3-0 e F5-0.

Na Figura 36, referente ao modelo F3-0, podemos observar que este consegue prever a diminuição e o aumento dos valores e em certos momentos os valores previstos conseguem praticamente sobrepor-se aos valores dos dados de teste. É de acrescentar que o modelo consegue identificar a sazonalidade, isso é percetível através da subida do número de casos de Pneumonia nos meses de inverno.

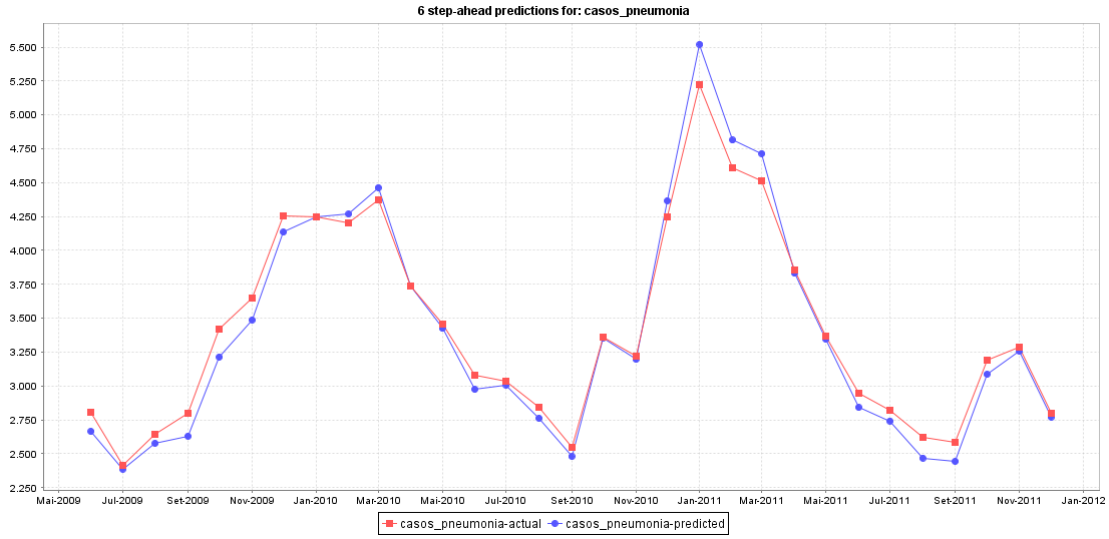


Figura 36 Gráfico comparativo dos valores da previsão sobre os dados de teste referente modelo F3-0

Na Figura 37, encontra-se representado o gráfico referente aos valores de previsão do número de vítimas mortais em comparação com os valores dos dados de teste. No que é possível observar,

os valores previstos pelo modelo aproximam-se bastante dos valores dos dados de teste. O modelo consegue identificar com rigor todas as variações significativas ao longo do tempo. É de destacar que existe uma relação entre o número de casos de pneumonia e o número de vítimas mortais quando comparadas a Figura 36 e a Figura 37.

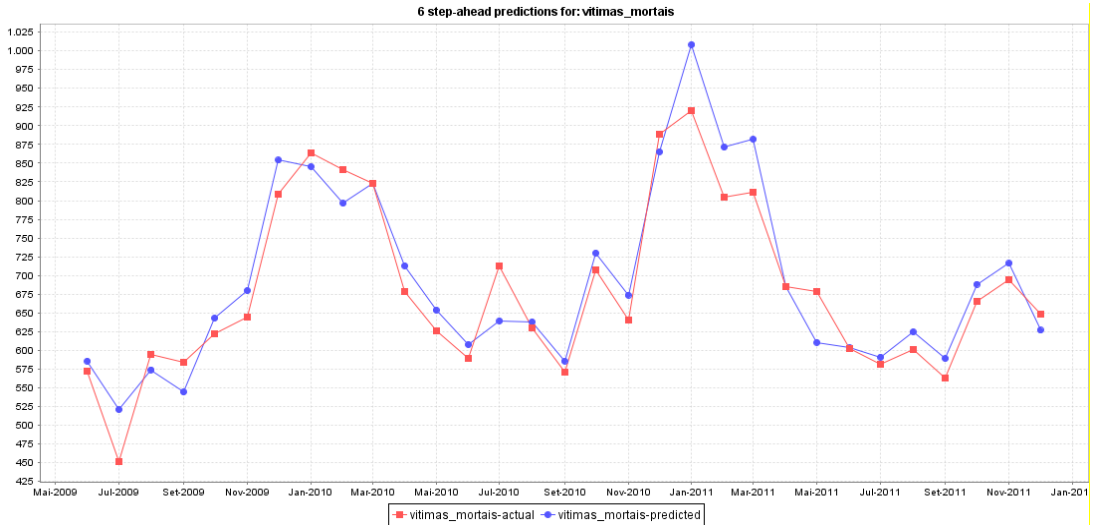


Figura 37 Gráfico comparativo dos valores da previsão sobre os dados de teste referente ao modelo F5-O

Após a análise dos resultados dos modelos criados para Portugal Continental foi feita a aplicação dos modelos para determinadas regiões, neste caso os distritos de Braga e Porto. O motivo que levou a juntar os dois distritos encontra-se explicado na secção 4.2. Pretende-se, assim, averiguar se os modelos aplicados aos dados de um *dataset* de menor dimensão mostram resultados próximos dos resultados obtidos nos modelos para Portugal Continental. Na Tabela 31 e na Tabela 33 são apresentados os valores médios obtidos nas previsões dos modelos definidos na secção 4.2 para o *dataset* referente a Braga e Porto.

Analisando os resultados referentes aos modelos preditivos de Pneumonias para Braga e Porto, na Tabela 31 é de destacar os modelos F10, F10-O, F11 e F11-O que utilização atributos em *overlay* apresentam uma melhoria significativa nos valores de erro na previsão de casos de Pneumonia. O modelo F11-O apresenta apenas um valor de erro superior de 0,2% em relação ao modelo F9-O, este também poderia ser considerado junto com o F9-O os melhores modelos.

Braga e Porto - Pneumonias					
Modelo	Métrica	Média dos 6 Steps	Modelo	Métrica	Média dos 6 Steps
F9	MAE	91,0	F9-O	MAE	26,3
	RMSE	124,8		RMSE	37,9
	MAPE	12,8		MAPE	3,3
F10	MAE	429,8	F10-O	MAE	64,2
	RMSE	479,7		RMSE	81,8
	MAPE	64,9		MAPE	10,2
F11	MAE	320,3	F11-O	MAE	24,5
	RMSE	360,4		RMSE	30,1
	MAPE	48,5		MAPE	3,5
F12	MAE	75,6	F12-O	MAE	44,9
	RMSE	96,8		RMSE	56,9
	MAPE	10,1		MAPE	6,1

Tabela 31 Resultados dos Modelos de Previsão de Pneumonia em Braga e Porto

O melhor modelo na previsão de pneumonias para Braga e Porto é o F9-O. O modelo e os resultados obtidos de F9-O podem ser consultados na Figura 38 e na Tabela 32. O atributo em *overlay* que apresenta o maior peso é a **idade**, além deste, o atributo **ano** apresenta também um elevado peso.

```
casos_pneumonia:
M5 pruned model tree:
(using smoothed linear models)
LM1 (84/4.909%)

LM num: 1
casos_pneumonia =
        -11.5967 * idade
        - 9.2977 * ano
        + 0.4237 * masculino
        + 0.6581 * feminino
        - 1.5952 * data-remapped
        - 0.0215 * Lag_casos_pneumonia-1
        + 19702.638

Number of Rules : 1
```

Figura 38 Fórmula do Modelo F9-O para a previsão de casos de Pneumonia em Braga e Porto

	Casos de Pneumonia	
Data	Valor Real	Valor Previsto
01/01/2009	1150	1210
01/02/2009	921	944
01/03/2009	929	952
01/04/2009	768	766
01/05/2009	768	746
01/06/2009	596	603

Tabela 32 Tabela de comparação entre os valores reais e previstos para os casos de Pneumonia em Braga e Porto do modelo F9-O

Como é possível observar na Tabela 33, de acordo com os resultados obtidos para os modelos preditivos de vítimas mortais em Portugal Continental, o melhor modelo continua a ser o que utiliza o algoritmo M5P, o modelo F13-O.

Braga e Porto – Vítimas Mortais					
Modelo	Métrica	Média dos 6 Steps	Modelo	Métrica	Média dos 6 Steps
F13	MAE	37,7	F13-O	MAE	10,6
	RMSE	42,4		RMSE	13,8
	MAPE	32,3		MAPE	8,9
F14	MAE	73,9	F14-O	MAE	29,1
	RMSE	83,6		RMSE	33,6
	MAPE	64,9		MAPE	25,3
F15	MAE	37,7	F15-O	MAE	14,5
	RMSE	42,4		RMSE	17,2
	MAPE	32,3		MAPE	12,3
F16	MAE	13,4	F16-O	MAE	12,8
	RMSE	17,1		RMSE	15,1
	MAPE	10,3		MAPE	10,5

Tabela 33 Resultados dos Modelos de Previsão Vítimas Mortais em Braga e Porto

Na Figura 55 encontra-se representado as regras que caracterizam o modelo F13-O. Este é constituído por 2 regras, a regra número 1 é utilizada quando o número total de indivíduos masculinos é inferior ou igual a 3173 e a regra número 2 quando o número de indivíduos do sexo masculino é superior a 3173. Os atributos em *overlay* que apresenta maior peso na formula da regra número 1 é o atributo **feminino** e na regra número 2 é o atributo **masculino**.

```
vitimas_mortais:
M5 pruned model tree:
(using smoothed linear models)

masculino <= 3173 : LM1 (63/21.061%)
masculino >  3173 : LM2 (21/21.749%)

LM num: 1
vitimas_mortais =
        0.0351 * masculino
        + 0.1338 * feminino
        + 0.0096 * n_reingressos
        + 0.0003 * dias_internamento
        - 0.0356 * casos_pneumonia
        + 1.1151 * data-remapped
        - 0.0986 * Lag_vitimas_mortais-3
        - 0.067 * Lag_vitimas_mortais-4
        + 0.0142 * Lag_vitimas_mortais-7
        - 0.0184 * Lag_vitimas_mortais-12
        - 0.0002 * data-remapped*Lag_vitimas_mortais-4
        + 0.001 * data-remapped*Lag_vitimas_mortais-5
        + 202.5697

LM num: 2
vitimas_mortais =
        0.1579 * masculino
        + 0.0011 * feminino
        + 0.0207 * n_reingressos
        + 0.0135 * dias_internamento
        - 0.077 * casos_pneumonia
        + 0.2333 * data-remapped
        - 0.0668 * Lag_vitimas_mortais-3
        + 0.0308 * Lag_vitimas_mortais-7
        - 0.0399 * Lag_vitimas_mortais-12
        - 0.0005 * data-remapped*Lag_vitimas_mortais-4
        + 0.0004 * data-remapped*Lag_vitimas_mortais-5
        - 119.8283

Number of Rules : 2
```

Figura 39 Fórmula do Modelo F13-O para a previsão de vítimas mortais em Braga e Porto

Os resultados obtidos de F13-O podem ser consultados na Tabela 34. Como esperado nos meses de inverno existe um maior número de casos previstos e que tendem a diminuir com o fim do inverno.

	Vítimas Mortais	
Data	Valor Real	Valor Previsto
01/01/2009	209	204
01/02/2009	159	153
01/03/2009	113	141
01/04/2009	119	125
01/05/2009	104	117
01/06/2009	93	97

Tabela 34 Tabela de comparação entre os valores reais e previstos para o número de vítimas mortais em Braga e Porto do modelo F13-O

O gráfico do modelo F9-O referente à previsão do número de casos de pneumonia nos distritos de Braga e Porto encontra-se representado na Figura 40. É possível observar através do gráfico que o modelo F9-O apresenta resultados muito próximos dos reais. Este consegue identificar todos os comportamentos de subida e descida dos valores dos dados de teste e obter valores muito próximos. A maior variação entre os valores reais e os valores previstos encontra-se nos meses de Dezembro de 2010 e Março de 2011. Pode-se concluir que F9-O é um bom modelo na previsão de pneumonias perante os dados referentes a Braga e Porto.

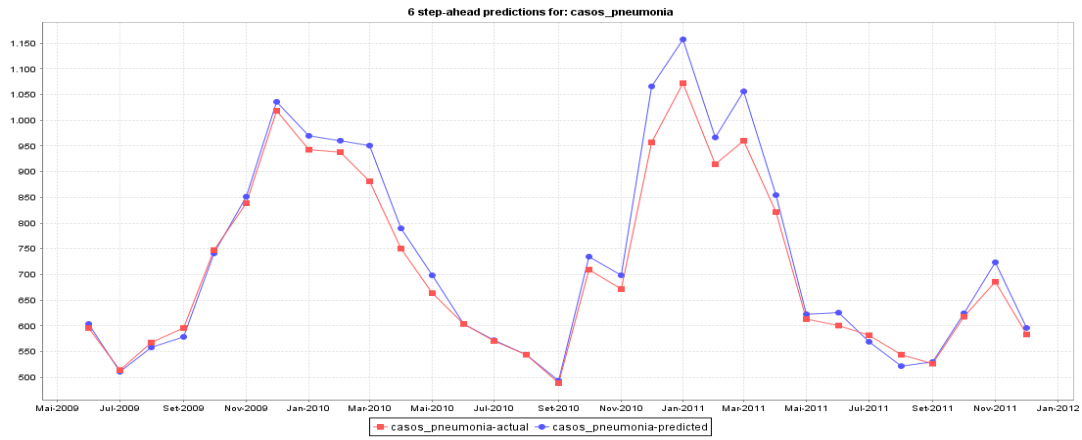


Figura 40 Gráfico comparativo dos valores da previsão sobre os dados de teste referente ao modelo F9-O

Na Figura 41, referente às vítimas mortais em Braga e Porto, encontra-se representado o gráfico do modelo F13-O. Neste é possível observar que existem folgas maiores entre os dados de teste e os dados previstos pelo *forecast*. Existem momentos em que o modelo não consegue identificar a tendência de descida ou subida dos valores, como por exemplo em Junho de 2010. Em outros momentos o comportamento dos resultados obtidos pelo modelo melhora significativamente,

apresentando um comportamento mais próximo dos valores de teste, apenas com alguma discrepância nos valores previstos, como por exemplo entre Novembro de 2009 e Maio de 2010. Apesar de não ser um modelo que consiga prever com precisão a evolução da doença ao longo dos meses, o F13-O não deixa de ser um bom modelo.

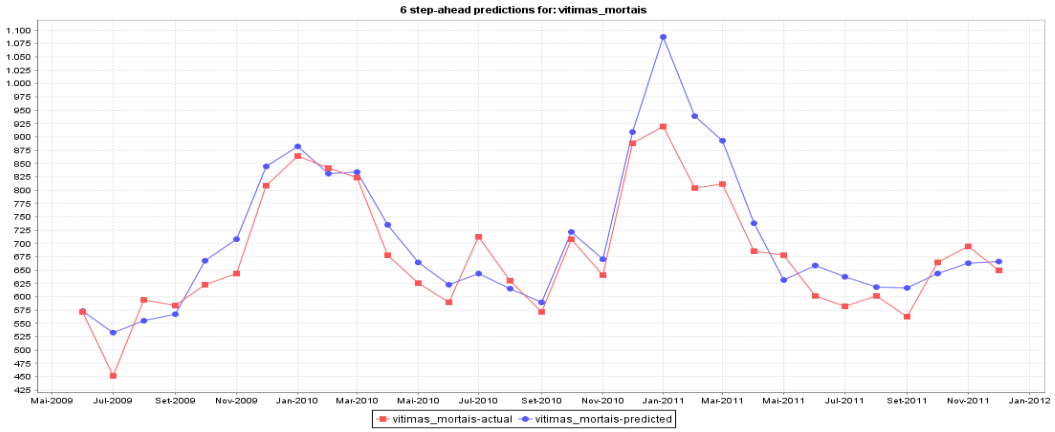


Figura 41 Gráfico comparativo dos valores da previsão sobre os dados de teste referente ao modelo F13-O

Após a análise dos resultados dos modelos criados para os distritos de Braga e Porto foi realizada a análise dos resultados obtidos nos modelos para o distrito de Lisboa. Podemos observar na Tabela 35, referente aos modelos de previsão de casos de Pneumonia em Lisboa, que o modelo que apresenta a menor percentagem de erro é o F17-O. Os modelos F17, F18 e F19, são os modelos que apresentam os piores resultados, mas quando aplicados com auxílio de atributos em *overlay* sofrem melhorias significativas na percentagem de erros obtidas.

Lisboa - Pneumonias					
Modelo	**Métrica**	**Média Steps**	**Modelo**	**Métrica**	**Média Steps**
F17	MAE	144,5	F17-O	MAE	27,1
	RMSE	201,2		RMSE	41,6
	MAPE	19,1		MAPE	3,5
F18	MAE	150,0	F18-O	MAE	45,1
	RMSE	193,6		RMSE	53,2
	MAPE	22,9		MAPE	7,0
F19	MAE	201,3	F19-O	MAE	32,9
	RMSE	232,4		RMSE	47,0
	MAPE	28,8		MAPE	4,6
F20	MAE	93,4	F20-O	MAE	42,4
	RMSE	133,9		RMSE	65,7
	MAPE	12,5		MAPE	6,0

Tabela 35 Resultados dos Modelos de Previsão de Pneumonia em Lisboa

Assim, o modelo F17-O é o melhor modelo para a previsão de pneumonias no distrito de Lisboa. A fórmula e os resultados da previsão a 6 meses do modelo F17-O podem ser consultadas na Figura 42 e na Tabela 36. O atributo em *overlay* que apresenta o maior peso no modelo F17-O é a **idade**. Tal como em nos casos anteriores, o maior número de casos verifica-se em meses de inverno com uma tendência para diminuir ao longo dos meses.

```
casos_pneumonia:
M5 pruned model tree:
(using smoothed linear models)
LM1 (84/6.748%)

LM num: 1
casos_pneumonia =
        -1.8297 * idade
        + 0.5027 * masculino
        + 0.5504 * feminino
        - 0.5254 * data-remapped
        - 0.0175 * Lag_casos_pneumonia-4
        - 0.0357 * Lag_casos_pneumonia-12
        + 193.7386

Number of Rules : 1
```

Figura 42 Fórmula do Modelo F22-O para a previsão de casos de Pneumonia em Lisboa

	Casos de Pneumonia	
Data	Valor Real	Valor Previsto
01/01/2009	1068	1141
01/02/2009	759	765
01/03/2009	772	773
01/04/2009	611	613
01/05/2009	607	583
01/06/2009	531	494

Tabela 36 Tabela de comparação entre os valores reais e previstos para os casos de Pneumonia em Lisboa do modelo F17-O

No geral os modelos com *overlay* para a previsão de casos de Pneumonias em Lisboa obtiveram bons resultados, com o destaque do modelo F17-O como já referido, com uma percentagem de erros compreendido entre os 3.5% e os 7%.

Na Tabela 37 encontram-se apresentados os resultados dos modelos para a previsão de vítimas mortais em Lisboa. Através dos resultados obtidos é possível verificar, tal como nos casos anteriores, que os modelos preditivos para a previsão de vítimas mortais que utilizam o algoritmo M5P com o auxílio de atributos em *overlay* apresentam os melhores resultados. É também de destacar que os modelos F22 e F23 apresentam melhorias com o uso de atributos em *overlay*.

Lisboa - Vítimas Mortais					
Modelo	Métrica	Média Steps	Modelo	Métrica	Média Steps
F21	MAE	36,9	F21-O	MAE	11,7
	RMSE	47,9		RMSE	17,1
	MAPE	26,0		MAPE	8,2
F22	MAE	33,6	F22-O	MAE	18,7
	RMSE	42,2		RMSE	23,7
	MAPE	24,0		MAPE	13,7
F23	MAE	23,5	F23-O	MAE	17,9
	RMSE	28,8		RMSE	25,2
	MAPE	16,6		MAPE	12,9
F24	MAE	22,9	F24-O	MAE	14,0
	RMSE	28,5		RMSE	17,7
	MAPE	17,7		MAPE	10,1

Tabela 37 Resultados dos Modelos de Previsão de Vítimas Mortais em Lisboa

Neste caso, o modelo F21-O é o melhor modelo, a fórmula e os resultados do modelo encontram-se representados na Figura 43 e na Tabela 38. No modelo F21-O o atributo em *overlay* com maior peso é também a idade, como no modelo F17-O. Através dos valores obtidos nas previsões a 6 meses verifica-se um maior número de vítimas nos meses de inverno, como seria de esperar, pois o número de casos nesses meses é mais elevado.

```
vitimas_mortais:
M5 pruned model tree:
(using smoothed linear models)
LM1 (84/27.076%)

LM num: 1
vitimas_mortais =
        3.3801 * idade
        + 1.728 * mes
        + 0.133 * feminino
        + 0.0074 * dias_internamento
        + 0.3286 * data-remapped
        + 0.1579 * Lag_vitimas_mortais-2
        - 0.0032 * data-remapped*Lag_vitimas_mortais-2
        - 239.5802
```

Figura 43 Fórmula do Modelo F21-O para a previsão de vitimas mortais em Lisboa

	Casos de Pneumonia	
Data	Valor Real	Valor Previsto
01/01/2009	1068	1141
01/02/2009	759	765
01/03/2009	772	773
01/04/2009	611	613
01/05/2009	607	583
01/06/2009	531	494

Tabela 38 Tabela de comparação entre os valores reais e previstos para os casos de Pneumonia em Lisboa do modelo F21-O

Como é possível observar na Figura 44, referente ao gráfico do modelo preditivo F17-O, este apresenta valores bastante próximos dos valores dos dados de teste. É de destacar que o comportamento do modelo perante a diminuição e o aumento dos valores enquadra-se quase na perfeição com as variações dos valores dos dados de teste. Este apresenta valores um pouco acima dos valores reais, um comportamento visível na maioria dos meses.

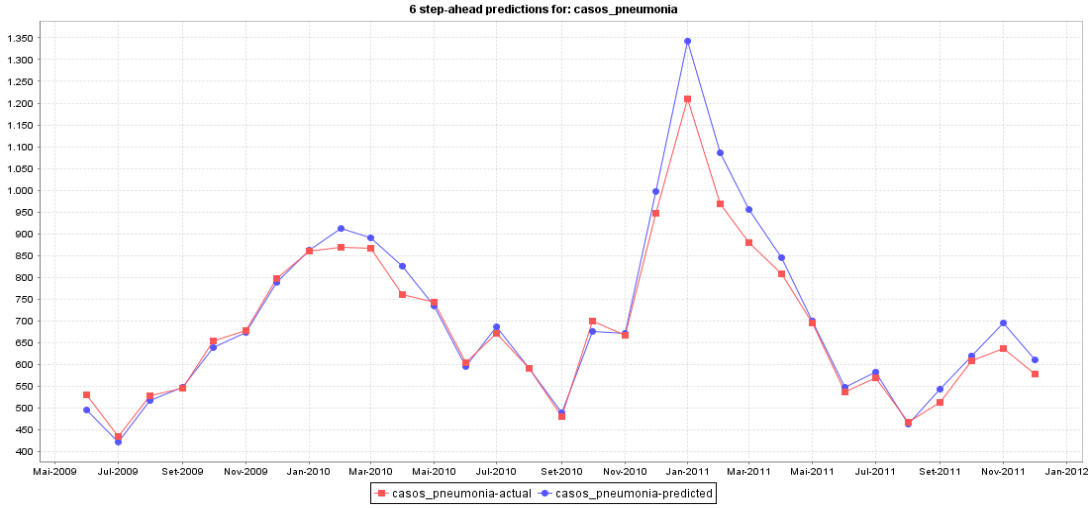


Figura 44 Gráfico comparativo dos valores da previsão sobre os dados de teste referente ao modelo F17-O

O gráfico gerado pelo modelo F21-O encontrasse representado na Figura 45. Neste é possível observar o comportamento dos valores obtidos no modelo F21-O, em comparação com os valores dos dados de teste. Pode-se verificar que o modelo consegue identificar o comportamento de diminuição e aumento dos valores ao longo do tempo, mas por vezes os valores ficam distantes dos reais, como por exemplo nos períodos de Novembro de 2009 a Maio de 2010. Podemos considerar F21-O um bom modelo, mas apresenta falta de precisão em determinados períodos.

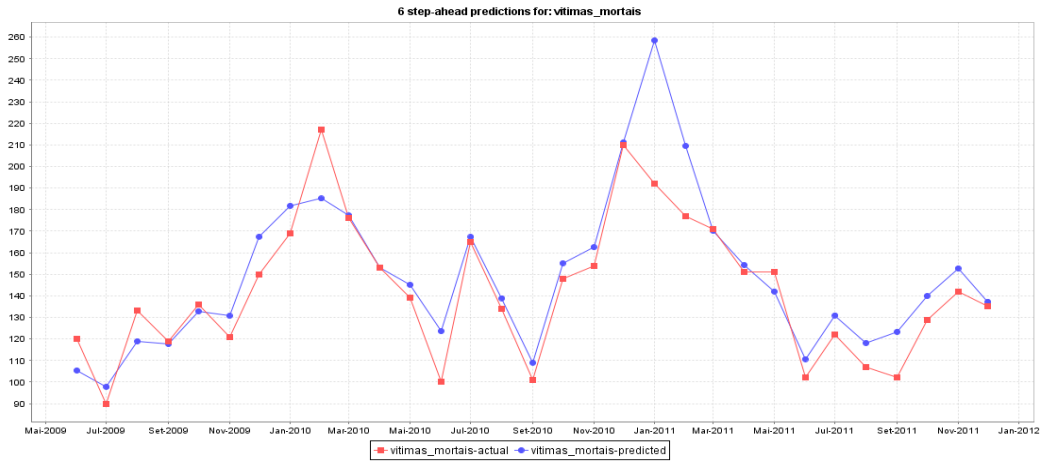


Figura 45 Gráfico comparativo dos valores da previsão sobre os dados de teste referente ao modelo F21-O

Perante os três casos apresentados podemos concluir que os modelos que utilizam os algoritmos M5P apresentam quase sempre os melhores resultados com exceção do Modelo F3-O que é considerado o melhor para a previsão de pneumonias em Portugal. Mas os valores obtidos nas medidas em F1-O, que utiliza o algoritmo M5P, são próximos dos obtidos em F23-O e podemos concluir que F1-O também seria um modelo viável. Assim, perante os resultados obtidos, podemos afirmar que se fosse realizada uma análise de outras regiões e de modo a não ser necessário executar vários algoritmos, possivelmente poderíamos utilizar modelos com o algoritmo M5P e obter bons resultados. No caso de não ser possível utilizar atributos em *overlay* é de destacar que os modelos que utilizam o algoritmo *RandomForest* apresentam os melhores resultados. Analisando os modelos, em concreto as fórmulas apresentadas anteriormente, podemos identificar um conjunto de atributos que apresentam maior peso nos modelos. Os atributos que se destacam são **mes**, **ano,** **feminino** e **masculino**.

Como referido no início da capítulo, será realizada avaliação a 12 e a 24 meses dos melhores modelos identificados para a previsão do número de casos de Pneumonia e de vítimas mortais por Pneumonia em Portugal Continental. A previsão será efetuada para os anos de 2009 e 2010, os valores previstos serão comparados com os valores reais presentes no DW. Para tal, serão apresentadas medidas de erro, mas também serão apresentados os valores previstos, mês a mês, com os valores reais que estão no DW. Além da comparação dos valores previstos com os valores do DW, serão também comparados com os valores totais que aparecem no relatório oficial (DGS, 2014). Os dados relativos a 2011 ainda não podem ser utilizados, pois o número total de casos presente no DW é muito inferior ao oficial registado, razão pela qual o DW precisa primeiro ser refrescado e só depois uma análise deste género pode ser efetuada.

Os valores totais para o número de casos de Pneumonia e o número de vítimas mortais do DW e dos relatórios oficiais encontram-se na Tabela 39.

	DW		Relatórios oficiais	
Ano	Casos Pneumonia	Vítimas Mortais	Casos Pneumonia	Vítimas Mortais
2009	42948	8082	42497	8499
2010	42368	8573	42712	8969

Tabela 39 Valores totais do DW e Relatório oficial de casos de Pneumonia e vítimas mortais em Portugal Continental nos anos 2009 e 2010

É possível verificar na Tabela 39 que existe algumas diferenças nos valores totais do DW e nos valores oficiais dos relatórios. Estas diferenças podem estar associadas à forma como são

contabilizados os casos, principalmente quando há transições de anos. No caso de um indivíduo que vai ao hospital nos últimos dias de 2009 e tem alta só em 2010, estes podem ser contabilizados em anos diferentes, no ano em que foi verificada a doença ou no ano que tem alta ou em em que é verificado o óbito. Assim, não existem garantias que todas as entidades usem os mesmos critérios, o que leva a variações na contabilização.

Em primeiro lugar, foi realizada a previsão do número de casos de Pneumonia para 2009 e 2010. Para este caso será utilizado o modelo F3-O, que utiliza o algoritmo SMOreg e atributos em *overlay*. Na Tabela 40 podemos observar os resultados obtidos para as medidas de erro do modelo F3-O com previsões a 12 meses, com a MAE= 85,3, o RMSE= 114,8 e a MAPE=2,4, e para 24 meses a MAE= 89,6, o RMSE= 123,6 e a MAPE=2,5. O modelo apenas apresenta uma diferença de 0,1% na medida MAE entre a previsão 12 meses e a de 24 meses. Como podemos verificar, os valores de erro obtidos são muito baixos, que permite afirmar que F3-O tem uma elevada precisão nos resultados, quer a 12 meses ou a 24 meses.

Em segundo lugar, foi realizada a previsão do número de vítimas mortais para 2009 e 2010. Para este caso será utilizado o modelo F5-O, que utiliza o algoritmo M5P e atributos em *overlay*. Na Tabela 40 podemos observar os resultados obtidos para as medidas de erro do modelo F5-O com previsões a 12 meses, com a MAE= 30,9, o RMSE= 38,2 e a MAPE=4,4, e para 24 meses a MAE= 32,8, o RMSE= 41,3 e a MAPE=4,5. Mais uma vez o modelo apresenta uma diferença de 0,1% na medida MAE entre a previsão 12 meses e a de 24 meses. Como podemos verificar, os valores de erro podem ser considerados baixos, sendo estes menores que 5%, pelo que permite afirmar que F5-O apresenta uma precisão considerável nos resultados, quer a 12 meses ou a 24 meses.

	Casos de Pneumonia		Vítimas Mortais	
	Média de erros			
	2009	2010	2009	2010
MAE	85,3	89,6	30,9	32,8
RMSE	114,8	123,6	38,2	41,3
MAE	2,4	2,5	4,4	4,5

Tabela 40 Resultados do erro na previsão de casos de Pneumonia e de vítimas mortais em Portugal Continental em 2009 e 2010

Na Tabela 41 encontram-se os valores previstos para os casos de Pneumonia em 2009 e 2010, mês a mês, e os valores disponíveis no DW. Uma vez que não existem os valores mensais no relatório oficial, apenas é apresentado o total. Com a análise da tabela é possível verificar que os valores mensais se encontram relativamente próximos, como seria de esperar com o baixo valor de erro

apresentado. Esta proximidade dos valores é também possível de ser observada com a ajuda da Figura 46.

Casos de Pneumonia 2009													
Mês	jan/09	fev/09	mar/09	abr/09	mai/09	jun/09	jul/09	ago/09	set/09	out/09	nov/09	dez/09	Total Ano
Previsão	5581	4164	4240	3632	3425	2672	2395	2604	2619	3241	3498	4146	42217
DW	5392	4123	4251	3675	3528	2803	2413	2641	2801	3421	3648	4252	42948
R. Oficial													42497
Casos de Pneumonia 2010													
Mês	jan/10	fev/10	mar/10	abr/10	mai/10	jun/10	jul/10	ago/10	set/10	out/10	nov/10	dez/10	Total Ano
Previsão	4251	4262	4477	3760	3445	2993	3022	2777	2487	3373	3224	4379	42450
DW	4246	4206	4376	3737	3458	3080	3039	2845	2550	3364	3222	4245	42368
R. Oficial													42712

Tabela 41 Tabela de comparação dos valores previstos com os valores do DW e dos relatórios oficiais para número de casos de Pneumonia

Comparando os valores totais para o ano de 2009, podemos verificar que o valor previsto fica 731 casos abaixo do valor real. Esta diferença ainda pode ser considerada significativa, mas não deixa de ser um bom resultado. Por outro lado, o valor total previsto de casos de Pneumonia apresenta uma diferença menor, de 280 casos, em relação aos dados do relatório oficial. Como o relatório oficial apenas contém os dados a partir de 2009, não são conhecidos os valores oficiais para todos os anos presentes no DW, elo que não podemos verificar se os totais para anos anteriores são próximos dos que estão no DW.

Em relação ao ano de 2010, é possível verificar que o valor total previsto de casos de Pneumonia fica apenas 82 casos acima do valor real do DW, o que é uma diferença muito baixa, comparado com 2009. Em relação ao valor de casos totais de Pneumonia do relatório oficial, o valor previsto apresenta um número de casos inferior de 262 casos.

Na Tabela 42 encontram-se os valores previstos para o número de vítimas mortais por Pneumonia em 2009 e 2010, mês a mês, com os valores os valores do DW. Tal como nos casos de Pneumonia, não existem os valores mensais no relatório oficial, deste modo apenas é apresentado o total. Com a análise da tabela é possível verificar que os valores mensais se encontram relativamente próximos, como seria de esperar com o baixo valor de erro apresentado. Esta proximidade dos valores é também possível ser observada com a ajuda da Figura 47.

Vítimas Mortais 2009													
Mês	jan/09	fev/09	mar/09	abr/09	mai/09	jun/09	jul/09	ago/09	set/09	out/09	nov/09	dez/09	Total Ano
Previsão	1048	771	756	626	644	586	521	574	543	644	681	853	8247
DW	1080	773	726	610	616	572	452	594	584	622	644	809	8082
R. Oficial													8499
Mês	jan/10	fev/10	mar/10	abr/10	mai/10	jun/10	jul/10	ago/10	set/10	out/10	nov/10	dez/10	Total Ano
Previsão	846	796	825	709	654	611	639	639	583	730	676	866	8574
DW	864	842	823	678	626	589	713	630	571	708	641	888	8573
R. Oficial													8970

Tabela 42 Tabela de comparação dos valores previstos com os valores do DW e dos relatórios oficiais para número de vítimas mortais por Pneumonia

Através da Tabela 42 é possível verificar que para 2009 o número de vítimas mortais previsto ultrapassou o valor real do DW em 165 vítimas mortais. Esta diferença pode ser considerada baixa tendo em conta o total de vítimas mortais registado. Em relação ao relatório oficial o valor previsto ficou abaixo com menos 252 vítimas mortais. No ano de 2010 a diferença do valor previsto e do valor real do DW foi de apenas 1 vítima mortal, o que é um bom resultado. Em relação ao relatório oficial o valor previsto ficou abaixo com menos 396 vítimas mortais.

Na Figura 46 e Figura 47, podemos ainda analisar o comportamento dos valores em termos de subidas e descidas dos mesmos ao longos dos meses. Assim, é possível verificar que os modelos utilizados conseguem identificar todos as descidas e subidas dos valores ao longo dos meses de forma precisa, o que permite aos modelos identificar a sazonalidade da doença ao longo dos meses.

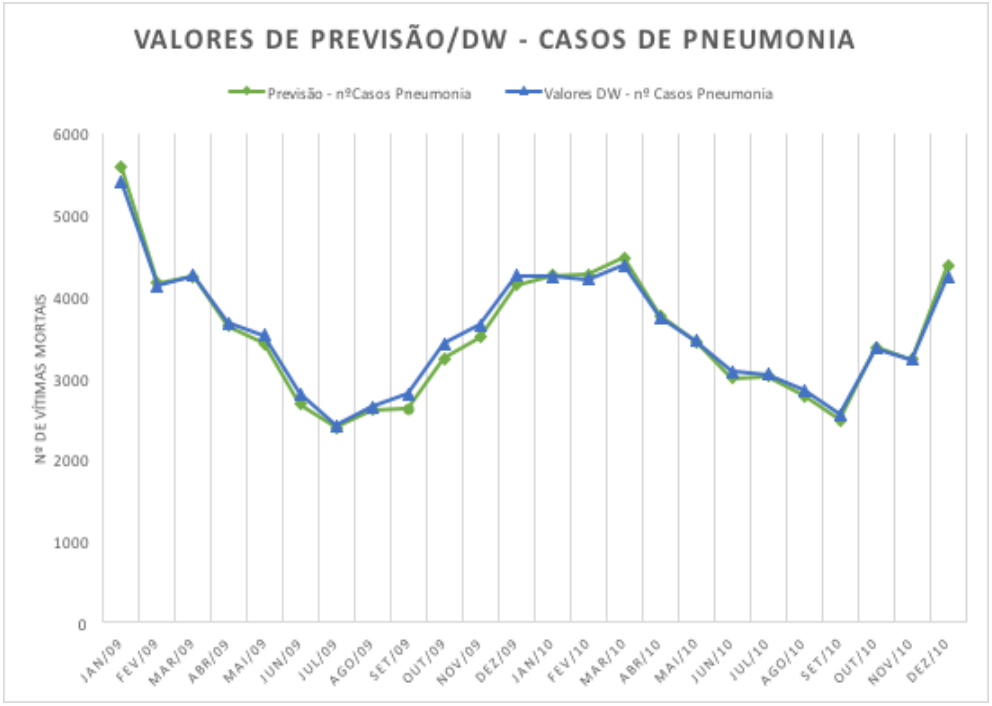


Figura 46 Representação do comportamento dos valores previstos e os valores do DW ao longo dos meses para o número de casos de Pneumonia

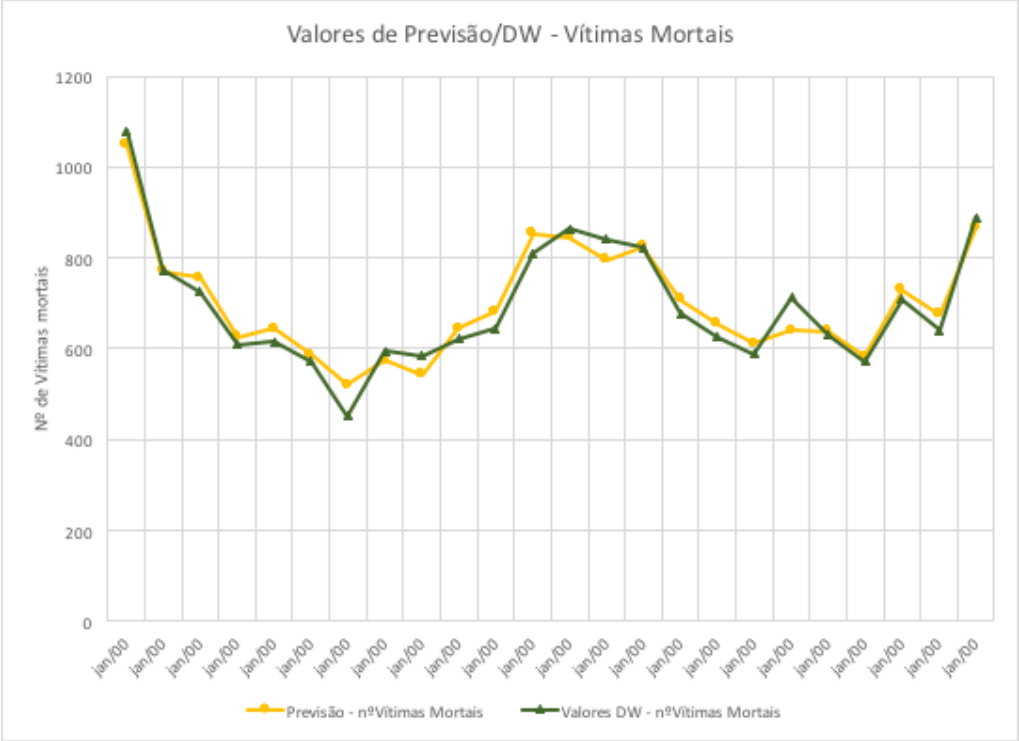


Figura 47 Representação do comportamento dos valores previstos e os valores do DW ao longo dos meses para o número de vítimas mortais

5. Conclusão

O presente trabalho de dissertação visou a aplicação de duas tarefas de *Data Mining*, para o estudo da incidência de Pneumonia em Portugal, sobre os dados armazenados num *Data Warehouse* que guarda a informação sobre pneumonias, referentes aos anos de 2002 a 2011. Em primeiro lugar, foi aplicada a tarefa de *Clustering* através de uma abordagem baseada em densidade, o F-SNN. O objetivo era o de encontrar os melhores modelos que caracterizassem geo-especialmente os casos de pneumonia em Portugal Continental. Em segundo lugar, foi utilizada a tarefa de *Time Series Forecasting* para encontrar os melhores modelos para prever o número casos de Pneumonia e o número vítimas mortais.

Para adquirir o conhecimento necessário, para abordar os temas presentes nesta dissertação, foi necessário explorar os vários temas envolvidos através da realização do enquadramento conceptual. O enquadramento conceptual é constituído por dois temas chave, a Descoberta de Conhecimento em Base de Dados e *Data Mining*. No tema de DCBD, foram estudados as fases e os desafios do processo de DCBD. No Tema de *Data Mining* foram exploradas as tarefas e técnicas necessárias para a concretização da dissertação. A realização do enquadramento conceptual permitiu deste modo adquirir e consolidar conhecimentos.

De modo a utilizar os dados armazenados no *Data Warehouse* foi necessário efetuar a seleção e extração dos dados, para *datasets,* de modo a satisfazer os requisitos para cada objetivo proposto. Este processo resultou em dois datasets, o dataset1 e o dataset2. O dataset1 é constituído por todos os registos referentes às pneumonias que apresentem coordenadas geográficas, para assim ser possível ao F-SNN calcular a distância entre os dois pontos. Deste modo, o dataset1 é constituído por 368 121 registos, que é um elevado número de registos e influenciou o tempo de processamento do F-SNN, devido à complexidade quadrática. O dataset2 é constituído por todos os registos referentes às pneumonias e que apresentem pelo menos uma patologia associada. Na obtenção deste *dataset* foram selecionados os registos que apresentassem pelo menos uma patologia associada à Pneumonia. Assim, o dataset2 sofreu uma redução elevada no número de registos, ficando com 212 788 registos. A redução afetou diretamente os resultados obtidos. No F-SNN, o dataset2 apresentou ligeiramente um menor número de *cluster*s significativos em relação ao dataset1. No espaço geográfico, os *clusters* do dataset2 encontram-se nas zonas identificadas no dataset1, indicando as zonas com maior foco de incidência. No caso das previsões, foi necessário

utilizar um novo *dataset*. Este foi criado de modo a conter todos os registos, ao contrário do dataset1 e dataset2, de modo a conseguir previsões com a maior precisão possível.

A aplicação da técnica de *clustering* espacial foi demorada mesmo com o uso da abordagem F-SNN, que é mais rápida que o SNN original. Foram necessárias várias horas para a execução dos vários modelos para três dimensões e quatro dimensões, devido ao elevado número de registos, como referido acima. De forma a avaliar a qualidade dos modelos identificados no F-SNN foi criada uma Métrica de Qualidade. Esta, através de medidas *intracluster* e do número de clusters significativos pelo total de clusters permite calcular a Métrica de Qualidade criada. Através desta métrica foi possível escolher os melhores modelos. É de acrescentar que os modelos apresentados, são modelos já bastantes otimizados, pelo que a utilização desta métrica permite-nos procurar os modelos mais refinados. Através da utilização de três e quatro dimensões foi possível averiguar que o processo de *clustering* apresenta valores mais baixos no *intracluster* com o uso de quatro dimensões. Os modelos com parametrização de K, MinPts e Eps diferentes, mas com pesos iguais para as dimensões referentes aos modelos de duas e três dimensões, obtiveram o mesmo número de clusters. Tal pareceu ser estranho, e de forma a verificar se este comportamento estava relacionado com o elevado número de registos e similaridade dos mesmos, e não algum erro de programação da função distância, foram executadas para o dataset2, que contêm menos registos, um conjunto de modelos com os respetivos K, MinPts e Eps recalculados. Neste caso, o número de clusters entre três dimensões e quatro dimensões foram diferentes.

Os modelos foram representados no espaço de forma a visualizar a incidência e características dos *clusters*. Os resultados obtidos neste objetivo foram interessantes uma vez que vários modelos conseguiram definir regiões de maior incidência. Foi possível concluir que os *clusters* significativos encontrados, estão localizados nos centros urbanos, principalmente na zona Litoral. Os distritos que apresentam uma maior ocupação pelos *clusters* e com maiores percentagens de vítimas mortais foram Braga, Porto, Viseu, Coimbra, Setúbal e Lisboa. Além das regiões afetadas nos clusters encontrados, os meses que se destacam apresentando o maior número de registos de casos de pneumonia são Janeiro, Fevereiro e Março.

Os indivíduos que apresentam a maior percentagem de mortalidade são aqueles com mais de 70 anos de idade, tal como já havia sido verificado em (Leite, 2014). Pelo contrário, os indivíduos que apresentam uma menor taxa de mortalidade são os que se encontram na faixa dos 0 aos 5

anos de idade. Em termos de vítimas mortais, também foi possível verificar que globalmente os indivíduos do sexo masculino são os mais afetados.

De forma a averiguar quais os melhores modelos para a previsão do número de casos de pneumonia e número de vítimas mortais, foi necessário definir um conjunto de algoritmos a utilizar na identificação dos modelos de forma a comparar resultados e descobrir quais os melhores modelos. Dos algoritmos selecionados na criação dos modelos foi percetível que o M5P, no geral, apresenta o menor erro em termos de previsão de casos de Pneumonia e vítimas mortais.

No caso da previsão de casos de Pneumonias e de vítimas mortais em Portugal, os resultados obtidos nos melhores modelos podem ser considerados bastante precisos, apresentado uma média de erros de 2,4% e 4,4%, respetivamente. Referente a Braga e Porto, os resultados obtidos para a previsão de casos de Pneumonia e de vítimas mortais em Portugal, foram também bastante satisfatórios na medida que as percentagens de erro se mantiveram abaixo dos 10%. Foram obtidas as seguintes percentagens, 3,3% para a previsão de casos e 8,9% para a previsão de vítimas mortais.

Os resultados obtidos para Lisboa foram de um erro de 3,5% na previsão de casos de Pneumonia e de 8,2% de erro na previsão de vítimas mortais. Tal como para Braga e Porto a percentagem de erro na previsão de vítimas mortais fica abaixo dos 10%. No geral, os resultados obtidos são satisfatórios.

Os resultados obtidos nos modelos de previsão a 12 e 14 meses comparados com a realidade do DW apresentaram ter uma elevada precisão, pelo que, os valores previstos foram sempre de encontro com os reais, com uma baixa percentagem de erro. De futuro seria interessante fazer o refrescamento do DW e aplicar os mesmos modelos para o ano de 2011 e 2012 de forma a analisar os resultados e o comportamento, para uma validação mais exaustiva dos modelos.

A minha proposta como trabalho futuro visa a parte tecnológica da dissertação, o F-SNN e o *Model Quality*. A proposta seria a integração da abordagem de avaliação de *clustering*, o *Model Quality*, no F-SNN. De forma a criar uma ferramenta única que permita ao utilizador correr os modelos e obter também a sua avaliação de imediato. Esta integração levaria a reduzir o tempo de processamento, pois as distâncias calculadas pelo F-SNN poderiam ser utilizadas pelo *Model Quality*. Por outro lado, tornaria a utilização dos dois mais intuitiva e amiga do utilizador.

No decorrer da dissertação existiram várias dificuldades. Desde dificuldades técnicas e pessoais que limitaram o tempo de trabalho na dissertação e que desviaram um pouco o foco no trabalho, contudo os objetivos foram alcançados.

Referências

Antunes, A., Moreira, A., & Santos, M. Y. (2014). A Fast SNN-based clustering approach for large geospatial data sets. *Connecting a Digital Europe Through Location and Place*, *Lecture No*, pp 179–195. Retrieved from http://ubicomp.algoritmi.uminho.pt/projects/f-snn/

Araújo, A. (2014). Diga não às Infeções Respiratórias Vacine-se! Proteja-se! Retrieved November 10, 2014, from http://www.fundacaoportuguesadopulmao.org/DIGA_NAO_AS_INFECOES_RESPIRATORIAS.html

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining*, 182–185. Retrieved from http://recipp.ipp.pt/handle/10400.22/136

Berkhin, P. (2006). Survey of Clustering Data Mining Techniques. *Grouping Multidimensional Data*, 25–71. Retrieved from http://link.springer.com/chapter/10.1007/3-540-28349-8_2

Chapman, P., Clinton, J., & Kerber, R. (2000). CRISP-DM 1.0. *CRISP-DM Consortium*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Crisp-dm+1.0#1

Choi, K., & Thacker, S. B. (1981). An evaluation of influenza mortality surveillance, 1962-1979. I. Time series forecasts of expected pneumonia and influenza deaths. *American Journal of Epidemiology*, *113*(3), 215–26. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6258426

DGS. (2014). Doenças Respiratórias em números – 2014, 101.

Dinis, R. M. G. (2011). *Padrões de Incidência Geográfica das Doenças Pulmonares Obstrutivas Crónicas e da Asma em Portugal*. Dissertação de Mestrado, Universidade do Minho.

Duarte, F. J. F. (2008). *Optimização da Combinação de Agrupamentos baseado na Acumulação de Provas pesadas por Índices de Validação e com uso de Amostragem*. Universidade de Trás-os-Montes e Alto Douro.

Ertöz, L., Steinbach, M., & Kumar, V. (2003). Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In *Proceedings of Second SIAM International Conference on Data Mining* (pp. 47–59). http://doi.org/doi:10.1137/1.9781611972733.5

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *KDD*. Retrieved from http://www.aaai.org/Papers/KDD/1996/KDD96-014

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. http://doi.org/10.1145/240455.240464

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996c, March 15). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*. Retrieved from http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992, September 15). Knowledge Discovery in Databases: An Overview. *AI Magazine*. Retrieved from http://www.aaai.org/ojs/index.php/aimagazine/article/view/1011

Fu, Y. (1997). Data Mining: Tasks, Techniques, and Applications. *IEEE Potentials*, *16*, 18–20. Retrieved from http://academic.csuohio.edu/fuy/Pub/pot97.pdf

Galvão, J. R. M. V. da C. (2014). *Segmentação de vastos volumes de dados com SNNagg*. Dissertação de Mestrado, Universidade do Minho, Guimarães.

Garner, S. R. (1995). WEKA: The Waikato Environment for Knowledge Analysis. Retrieved October 23, 2015, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.54.3371&rep=rep1&type=pdf

Guo, D., & Gahegan, M. (2006). Spatial ordering and encoding for geographic data mining and visualization. *Journal of Intelligent Information Systems*, *27*(3), 243–266. http://doi.org/10.1007/s10844-006-9952-8

Han, J., Kamber, M., & Pei, J. (2006). *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Retrieved from https://www.google.pt/books?hl=pt-PT&lr=&id=AfL0t-YzOrEC&pgis=1

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier. Retrieved from https://books.google.com/books?id=pQws07tdpjoC&pgis=1

Hodzic, N., Konjic, T., & Miranda, V. (2006). Artificial Neural Networks Applied To Short Term Load Diagram Prediction. In *2006 8th Seminar on Neural Network Applications in Electrical Engineering* (pp. 219–223). IEEE. http://doi.org/10.1109/NEUREL.2006.341217

Hyndman, R. J., & Athanasopoulus, G. (2013). Forecasting: principles and practice. Retrieved October 19, 2015, from https://play.google.com/store/books/details?id=gDuRBAAAQBAJ&source=ge-web-app

Jain, a. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323. http://doi.org/10.1145/331499.331504

Jarvis, R. a, & Patrick, E. a. (1973). Clustering Using a Similarity Measure Based on Shared Near Neighbors. *Ieee Transactions on Computers*, *C-22*(11), 1025–1034. http://doi.org/10.1109/T-C.1973.223640

Jonge, E. de, Pelt, M. van, & Roos, M. (2012). Time patterns , geospatial clustering and mobility statistics based on mobile phone network data. *Statistics Netherlands*, 23. Retrieved from http://www.cbs.nl/NR/rdonlyres/010F11EC-AF2F-4138-8201-2583D461D2B6/0/201214x10pub.pdf

Kalogirou, S. A. (2001). Artificial neural networks in renewable energy systems applications: a review. *Renewable and Sustainable Energy Reviews*, *5*(4), 373–401. http://doi.org/10.1016/S1364-0321(01)00006-5

Kotsiantis, S. B., & Pintelas, P. E. (2004). Recent Advances in Clustering : A Brief Survey. *Methods*, *1*, 73–81. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.1130&rep=rep1&type=pdf

Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine*, *14*(8), 799–810. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7644860

Leite, V. (2014). *Business Intelligence no estudo das Pneumonias e da sua incidência em Portugal*. Dissertação de Mestrado, Universidade do Minho.

Lek, S., & Guégan, J. F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, *120*(2-3), 65–73. http://doi.org/10.1016/S0304-3800(99)00092-7

Levent Ertöz, Michael Steinbach, V. K. (2002). A New Shared Nearest Neighbor Clustering Algorithm and its Applications. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.123.4729

Martin Ester, Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*. Retrieved from

http://www.dbs.ifi.lmu.de/Publikationen/Papers/KDD-96.final.frame.pdf

Miller, H. J., & Han, J. (2009). *Geographic Data Mining and Knowledge Discovery*. (U. ©. Taylor & Francis, Inc. Bristol, PA, Ed.) (Second Edi). CRC Press.

Moreira, A., Santos, M. Y., & Carneiro, S. (2005). Density-based clustering algorithms–DBSCAN and SNN, 1–18. Retrieved from http://denis.kraynov.2009.homepage.auditory.ru/2006/Ivan.Ignatyev/AD/snn&dbscan.pdf

Moreira, J. (2013). *Input Parameters Self-tuning on the SNN Algorithm (Shared Nearest Neighboor)*. Dissertação de Mestrado, Universidade do Minnho.

Nath, S. (2006). Crime pattern detection using data mining. *Web Intelligence and Intelligent Agent Technology ...*, *1*(954). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4053200

Oliveira, R., Santos, M. Y., & Pires, J. M. (2013). 4D+SNN: A Spatio-Temporal Density-Based Clustering Approach with 4D Similarity. *2013 IEEE 13th International Conference on Data Mining Workshops*, 1045–1052. http://doi.org/10.1109/ICDMW.2013.119

Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved February 10, 2015, from http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Pujari, A. K. (2001). *Data Mining Techniques*. Universities Press. Retrieved from https://books.google.com/books?id=dH2KQhJboSYC&pgis=1

Rokach, L., & Maimon, O. (2010). Chapter 15— Clustering methods. *The Data Mining and Knowledge Discovery Handbook*, 32. http://doi.org/10.1007/0-387-25465-X_15

Santos, M. Y. (2001). *Padrão: um sistema de descoberta de conhecimento em bases de dados geo-referenciadas*. Universidade do Minho. Retrieved from http://repositorium.sdum.uminho.pt/handle/1822/202

Santos, M. Y., Moreira, A., & Carneiro, S. (2005). Clustering in the Identification of Space Models. In J. Wang (Ed.), *Encyclopedia of Data Warehousing and Mining* (pp. 165–172). Idea Group Reference.

Torrence, C., & Compo, G. P. (1998). A practical guide to wavelet analysis. *Bams*, *79*, 61. http://doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2

Usama Fayyad, Gregory Piatetsky-shapiro, P. S. (1996). From data mining to knowledge discovery in databases. *AL MAGAZINE*, 54. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.116.3382&rank=2

Vinnakota, S., & Lam, N. S. N. (2006). Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach. *International Journal of Health Geographics*, *5*, 9. http://doi.org/10.1186/1476-072X-5-9

Wang, W., Yang, J., & Muntz, R. (1997). STING: A statistical information grid approach to spatial data mining. *Proceedings of International Conference on Very Large Data*, 1–18. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.3370&rep=rep1&type=pdf

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques*. (Elsevier, Ed.) (Third Edit). Morgan Kaufmann.

Anexos

A. Código do Script em R

```
#############################################################################
#                          2 Dimensões                                      #
#############################################################################
An2D <- function(fileName, pathName, xPoints,Sd){
#defenir o  path
fileName= fileName
#path= "/Users/Rui/Dropbox/TESE/SNN/dados/Runs-idade-meses/Runs/K-val 2d/"
path= pathName

fullPath = paste0(path,fileName)
fullPath

#load CSV
ds=read.csv2(fullPath,sep = ";",dec = ".", header = FALSE)
colnames(ds) <- c("idade","latitude","longitude","mes","sexo","reingressos","dias_internamento","flag_vitima","distrito","concelho"
,"hospital","cluster","tipo_cluster")
max(ds$cluster)
#calculo do numero de pontos por cada cluster e seleccao dos que apresentam um valor igual ou superior ao defenido. neste caso
maior que 1500

  df= data.frame()
  points=xPoints
  points
  for (i in 1:max(ds$cluster)) {
    cs=ds[ds[12]==i,]
    fv= cs$flag_vitima
    bcs=cs$idade
    bcs=ordered(bcs)
    idade=cs$idade

    if (length(bcs)>points) {
      cat("Cluster = ",i, "::Numero de Pontos = ", length(bcs), "::Media =",signif(mean(idade), digits = 6), "::Desvio-
padrao =", signif(sd(idade),digits = 6), "::Minimmo =", min(idade), "::Maximo =", max(idade), ":: FlagVitima =",sum(fv), "\n")
      df= rbind(df, data.frame("cluster"=i, "NumeroDePontos" = length(bcs), "Media"= round(mean(idade), digits = 6), "DesvioPadrao
"=round(sd(idade),digits = 6),"Minimmo"= min(idade), "Maximo"= max(idade), "FlagVitima"=sum(fv)))
    }
  }
  # criaÃ§Ã£o de csv com os clusters selecionados e referentes dados estatisticos
  write.table(df,file = paste0(pathName,paste0(xPoints,paste0("_pontos_",fileName))), sep = ";", dec = ".",fileEncoding = "utf8", row
.names=FALSE)


#########################################
#########################################

#filterByDP2D<- function( fileName,patchName, Sd){

  #Filtragem dos clusters segundo o valor do desvio padrao. neste caso dp <= 10

df=df
df2 = data.frame()
numrows= nrow(df)
numrows

saveP = paste0(paste0(xPoints,"_pontos_"),fileName)
saveP = paste0(paste0(Sd,"_DP_"), saveP)
for (i in 1:numrows) {
  dp= df$DesvioPadrao[i]
  min = df$Minimmo[i]
```

```
  max= df$Maximo[i]
  media = df$Media[i]
  nP= df$NumeroDePontos[i]
  fv = df$FlagVitima[i]
  percentVitimas= signif(((fv/nP)*100), digits = 4)


  if (dp <= Sd ) {

    cat("cluster :", df$cluster[i],"Media : ", media, "desvio-
padrao :", dp, "min : ", min, "max : " ,max, "Np :", nP,"FVitima :", fv,"PercentVitimas",percentVitimas , "\n")
    df2= rbind(df2, data.frame("cluster"=df$cluster[i],"Media"= media, "desvio-
padrao"= dp, "min "= min, "max"= max, "Np"= nP,"FVitima"= fv, "PercentVitimas" = percentVitimas))
  }

  write.table(df2,file = paste0(pathName,saveP), sep = ";", dec = ".",fileEncoding = "utf8", row.names=FALSE)
}


df2=df2
df3 = ds[ds$cluster %in% df2$cluster, ]
df3
max(df3$cluster)
cluster = df3$cluster
cluster
library(scatterplot3d)
library(Rcmdr)
library(rgl)
options(rgl.useNULL=FALSE)
open3d()
par3d(windowRect = c(10, 10, 1200, 750))
grafico=plot3d(df3$longitude, df3$latitude, df3$idade, size= 2, xlab="Lon",ylab="Lat",zlab="Idade",axes=TRUE,col=df3$cluster, ma
in = "", pch = 20)
legend3d("topright", legend = paste('Cluster', df2$cluster), col=df2$cluster, pch = 20, cex=1.5, inset=c(0.02))

write.table(df3,file = paste0(pathName,paste0("Full_data_",saveP)), sep = ";", dec = ".",fileEncoding = "utf8", row.names=FALSE)



library(maptools)
library(zoom)

##load csv com os pontos
df4=df3
##load shapefile
mapa<- readShapePoly("/Users/Rui/Dropbox/TESE/Shapes_PT/mapa.shp" ,proj4string=CRS("+proj=longlat +datum=WGS84 +no
_defs"))
##desenhar mapa e pontos
plot(mapa , axes=FALSE)
towns <- SpatialPoints(df4[,c("longitude","latitude")],proj4string=CRS("+proj=longlat +datum=WGS84 +no_defs"))
plot(towns, add = TRUE, pch=20, col = df4$cluster, cex=0.5, size= 1, labels=df4$cluster)
legend("topright", legend = paste0("Cluster ",unique(df4$cluster)), col=c(unique(df4$cluster)), pch = 20, cex=1.5, inset=c(0.02)) }

#############################################################################
#                          3 Dimensõµes                                     #
#############################################################################
An3D <- function(fileName, pathName, xPoints,Sd){
  #defenir o  path
  fileName= fileName
  #path= "/Users/Rui/Dropbox/TESE/SNN/dados/Runs-idade-meses/Runs/K-val 2d/"
  path= pathName

  fullPath = paste0(path,fileName)
  fullPath
```

```
#load CSV
ds=read.csv2(fullPath,sep = ";",dec = ".", header = FALSE)
colnames(ds) <- c("idade","latitude","longitude","mes","sexo","reingressos","dias_internamento","flag_vitima","distrito","concelho","hospital","cluster","tipo_cluster")
max(ds$cluster)
#calculo do numero de pontos por cada cluster e seleção dos que apresentam um valor igual ou superior ao defenido. neste caso maior que 1500

df= data.frame()
points=xPoints
points
for (i in 1:max(ds$cluster)) {
  cs=ds[ds[12]==i,]
  fv= cs$flag_vitima
  bcs=cs$idade
  bcs=ordered(bcs)
  idade=cs$idade
  mes = cs$mes
  if (length(bcs)>points) {
    cat("Cluster = ",i, "::Numero de Pontos = ", length(bcs), "::MediaIdade =",signif(mean(idade), digits = 6), "::Desvio-padraoIdade =", signif(sd(idade),digits = 6), "::MinimoIdade =", min(idade), "::MaximoIdade =", max(idade),
        "::MediaMes =",signif(mean(mes), digits = 6), "::Desvio-padraoMes =", signif(sd(mes),digits = 6), "::MinimoMes =", min(mes), "::MaximoMes =", max(mes),
        ":: FlagVitima =",sum(fv),"\n")
    df= rbind(df, data.frame("cluster"=i, "NumeroDePontos" = length(bcs), "MediaIdade"= round(mean(idade), digits = 6), "DesvioPadraoIdade"=round(sd(idade),digits = 6),"MinimoIdade"= min(idade), "MaximoIdade"= max(idade)
                             , "MediaMes"= round(mean(mes), digits = 6), "DesvioPadraoMes"=round(sd(mes),digits = 6),"MinimoMes"= min(mes), "MaximoMes"= max(mes), "FlagVitima"=sum(fv)   ))
  }
}
# criação de csv com os clusters selecionados e referentes dados estatisticos
write.table(df,file = paste0(pathName,paste0(xPoints,paste0("_pontos_",fileName))), sep = ";", dec = ".",fileEncoding = "utf8", row.names=FALSE)



########################################
########################################

#filterByDP2D<- function( fileName,patchName, Sd){

#Filtragem dos clusters segundo o valor do desvio padrao. neste caso dp <= 10

#df= read.csv(paste0(patchName,fileName), header = TRUE, dec = ".", sep = ";")
df=df
df2 = data.frame()
numrows= nrow(df)
numrows

saveP = paste0(paste0(xPoints,"_pontos_"),fileName)
saveP = paste0(paste0(Sd,"_DP_"), saveP)
for (i in 1:numrows) {
  dpIdade= df$DesvioPadraoIdade[i]
  minIdade = df$MinimoIdade[i]
  maxIdade= df$MaximoIdade[i]
  mediaIdade = df$MediaIdade[i]

  dpMes = df$DesvioPadraoMes[i]
  minMes = df$MinimoMes[i]
  maxMes= df$MaximoMes[i]
  mediaMes = df$MediaMes[i]

  nP= df$NumeroDePontos[i]
```

```
    fv = df$FlagVitima[i]
    percentVitimas= round(((fv/nP)*100), digits = 4)


    if (dpIdade <= Sd ) {

      cat("cluster :", df$cluster[i],"MediaIdade : ", mediaIdade, "desvio-
padraoIdade :", dpIdade, "minIdade : ", minIdade, "maxIdade : " ,maxIdade,
          "MediaIdade : ", mediaMes, "desvio-
padraoMes :", dpMes, "minMes : ", minMes, "maxMes : " ,maxMes, "Np :", nP,"FVitimaIdade :", fv,"PercentVitimas",percentVitima
s, "\n")

      df2= rbind(df2, data.frame("cluster"=df$cluster[i],"MediaIdade"= mediaIdade, "desvio-
padraoIdade"= dpIdade, "minIdade"= minIdade, "maxIdade"= maxIdade,
                                 "MediaMes"= mediaMes, "desvio-
padraoMes"= dpMes, "minMes"= minMes, "maxMes"= maxMes,"Np"= nP,"FVitima"= fv, "PercentVitimas" = percentVitimas))
    }

    write.table(df2,file = paste0(pathName,saveP), sep = ";", dec = ".",fileEncoding = "utf8", row.names=FALSE)
  }


  df2=df2
  df3 = ds[ds$cluster %in% df2$cluster, ]
  df3
  max(df3$cluster)
  cluster = df3$cluster
  cluster
  library(scatterplot3d)
  library(Rcmdr)
  library(rgl)
  options(rgl.useNULL=FALSE)
  open3d()
  par3d(windowRect = c(10, 10, 1200, 750))
  grafico=plot3d(df3$longitude, df3$latitude, df3$idade, size= 2, xlab="Lon",ylab="Lat",zlab="Idade",axes=TRUE,col=df3$cluster, m
ain = "", pch = 20)
  legend3d("topright", legend = paste('Cluster', df2$cluster), col=df2$cluster, pch = 20, cex=1.5, inset=c(0.02))
  open3d()
  par3d(windowRect = c(10, 10, 1200, 750))
  grafico2 = plot3d(df3$longitude, df3$latitude, df3$mes, size= 1, xlab="Lon",ylab="Lat",zlab="MÃªs",axes=TRUE,col=df3$cluster,
main = "", pch = 20)
  legend3d("topright", legend = paste('Cluster', df2$cluster), col=df2$cluster, pch = 20, cex=1.5, inset=c(0.02))
  write.table(df3,file = paste0(pathName,paste0("Full_data_",saveP)), sep = ";", dec = ".",fileEncoding = "utf8", row.names=FALSE)




  library(maptools)
  library(zoom)

  ##load csv com os pontos
  df4=df3
  ##load shapefile
  mapa<- readShapePoly("/Users/Rui/Dropbox/TESE/Shapes_PT/mapa.shp" ,proj4string=CRS("+proj=longlat +datum=WGS84 +n
o_defs"))
  ##desenhar mapa e pontos
  plot(mapa , axes=FALSE)
  towns <- SpatialPoints(df4[,c("longitude","latitude")],proj4string=CRS("+proj=longlat +datum=WGS84 +no_defs"))
  plot(towns, add = TRUE, pch=20, col = df4$cluster, cex=0.5, size= 1, labels=df4$cluster)
  legend("topright", legend = paste0("Cluster ",unique(df4$cluster)), col=c(unique(df4$cluster)), pch = 20, cex=0.9, inset=c(0.02))

}

#######################################
```

```
#######################################
draw3dIdade<- function(filePath,filePath2){
# desenhar em 3d os clusters selecionados anteriormente e criação de um csv com apenas esses clusters para ser usado no tablea
u
#fullPath = paste0(pathName,fileName2)

ds=read.csv2(filePath,sep = ";",dec = ".", header = TRUE)
ds2= read.csv2(filePath2,sep = ";", dec = ".",header = TRUE)

df3 = ds2[ds2$cluster %in% ds$cluster, ]

max(df3$cluster)
cluster = df3$cluster
cluster
library(scatterplot3d)
library(Rcmdr)
library(rgl)
options(rgl.useNULL=FALSE)
open3d()
par3d(windowRect = c(10, 10, 1200, 750))
grafico=plot3d(df3$longitude, df3$latitude,df3$idade , size= 2, xlab="Lon",ylab="Lat",zlab="Idade",axes=TRUE,col=df3$cluster, ma
in = "", pch = 20)
legend3d("topright", legend = paste0("Cluster ",unique(df3$cluster)), col=c(unique(df3$cluster)), pch = 20, cex=1.5, inset=c(0.02))

}
#draw3dIdade("/Users/Rui/Dropbox/TESE/SNN/dados/Runs-idade-meses/Runs/80-75-13 3d/10_DP_1000_pontos_dados-
Clusters-80-80-13-75-8000.20.00.40.4.csv")

draw3dMes<- function(filePath, filePath2){
  # desenhar em 3d os clusters selecionados anteriormente e criação de um csv com apenas esses clusters para ser usado no table
au
  #fullPath = paste0(pathName,fileName2)
  ds=read.csv2(filePath,sep = ";",dec = ".", header = TRUE)
  ds2= read.csv2(filePath2,sep = ";", dec = ".",header = TRUE)

  df3 = ds2[ds2$cluster %in% ds$cluster, ]

  max(df3$cluster)
  cluster = df3$cluster
  cluster
  library(scatterplot3d)
  library(Rcmdr)
  library(rgl)
  options(rgl.useNULL=FALSE)
  open3d()
  par3d(windowRect = c(10, 10, 1200, 750))
  grafico=plot3d(df3$longitude, df3$latitude,df3$mes, size= 2, xlab="Lon",ylab="Lat",zlab="MÃªs",axes=TRUE,col=df3$cluster, mai
n = "", pch = 20)
  legend3d("topright", legend = paste0("Cluster ",unique(df3$cluster)), col=c(unique(df3$cluster)), pch = 20, cex=1.5, inset=c(0.02)
)
}
drawMapWithPoints <- function(filePath, filePath2) {

  ds=read.csv2(filePath,sep = ";",dec = ".", header = TRUE)
  ds2= read.csv2(filePath2,sep = ";", dec = ".",header = TRUE)
  #ds=read.csv2("/Users/Rui/Dropbox/TESE/SNN/dados/Runs-idade-meses/Runs/80-75-
13 3d/filtrado_10_DP_1000_pontos_dados-Clusters-80-80-13-75-8000.20.00.40.4.csv",sep = ";",dec = ".", header = TRUE)
  #ds2= read.csv2("/Users/Rui/Dropbox/TESE/SNN/dados/Runs-idade-meses/Runs/80-75-
13 3d/Full_data_10_DP_1000_pontos_dados-Clusters-80-80-13-75-8000.20.00.40.4.csv",sep = ";", dec = ".",header = TRUE)
  df4 = ds2[ds2$cluster %in% ds$cluster, ]
  ##### VisualizaÃ§Ã£o clusters no mapa de Portugal

  library(maptools)
  library(zoom)
```

```
  ##load shapefile
  mapa<- readShapePoly("shp/mapa.shp" ,proj4string=CRS("+proj=longlat +datum=WGS84 +no_defs"))
  ##desenhar mapa e pontos
  plot(mapa , axes=FALSE)
  towns <- SpatialPoints(df4[,c("longitude","latitude")],proj4string=CRS("+proj=longlat +datum=WGS84 +no_defs"))
  plot(towns, add = TRUE, pch=20, col = df4$cluster, cex=0.5, size= 1, labels=df4$cluster)
  legend("topright", legend = paste0("Cluster ",unique(df4$cluster)), col=c(unique(df4$cluster)), pch = 20, cex=1, inset=c(0.02))
zm()
}
```

B. Resultados dos Modelos de Previsão

As seguintes tabelas apresentam todos os resultados de cada mês dos modelos referenciados na secção 4.

Portugal - Vítimas Mortais								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Steps
F1	MAE	104,1	159,7	214,5	257,4	292,2	321,2	224,9
	RMSE	135,5	216,4	270,8	331,2	369,3	401,6	287,5
	MAPE	15,2	22,8	31,0	37,7	42,8	47,3	32,8
F2	MAE	95,7	134,3	154,6	164,4	163,6	157,5	145,0
	RMSE	123,5	175,8	200,3	212,7	213,6	202,5	188,1
	MAPE	14,3	19,7	22,2	23,7	23,6	22,5	21,0
F3	MAE	64,6	74,1	71,7	73,8	77,4	75,2	72,8
	RMSE	88,6	96,8	99,8	102,2	105,5	103,2	99,3
	MAPE	9,8	11,1	10,6	10,9	11,5	11,1	10,8
F4	MAE	61,5	57,3	57,9	60,0	60,2	61,4	59,7
	RMSE	83,5	76,0	77,2	78,6	79,0	79,4	78,9
	MAPE	8,9	8,8	8,9	9,3	9,3	9,5	9,1

Tabela 43 Tabela de Resultados para os Modelos de previsão de Vítimas mortais em Portugal

Portugal - Vítimas Mortais overlayed								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Meses
F5	MAE	29,1	29,0	29,8	29,8	30,3	31,8	29,1
	RMSE	36,1	36,2	36,7	36,9	37,5	39,3	36,1
	MAPE	4,2	4,3	4,4	4,4	4,5	4,7	4,2
F6	MAE	39,8	37,5	37,0	37,0	37,7	38,9	39,8
	RMSE	51,8	48,9	48,6	49,1	49,7	50,6	51,8
	MAPE	5,7	5,4	5,4	5,4	5,5	5,6	5,7
F7	MAE	55,6	60,1	59,0	58,8	55,9	61,2	55,6
	RMSE	68,6	74,2	72,1	71,9	68,4	73,3	68,6
	MAPE	8,1	8,8	8,6	8,6	8,2	9,0	8,1
F8	MAE	48,1	46,5	47,3	47,3	45,9	45,1	48,1
	RMSE	59,2	56,3	57,1	57,4	55,9	55,8	59,2
	MAPE	7,1	7,1	7,3	7,3	7,1	6,9	7,1

Tabela 44 Tabela de Resultados para os modelos de previsão de Vítimas mortais overlayed em Portugal

Portugal - Pneumonias								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Meses
F9	MAE	348,2	494,5	538,1	549,8	564,6	587,6	513,8
	RMSE	470,7	673,8	686,5	698,8	713,8	737,4	663,5
	MAPE	9,8	14,1	15,6	16,1	16,5	17,2	14,9
F10	MAE	393,5	634,1	679,3	683,6	747,1	815,1	658,8
	RMSE	565,2	814,2	861,2	863,0	944,4	1027,8	846,0
	MAPE	11,6	19,1	20,8	20,8	22,6	24,7	20,0
F11	MAE	377,5	485,9	490,3	466,1	466,2	463,2	458,2
	RMSE	484,7	599,1	607,3	602,3	621,8	624,6	590,0
	MAPE	11,3	14,9	15,1	14,2	14,1	14,1	13,9
F12	MAE	326,7	317,0	324,2	321,1	319,2	307,1	319,2
	RMSE	443,6	426,3	432,6	432,5	432,7	426,5	432,4
	MAPE	8,8	8,9	9,1	9,1	9,0	8,7	8,9

Tabela 45 Tabela de Resultados para os modelos de previsão de Pneumonias em Portugal

Portugal - Pneumonias overlayed								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Meses
F13	MAE	111,8	108,1	111,1	112,9	116,3	117,2	112,9
	RMSE	149,0	145,7	147,9	149,9	152,2	153,8	149,7
	MAPE	3,0	2,9	3,0	3,0	3,1	3,2	3,0
F14	MAE	279,3	291,9	290,3	304,3	312,1	310,3	298,0
	RMSE	327,3	340,8	336,0	349,2	355,1	353,4	343,6
	MAPE	8,8	9,2	9,2	9,6	9,9	9,9	9,4
F15	MAE	83,4	83,6	83,8	86,0	87,4	87,0	85,2
	RMSE	111,3	112,5	112,6	114,4	116,0	116,4	113,9
	MAPE	2,3	2,4	2,4	2,5	2,5	2,5	2,4
F16	MAE	226,2	210,9	211,2	212,1	208,4	205,3	212,4
	RMSE	310,5	284,2	286,0	288,8	286,6	281,8	289,7
	MAPE	6,2	6,0	6,0	6,1	6,0	5,9	6,0

Tabela 46 Tabela de Resultados para os modelos de previsão de Pneumonia em Portugal

Braga e Porto - Vítimas Mortais								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Meses
F17	MAE	16,1	15,4	14,9	14,4	14,2	14,3	37,7
	RMSE	20,2	20,4	19,8	19,3	19,2	19,6	42,4
	MAPE	12,8	12,8	12,5	12,1	11,9	12,0	32,3
F18	MAE	45,4	68,5	83,2	92,6	83,0	70,9	73,9
	RMSE	52,2	78,4	94,2	103,5	93,1	80,0	83,6
	MAPE	38,9	60,1	73,1	81,5	73,3	62,7	64,9
F19	MAE	26,0	35,8	40,3	44,1	39,6	40,4	37,7
	RMSE	31,9	40,2	45,2	48,5	44,0	44,7	42,4
	MAPE	21,4	30,5	34,7	38,0	34,2	34,8	32,3
F20	MAE	13,6	13,4	13,6	13,6	13,2	12,9	13,4
	RMSE	17,8	16,9	17,1	17,2	16,8	16,5	17,1
	MAPE	10,2	10,4	10,6	10,6	10,2	9,9	10,3

Tabela 47 Tabela de Resultados para os Modelos de previsão de Vítimas mortais Braga e Porto

Braga e Porto - Vítimas Mortais overlayed								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Meses
F21	MAE	10,7	10,9	10,9	10,2	10,4	10,3	10,6
	RMSE	13,9	14,1	14,1	13,4	13,6	13,6	13,8
	MAPE	9,0	9,2	9,3	8,6	8,7	8,6	8,9
F22	MAE	30,3	32,6	26,6	28,1	30,2	26,8	29,1
	RMSE	35,0	37,5	31,2	32,5	34,6	31,1	33,6
	MAPE	26,2	28,4	23,2	24,5	26,3	23,4	25,3
F23	MAE	14,2	14,9	14,8	14,8	14,3	14,0	14,5
	RMSE	16,8	17,4	17,4	17,4	17,1	17,1	17,2
	MAPE	11,9	12,6	12,7	12,6	12,2	11,8	12,3
F24	MAE	13,3	13,0	12,8	12,4	12,5	12,9	12,8
	RMSE	15,6	15,3	15,2	14,8	14,9	15,1	15,1
	MAPE	10,6	10,6	10,6	10,2	10,2	10,6	10,5

Tabela 48 Tabela de Resultados para os Modelos de previsão de Vítimas mortais overlayed Braga e Portugal

Braga e Porto - Pneumonias								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Meses
F25	MAE	85,0	89,4	90,8	92,9	93,9	93,9	91,0
	RMSE	110,7	124,3	126,0	128,2	129,2	130,6	124,8
	MAPE	11,4	12,6	12,9	13,2	13,3	13,3	12,8
F26	MAE	311,0	511,3	574,2	491,1	377,2	314,2	429,8
	RMSE	350,5	570,4	635,6	540,7	420,8	360,2	479,7
	MAPE	46,1	77,2	86,9	74,3	57,2	48,0	64,9
F27	MAE	190,1	302,3	383,9	377,0	353,1	315,6	320,3
	RMSE	220,9	343,2	427,8	418,3	393,2	359,2	360,4
	MAPE	27,9	45,6	58,3	57,3	53,8	48,3	48,5
F28	MAE	79,6	75,0	78,8	76,6	74,5	69,3	75,6
	RMSE	99,9	94,7	99,1	97,6	96,7	92,5	96,8
	MAPE	10,3	10,0	10,5	10,3	10,1	9,5	10,1

Tabela 49 Tabela de Resultados para os Modelos de previsão de Pneumonias em Braga e Porto

Braga e Porto - Pneumonias overlayed								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Meses
F29	MAE	26,8	25,8	25,9	26,0	26,4	26,8	26,3
	RMSE	38,0	37,2	37,5	37,8	38,1	38,7	37,9
	MAPE	3,3	3,2	3,2	3,2	3,3	3,3	3,3
F30	MAE	61,0	61,5	63,2	65,1	66,5	68,0	64,2
	RMSE	79,2	79,9	81,1	82,3	83,5	84,8	81,8
	MAPE	9,6	9,7	10,0	10,3	10,6	10,8	10,2
F31	MAE	23,4	24,3	24,2	24,6	25,2	25,3	24,5
	RMSE	28,7	29,9	29,9	30,3	30,9	30,9	30,1
	MAPE	3,3	3,4	3,5	3,5	3,6	3,6	3,5
F32	MAE	46,6	44,4	44,9	45,3	44,6	43,6	44,9
	RMSE	59,6	55,7	56,7	57,3	56,7	55,7	56,9
	MAPE	6,1	6,0	6,1	6,2	6,1	6,0	6,1

Tabela 50 Tabela de Resultados para os Modelos de previsão de Pneumonias overlayed em Portugal

Lisboa - Pneumonias								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Meses
F33	MAE	110,7	151,2	152,1	153,3	150,4	149,1	144,5
	RMSE	157,4	222,2	206,7	211,6	204,3	205,1	201,2
	MAPE	14,7	19,8	20,2	20,6	19,9	19,5	19,1
F34	MAE	103,8	132,0	145,8	160,3	172,9	185,2	150,0
	RMSE	128,3	183,8	184,8	209,2	226,5	228,6	193,6
	MAPE	15,3	19,6	21,5	24,4	27,2	29,1	22,9
F35	MAE	127,3	204,0	243,7	241,8	204,5	186,3	201,3
	RMSE	157,4	234,1	273,9	272,3	237,0	219,4	232,4
	MAPE	17,8	29,4	35,3	35,1	29,0	26,0	28,8
F36	MAE	90,3	92,3	92,1	95,0	96,6	94,4	93,4
	RMSE	126,4	131,6	130,3	135,8	138,7	140,9	133,9
	MAPE	12,0	12,4	12,4	12,8	12,9	12,6	12,5

Tabela 51 Tabela de Resultados para os Modelos de previsão de Pneumonias em Lisboa

Lisboa - Pneumonias overlayed								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Meses
F37	MAE	27,3	25,9	26,5	27,3	27,9	28,0	27,1
	RMSE	41,7	40,4	41,0	41,6	42,2	42,6	41,6
	MAPE	3,5	3,4	3,5	3,6	3,6	3,6	3,5
F38	MAE	43,6	44,7	44,8	44,9	46,1	46,6	45,1
	RMSE	51,8	53,1	52,8	53,0	53,9	54,4	53,2
	MAPE	6,7	7,0	7,0	7,0	7,2	7,3	7,0
F39	MAE	32,7	32,3	32,7	32,3	32,8	34,8	32,9
	RMSE	45,9	46,2	47,0	46,8	47,3	49,0	47,0
	MAPE	4,5	4,5	4,6	4,5	4,5	4,8	4,6
F40	MAE	41,7	41,9	41,1	42,1	42,6	44,7	42,4
	RMSE	63,5	64,2	64,6	65,6	67,2	69,2	65,7
	MAPE	5,8	6,0	5,9	6,0	6,1	6,3	6,0

Tabela 52 Tabela de Resultados para os Modelos de previsão de Pneumonias overlayed em Lisboa

Lisboa - Vítimas Mortais								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Meses
F41	MAE	31,2	36,9	36,8	37,3	39,9	39,0	36,9
	RMSE	37,6	47,6	47,8	50,1	52,1	52,0	47,9
	MAPE	23,0	26,7	25,3	26,1	28,0	27,0	26,0
F42	MAE	26,9	32,3	35,0	36,0	37,1	34,4	33,6
	RMSE	33,9	42,1	42,7	44,6	45,5	44,5	42,2
	MAPE	19,9	23,6	24,5	25,6	26,4	24,0	24,0
F43	MAE	22,0	23,7	23,7	24,0	24,8	22,8	23,5
	RMSE	27,9	29,3	29,6	29,0	29,4	27,6	28,8
	MAPE	16,0	17,1	16,6	16,8	17,2	15,7	16,6
F44	MAE	21,6	21,9	22,8	23,4	23,5	23,8	22,9
	RMSE	26,7	27,9	28,7	29,1	29,3	29,4	28,5
	MAPE	16,5	17,0	17,8	18,2	18,3	18,5	17,7

Tabela 53 Tabela de Resultados para os Modelos de previsão de Vítimas Mortais em Lisboa

Lisboa - Vítimas Mortais overlayed								
Modelo	Métrica	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Média Meses
F45	MAE	12,5	12,3	11,0	11,2	11,4	11,7	11,7
	RMSE	17,2	17,2	16,7	16,9	17,2	17,4	17,1
	MAPE	8,7	8,7	7,8	7,9	8,1	8,3	8,2
F46	MAE	19,2	19,2	18,9	19,4	17,0	18,7	18,7
	RMSE	24,0	24,0	23,9	24,3	22,2	24,0	23,7
	MAPE	13,9	14,0	13,8	14,2	12,4	13,8	13,7
F47	MAE	18,4	17,6	18,1	18,1	16,7	18,4	17,9
	RMSE	26,0	25,3	25,7	25,3	23,7	25,2	25,2
	MAPE	13,1	12,8	13,2	13,2	12,0	13,3	12,9
F48	MAE	14,0	14,3	13,9	13,8	14,0	14,2	14,0
	RMSE	17,3	17,9	17,6	17,6	17,8	18,1	17,7
	MAPE	10,0	10,3	10,1	10,0	10,1	10,2	10,1

Tabela 54 Tabela de Resultados para os Modelos de previsão de Vítimas mortais overlayed em Lisboa