isla
instituto politécnico de gestão e tecnologia

ISLA
Santarém

MESTRADO EM ENGENHARIA DE TECNOLOGIAS E

SISTEMAS WEB

Data Mining do modelo de previsão dos métodos de controlo da diabetes

Tiago Fernandes Godinho

Professor Jorge Duque

DISSERTAÇÃO

VILA NOVA DE GAIA

Abril | 2024

INSTITUTO POLITÉCNICO DE GESTÃO E TECNOLOGIA

Data Mining do modelo de previsão dos métodos de controlo da diabetes

Tiago Fernandes Godinho

Aprovado em 28/06/2024

Composição do Júri

Firmino Silva

Presidente

José Vasconcelos

Arguente

Jorge Duque

Orientador/a

Vila Nova de Gaia

2022

Dissertação de Mestrado submetida para satisfação parcial dos requisitos do grau de Mestre realizada sob a orientação do Prof. Doutor Jorge Manuel Pereira Duque apresentada ao ISLA - Instituto Politécnico de Gestão e Tecnologia de Vila Nova de Gaia para obtenção do grau de Mestre em Engenharia de Tecnologias e Sistemas Web, conforme o Despacho n.º 9371/2020.

Agradecimentos

Ao Professor Jorge Duque, o meu obrigado, pelo apoio, orientação e disponibilidade, no desenvolvimento do trabalho de dissertação.

A todos os docentes que tive o prazer de conhecer no decorrer dos dois anos de mestrado, agradeço pelo conhecimento transmitido, contribuíram para melhorar de alguma forma o meu percurso.

Aos meus amigos especialmente aos meus companheiros da sede, por estarem sempre presentes para me apoiar, motivar a finalizar este grande passo da minha vida.

Ao meu irmão, um obrigado pela motivação, confiança e amizade.

Por último, o agradecimento especial e mais importante aos meus pais, pois sem eles nada disto seria possível, por tentarem sempre dar o seu melhor para o meu futuro e me ajudarem a tornar uma melhor pessoa. Obrigado por toda a motivação, confiança, apoio, paciência, amizade e amor dado durante este longo percurso.

Resumo

A "Era da informação digital" é uma realidade nos tempos atuais, permitindo que a informação esteja acessível a todos, disponível a qualquer hora e em qualquer lugar. Nas organizações, continuam a existir várias dificuldades para tratar o volume de dados recolhidos nas diversas iterações e níveis, bem como nas vertentes de atuação.

O processo de extração de dados é importante para o Data Mining, para a recolha de dados de forma organizada e orientada, para o desenvolvimento de algoritmos.

Neste contexto, a gestão do conhecimento representa para as organizações a possibilidade de ter informação orientada para o desenvolvimento de estratégias e a obtenção de vantagens competitivas. Para as organizações, em particular na área da saúde, é importante a utilização de tecnologias e técnicas para processar os dados de forma rápida e automatizada, auxiliando as organizações no processo de tomada de decisão. Na presente dissertação propõe-se apresentar um modelo de suporte para a análise do registo de utentes com Diabetes Mellitus tipo 2 com recurso ao Data Mining. Esta tecnologia permite tratar volumes de dados de forma organizada, localizando padrões, para realizar conexões, correlações ou anomalias numa grande quantidade de dados, permitindo encontrar problemas, hipóteses e oportunidades de forma coerente, bem como gerar insights vantajosos para a classificação e previsão do registo de utentes com Diabetes Mellitus tipo 2 nos Agrupamentos de Centros de Saúde (ACES).

A metodologia de investigação utilizada, Design Science Research, conduziu à formalização de uma estratégia para o tratamento de dados recolhidos e a gestão do conhecimento com recurso ao Data Mining para a construção de um modelo de suporte à análise da diabetes.

Palavras-chave : Data Mining, Design Science Research

Abstract

The “Digital Information Age” is a reality today, allowing information to be accessible to everyone, available anytime and anywhere. In organizations, several difficulties remain in processing the volume of data collected in different iterations and at different levels and aspects of action.

The data extraction process is important for Data Mining, collecting data in an organized and targeted way, to use in the development of algorithms. In this context, knowledge management represents for organizations the possibility of having information aimed at developing strategies and obtaining competitive advantages.

For organizations, particularly in the healthcare sector, it is important to use technologies and techniques to process data quickly and automatically, assisting organizations in the decision-making process. In this dissertation we propose the presentation of a support model for analyzing the registry of users with type 2 Diabetes Mellitus with support for Data Mining. This technology allows the processing of data volumes in an organized way, to locate patterns, making connections, correlations, or anomalies in a large amount of data, allowing the discovery of problems, hypotheses and opportunities in a coherent way, as well as generating advantageous insights for classifying and forecasting the registration of users with type 2 Diabetes Mellitus in Health Center Groups (ACES).

The research methodology used, Design Science Research (DSR), led to the formalization of a strategy for processing collected data and managing knowledge using Data Mining to build a model to support diabetes analysis.

Keywords : Data Mining, Design Science Research

Índice

Índice de figuras

Índice de tabelas

Lista de abreviaturas

ACES – Agrupamento de Centros de Saúde

BD – Bases de Dados

BI - Business Intelligence

DM – Data Mining

DMe – Diabetes Mellitus

DMT2 – Diabetes Mellitus tipo 2

DS – Data Science

DW – Data Warehouse

DCDB – *Descoberta de Conhecimento em Bases de Dados*

ETL – Extract, Transform and Load

KDD – Knowledge Discovery in Databases

OLAP – Online Analytical Processing

OLTP – Online Transactional Processing

SAD – Sistemas de Apoio á Decisão

SI – Sistemas de Informação

SNS – Serviço Nacional de Saúde

SSAS – SQL Analysis Services

SSIS – SQL Server Integration Service

WWW – World Wide Web

1. Introdução

Este capítulo está organizado em três secções, abrangendo o enquadramento, objetivo e a estrutura do documento. Na primeira secção "Enquadramento" é explicada a relevância desta investigação. Na segunda secção "Objetivo" apresentam-se os objetivos para este estudo. Na terceira secção é delineada a modelagem e a identificação do conteúdo de cada capítulo desta investigação.

1.1 Enquadramento

O seguinte trabalho de investigação integra-se no projeto de dissertação de mestrado em Engenharia de Tecnologias e Sistemas Web, do ISLA - Instituto Politécnico de Gestão e Tecnologia de V. N. de Gaia.

A quantidade de informação digital tem vindo a aumentar e assim, surge a necessidade de retirar conhecimento da informação. Ao longo dos anos foram criadas diversas técnicas e tecnologias para trabalhar com a informação digital, sendo Data Science (DS) e Data Mining (DM) algumas das tecnologias que permitem a análise profunda de dados que contribuem para a descoberta de padrões para serem utilizados pelas organizações. Estas tecnologias são essenciais para a sustentabilidade e o desenvolvimento das organizações. Algumas das áreas de aplicação das tecnologias, a título de exemplo, são: o marketing; a gestão; a agricultura; a segurança pública; e a saúde.

No decorrer da dissertação aprofundam-se os conhecimentos científicos com o objetivo de identificar os problemas registados no Serviço Nacional de Saúde (SNS), com a evolução da Diabetes Mellitus tipo 2. Assim, a utilização de tecnologias de DM, pode exercer um papel fundamental nos estudos relacionados com a área da saúde, devido à organização e disponibilidade da qualidade da informação para o aumento da qualidade de vida e bem-estar das pessoas. A investigação permite analisar e compreender os dados disponibilizados, e entender como as técnicas de DM podem apoiar o processo de tomada de decisão nas organizações.

Este documento está dividido em seis partes, uma de introdução, duas de desenvolvimento teórico, duas de desenvolvimento prático e a conclusão. A introdução tem como objetivo contextualizar o tema e a motivação para o desenvolvimento da

investigação. As duas fases de desenvolvimento teórico desta investigação serviram para adquirir conhecimento científico relevante para o tema sobre: Saúde, Diabetes Mellitus, Data Science e Data Mining, sendo estes os principais objetos de estudo, serviram, de igual modo, estas fases, para aquisição de conhecimento sobre a metodologia utilizada para o desenvolvimento da solução. Nas duas partes do desenvolvimento prático desta investigação foram obtidos conhecimentos sobre a estrutura de dados utilizada para o estudo e aplicados conhecimentos sobre Data Science e Data Mining adquiridos na primeira fase de desenvolvimento. Estes conhecimentos foram obtidos a partir de uma grande quantidade de dados relacionados com a Diabetes Mellitus tipo 2, para a formulação de conclusões relativamente ao objetivo do estudo.

1.2 Objetivo

A Diabetes Mellitus tipo 2 é considerada uma "epidemia" devido à falta de qualidade de vida das pessoas. O sedentarismo e a falta de qualidade na alimentação contribuem para o desenvolvimento da doença. Este trabalho pretende analisar e contribuir com uma visão sobre a previsão do número de doentes afetados pela Diabetes Mellitus tipo 2 nos próximos anos. O objeto de estudo é a utilização de um dataset do Serviço Nacional de Saúde (SNS), com informações atualizadas a partir do ano de 2014, distribuído por regiões e ACES em Portugal Continental.

O objetivo principal deste trabalho de dissertação é analisar o uso de técnicas de Data Mining para prever a evolução da Diabetes Mellitus tipo 2 e dar suporte à tomada de decisão nas organizações. Subsequentemente, consideram-se dois objetivos para dar suporte ao objetivo principal:

- Analisar e avaliar a aplicação eficaz de técnicas de DS;
- Desenvolver e avaliar um modelo de previsão da Diabetes Mellitus tipo 2 por região e ACES, informando sobre o estado da doença em Portugal;

1.3 Organização do Relatório

O documento está subdividido em 6 capítulos. O capítulo 1 considera o enquadramento do tema e os objetivos da investigação. O capítulo 2 considera a revisão da literatura com o desenvolvimento da componente teórica do tema que apresenta os diferentes tópicos

em estudo (Data Science, Data Mining, Diabetes Mellitus, etc.), com recurso a documentação científica. No capítulo 3 apresentam-se os métodos e as técnicas utilizadas para o desenvolvimento da parte prática da dissertação, Design Science Research (DSR). O capítulo 4 descreve a primeira fase da parte prática da dissertação. Nesta primeira parte é obtido conhecimento relevante sobre o problema em estudo, procede-se à seleção de um conjunto de dados adequados ao estudo, é classificado o problema e são selecionados e agrupados todos os dados. No capítulo 5 é desenvolvida e explicada a segunda parte prática da dissertação. Nesta segunda parte procede-se à preparação dos dados, realiza-se o processo de ETL (Extract, Transform and Load), constrói-se o cubo OLAP (Online Analytical Processing), desenvolvem-se as dashboards a partir do Power BI e são aplicadas técnicas de DM. No capítulo 6 são apresentadas as considerações finais do projeto e limitações no desenvolvimento.

2. Estado da Arte

Este capítulo compreende uma revisão bibliográfica indispensável para o desenvolvimento eficaz deste projeto. Este, subdivide-se em secções relevantes para a abordagem ao tema, como: a estratégia de pesquisa; Big Data; Diabetes Mellitus tipo 2; Descoberta de conhecimento em Bases de Dados; Sistemas de Apoio á decisão; Data Science e Data Mining.

2.1 Estratégia de investigação

A estratégia utilizada na investigação para a dissertação teve em consideração plataformas de indexação de documentos, destacando-se: Google Scholar, Web of Science, Research Gate, Science Direct, Scopus e Springer, para complementar alguns conceitos e termos relativamente a Diabetes Mellitus, BigData, Data Mining e Data Science, Descoberta de Conhecimento em Bases de Dados e Sistemas de Apoio à decisão. Também, foram utilizados o Google e Cambridge Dictionary para complementar e esclarecer alguns conceitos.

A revisão de literatura está limitada a documentos redigidos em português e em inglês e toda a literatura científica está disponível para qualquer utilizador.

2.2 Descoberta de conhecimento em bases de dados

O crescente aumento de dados tem colocado algumas dificuldades aos utilizadores para os converter em informação de qualidade, bem como extrair conhecimento para suporte à tomada de decisão. No âmbito de uma convenção de 1989, surge um novo conceito (Piatetsky-Shapiro & Frawley, 1991), descoberta de conhecimento em bases de dados (DCBD) ou KDD (Knowledge Discovery in Data) (Fayyad et al., 1996). O conceito delineado por Frawley et al. (1992), e também explorado por Zhong et al. (1997), descreve o processo de extração de conhecimento valioso a partir de bases de dados. Com a utilização de técnicas e algoritmos de Data Mining, torna-se viável a identificação de padrões e relações entre os dados. DCBD possui as fases: seleção; pré-processamento dos dados; transformação; Data Mining; e avaliação/interpretação. Estas consideram-se as cinco fases principais do DCBD, com o propósito de extrair insights elementares para dar suporte em processos decisórios (Frawley et al., 1992) (Zhong et al., 1997). Também, é

caracterizado por ser um processo iterativo, permitindo retroceder para etapas anteriores como demonstrado na figura 1, sendo interativo e necessária a participação do utilizador para o processo de tomada de decisão (Frawley et al., 1992) (Zhong et al., 1997).

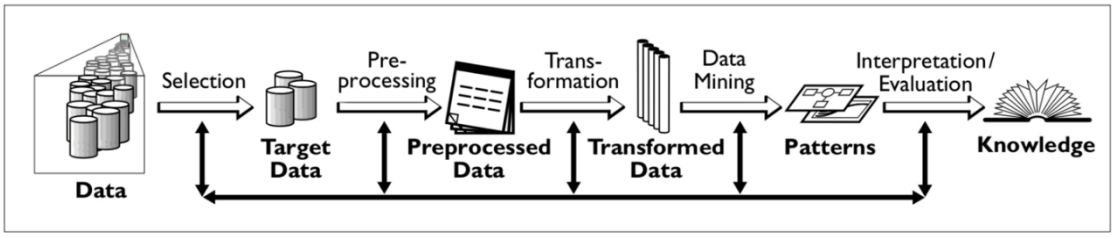

Figura 1 - Processo de DCBD (retirado de (Fayyad et al., 1996))

Os autores Alnoukari e El Sheikh (2012), apresentam as cinco principais etapas do DCDB da seguinte forma (Alnoukari, & El Sheik, 2012):

- **Seleção** - Consiste no desenvolvimento do domínio de aplicação e na criação de um conjunto de dados-alvo a partir de uma base de dados maior. Nesta fase, o objetivo é identificar os dados relevantes para o processo de descoberta de conhecimento.
- **Pré-processamento** - etapa em que é necessário lidar com dados incoerentes, para assegurar que os dados estão preparados para serem analisados, removendo inconsistências e preparando-os para a etapa seguinte.
- **Transformação** - Nesta etapa final do processamento de dados antes da aplicação das técnicas de análise procura-se encontrar atributos úteis e representações invariantes dos dados. É realizado, através da aplicação de métodos de redução de dimensão e transformação dos dados.
- **Data Mining** - O cerne do processo, composto por três etapas. Na primeira etapa, procede-se à escolha da tarefa de prospeção de dados mais adequada aos objetivos definidos. Segue-se a escolha dos algoritmos de Data Mining e definem-se os parâmetros necessários para a procura de padrões nos dados. Por último, aplica-se os algoritmos selecionados para gerar padrões e modelos nos dados.
- **Avaliação/Interpretação** – A etapa que conclui o DCBD. É composta por duas etapas. No primeiro momento, interpreta-se os padrões encontrados, podendo retroceder a qualquer uma das etapas anteriores para uma nova iteração. Esta etapa, também, pode incluir a visualização dos padrões e modelos extraídos, ou dos dados extraídos dos modelos. Posteriormente, consolida-se o conhecimento descoberto, incorporando-o no sistema de desempenho ou documentando-o para uso futuro.

2.3 Sistemas de apoio à decisão

A decisão é o processo de escolha de diferentes alternativas que as organizações enfrentam diariamente. O conceito de Sistemas de Apoio à Decisão (SAD), deriva do termo inglês Decision Support Systems (DSS), apresentado por Michael Scott Morton no início da década de 70, definindo como um sistema computacional que suporta a tomada de decisão através de dados e modelos para resolver problemas não estruturados. No entanto, desde a fase inicial, este novo conceito despertou interesse e foram surgindo diferentes definições do termo (Sprague,1980; Vercellis,2009).

A nível organizacional foram criados modelos de suporte para a tomada de decisão. O primeiro modelo a ser reconhecido foi o de Simon (1977), contempla três fases principais: a inteligência, a conceção e a escolha. Mais tarde, o autor adicionou uma quarta fase ao processo, a implementação. Turban (2011) adiciona uma nova fase – controlo, que permite o modelo ser o mais atual, conforme a figura 2 (Turban, Sharda & Delen, 2011).

Figura 2 - Fases do processo de Tomada de decisão (adaptado de (Vercellis, 2009))

Vercellis (2009) caracteriza as 5 fases do processo de tomada de decisão da seguinte forma:

- **Inteligência** – fase caracterizada pela identificação do problema por parte do tomador de decisão. Esta fase, considera a realização de uma análise do contexto e das informações para entender o problema;
- **Conceção** – nesta fase, são desenvolvidas e previstas as medidas para solucionar o problema. A experiência do tomador de decisão atua diretamente, para criar mais soluções possíveis;

- **Escolha** – após a identificação das possíveis soluções é necessário avaliar as mesmas, com base em critérios significativos. Nesta fase, são usados modelos e métodos de otimização que permitem encontrar a melhor solução para os SAD;
- **Implementação** – quando a melhor solução for identificada pelo tomador de decisão, é criado um plano de ação;
- **Controlo** – após a identificação da solução é necessário confirmar se as expectativas foram alcançadas com a ação da solução. Nos SAD, o resultado das avaliações é transformado em conhecimento que é guardado em um Data Warehouse (DW), para suporte futuro às decisões.

O termo SAD é definido por Bonczek et al. (2014) como um programa de computador interativo que assiste diretamente o utilizador na tomada de decisão (Bonczek, Holsapple, & Whinston, 2014). Para Moore e Chang, SAD é um sistema extensível de diferentes programas que funcionam como um todo para ajudar o utilizador na análise de dados, redução de dados não necessários e modelagem da decisão (Bonczek, Holsapple, & Whinston, 2014). Vercellis (2009) caracteriza SAD como uma aplicação computacional que combina dados com modelos matemáticos, com o intuito de apoiar no processo de decisão das organizações públicas ou privadas (Vercellis, 2009).

A utilização de SAD tem vindo a aumentar, devido à velocidade de processamento e análise de informação orientada para os problemas estruturados pelo utilizador. Assim, os SAD têm um grande potencial de se tornarem uma ferramenta de suporte na resolução de problemas (Vercellis, 2009). Permitem, também, potenciar ao máximo uma empresa, apoiando a eficácia na realização de tarefas, a redução da ação dos funcionários, a organização da empresa e a gestão estratégica de custos (Power, 2002).

Em conformidade com as novas definições do termo, também, novas ferramentas são concebidas, para apoiar e facilitar o processo de apoio à decisão. No início da década de 90, as tecnologias de DW, OLAP servers e técnicas de DM permitiram a evolução dos SAD (Rashidi et al., 2018).

2.4 Big Data

O termo Big Data, é caraterizado de forma genérica como um grande volume de dados. No entanto, a sua definição não é tão simples e tem sido alterada ou melhorada ao longo

dos anos (Russom, 2011). A primeira definição surgiu com os 3 V's, Volume; Velocidade; Variedade, definidos por ((Russom, 2011) e (McAfee et al., 2012)) e mais tarde com o aditamento de mais 2 V's, Veracidade e Valor (Oracle, 2012), para complementar a definição (Younas, 2019).

A caraterização dos V's em Big Data:

- Volume: caracteriza-se por uma grande quantidade de dados que são gerados e processados, com dimensões de exabytes e zettabytes;
- Velocidade: caracteriza-se pela velocidade de processamento dos dados e pela velocidade de geração de dados;
- Variedade: caracteriza-se pela quantidade de tipos de dados diferentes e que assim complementam da melhor forma o Big Data;
- Veracidade: caracteriza-se pela credibilidade e confiança na fonte de dados, pela segurança e consistência das mesmas;
- Valor: caracteriza-se pelo valor que os dados podem acrescentar para o assunto em estudo.

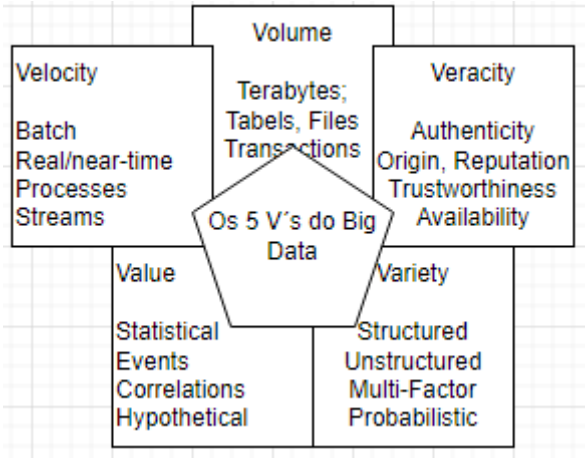

Figura 3 - Os V's do Big Data (adaptado de ("IBM big data plataform - Bringing big data to the Enterprise" 2014))

Assim, Big Data tornou-se uma peça fundamental no cenário atual, transformando a forma como as organizações e as pessoas compreendem e utilizam de forma massiva a quantidade de dados disponíveis.

2.5 Ferramentas ETL

O ETL é um sistema de ferramentas entre a base de dados operacional e o Data Warehouse (DW), e o Business Intelligence (BI). Estas ferramentas são a base do DW, concebendo um modelo estruturado e organizado para o mesmo (Kimbal et al., 2013; Kherdekar, & Metkewar, 2016).

O ETL contempla três tarefas:

- Extract – realizar a extração de dados das diversas fontes de dados, com uma leitura e compreensão dos mesmos;
- Transform – transformar os dados, com a utilização de ferramentas que permitem combinar dados das diversas fontes de dados, eliminando informação duplicada, corrigindo pequenos problemas de conteúdo e proceder à organização dos dados
- Load – realizar o carregamento da informação processada para o DW

A figura 4 apresenta a estrutura do processo de ETL.

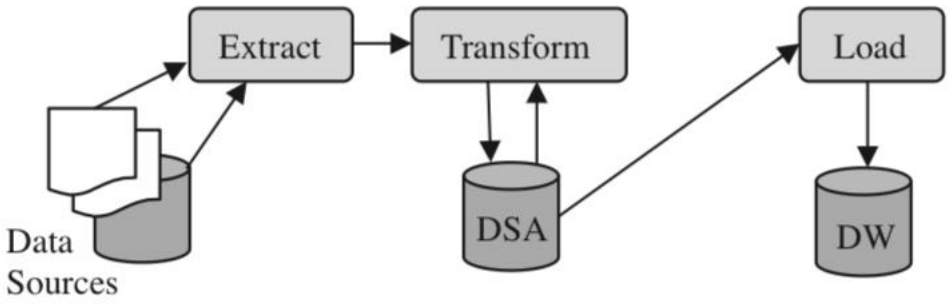

Figura 4 - Estrutura processo ETL

No processo de ETL os dados são extraídos das BD OLTP, ou de ficheiros externos. Este é um processo dispendioso e em constante evolução, para garantir a integridade no processo de descoberta de conhecimento nas organizações. Após o carregamento dos dados são realizadas transformações dos dados originais para a estrutura do DW, carregando informações para as tabelas dimensões e na tabela de factos (El-Sappagh, Hendawi, & El Bastawissy, 2011).

2.6 Data Warehouse

Um sistema de DW, ou armazém de dados, consiste no agrupamento de dados de diferentes bases de dados (BD), ou outras fontes de dados. Assim, um DW é caracterizado por armazenar um grande volume de dados, que posteriormente são tratados e formatados para constituírem uma estrutura única bem organizada, que possa ser utilizada em processos de descoberta de conhecimento em BD (Ferreira et al., 2010). O DW é também uma BD de suporte à tomada de decisão, permitindo a análise de dados, encontrando-se à parte da BD operacional. No entanto, a BD operacional com as fontes externas é responsável por acoplar o DW com informação útil para a direção de executivos das organizações e outros recursos que façam uso das tecnologias (Sethi, 2012).

Neste contexto, as ferramentas ETL, são responsáveis por extrair, transformar e carregar a informação para o DW.

O DW armazena um grande volume de informação organizada por tópicos, integrada, variável no tempo e imutável. As ferramentas, Data cleansing, Data Integration e OLAP são utilizadas no âmbito do DW, sendo a ferramenta OLAP responsável por organizar e realizar uma análise detalhada dos dados, permitindo a criação do modelo multidimensional (Fayyad et al.,1996; Saagari et al., 2013).

2.7 Cubo OLAP

O OLAP surgiu nos anos 90. Refere-se a diferentes técnicas desenvolvidas para analisar informação de um DW, sendo um processo essencial para a tomada de decisão. Assim, a sua utilização a nível empresarial tem vindo a aumentar, caracterizando-se por ser um programa interativo, permitindo a visualização em cubo (multidimensional) da informação de forma lógica (Fayyad et al., 2013).

Para a conceção do processo OLAP é importante o processo de Online Transactional Processing (OLTP), que apresenta informações para fins operacionais. O OLTP é utilizado em BD operacionais, as suas ferramentas permitem a alteração dos dados na BD, ao contrário das tecnologias OLAP, cuja característica é a imutabilidade dos dados. As ferramentas de OLAP funcionam no âmbito dos DW, encontram-se à parte da BD

operacional, permitem colmatar o baixo desempenho das queries[1] aquando da amostragem da informação estruturada e extensa (Reddy et al., 2010; Burstein et al., 2008).

2.8 Data Mining

O DM é o termo mais conhecido e utilizado como sinónimo do método de DCBD. Esta etapa consiste na aplicação de análises e algoritmos em grande quantidade de dados, com suporte computacional, produzindo um número particular de padrões sobre os dados (Fayyad et al., 1996). Assim, é possível extrair conhecimento dos padrões encontrados (Han et al., 2012). Os autores Goebel e Gruenwald (1999) comparam o termo DCBD e DM, definindo DCBD como o processo responsável por transformar low-level data em high-level knowledge e DM como o processo de extração de modelos, padrões e relações entre os dados em estudo (Goebel, & Gruenwald, 1999). Segundo Algarni (2016) a precisão na descoberta desses padrões depende da utilização das técnicas de DM, para tratar o volume e integridade dos dados e o conhecimento na área de estudo, para gerar e analisar padrões, selecionando os mais relevantes (Algarni, 2016).

Segundo H.Witten & E.Frank (2002) o processo de DM é prático e permite a aprendizagem, com a automatização da descoberta de padrões para solucionar problemas. Os padrões encontrados permitem uma análise descritiva ou previsível (Witten & Frank, 2002). Han & Kamber caracterizam a classificação dos padrões como tarefas, podendo ser descritivas, caracterizando as propriedades do dataset, ou previsíveis, utilizando os dados do dataset para fazer previsões (Han el al., 2012).

Os objetivos do DM e do processo de descoberta de conhecimento contemplam duas categorias: as técnicas de previsão e as técnicas de descrição. Estas consideram objetivos diferentes, mas ambas contribuem com conhecimento para o utilizador. Estas técnicas estão divididas em subcategorias, como apresentado na figura 5 (Fayyad et al., 1996).

[1] Query – " é um pedido realizado para consultar/alterar informação de uma determinada BD"

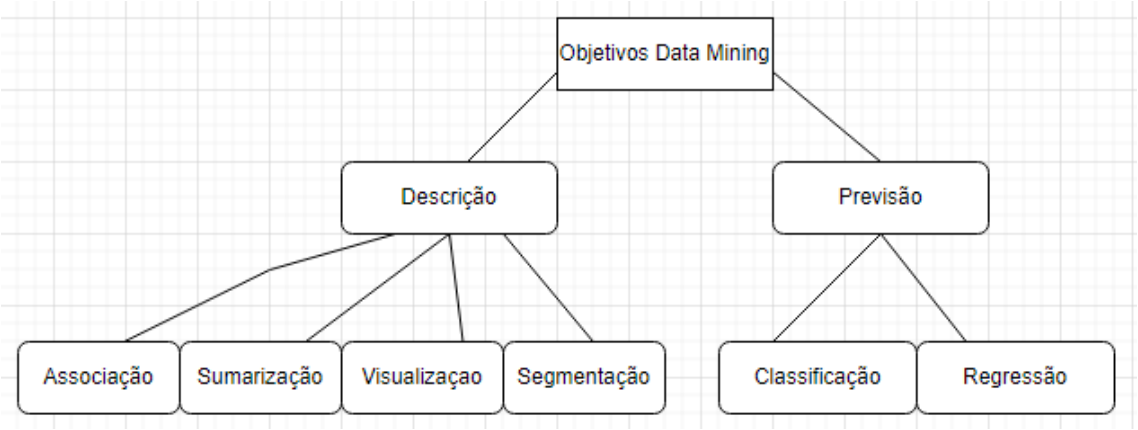

Figura 5 - Técnicas de previsão e descrição em Data Mining(adaptado de (Pereira, 2005))

Classificação – é a função que atribui classes a objetos, criando um modelo que caracteriza os dados, sendo possível organizar um grande volume de dados não classificados. Decision Trees e Bayesian network são alguns dos algoritmos utilizados para a classificação dos dados (Fayyad et al., 1996).

- Decision Tree – este algoritmo é utilizado para obter conhecimento para o processo de tomada de decisão. Este tipo de algoritmo, quando comparado com os outros modelos de classificação, é de fácil interpretação e compreensão. Permite trabalhar com dados numéricos e categóricos, não necessita de relações entre as classes de dados e consegue trabalhar com dados em falta. Este procedimento divide um grande volume de dados inicial (raiz), em grupos de dados menores (folhas), após a aplicação de testes para perceber como dividir o grupo de dados, C4.5, ID3 e CART são alguns dos algoritmos de árvore de decisão mais utilizadas (Priyam et ., 2013).
- Bayesian Network – é um modelo gráfico de probabilidade da relação entre variáveis, sendo também um modelo de classificação, tendo como objetivo a classificação de dados de um dataset em categorias e classes, permitindo compreender dados discretos e contínuos (Phyu, 2009) (Bielza, & Larragana, 2014).

Regressão – é utilizada para fazer a previsão de dados em falta, fazendo previsões numéricas; ao contrário da classificação, a regressão avalia dados contínuos e permite fazer a previsão de valores futuros (Han et al., 2012).

Associação – é a função utilizada para encontrar associações recorrentes no dataset. Este tipo de função é utilizado no retalho, para organizar a disposição dos produtos (Vercellis, 2011).

Segmentação/Clustering – é a função que analisa e caracteriza sem consultar as classes definidas, contrariamente à classificação e à regressão. Clustering é utilizado para criar classes de dados, partilhando uma grande quantidade de dados em grupos menores e homogéneos (clusters), criando assim as diferentes classes e descrevendo o dataset (Han et al., 2012).

Sumarização – é a técnica utilizada para encontrar a descrição correta para um subset de dados do dataset principal, permitindo a visualização da relação entre variáveis e visualização das mesmas (Fayyad et al., 1996).

Visualização – é a tarefa que faz uma descrição simples e concisa do dataset, apresenta os padrões nos dados e os resultados das técnicas aplicadas em gráficos (Fayyad et al., 1996).

As tarefas de DM descritas, são apresentadas resumidamente na tabela 1.

Tabela 1 - Técnicas de Data Mining

Técnicas	Descrição	Exemplo
Classificação	Atribuir classes a dados com base num modelo	Classificar email como spam
Regressão	Prever um valor numérico com o suporte aos dados anteriores	Prever o preço de uma casa com certas caracteristicas
Associação	Identificar padrões de associação entre variáveis	Organizar produtos em supermercados
Segmentação	Agrupamento de dados em clusters com caracteristicas comuns	Agrupar clientes da mesma região
Sumarização	Gerar a descrição concisa dos dados	Apresentar resumos das vendas de uma loja

Visualização	Representação gráfica dos dados	

Após a apresentação das técnicas é possível perceber que as técnicas de DM, constituem uma ferramenta essencial para otimizar o funcionamento de uma organização. Estas técnicas podem ser aplicadas em diferentes contextos, como a detenção de fraudes, diagnósticos médicos, avaliação de riscos e marketing, nesta última as técnicas podem apoiar na identificação das melhores campanhas por utilizador, bem como na disposição de produtos na loja (Fayyad et al., 1996).

2.9 Data Science

O DS contempla um conjunto de fundamentos que guiam o processo de extração de conhecimento de dados. O DM é frequentemente confundido com DS. O DM é o processo técnico de extração de conhecimento de dados. O DS não é apenas a aplicação de algoritmos de DM, mas também a análise do negócio através de dados, permitindo a visualização de problemas, soluções e oportunidades futuras. Assim, é utilizado nas áreas do marketing e financeira das organizações para melhorar o processo de tomada de decisão. Compete à organização garantir um analista competente que, estrategicamente, com o apoio do DS consiga tirar partido das análises, e com o seu pensamento critico ajudar a organização em diferentes vetores (Provost, & Fawcett, 2013). A evolução exponencial da DS levou à criação de diversas ferramentas, sendo necessário distinguir as tecnologias de DS. Os processos como Data processing e Data engeneering são bastante utilizados, mas apenas servem como suporte para o DS (Provost, & Fawcett, 2013).

2.9.1 Business Intelligence

A evolução contínua no mundo empresarial e os seus processos cada vez mais complexos são um desafio para os gestores que constatam uma dificuldade na realização do seu trabalho organizacional. Neste sentido, recorrem aos SAD para apoio na análise dos dados e consequente tomada de decisão. O Business Intelligence (BI) é uma das áreas dos SAD, que recorre aos Sistemas de Informação (SI) para ajudar no processo de decisão e solucionar problemas organizacionais (Khan & Quadri, 2012).

O BI é um processo que utiliza métodos e técnicas para obter informação e conhecimento, tendo como objetivo prever comportamentos dos clientes, fornecedores, da concorrência e do mercado (Vedder et al., 1999).

Os sistemas de BI estão ligados a três tecnologias apresentadas neste projeto, DW, OLAP e DM. Estas tecnologias colocam ao dispor das organizações informações que melhoram a inteligência da organização, a capacidade de aprendizagem e a criatividade organizacional permitindo a evolução de certos processos organizacionais (Santos, & Ramos, 2009).

Segundo Negash (2004) os dados utilizados pelas organizações permitem realizar analises para o suporte à tomada de decisão. Os dados podem ser estruturados (OLAP, DM, DW, etc.…), ou não estruturados (Tabelas, Texto, Gráficos), como é possível visualizar na figura 6 (Negash, 2004).

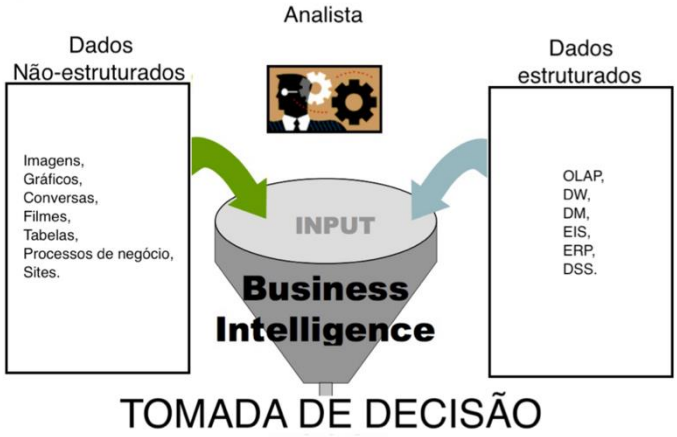

Figura 6 - Tipos de dados usados pelo BI (retirado de (Negash, 2004))

3 Metodologias

Este capítulo reflete a metodologia e as ferramentas utilizadas para o desenvolvimento deste projeto, bem como a organização das tarefas. O capítulo está dividido em três secções. A primeira secção explica a metodologia utilizada. A segunda secção indica as ferramentas utilizadas para o desenvolvimento do projeto, com suporte DSR, a metodologia utilizada para o desenvolvimento deste projeto. A terceira secção refere-se à organização das tarefas, com o desenvolvimento do projeto a considerar dois tipos de tarefas a serem realizadas, as tarefas de aquisição de conhecimento e as tarefas de desenvolvimento da solução.

3.1 Design Science Research

O DSR é uma metodologia adotada em várias áreas do conhecimento, como a ciência da computação, a engenharia e administração, com o propósito de desenvolver e validar soluções inovadoras para problemas complexos. À medida que os problemas enfrentados por organizações e investigadores se tornam mais complexos, a necessidade de uma abordagem sistemática e robusta para a criação de soluções práticas e eficazes torna-se cada vez mais necessário (Peffers et al., 2007).

O modelo processual da DSR é caracterizado por uma sequência de etapas iterativas e interconectadas, incluindo a identificação do problema, a definição dos objetivos da solução, o design e a conceção do artefacto, a demonstração, a avaliação e a comunicação dos resultados. Estas etapas colaborativas asseguram que o desenvolvimento de artefactos seja conduzido de maneira sistemática e que os resultados gerados sejam pertinentes e

aplicáveis tanto para a prática quanto para a teoria (Hevner et al., 2004). A figura 7, apresenta o modelo definido pelos diferentes autores.

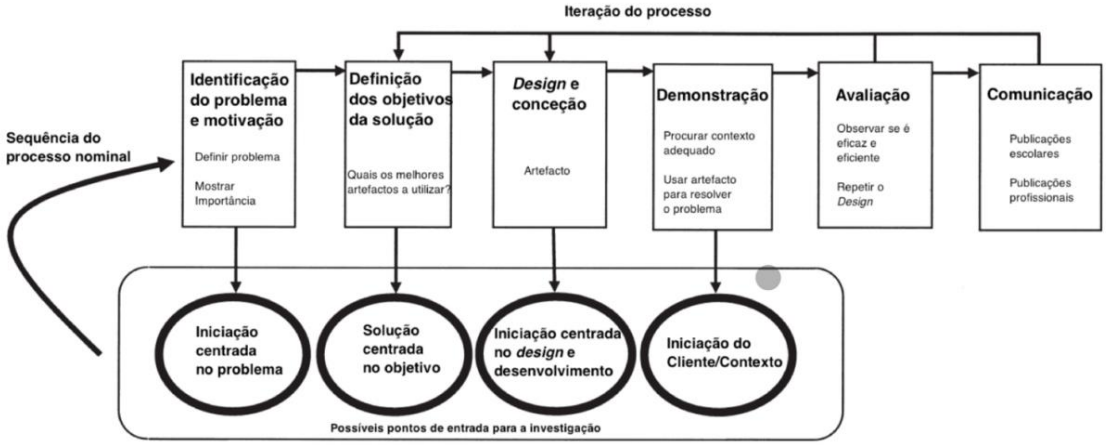

Figura 7 - Modelo processual da DSR (retirado de (Peffers, 2007))

3.1.1 Identificação do Problema e Motivação

A identificação do problema e a motivação permitem caracterizar o problema a ser abordado, e compreender as razões que justificam a necessidade de uma solução. É crucial entender o contexto e as necessidades para garantir que o artefacto desenvolvido seja relevante e útil (Peffers et al., 2007).

Esta etapa foi utilizada como referência para a identificação do problema em estudo neste projeto, tal como foi descrita a motivação para a realização do mesmo.

3.1.2 Definição dos Objetivos da Solução

Nesta etapa, os objetivos específicos da solução são estabelecidos com base na análise do problema e nas necessidades identificadas. Os critérios de sucesso e as métricas de avaliação são definidos para orientar o desenvolvimento do artefacto e garantir que atenda às expectativas (Hevner et al., 2004).

3.1.3 Design e Conceção

O design e a conceção do artefacto ocorrem nesta etapa, onde são desenvolvidas soluções concretas para o problema identificado. As melhores práticas de design são aplicadas para garantir a eficácia, a eficiência e a usabilidade do artefacto proposto (Peffers et al., 2007).

Esta etapa da metodologia foi utilizada para desenhar a arquitetura que vai levar à realização do projeto.

3.1.4 Demonstração

Após o desenvolvimento do artefacto, é demonstrado num ambiente controlado ou em um contexto real para mostrar como opera na prática. Esta etapa permite validar a funcionalidade e a utilidade do artefacto, bem como obter feedback dos stakeholders (Hevner et al., 2004).

3.1.5 Avaliação

A avaliação do artefacto ocorre para avaliar a sua eficácia em resolver o problema identificado e alcançar os objetivos estabelecidos. Métodos de avaliação, como testes de usabilidade e análise de desempenho, são empregues para garantir a qualidade e a adequação do artefacto (Peffers et al., 2007).

3.1.6 Comunicação

Por último, os resultados do estudo são comunicados por meio de relatórios técnicos, publicações académicas e apresentações em conferências. A comunicação eficaz dos resultados é essencial para compartilhar as descobertas, lições aprendidas e contribuições teóricas e práticas do estudo (Hevner et al., 2004).

3.2 Ferramentas utilizadas

No desenvolvimento da investigação, foi necessário a utilização de diferentes ferramentas. Na tabela 2, estão descritas as ferramentas utilizadas para o desenvolvimento com a respetiva descrição.

Tabela 2 - Ferramentas utilizadas para o desenvolvimento de uma solução

Ferramentas	Descrição

Microsoft Excel	O Microsoft Excel permite a exploração de dados e a sua análise. Esta ferramenta foi utilizada para a exploração dos dados (dataset) para o estudo.
Microsoft SQL Server	O Microsoft SQL Server permite a criação e a gestão de bases de dados. Esta ferramenta foi utilizada para armazenar e gerir os dados do estudo.
Microsoft Visual Studio	O Microsoft Visual Studio é uma ferramenta que permite o desenvolvimento de software da framework .net, com recurso ao pacote de BI permite a criação do Cubo OLAP e a aplicação de técnicas de Data Mining.
Microsoft Power BI	O Microsoft Power BI é uma ferramenta que permite a criação de Dashboards de forma a ajudar na análise dos dados. Esta ferramenta foi utilizada para criar dashboards sobre os dados e um relatório sobre os mesmos.

As ferramentas utilizadas para o desenvolvimento da solução foram desenvolvidas pela Microsoft. A escolha destas ferramentas deriva da utilização anterior no desenvolvimento de outros projetos e da universidade disponibilizar o Microsoft office 365. Estas ferramentas apresentam uma boa compatibilidade, facilitando os processos no desenvolvimento da solução.

3.3 Organização das tarefas

Nesta fase da dissertação é descrita a organização da parte prática, com base na abordagem metodológica descrita anteriormente (DSR). Esta metodologia está dividida em 7 etapas iterativas. Na parte teórica foi realizada a primeira etapa da metodologia, Identificação do Problema e Motivação. Na parte prática serão desenvolvidas as etapas de definição dos objetivos, design e conceção, demonstração e avaliação. Assim, como é possível visualizar na tabela 3 a parte prática encontra-se dividida em duas componentes. Na primeira componente é obtido o conhecimento sobre os dados em estudo. Na segunda

componente pretende-se apresentar a solução do projeto. Esta componente contém as etapas 3, 4 e 5.

Tabela 3 - Orientação das Tarefas

Etapa	Descrição	Aquisição de conhecimento	Desenvolvimento da Solução
2	Definição dos Objetivos	-Recolha e estudo dos dados -Classificação do Problema -Seleção e agregação dos dados	
3	Design e Conceção		-Preparação dos dados -Processo ETL -Cubo OLAP -Power BI -Data Mining
4	Demonstração		-Criação Dashboards
5	Avaliação		-Análise Dashboards

A metodologia DSR é iterativa, permitindo avançar e recuar nos processos. No decorrer do projeto foram desenvolvidas e avaliadas diferentes soluções, até o modelo final estar estruturado. Por exemplo, os dados no início tinham valores nulos devido ao tamanho da variável na base de dados. Após uma avaliação foi possível entender que havia a necessidade de refazer o processo de design e conceção.

4 Aquisição de conhecimento

Este capítulo reflete sobre o primeiro contacto com os dados (dataset) disponíveis para o desenvolvimento do projeto, sendo a primeira parte prática do projeto. O capítulo encontra-se dividido em 4 secções. A primeira secção consiste na pesquisa e obtenção de conhecimento relativamente à doença Diabetes Mellitus tipo 2, para um correto desenvolvimento da solução. A segunda secção, consiste na descrição da primeira análise ao dataset. Na terceira secção, é apresentada a classificação do problema em estudo. Na quarta secção, apresentam-se as alterações ao dataset e a estrutura final dos dados para o desenvolvimento da solução.

4.1 Obtenção de conhecimento

No desenvolvimento deste projeto é necessário fazer uma investigação sobre o tema do dataset, de forma a entender os dados que compõem o dataset.

4.1.1 Diabetes Mellitus

A Diabetes Mellitus (DMe) é classificada como um conjunto de doenças metabólicas de origem múltipla, problemas relacionados com as secreções da insulina e/ou ação da insulina, que poderão estar no aparecimento da doença (Negash, 2004). A DMe, como referido, é um conjunto de doenças, mas apenas uma será relevante para este estudo a DMe Tipo 2.

A DMe não sendo acompanhada corretamente pode trazer consequências futuras para o paciente. Assim, há necessidade de um acompanhamento constante dos níveis de hiperglicemia. A falta do referido acompanhamento pode trazer complicações como "danos a longo prazo, disfuncionalidade e falência de diferentes órgãos, especialmente olhos, rins, nervos, coração e vasos sanguíneos, ou até mesmo, a morte" (American Diabetes Association, 2014).

4.1.1.1 Tipo 2

A DMT2 é um problema global de saúde pública, caracterizando-se pelo facto de o doente não ser insulina-dependente, tornando-se resistente à insulina (as células pancreáticas não conseguem acompanhar os níveis de produção necessários pedido pelo corpo do doente). Sendo o tipo mais comum de DM estima-se que cerca de 9%-10% da população é afetada

por esta epidemia, sendo que 80% dos casos estão em países mais desfavorecidos. Fatores como a nutrição pobre na gestação, o sedentarismo e a obesidade são considerados como as maiores causas para a resistência à insulina (Defronzo et al., 2015; Kharroubi & Darwish, 2015).

4.1.1.2 Exame do pés

Anualmente, 2% a 4% dos utentes com DMT2 apresentam casos de úlceras nos pés. Este problema, nos países em desenvolvimento, representa um grande número dos casos hospitalares urgentes. A falta de controlo pode trazer graves complicações futuras, como a amputação do pé, ou a morte. O exame aos pés serve de controlo e para deteção da doença. No exame é realizado um teste de sensibilidade e são analisadas certas condições dermatológicas. A falta de sensibilidade, pele seca, rachaduras na pele, unhas encravadas, calosidades e temperatura dos pés, constituem condições pré-ulcerativas (Sociedade Brasileira de Diabetes, 2016).

4.1.1.3 Exame ao HgbA1c

A Hemoglobina A1C (HgbA1c) é um conjunto de substâncias formadas através da hemoglobina A e os açúcares ingeridos. Esta hemoglobina é responsável por indicar o nível de glicose no sangue. A prevalência de um nível de glicemia elevado pode afetar irreversivelmente os órgãos, como olhos e rins, vasos sanguíneos e nervos. Os adultos não-diabéticos tendem a ter cerca de 4% a 6 % de glicose no sangue. Os níveis de glicemia devem-se manter abaixo dos 7%. Após este nível, existe a probabilidade de desenvolver DMT2, entre outros problemas de saúde (Netto et al., 2009). Na figura 8, é visível o quanto os níveis altos de glicose correlacionam.se com uma menor probabilidade de sobrevivência.

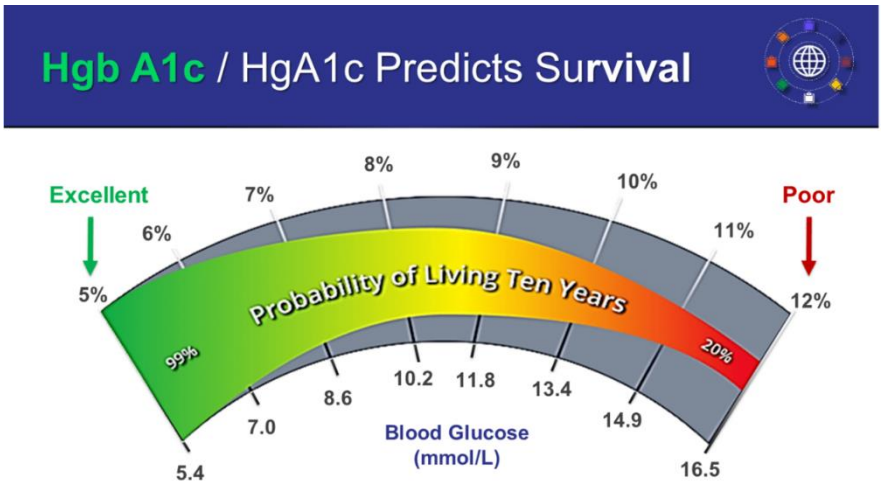

Figura 8 - Níveis de glicose e o risco de vida (retirado de Elisa/Act Biotechnologies)

4.2 Recolha e estudo dos dados

Para o desenvolvimento deste projeto é utilizado um dataset disponível no portal do Serviço Nacional de Saúde, disponível em transparencia.sns.gov.pt. O local referido é um repositório de dados do Sistema Nacional de Saúde português com um conjunto de datasets para consulta pública, em conformidade com as regras da transparência. Para o desenvolvimento da presente dissertação foi escolhido um dataset relativo à saúde dos portugueses, "atividade do programa da diabetes". Os dados do dataset estão disponíveis por diferentes agrupamentos de centros de saúde (ACES) de Portugal. Os ACES têm informação disponível através de datasets desde o ano de 2014, com informação atualizada até ao presente ano.

O dataset relativo à atividade do Programa da Diabetes possui informações sobre pacientes com DMT2. É composto por informações sobre o número de exames dos pés realizado aos doentes inscritos com diabetes em cada ano e a proporção relativamente ao ano anterior. Também, disponibiliza informação sobre o número de utentes inscritos que tiveram 8% ou menos no exame de HgbA1C em proporção ao ano anterior.

Após uma primeira análise é possível verificar que o dataset tem disponível informação útil para o estudo a realizar. No decorrer de uma primeira análise foi possível verificar alguns problemas de estruturação. Na tabela 4, podemos verificar a estrutura que compõe o dataset, sendo apresentadas as diversas variáveis do estudo.

Tabela 4 - Descrição do dataset inicial

Tabela	Descrição	Nº de Registos	Campos	Tipo
Atividade do Programa de Diabetes	Monotorização do programa de controlo da diabetes	6620	ID_Registo Período Região ACES Localização geográfica Utentes c/exame pés Proporção exame pés Utentes c/exame <= 8 Proporção exame <= 8	Int Varchar (50) Varchar (50) Varchar (50) Varchar (50) Int Int Int Int

Nesta primeira análise, foi possível perceber que o dataset deveria ser estudado e dividido por diferentes regiões, com diferentes ACES. Na tabela 5, é possível verificar as ACES distribuídas por regiões.

Tabela 5 - Descrição dos ACES por região

Região	Região - Descrição	ACES
1	Região de Saúde do Alentejo	ACES Baixo Alentejo ACES Alentejo Central ACES São Mamede ACES Alentejo Litoral
2	Região de Saúde do Algarve	ACES Algarve I

		ACES Algarve II ACES Algarve III
3	Região de Saúde do Centro	ACES Baixo Mondego ACES Baixo Vouga ACES Beira Interior Sul ACES Cova da Beira ACES Dão Lafões ACES Guarda ACES Pinhal Interior Norte ACES Pinhal Interior Sul ACES Pinhal Litoral
4	Região de Saúde LVT	ACES Almada -Seixal ACES Amadora ACES Arco Ribeirinho ACES Arrábida ACES Cascais ACES Estuário do Tejo ACES Lezíria ACES Lisboa Central ACES Lisboa Norte ACES Lisboa Ocidental e Oeiras

		ACES Loures – Odivelas ACES Médio Tejo ACES Oeste Norte ACES Oeste Sul ACES Sintra
5	Região de Saúde Norte	ACES Alto Ave ACES Alto Minho ACES Alto Trás-os-Montes ACES Ave ACES Cávado I ACES Cávado II ACES Cávado III ACES Douro I ACES Douro II ACES Entre Douro e Vouga I ACES Entre Douro e Vouga II ACES Grande Porto I ACES Grande Porto II ACES Grande Porto III ACES Grande Porto IV ACES Grande Porto V

		ACES Grande Porto VI ACES Grande Porto VII ACES Grande Porto VIII ACES Matosinhos ACES Tâmega I ACES Tâmega II ACES Tâmega III

4.3 Classificação do problema

O problema em estudo na presente dissertação está relacionado com a doença da diabetes. Com os dados disponíveis no portal do SNS é possível consultar a evolução da diabetes e a monitorização da doença, através das informações obtidas do exame realizado aos pés e do exame de controlo da HgbA1C. Assim, o problema central é como prever o número de utentes inscritos com a DMT2 nos próximos anos, utilizando informações disponíveis no dataset. A evolução no controlo da doença e eficácia das tecnologias de previsão como DM são problemas que foram encontrados ao longo da interpretação dos dados e do desenvolvimento do projeto, permitindo entender outros fatores importantes no desenvolvimento da solução.

Com base no problema central do trabalho, é possível concluir que este tema é de elevada relevância, permitindo visualizar o período dos exames de controlo do avanço da doença, para que esta não se agrave e prejudique a vida dos doentes. Este tema é também, relevante para os doentes da DMe, conferindo maior confiança e que os mesmos se sintam mais confortáveis a conviver com a doença.

4.4 Seleção e agregação do dados

Após diversas análises efetuadas aos dados, foram estabelecidas algumas alterações ao modelo extraído do SNS. Na figura 9, é apresentada a estrutura do dataset fornecido, apresenta uma estrutura de difícil análise e interpretação dos dados. Assim, o novo

modelo definido apresenta uma nova estrutura permitindo ter a informação do dataset organizada, facilitando a consulta da mesma.

Column Name	Data Type
Período	varchar(50)
Região	varchar(50)
ACES	varchar(50)
[Localização Geográfica]	varchar(50)
[Utentes inscritos com diabetes com exame dos pés realizado no último ano]	varchar(50)
[Proporção DM c exame dos pés no último ano]	varchar(50)
[Utentes inscritos com diabetes com último resultado de HgbA1c inferior ou igual a 8,0%]	varchar(50)
[Proporção DM c última HgbA1c < 8,0 %]	varchar(50)

Figura 9 - Estrutura do dataset inicial

A partir do campo "período" do dataset original foram concebidos os campos "ano", "mês", "Mês_número" e "trimestre", e eliminado o campo "período", sendo criada uma tabela dimensão para guardar estes dados. Foram excluídos do dataset os dados relativos ao ano de 2024, por ter registado apenas os dados relativos ao mês de janeiro, conduzindo a uma descontextualização do processo de análise e da representação das previsões. A partir dos campos Localização geográfica, Região e ACES foram criadas três tabelas dimensões onde foram guardados os dados. A figura 10, apresenta a estrutura final da tabela Facto.

Column Name	Data Type
IDFacto	int
IDPeriodo	int
IDRegiao	int
IDAces	int
IDLocalizacao	int
Utentes_Insc_c_Exm_Ult_Ano	int
Prop_DM_c_Exam_Ult_Ano	int
Utentes_insc_c_Diabetes_Ult_res_HgbA1c	int
Proporcao_DM_Ult_HgbA1c	int

Figura 10 - Estrutura da tabela facto dataset final

Nas figuras 11 , 12, 13, 14 e 15, é possível consultar as estruturas das tabelas dimensão criadas para armazenar os dados das regiões, localizações, ACES e períodos dos registos. Na figura 15 é possível consultar o novo modelo estruturado para o dataset e as ligações entre as diferentes tabelas de dimensão com a tabela de facto, utilizando foreigns keys. A

informação IDLocalizacao, IDRegiao, IDAces e IDPeriodo é partilhada com a tabela facto através das foreigns keys, utilizando o ID das tabelas dimensão.

Column Name	Data Type
IDAces	int
ACES	varchar(50)

Figura 11 - Estrutura da tabela ACES dataset final

Column Name	Data Type
IDLocalizacao	int
Localizacao	varchar(50)

Figura 12 - Estrutura da tabela Localizacao dataset final

Column Name	Data Type
IDPeriodo	int
Ano	varchar(50)
mes	varchar(50)
mes_num	varchar(50)
trim	varchar(50)

Figura 13 - Estrutura da tabela Periodo dataset final

Column Name	Data Type
IDRegiao	int
Regiao	varchar(50)

Figura 14 - Estrutura da tabela Regiao dataset final

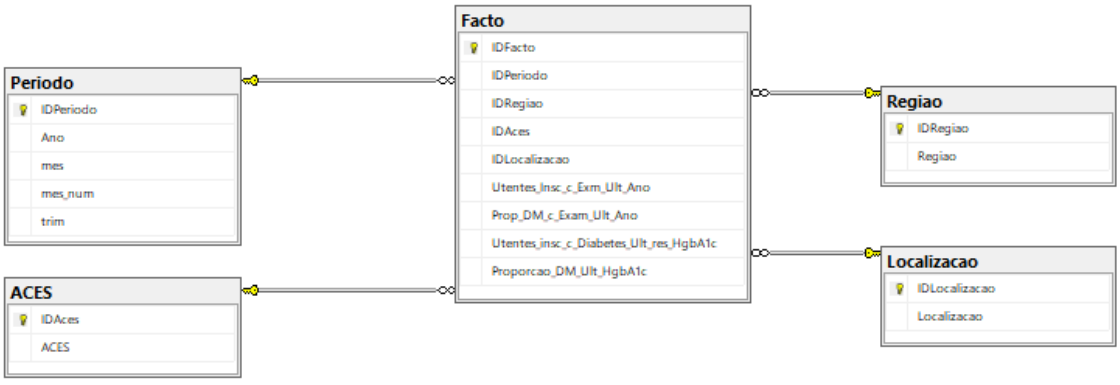

Figura 15 - Estrutura do dataset final

Após estabilizar o dataset, procedeu-se à construção do modelo de base de dados para suportar o DW, constituído pelas tabelas de dimensão e facto. As tabelas de dimensão,

classificadas como tabelas auxiliares que identificam de forma unívoca os registos, que servem de suporte à tabela principal - facto, onde os factos se concretizam.

Com a estrutura e os dados definidos para o estudo, estão reunidas as condições para o desenvolvimento da solução deste projeto.

5 Desenvolvimento da solução

Este capítulo reflete o desenvolvimento prático do projeto. O capítulo encontra-se dividido em 5 secções, com a preparação dos dados, o processo de ETL, o Cubo OLAP, o Data Mining e o Power BI, com a criação de dashboards para a apresentação dos dados.

Na primeira secção, decorre o processo de identificação das etapas da preparação dos dados. Na secção seguinte, é apresentado o processo de ETL, com as alterações e tratamento dos dados. Na terceira secção, é apresentada a conceção do cubo OLAP, responsável pela conceptualização dos dados para análise. Na quarta secção, decorre o processo de conceptualização dos algoritmos de DM através da regressão e classificação, para previsão de valores futuros. Na quinta e última secção, é apresentado o processo de criação de dashboards e apresentação dos dados.

5.1 Preparação dos dados

A preparação dos dados é o primeiro processo no desenvolvimento da solução. Assim, é importante que este processo seja realizado corretamente para o sucesso da solução. Esta secção encontra-se dividida em duas partes. A primeira parte, faz referência à organização do dataset, para realizar a correta extração dos dados para o DW. A seguinte secção, refere-se à criação do DW, local de destino dos dados através do processo de ETL.

5.1.1 Organização do Dataset

Nesta parte é realizado o processo de organização do dataset para futuramente ser realizada a extração, transformação e carregamento de dados no DW, dando assim início ao processo de DCBD.

O dataset foi extraído do site do portal do SNS, em formato *csv.* Assim, estas primeiras análises e a organização dos dados foram realizadas com recurso à ferramenta auxiliar Power Querie. Esta ferramenta permite a transformação dos dados em diferentes formatos.

Nas tabelas 6 e 7, é possível consultar as alterações realizadas aos campos de informação do dataset, como foi explicado no processo de seleção e agregação dos dados, sendo possível comparar os modelos. A tabela 6, apresenta a estrutura do dataset sem análise.

Na tabela 7, é apresentada a estrutura atualizada do dataset para o desenvolvimento da solução.

Tabela 6 - Estrutura do dataset inicial

Dataset	Tabela	Campos	Tipo	N º de Registos
DataSet s/análise	Facto	• ID_Registo • Período • Região • ACES • Localização geográfica • Utentes c/exame pés • Proporção exame pés • Utentes c/exame <= 8 • Proporção exame <= 8	• Int • Varchar(50) • Varchar(50) • Varchar(50) • Varchar(50) • Varchar(50) • Varchar(50) • Varchar(50) • Varchar(50)	6620

Tabela 7 - Estrutura do dataset final

Dataset	Tabela	Campos	Tipo	N º de Registos
DataSet C/análise	Facto	ID_Registo IDPeriodo IDRegiao IDACES IDLocalizacao Utentes_Insc_c_Exm_Ult_Ano Prop_DM_c_exam_Ult_Ano Utentes_insc_c_Diabetes_Ult_res_HgbA1c Proporcao_DM_Ult_HgbA1c	Int Int Int Int Int Int Int Int Int	6581

DataSet c/análise	Regiao	IDRegiao Regiao	Int Varchar(50)	0
DataSet c/análise	Localizacao	IDLocalizacao Localizacao	Int Varchar(50)	0
DataSet c/análise	ACES	IDAces ACES	Int Varchar(50)	0
DataSet c/análise	Periodo	IDPeriodo Trim Ano Mes Mês_num	Int Varchar(50) Varchar(50) Varchar(50) Varchar(50)	0

Após a seleção e estruturação correta dos dados, estes foram guardados em formato *xls*[2], para utilização no processo de ETL , a fase seguinte do processo de desenvolvimento.

5.1.2 Conceção do Data Warehouse

A estrutura do dataset foi estabilizada e a conversão em formato próprio para o processo de carregamento dos dados foi executado. Segue-se a conceção do modelo de DW, com a criação da base de dados com as respetivas tabelas de dimensão e facto. Para a conceção da estrutura do DW foi utilizada a plataforma da Microsoft SQL Server.

Para melhor perceção da nova estrutura foi desenvolvido um modelo multidimensional, como mostrado na figura 16, onde é possível visualizar a tabela facto e as tabelas dimensão.

[2] xls– " formato dos documentos utilizados no Microsoft Excel"

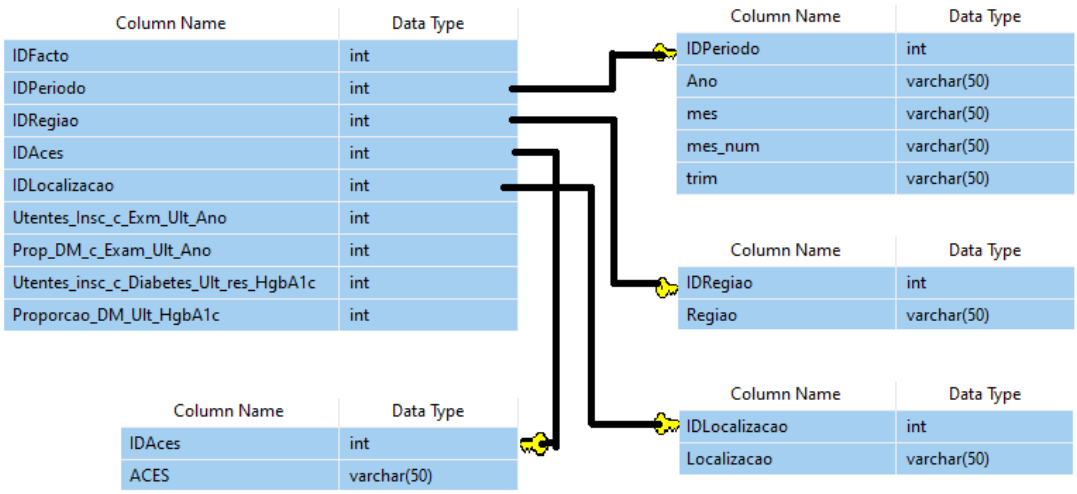

Figura 16 - Modelo multidimensional dos dados

A tabela principal Facto, contempla as alterações com os atributos em definitivo. Como é possível verificar IDPeriodo, IDRegiao, IDACES e IDLocalizacao constituem as chaves estrangeiras que permitem a integração com as respetivas tabelas de dimensão através das chaves primárias. Também os campos Utentes_Insc_c_Exm_Ult_Ano, Prop_DM_c_Exam_Ult_Ano, Utentes_insc_c_Diabetes_Ult_res_HgbA1c e Proporcao_DM_Ult_HgbA1c fazem parte da tabela Facto.

O modelo multidimensional contempla as 4 novas tabelas de dimensão adicionadas à estrutura, Regiao, ACES, Localizacao e Periodo.

- A tabela "Regiao" com os campos IDRegiao (Int) e a Regiao (Varchar(50)) regista as regiões existentes no dataset.
- A tabela "ACES" com os campos IDAces (Int) e ACES (Varchar(50)) regista as diversas ACES em Portugal.
- A tabela Localizacao com os campos IDLocalizacao e Localizacao regista as coordenadas da localização dos diferentes ACES.
- A tabela Periodo com os campos IDPeriodo, ano, mês, trimestre e mes_num, regista o período dos dados.

Para a criação da base de dados foi utilizada uma querie de suporte que permitiu a conceção das tabelas com o tipo de variáveis a receber e com as devidas ligações entre si, ou seja, foram definidas as chaves primárias de cada tabela de dimensão e as chaves estrangeiras na tabela facto para permitir a conexão entre tabelas.

5.2 Processo ETL

O processo de ETL, é a segunda etapa no desenvolvimento da solução. Esta etapa é responsável por extrair dados de diversas fontes de dados, transformar os dados extraídos, realizar a limpeza e à posteriori carregar no DW.

O processo de ETL, contempla o carregamento dos dados e a conexão com o DW através da utilização do data flow task, utilitário disponível no Integration Services disponível no Visual Studio.

Para o desenvolvimento deste processo é necessário ter disponível o modulo de SQL Server Integration Services (SSIS) que permite a integração do SQL Server com o Visual Studio, para que decorra o processo de ETL com a extração, a transformação e o carregamento dos dados na respetiva base de dados do DW.

5.2.1 Carregamento dos dados e a conexão ao DW

A primeira etapa, o carregamento dos dados do ficheiro xls extraído do portal do SNS e respetiva conexão com o DW é possível após ter sido realizado o processo de preparação dos dados, tendo sido cruzados dois modelos iniciais para resultar numa melhor análise futura dos dados. Nesta etapa, é realizada a criação do projeto de integração de dados, o carregamento dos dados para o sistema responsável pelo tratamento dos dados (SSIS), e a criação da conexão para o armazenamento no DW.

A primeira tarefa na criação do projeto designado Diabetes é o desenvolvimento do processo de ETL através do processo Integration Services Project, que permite o carregamento dos dados para o DW. Na figura 17, é possível visualizar a primeira etapa do processo de ETL.

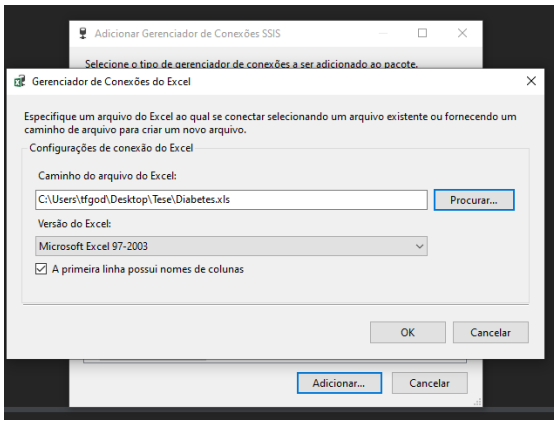

Figura 17 - Conexão com a fonte de dados

Para terminar a primeira parte do processo, é criada uma conexão com o DW que vai receber os dados. Na Figura 18, é possível visualizar a criação da conexão com o DW.

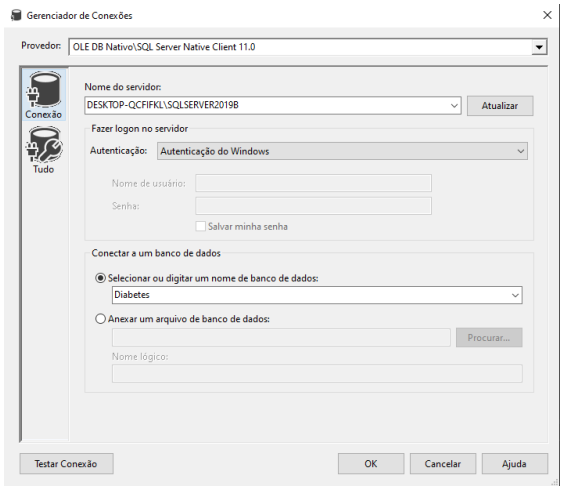

Figura 18 - Conexão com o destino dos dados (DW)

5.2.2 Data Flow Task

Na segunda etapa é realizado o processo de tratamento dos dados previamente carregados. Neste processo é realizado o tratamento dos dados, retirando do estudo dados nulos, dados repetidos e dados inválidos. É também, realizado o processo de inserção dos dados transformados nas tabelas do DW.

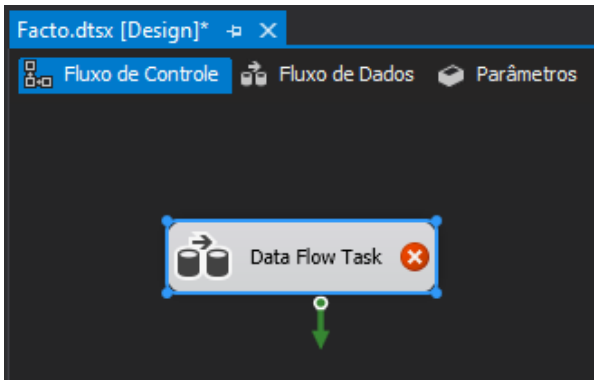

Figura 19 - Exemplo Data Flow Task

Para a concretização do processo de Data Flow Task foram criados packages no Integration Services, correspondendo ao mesmo número das tabelas de dimensão e facto representados no DW. Em cada package está registado o processo de carregamento do dataset, a transformação (formatação e ordenação) e o carregamento dos dados para as respetivas tabelas do DW, figura 20.

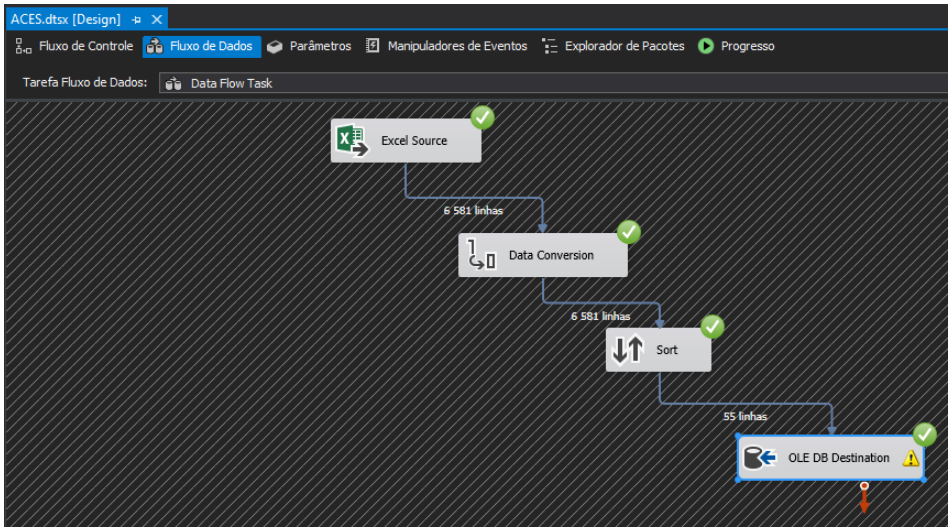

Figura 20 - Exemplo estrutura Data Flow Task

Na figura 21, é possível consultar os diferentes packages criados, no desenvolvimento da solução.

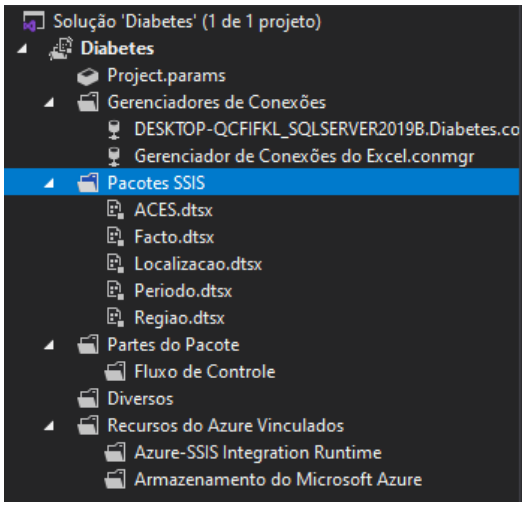

Figura 21 - Packages criados para o processo de ETL

Os packages no fluxo de controlo são executados através da ferramenta Data Flow Task, responsável pelo carregamento dos dados do dataset no formato xls, para posterior transformação, organização e carregamento nas respetivas tabelas do DW.

Na figura 19, é apresentado o package da tabela Facto com a tarefa de Data Flow Task a ser realizada no fluxo de controlo. Os restantes packages apresentam a mesma estrutura de configuração e desenvolvimentos.

Apesar de todos os packages realizarem uma única data flow task, a sua configuração e representação é diferente relativamente às tabelas de dimensão e Facto. Assim, esta segunda etapa é dividida em duas secções, Data Flow Task Dimensões e Data Flow Task Factos. A tarefa do package de Facto é a última a ser realizada, para garantir que todos os dados estão previamente transformados e carregados nas tabelas de dimensão. A tabela Facto vai receber os dados das tabelas de dimensão para o processamento.

5.2.2.1 Data Flow Task Dimensões

As figuras 22, 23, 24 e 25, apresentam os diferentes processos das tarefas de Data Flow Task nos packages relacionados com as quatro tabelas de dimensões e os respetivos resultados do processo.

A figura 22, apresenta a tarefa Data Flow Task da tabela dimensão Regiao, permite a visualização das quatro tarefas a ser realizadas na tarefa e o respetivo resultado.

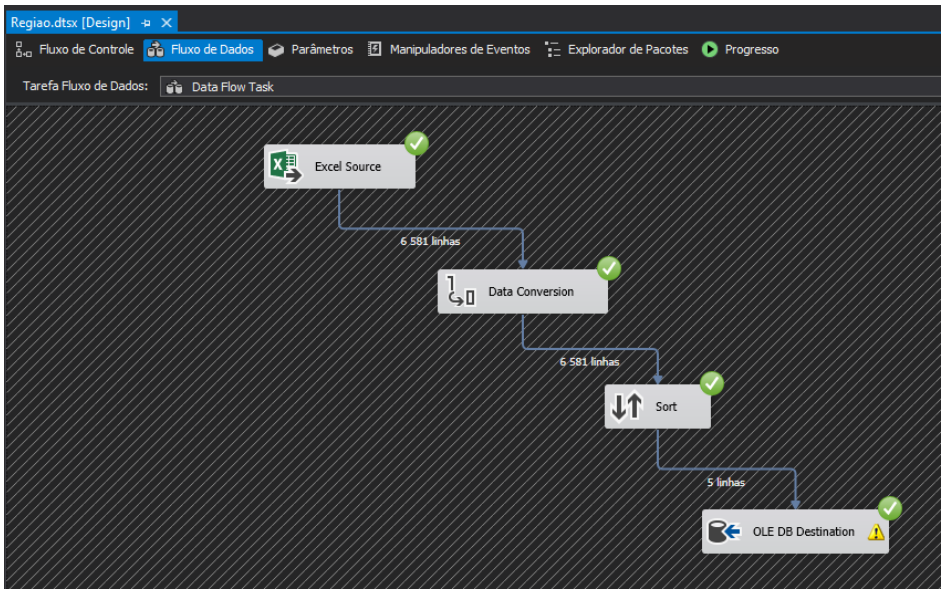

Figura 22 - Processos e resultados do Data Flow Task do package Regiao

A figura 23, apresenta a tarefa Data Flow Task da tabela dimensão ACES, permite a visualização das quatro tarefas a ser realizadas na tarefa e o respetivo resultado.

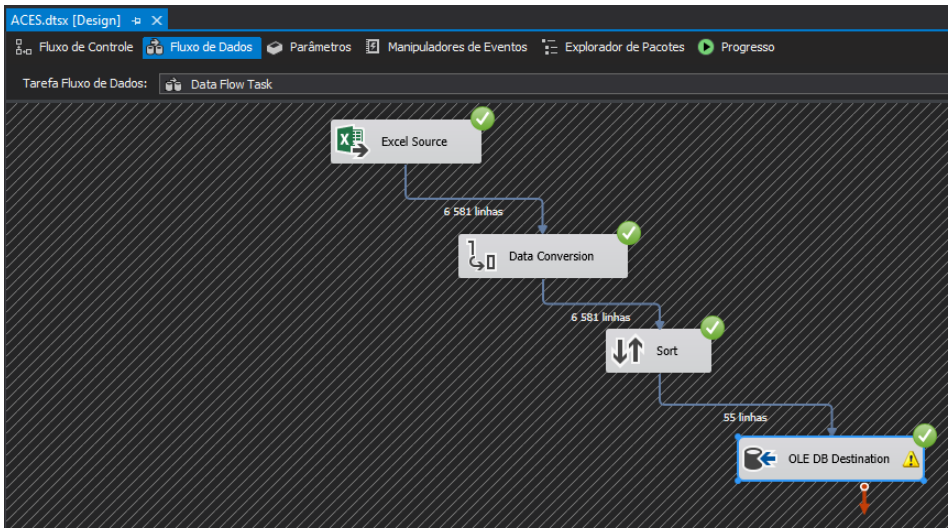

Figura 23 - Processos e resultados do Data Flow Task do package ACES

A figura 24, apresenta a tarefa Data Flow Task da tabela dimensão Periodo, permite a visualização das quatro tarefas a ser realizadas na tarefa e o respetivo resultado.

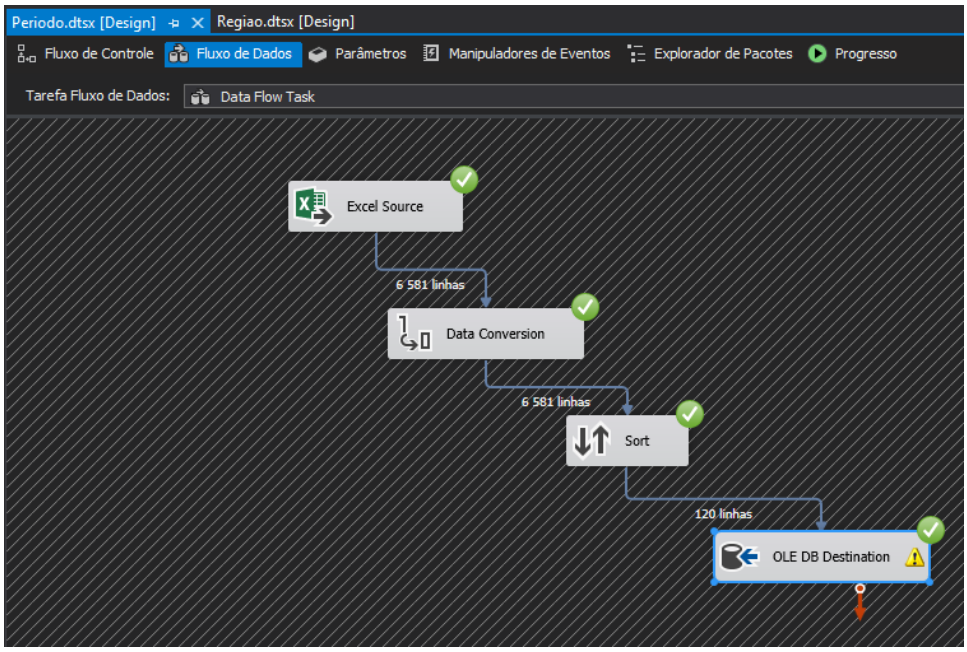

Figura 24 - Processo e resultados do Data Flow Task do package Periodo

A figura 25, apresenta a tarefa Data Flow Task da tabela dimensão Localizacao, permite a visualização das quatro tarefas a ser realizadas na tarefa e o respetivo resultado.

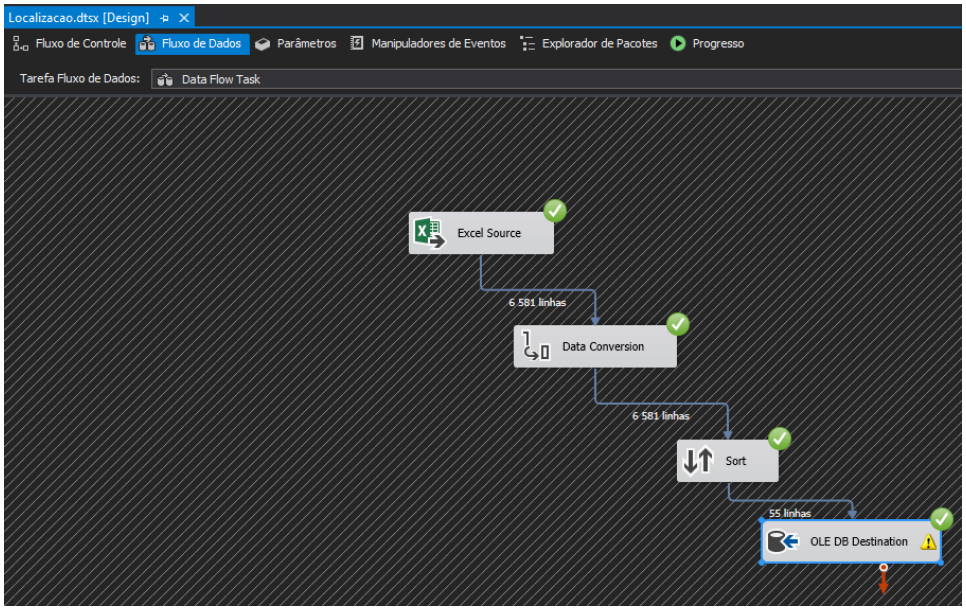

Figura 25 - Processos e resultados do Data Flow Task do package Localizacao

As figuras 22, 23, 24 e 25, representam as tarefas de Data Flow Task divididas em 4 etapas:

- Excel Source

- Data Conversion
- Sort
- OLE DB Destination

5.2.2.1.1 Excel Source

A primeira etapa da tarefa do Data Flow Task é responsável pelo carregamento dos dados do dataset, a partir do formato xls, para serem analisados e tratados no sistema de ETL. Foram carregados 6581 registos para serem tratados nas respetivas packages. Esta tarefa permite selecionar os dados relevantes, por exemplo, no package Regiao foi apenas selecionada a informação do dataset original relativa à coluna Regiao, permitindo trabalhar dados relevantes para esta tabela. O mesmo procedimento foi aplicado aos restantes packages.

5.2.2.1.2 Data Conversion

A etapa seguinte "Data Conversion" é responsável por converter o tipo de dados da fonte (dataset) para o tipo de dados do recetor dos dados (DW), com a conversão dos dados de xls para sql statements. Com este procedimento, os dados nulos e inválidos foram desconsiderados. No dataset objeto de estudo, os 6581 registos utilizados estão em conformidade para serem utilizados no processo de carregamento para o DW.

5.2.2.1.3 Sort

A terceira etapa "Sort" é responsável por realizar limpezas aos dados, eliminando dados repetidos. No final desta etapa é possível verificar que todos os packages apresentam valores diferentes.

O Package "Regiao", apresenta os 5 registos a serem integrados na tabela Regiao do DW. Este valor significa que os dados se encontram divididos em 5 regiões principais. Na figura 26, é possível consultar as 5 regiões.

	IDRegiao	Regiao
1	1	Região de Saúde do Alentejo
2	2	Região de Saúde do Algarve
3	3	Região de Saúde do Centro
4	4	Região de Saúde LVT
5	5	Região de Saúde Norte

Figura 26 - Representação dos registos da tabela Região no DW

O Package "ACES" apresenta 55 registos a serem integrados na tabela ACES no DW. Este valor significa que existem 55 agrupamentos de centros de saúde. Na figura 27, é possível consultar as 55 ACES.

	IDAces	ACES		IDAces	ACES
1	1	ACES Alentejo Central	30	30	ACES Grande Porto I - Santo Tirso e Trofa
2	2	ACES Alentejo Litoral	31	31	ACES Grande Porto II - Gondomar
3	3	ACES Algarve I - Algarve Central	32	32	ACES Grande Porto III - Maia e Valongo
4	4	ACES Algarve II - Algarve Barlavento	33	33	ACES Grande Porto IV - Póvoa do Varzim e Vila d...
5	5	ACES Algarve III - Algarve Sotavento	34	34	ACES Grande Porto V - Porto Ocidental
6	6	ACES Almada-Seixal	35	35	ACES Grande Porto VI - Porto Oriental
7	7	ACES Alto Ave - Guimarães, Vizela e Terras de Bast	36	36	ACES Grande Porto VII - Gaia
8	8	ACES Alto Minho	37	37	ACES Grande Porto VIII - Espinho/Gaia
9	9	ACES Alto Trás-os-Montes - Alto Tâmega e Barroso	38	38	ACES Guarda
10	10	ACES Alto Trás-os-Montes - Nordeste	39	39	ACES Leziria
11	11	ACES Amadora	40	40	ACES Lisboa Central
12	12	ACES Arco Ribeirinho	41	41	ACES Lisboa Norte
13	13	ACES Arrábida	42	42	ACES Lisboa Ocidental e Oeiras
14	14	ACES Ave - Famalicão	43	43	ACES Loures-Odivelas
15	15	ACES Baixo Alentejo	44	44	ACES Matosinhos
16	16	ACES Baixo Mondego	45	45	ACES Médio Tejo
17	17	ACES Baixo Vouga	46	46	ACES Oeste Norte
18	18	ACES Beira Interior Sul	47	47	ACES Oeste Sul
19	19	ACES Cascais	48	48	ACES Pinhal Interior Norte
20	20	ACES Cávado I - Braga	49	49	ACES Pinhal Interior Sul
21	21	ACES Cávado II - Gerês e Cabreira	50	50	ACES Pinhal Litoral
22	22	ACES Cávado III - Barcelos e Esposende	51	51	ACES São Mamede
23	23	ACES Cova da Beira	52	52	ACES Sintra
24	24	ACES Dão Lafões	53	53	ACES Tâmega I - Baixo Tâmega
25	25	ACES Douro I - Marão e Douro Norte	54	54	ACES Tâmega II - Vale do Sousa Sul
26	26	ACES Douro II - Douro Sul	55	55	ACES Tâmega III - Vale do Sousa Norte
27	27	ACES Entre Douro e Vouga I - Feira e Arouca			
28	28	ACES Entre Douro e Vouga II - Aveiro Norte			
29	29	ACES Estuário do Tejo			

Figura 27 - Representação dos registos da tabela ACES no DW

O Package "Localizacao", apresenta 55 registo a serem guardados. Este valor corresponde às coordenadas geográficas dos anteriormente apresentados 55 agrupamentos de centros de saúde. Na figura 28, é possível consultar as 55 localizações.

	IDLocalizacao	Localizacao		IDLocalizacao	Localizacao
1	1	37.0274264, -7.9395984	29	29	40.5353285, -7.2724426
2	2	37.1387554, -8.5445093	30	30	40.6405055, -8.6537539
3	3	37.383008, -7.7293275	31	31	40.6565861, -7.9124712
4	4	38.0153039, -7.8627308	32	32	40.8472225, -8.4656674
5	5	38.3841359, -8.5132031	33	33	40.9263169, -8.5478813
6	6	38.5324373, -8.8627453	34	34	41.0022421, -8.642018
7	7	38.6672076, -9.0543329	35	35	41.0953745, -7.8123805
8	8	38.6678522, -9.1875777	36	36	41.1278433, -8.580396
9	9	38.6914388, -9.3184304	37	37	41.1548155, -8.2163769
10	10	38.69904, -9.3818008	38	38	41.1694498, -8.6126428
11	11	38.7306317, -9.1510414	39	39	41.1738444, -8.5612238
12	12	38.7396924, -9.168506	40	40	41.1741353, -8.6691662
13	13	38.7519209, -9.2753636	41	41	41.1918541, -8.6590693
14	14	38.7573081, -9.23429	42	42	41.2333945, -8.6187474
15	15	38.8012718, -9.1137868	43	43	41.2719719, -8.0791631
16	16	38.8442031, -7.5826619	44	44	41.2793583, -8.2787457
17	17	38.9528113, -8.9911154	45	45	41.2968711, -7.7483727
18	18	39.0917759, -9.2600341	46	46	41.3469246, -8.4822648
19	19	39.2371689, -8.6908895	47	47	41.355519, -8.746653
20	20	39.2967086, -7.4284755	48	48	41.3835805, -8.4157518
21	21	39.410844, -9.1396215	49	49	41.4417064, -8.1722408
22	22	39.478072, -8.5404429	50	50	41.5317419, -8.6178922
23	23	39.7495331, -8.807683	51	51	41.5518714, -8.4126362
24	24	39.7510898, -7.9203657	52	52	41.6298661, -8.3596897
25	25	39.8211367, -7.5036651	53	53	41.7056054, -8.8252713
26	26	40.1151262, -8.2473136	54	54	41.741781, -7.4731648
27	27	40.2150023, -8.4071805	55	55	41.8069684, -6.7587977
28	28	40.2677741, -7.4995083			

Figura 28 - Representação dos registos da tabela Localizacao no DW

O Package "Periodo", apresenta 120 registos a serem integrados na tabela Periodo no DW. Este valor significa que foram realizados 120 movimentos de registo de informação sobre o programa da diabetes. Na figura 29, é possível consultar os períodos dos registos.

	IDPeriodo	Ano	mes	mes_num	trim		IDPeriodo	Ano	mes	mes_num	trim		IDPeriodo	Ano	mes	mes_num	trim
1	1	2014	1	janeiro	1	41	41	2014	5	maio	2	81	81	2014	9	setembro	3
2	2	2015	1	janeiro	1	42	42	2015	5	maio	2	82	82	2015	9	setembro	3
3	3	2016	1	janeiro	1	43	43	2016	5	maio	2	83	83	2016	9	setembro	3
4	4	2017	1	janeiro	1	44	44	2017	5	maio	2	84	84	2017	9	setembro	3
5	5	2018	1	janeiro	1	45	45	2018	5	maio	2	85	85	2018	9	setembro	3
6	6	2019	1	janeiro	1	46	46	2019	5	maio	2	86	86	2019	9	setembro	3
7	7	2020	1	janeiro	1	47	47	2020	5	maio	2	87	87	2020	9	setembro	3
8	8	2021	1	janeiro	1	48	48	2021	5	maio	2	88	88	2021	9	setembro	3
9	9	2022	1	janeiro	1	49	49	2022	5	maio	2	89	89	2022	9	setembro	3
10	10	2023	1	janeiro	1	50	50	2023	5	maio	2	90	90	2023	9	setembro	3
11	11	2014	2	fevereiro	1	51	51	2014	6	junho	2	91	91	2014	10	outubro	4
12	12	2015	2	fevereiro	1	52	52	2015	6	junho	2	92	92	2015	10	outubro	4
13	13	2016	2	fevereiro	1	53	53	2016	6	junho	2	93	93	2016	10	outubro	4
14	14	2017	2	fevereiro	1	54	54	2017	6	junho	2	94	94	2017	10	outubro	4
15	15	2018	2	fevereiro	1	55	55	2018	6	junho	2	95	95	2018	10	outubro	4
16	16	2019	2	fevereiro	1	56	56	2019	6	junho	2	96	96	2019	10	outubro	4
17	17	2020	2	fevereiro	1	57	57	2020	6	junho	2	97	97	2020	10	outubro	4
18	18	2021	2	fevereiro	1	58	58	2021	6	junho	2	98	98	2021	10	outubro	4
19	19	2022	2	fevereiro	1	59	59	2022	6	junho	2	99	99	2022	10	outubro	4
20	20	2023	2	fevereiro	1	60	60	2023	6	junho	2	100	100	2023	10	outubro	4
21	21	2014	3	março	1	61	61	2014	7	julho	3	101	101	2014	11	novem...	4
22	22	2015	3	março	1	62	62	2015	7	julho	3	102	102	2015	11	novem...	4
23	23	2016	3	março	1	63	63	2016	7	julho	3	103	103	2016	11	novem...	4
24	24	2017	3	março	1	64	64	2017	7	julho	3	104	104	2017	11	novem...	4
25	25	2018	3	março	1	65	65	2018	7	julho	3	105	105	2018	11	novem...	4
26	26	2019	3	março	1	66	66	2019	7	julho	3	106	106	2019	11	novem...	4
27	27	2020	3	março	1	67	67	2020	7	julho	3	107	107	2020	11	novem...	4
28	28	2021	3	março	1	68	68	2021	7	julho	3	108	108	2021	11	novem...	4
29	29	2022	3	março	1	69	69	2022	7	julho	3	109	109	2022	11	novem...	4
30	30	2023	3	março	1	70	70	2023	7	julho	3	110	110	2023	11	novem...	4
31	31	2014	4	abril	2	71	71	2014	8	agosto	3	111	111	2014	12	dezembro	4
32	32	2015	4	abril	2	72	72	2015	8	agosto	3	112	112	2015	12	dezembro	4
33	33	2016	4	abril	2	73	73	2016	8	agosto	3	113	113	2016	12	dezembro	4
34	34	2017	4	abril	2	74	74	2017	8	agosto	3	114	114	2017	12	dezembro	4
35	35	2018	4	abril	2	75	75	2018	8	agosto	3	115	115	2018	12	dezembro	4
36	36	2019	4	abril	2	76	76	2019	8	agosto	3	116	116	2019	12	dezembro	4
37	37	2020	4	abril	2	77	77	2020	8	agosto	3	117	117	2020	12	dezembro	4
38	38	2021	4	abril	2	78	78	2021	8	agosto	3	118	118	2021	12	dezembro	4
39	39	2022	4	abril	2	79	79	2022	8	agosto	3	119	119	2022	12	dezembro	4
40	40	2023	4	abril	2	80	80	2023	8	agosto	3	120	120	2023	12	dezembro	4

Figura 29 - Representação dos registo da tabela Periodo no DW

5.2.2.1.4 OLE DB Destination

Por último, o OLE DB Destination, é responsável por estabelecer a conexão com a base de dados e integrar os dados previamente transformados nas tabelas respetivas do DW.

Concluídas as quatro tarefas realizadas no Integration Services relativas aos packages das tabelas de dimensão, decorre a segunda etapa, com a realização do Data Flow Task, aplicado ao package "Facto".

5.2.2.2 Data Flow Task Facto

Na figura 30 podemos observar as etapas da tarefa de Data Flow Task para a tabela principal "Facto".

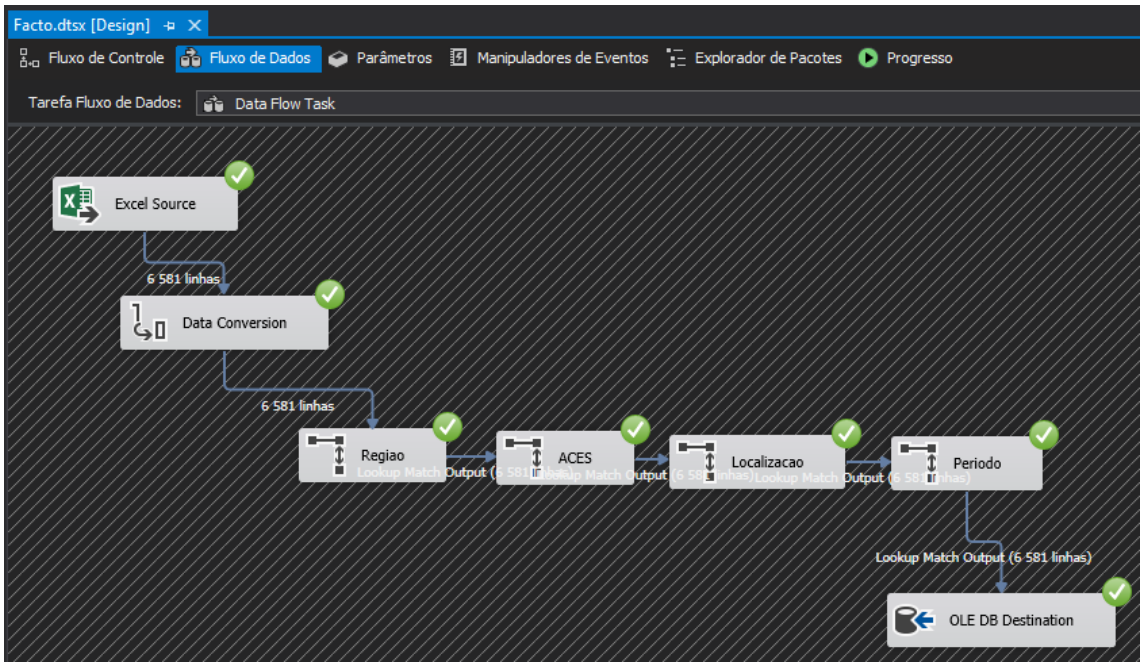

Figura 30 - Processo de fluxo de dados do package Facto

O Data Flow Task do package Facto é constituido pelos seguintes processos:

- Excel Source
- Data Conversion
- Regiao
- ACES
- Localizacao
- Periodo
- OLE DB Destination

5.2.2.2.1 Excel Source

A primeira etapa da tarefa do Data Flow Task é responsável pelo carregamento dos dados do dataset, a partir do formato xls, para serem analisados e tratados no sistema de ETL. Foram carregados 6581 registos para serem tratados nas respetivas packages. Esta tarefa permite selecionar os dados relevantes, por exemplo, no package Regiao foi apenas

selecionada a informação do dataset original relativa à coluna Regiao, permitindo trabalhar dados relevantes para esta tabela. O mesmo procedimento foi aplicado aos restantes packages.

5.2.2.2.2 Data Conversion

A etapa seguinte "Data Conversion" é responsável por converter o tipo de dados da fonte (dataset), para o tipo de dados do recetor dos dados (DW), com a conversão dos dados de xls para SQL statements. Com este procedimento, os dados nulos e inválidos foram desconsiderados. No dataset objeto de estudo, os 6581 registos utilizados estão em conformidade, para serem utilizados no processo de carregamento para o DW.

5.2.2.2.3 Regiao

Nesta etapa é utilizada a ferramenta "Lookup". Esta ferramenta permite realizar transformações aos dados do dataset. A integração de dados, melhoria no desempenho no processo de integração de dados e tratamento de dados duplicados são algumas das tarefas que esta ferramenta permite realizar. No desenvolvimento deste projeto esta ferramenta é utilizada para realizar o processo de integração de dados.

É realizada a substituição dos valores da coluna IDRegiao da tabela Facto pelos valores da coluna IDRegiao da tabela de dimensão Regiao. Esta substituição dos dados permite que os dados na tabela Facto na coluna IDRegiao apenas contenha valores inteiros. Nesta etapa foram carregados 6581 valores, e após o processo realizado pela ferramenta, manteve-se a quantidade de informação. As figuras 31 e 32, permitem visualizar o exemplo das alterações nos dados.

Na figura 31, é possível consultar os dados relativamente a coluna Regiao no dataset inicial, antes da utilização da ferramenta.

Região
Região de Saúde do Alentejo
Região de Saúde do Centro
Região de Saúde LVT
Região de Saúde Norte
Região de Saúde Norte
Região de Saúde Norte
Região de Saúde Norte
Região de Saúde Norte
Região de Saúde do Algarve
Região de Saúde do Centro

Figura 31 - Dados coluna Regiao dataset inicial

Na figura 32, é possível visualizar como os dados são apresentados na tabela Facto após a utilização da ferramenta.

IDRegiao
1
1
1
1
1
1
1
1

Figura 32 - Dados coluna IDRegiao tabela Facto após processo

5.2.2.2.4 ACES

Nesta etapa é utilizada a ferramenta "Lookup". É realizada a substituição dos valores da coluna IDACES da tabela Facto pelos valores da coluna IDAces da tabela de dimensão ACES. Esta substituição dos dados permite que os dados na tabela Facto na coluna IDACES apenas contenha valores inteiros. Nesta etapa foram carregados 6581 valores, e após o processo realizado pela ferramenta, manteve-se a quantidade de informação. As figuras 33 e 34, permitem visualizar o exemplo das alterações nos dados.

Na figura 33, é possível consultar os dados relativamente a coluna ACES no dataset inicial, antes da utilização da ferramenta.

ACES
ACES Baixo Alentejo
ACES Dão Lafões
ACES Cascais
ACES Ave - Famalicão
ACES Cávado I - Braga
ACES Cávado III - Barcelos e Esposende
ACES Tâmega I - Baixo Tâmega
ACES Tâmega III - Vale do Sousa Norte
ACES Algarve III - Algarve Sotavento

Figura 33 - Dados coluna ACES dataset inicial

Na figura 34, é possível visualizar como os dados são apresentados na tabela Facto após a utilização da ferramenta.

IDAces
2
1
51
51
2
15
51
15

Figura 34 - Dados coluna IDACES tabela Facto após processo

5.2.2.2.5 Localizacao

Nesta etapa é utilizada a ferramenta “Lookup”. É realizada a substituição dos valores da coluna IDLocalizacao da tabela Facto pelos valores da coluna IDLocalizacao da tabela de dimensão Localizacao. Esta substituição dos dados permite que os dados na tabela Facto na coluna IDLocalizacao apenas contenha valores inteiros. Nesta etapa foram carregados 6581 valores, e após o processo realizado pela ferramenta, manteve-se a quantidade de informação. As figuras 35 e 36, permitem visualizar o exemplo das alterações nos dados.

Na figura 35, é possível consultar os dados relativamente a coluna IDLocalizacao no dataset inicial, antes da utilização da ferramenta.

Localização Geográfica
38.0153039, -7.8627308
40.6565861, -7.9124712
38.69904, -9.3818008
41.3835805, -8.4157518
41.5518714, -8.4126362
41.5317419, -8.6178922
41.2719719, -8.0791631
41.2793583, -8.2787457

Figura 35 - Dados coluna Localização Geográfica dataset inicial

Na figura 36, é possível visualizar como os dados são apresentados na tabela Facto após a utilização da ferramenta.

IDLocalizacao
5
16
20
20
5
4
20
4

Figura 36 - Dados coluna IDLocalizacao tabela Facto após processo

5.2.2.2.6 Periodo

Nesta etapa é utilizada a ferramenta "Lookup". É realizada a substituição dos valores da coluna IDPeriodo da tabela Facto pelos valores da coluna IDPeriodo da tabela de dimensão Periodo. Esta substituição dos dados permite que os dados na tabela Facto na coluna IDPeriodo apenas contenha valores inteiros. Nesta etapa foram carregados 6581 valores, e após o processo realizado pela ferramenta, manteve-se a quantidade de informação. As figuras 37 e 38, permitem visualizar exemplo das alterações nos dados.

Na figura 37, é possível consultar os dados relativamente a coluna IDPeriodo no dataset inicial, antes da utilização da ferramenta.

Período
2014-01
2014-01
2014-01
2014-01
2014-01
2014-01
2014-01

Figura 37 - Dados coluna ACES dataset inicial

Na figura 38, é possível visualizar como os dados são apresentados na tabela Facto após a utilização da ferramenta.

IDPeriodo
1
21
41
51
61
71
81
101

Figura 38 - Dados coluna IDPeriodo tabela Facto após processo

5.2.2.2.7 OLE DB Destination

O OLE DB Destination, é responsável por estabelecer a conexão com a base de dados e integrar os dados provenientes das tabelas de dimensão na tabela Facto do DW.

O processo de ETL está concluído com o OLE DB Destination. Ou seja, os dados foram extraídos a partir do dataset, transformados, para serem carregados nas respetivas tabelas da base de dados do DW.

5.3 Cubo OLAP

O Cubo OLAP é uma ferramenta multidimensional de análise e integração do SQL Server Analysis Services (SSAS), utilizado para realizar análises e construir indicadores, a partir de um grande volume de dados do DW. Também, permite a estruturação de modelos de dados multidimensionais, como o modelo em estudo. A utilização desta ferramenta permite estruturar corretamente os dados e prosseguir para os processos seguintes do processo de DCBD.

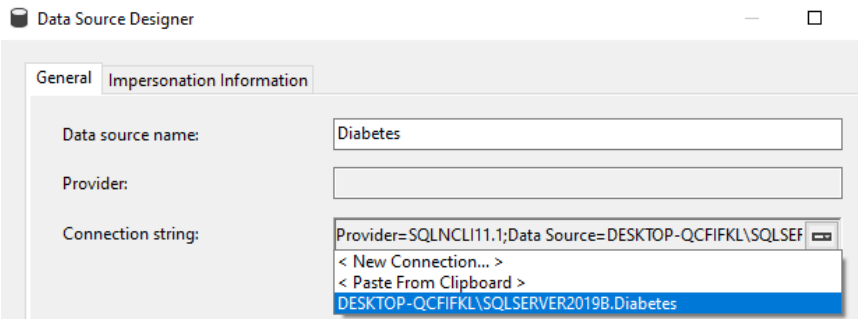

Figura 39 - Criação da conexão com a fonte de dados (DW)

Na figura 39, é possível observar a primeira etapa para o desenvolvimento do Cubo OLAP com a conexão com o DW (é utilizado o mesmo DW definido no processo ETL), para a utilização dos dados. Após a conexão decorre a conceção da estrutura do cubo multidimensional, com a integração das tabelas de dimensão com a tabela Facto, para a partilha dos dados dos registos. Na figura 40, é possível observar a estrutura multidimensional do Cubo OLAP.

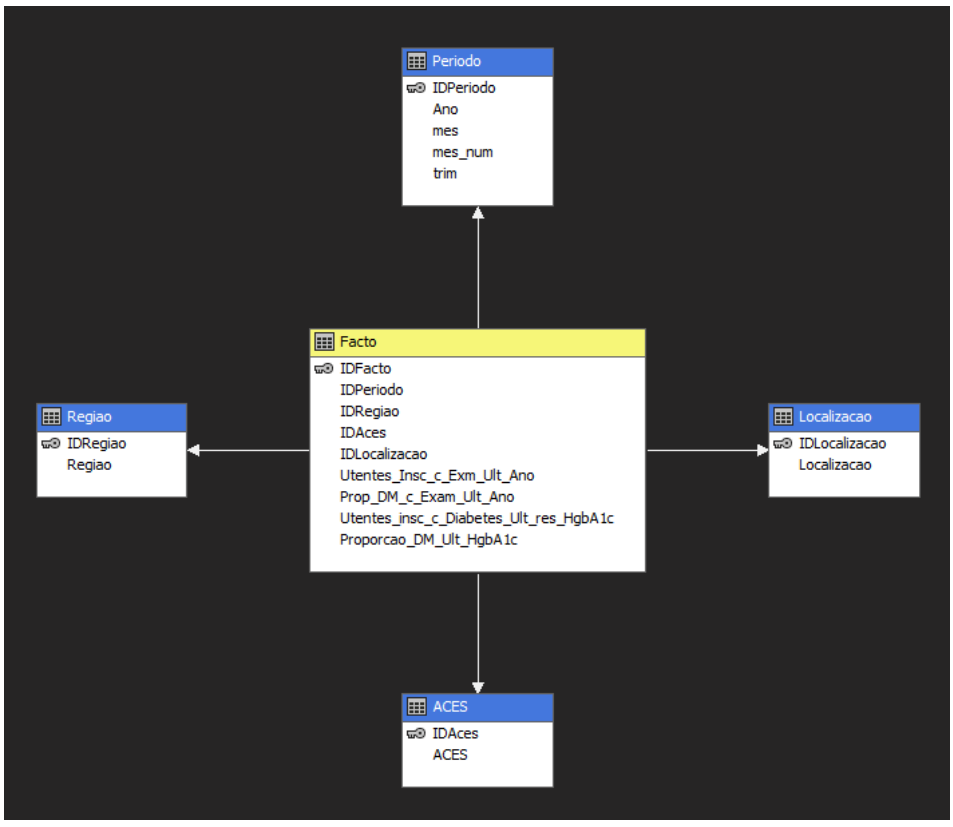

Figura 40 - Modelo estruturado pela ferramenta cubo OLAP

A figura 40, representa o modelo multidimensional desenvolvido pela ferramenta cubo OLAP. O modelo apresenta cinco tabelas interrelacionadas, quatro a cor azul e uma a cor amarela. As quatro tabelas apresentadas a cor azul são tabelas de dimensão, a tabela principal a cor amarela é a tabela de Facto. As tabelas de dimensão contêm informações sobre as regiões, localizações, ACES e período dos registos. A tabela "Regiao" é composta por dois campos, o IDRegiao e a Regiao, a tabela "Periodo" é composta por cinco campos IDPeriodo, Ano representando o ano em que foi realizado o registo, mes e mes_num representam o mês em que foi realizado o registo e trim representando o trimestre do registo. A tabela de dimensão ACES é composta por dois campos IDACES e ACES que guardam informações sobre os ACES, a tabela de Localizacao é também composta por dois campos IDLocalizacao e Localizacao.

5.4 Power BI

O Power BI é uma plataforma de BI para análise de dados. Disponibiliza um conjunto de ferramentas que permitem gerar relatórios e dashboards visualmente criativos e interativos. No contexto deste projeto o Power BI foi a aplicação escolhida, pelo contato anterior com a ferramenta. O facto de a aplicação ser intuitiva e o carregamento de dados para a ferramenta ser facilitado, devido aos desenvolvimentos dos processos anteriores com ferramentas da Microsoft, foram fatores que conduziram à escolha da plataforma.

5.4.1 Processo de criação e análise de dashboards

Neste processo foram desenvolvidas dashboards para representar as análises gerais dos registos, dos padrões de evolução da doença e dos padrões de evolução dos exames de controlo, para facilitar a usabilidade dos dados. Para a criação destas dashboards é necessário realizar a conexão com o cubo OLAP criado no processo anterior. Esta conexão cria uma ligação estruturada entre as diferentes fontes de dados e o Power BI, permitindo que as dashboards estejam em constante atualização com a entrada de novos dados. Na figura 41, é possível consultar o cubo OLAP criado no DW Analysis.

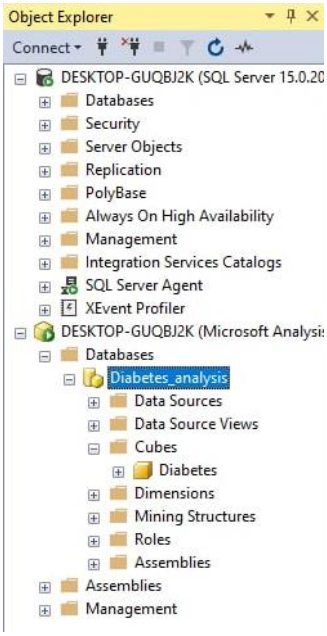

Figura 41 - Cubo OLAP no SSMS

Após a conexão com o cubo OLAP é possível a realização de alguns testes aos dados e a conceção de dashboards de teste para entender melhor algumas ligações entre os dados, tendo sido concebidas quatro dashboards para apoio e suporte à interpretação dos dados, bem como para perceber a utilidade destes para o desenvolvimento do estudo. As dashboards criadas dividem este capítulo em quatro secções:

- Análises de Registos
- Análises por ACES
- Análises por Ano
- Análises por Região

5.4.1.1 Análises de registos

Nesta secção, procedemos à apresentação e análise das dashboards que contêm a informação sobre os registos do dataset. Através da figura 42, é apresentada a dashboard dos registos, com a aplicação de filtros temporais, entre os anos de 2014 e 2023. Os gráficos apresentam informações sobre o número de registo totais, registos por região e registos por ACES. O total de registos é de 6,58 mil registos, sendo a "Região de Saúde Norte" com mais registos, representando cerca de 44% (2,88 mil registos) do total dos registos. Segue-se a "Região de Saúde LVT", que representa 27% (1,79 mil registos) do total dos registos. A "Região de Saúde do Centro", representa 17% (1,08 mil registos) do

total de registos. A "Região de Saúde do Alentejo", representa 7% (0,48 mil registos) do total dos registos. A "Região de Saúde do Algarve", representa 5% (0,36 mil registos) do total de registos. Assim, é possível verificar que o número de registos por ACES é bastante semelhante, ou seja, situa-se entre os 119 e os 120 registos.

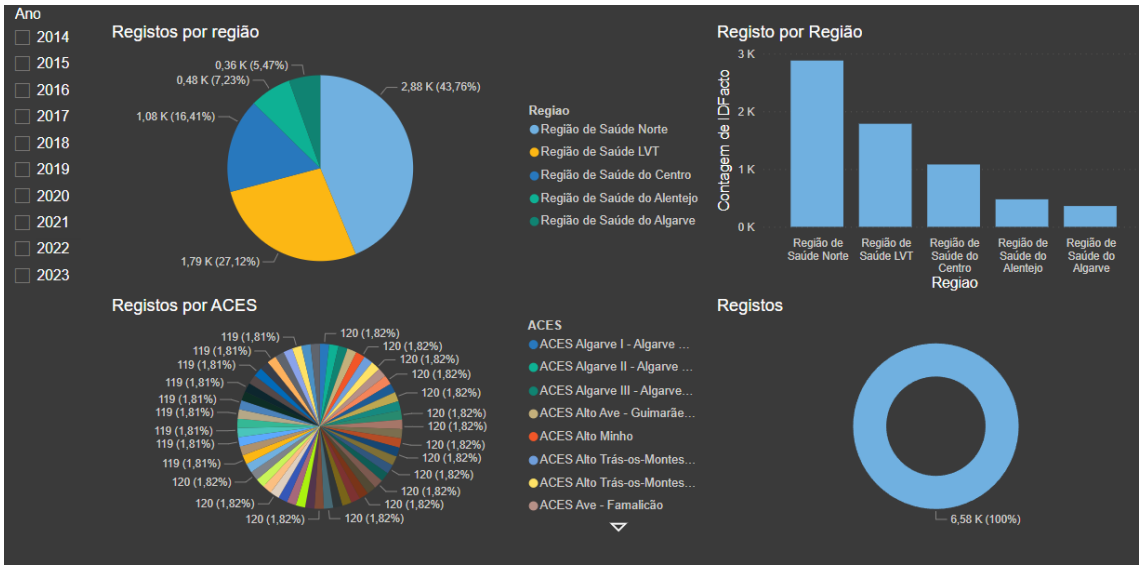

Figura 42 - Dashboard 1 - representação dos registos por região e ACES

A tabela 8, permite consultar o número de registos realizados em cada ano.

Tabela 8 - Número de registos por ano

Ano	Registos
2014	660
2015	660
2016	645
2017	660
2018	660
2019	656
2020	660

2021	660
2022	660
2023	660

A dashboard com o recurso aos filtros temporais permite a consulta de informação relativa aos registos de um determinado ano, para uma pesquisa mais aprofundada sobre os registos de um determinado ano. Para a pesquisa foram escolhidos os anos de 2019 e 2023. O ano de 2019 foi escolhido para análise por ter sido o ano anterior à pandemia Covid-19 e por ter menos registos realizados relativamente aos outros anos. O segundo ano de 2023 foi escolhido pois a pandemia já tinha terminado.

Nas figuras 43 e 44, é possível consultar a mesma tipologia de dashboard, com a aplicação de filtros temporais. A figura 43, apresenta a informação relativa ao ano de 2019, na qual é possível verificar 656 registos, sendo que as regiões mantêm as percentagens de registos. A região norte tem 288 registos, a região LVT tem 180 registos, a região do Centro tem 108 registos, a região do Alentejo tem 44 registos e região do Algarve tem 36 registos. Cada ACES foi responsável por 12 registos, com a exceção das ACES São Mamede, ACES Alentejo Litoral, ACES Alentejo Central e ACES Baixo Alentejo que apenas realizaram 11 registos.

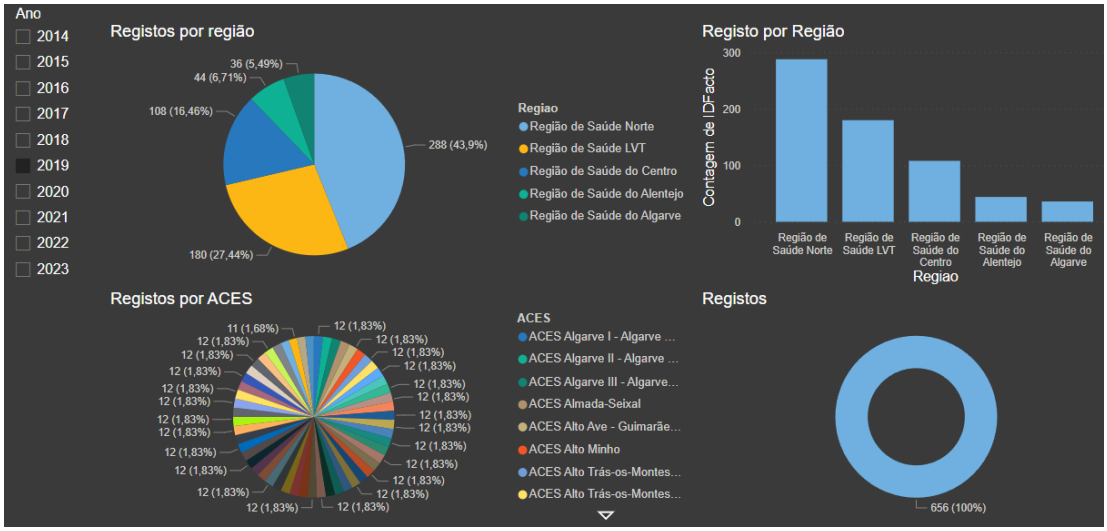

Figura 43 – Dashboard 1 - análise dos registos por região e ACES no ano 2019

A figura 44, contém informação sobre o ano de 2023 onde é possível consultar que foram realizados 660 registos, sendo que as regiões mantêm as percentagens de registos por região, o número de registo por região foi igual ao ano de 2019, com a exceção da região do Alentejo que teve 48 registos. Cada ACES foi responsável por 12 registos.

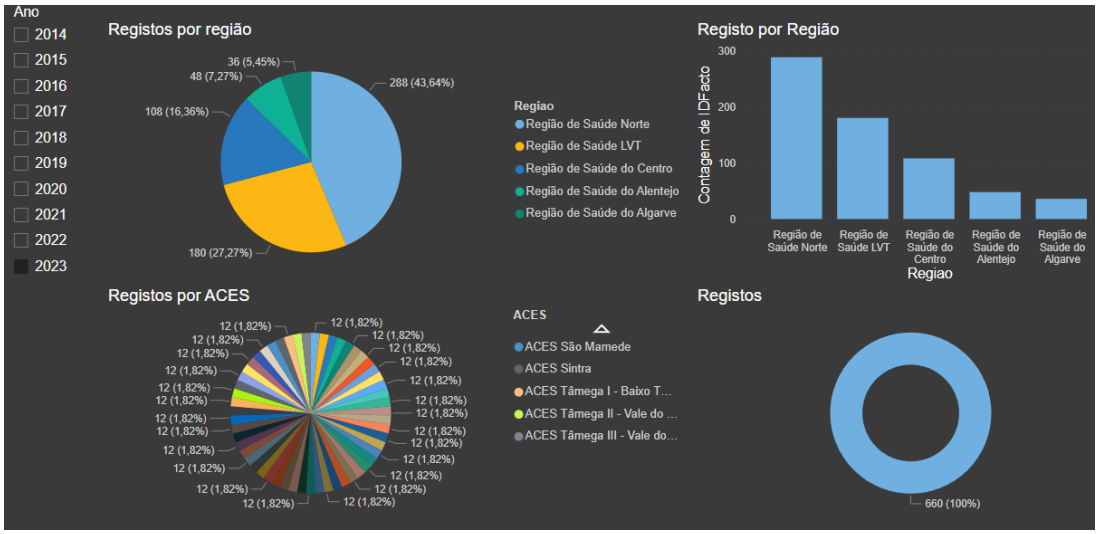

Figura 44 - Dashboard 1 - análise dos registos por região e ACES no ano 2023

5.4.1.2 Análise das regiões

Nesta secção, é apresentada e analisada a dashboard que contém informações sobre a média do número de utentes inscritos com o exame do pezinho e o exame à HgbA1c por ACES em cada região, bem como a média da proporção de utentes inscritos com o exame dos pés e a HgbA1c por ACES em cada região. Na figura 45, é possível visualizar sem filtros temporais os quatros gráficos com a média dos valores dos dois exames e das mesmas proporções por Região.

No gráfico de "Utentes inscritos c/exame pezinho" é possível verificar que a "Região de Saúde do Centro" tem a maior média de utentes inscritos com o exame aos pés (7,8 mil utentes por ACES), seguido da "Região de Saúde do Norte" (7,5 mil utentes por ACES), da "Região de Saúde LVT" (6,4 mil utentes por ACES), da "Região de Saúde do Alentejo" (5,4 mil utentes por ACES) e da "Região de Saúde do Algarve" (3,9 mil utentes por ACES).

No gráfico de "Utentes inscritos c/exame HgbA1c" é possível verificar que a "Região de Saúde do Centro" tem a maior média de utentes inscritos com o exame ao HgbA1c (6,5

mil utentes por ACES), seguido da "Região de Saúde LVT" (5,1 mil utentes por ACES), da "Região de Saúde do Norte" (5 mil utentes por ACES), da "Região de Saúde do Alentejo" (3,9 mil utentes por ACES) e da "Região de Saúde do Algarve" (3,4 mil utentes por ACES).

No gráfico de "Proporção de utentes inscritos c/exame pezinho" é visível que a "Região de Saúde do Norte" (58,44%) tem a maior média de proporção de utentes inscritos com exames aos pés realizado em relação às restantes regiões, segue-se a "Região de Saúde do Alentejo" (44,13%), a "Região de Saúde do Centro" (40,28%), a "Região de Saúde do Algarve" (37,86%) e por último a "Região de Saúde LVT" (35,61%). Assim, é possível verificar que o "Região de Saúde Norte" e a "Região de Saúde do Alentejo" apesar de terem menos exames realizados que a "Região de saúde Centro", têm uma maior percentagem de controlo da doença nos utentes inscritos.

No gráfico de "Proporção de utentes inscritos c/exame HgbA1c" é visível que a "Região de Saúde do Norte" (39,09%) tem a maior média de proporção de utentes inscritos com exames ao HgbA1c realizados em relação ao resto das regiões, segue-se a "Região de Saúde do Centro" (34,66%), a "Região de Saúde do Algarve" (32,71%), a "Região de Saúde do Alentejo" (32,43%) e por último a "Região de Saúde LVT" (28,36%). Após a análise do gráfico da figura 45 é também possível verificar que a "Região de Saúde do Norte" tem uma maior percentagem de controlo da doença com a realização do exame ao HgbA1c. A "Região de Saúde LVT" na mesma tipologia de exames, tem muitos utentes com o exame realizado, mas é visível que é a região que deveria aumentar o número de exames aos utentes.

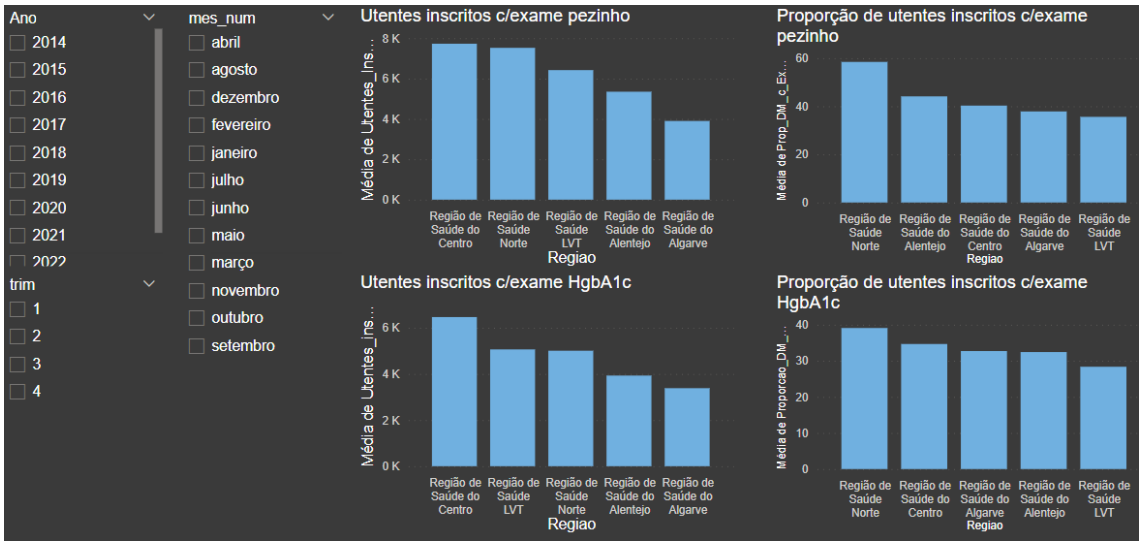

Figura 45 - Dashboard 2 - análise das regiões

As figuras 46, 47 e 48, apresentam a dashboard da figura 45, com filtros temporais. Nesta análise foram escolhidos três anos, permitindo uma melhor compreensão dos dados e visualizar como a pandemia Covid-19 afetou a evolução destes exames de controlo em cada região. São apresentadas informações relativas ao ano de 2017, anterior à pandemia, o ano de 2020 início da pandemia Covid-19 em Portugal e o ano de 2023 posterior à pandemia. Na figura 46, é possível consultar informação relativa ao ano de 2017 no mês de dezembro. A ordem das Regiões nos diferentes gráficos no ano de 2017 manteve-se, à exceção do gráfico "proporção de utentes inscritos c/exame HgbA1c" onde a "Região de Saúde do Alentejo" assume a segunda posição. No entanto, os valores observados são superiores aos valores médios. É possível verificar que a "Região de Saúde do Centro" ao contrário dos 7,8 mil utentes com exame dos pés por ACES teve cerca de 10,7 mil utentes por ACES no mês de dezembro de 2017, e cerca de 10 mil utentes com o exame ao HgbA1c realizado. No gráfico de "Proporção de utentes inscritos c/exame pezinho" é visível que a "Região de Saúde do Norte" tem uma proporção de 81,38% dos utentes inscritos com o exame dos pés realizado. No exame ao HgbA1c é possível verificar um aumento em relação à média (39,09%) tendo sido registado 65,04% dos utentes com este exame realizado.

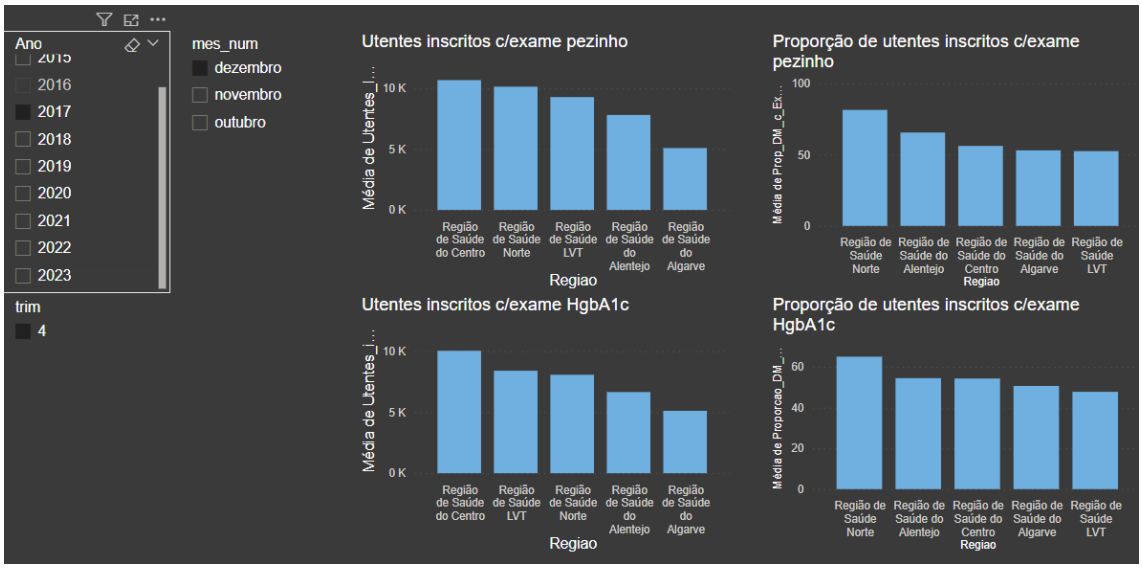

Figura 46 – Dashboard 2 - análise das regiões ano 2017

Na figura 47, é possível consultar informação sobre os registos de dezembro de 2020, em comparação com o ano de 2017 é possível verificar que o número de utentes inscritos com exame dos pés e utentes inscritos com exame à HgbA1c é semelhante. Este valor significa que apesar da pandemia instalada foi possível manter o controlo na evolução da doença. Na proporção de utentes inscritos c/exame pezinho e a proporção de utentes inscritos com exame à HgbA1c é possível verificar uma redução nos valores em relação ao ano de 2017. Esta redução significa que apesar de serem detetados novos casos, houve uma queda no controlo da doença com a pandemia.

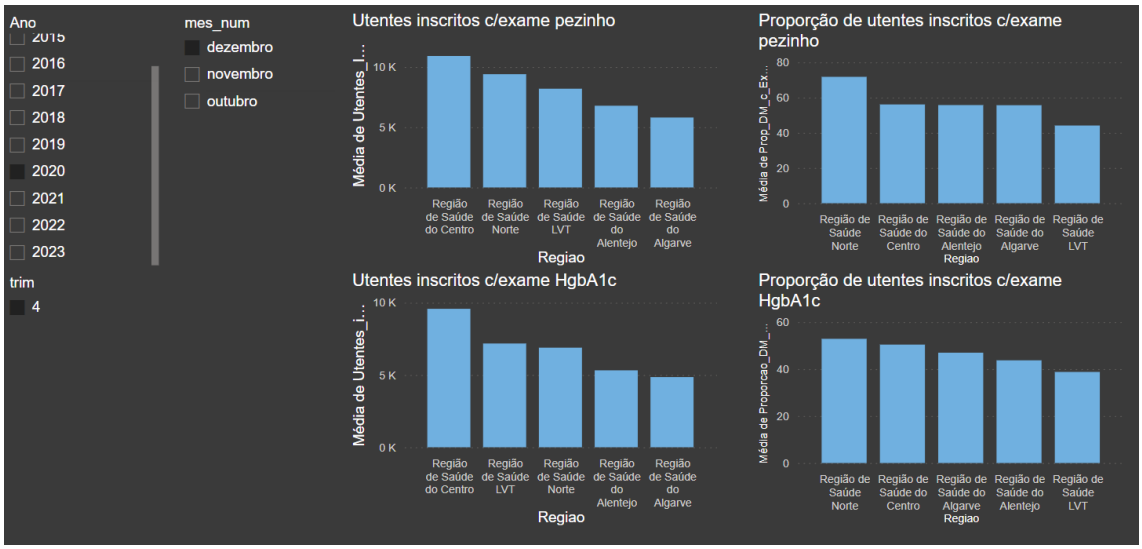

Figura 47 – Dashboard 2 - análise das regiões ano 2020

Na figura 48, é possível consultar informação relativa a dezembro de 2023, em comparação com o ano de 2017 é possível verificar que a doença evoluiu consideravelmente, tendo sido registados cerca de 14,5 mil utentes inscritos com o exame aos pés e cerca de 12,5 mil utentes inscritos com o exame ao HgbA1c realizado por ACES na "Região de Saúde Centro". Também o controlo da doença aumentou, sendo possível verificar que em média cerca 91,04 % dos utentes com DMe realizaram o exame dos pés e 73,33% dos utentes inscritos têm o exame ao HgbA1c realizado nos ACES da "Região de Saúde Norte".

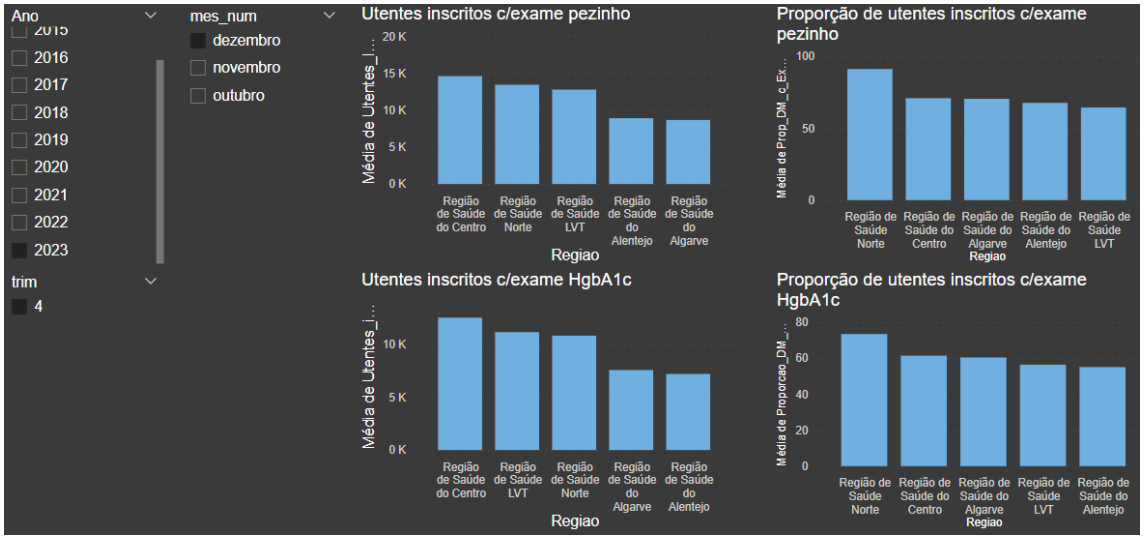

Figura 48 - Dashboard 2 - análise das regiões ano 2023

5.4.1.3 Análises das ACES por região

Nesta última secção, é apresentada e analisada a dashboard com informações sobre a proporção de utentes inscritos com o exame aos pés realizado aos utentes inscritos com o exame à HgbA1c, por ACES. Nas figuras 49 e 50, é possível visualizar os dois gráficos presentes na dashboard sem a aplicação de filtros, no entanto, estão disponíveis dois filtros que podem ser utilizados para visualizar apenas informação sobre uma determinada região num determinado ano. Assim, é possível consultar os ACES por região. A figura 49, representa a média anual da proporção de utentes inscritos com o exame em todos ACES, com a ordenação do ACES com a maior proporção de utentes inscritos com exame dos pés, para a ACES com menor proporção. O ACES Grande Porto I é a que tem a maior proporção (66,78%) de utentes inscritos com o exame realizado e o ACES Cova da Beira a menor (21,90%).

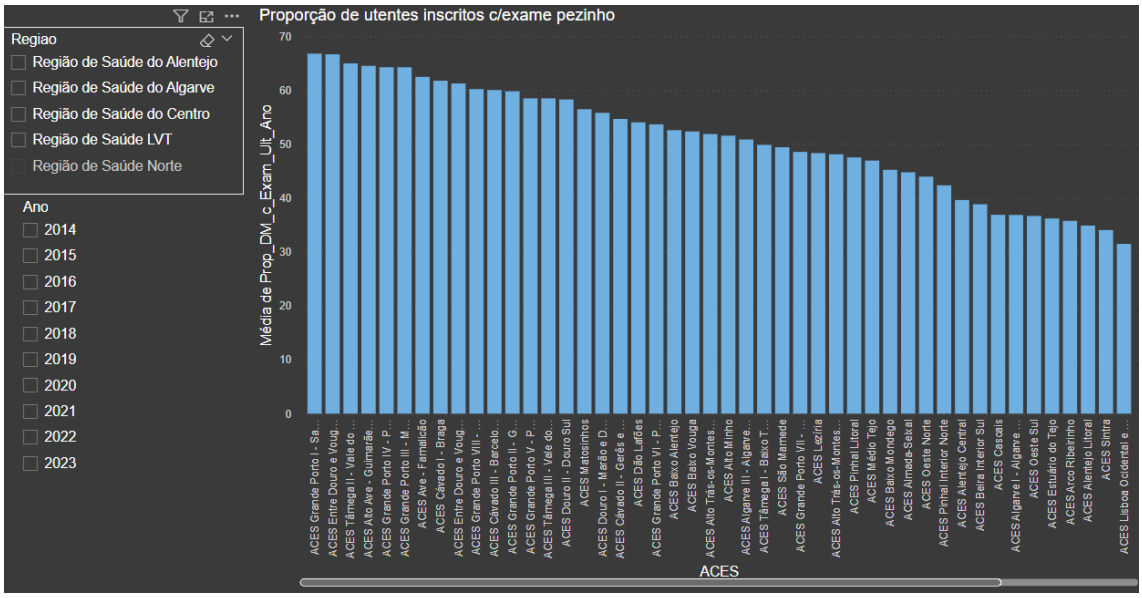

Figura 49 - Dashboard 3 - Análises ACES por região, proporção exame pés

A figura 49, apresenta um gráfico com a média anual da proporção de utentes inscritos com exame ao HgbA1c em todas os ACES. A média da proporção encontra-se ordenada do ACES com a maior proporção para o ACES com menor proporção. Na média da proporção de utentes inscritos com exame a HgbA1c o ACES Cávado I apresenta a maior proporção (47,93%) e o ACES Lisboa Norte apresenta a menor proporção (23,17%).

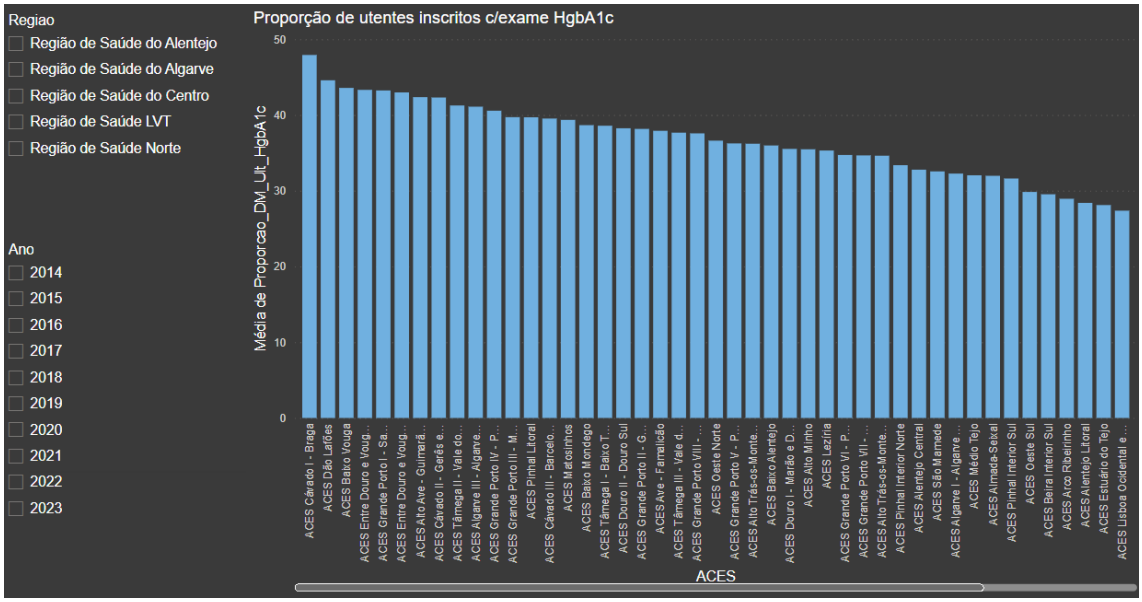

Figura 50 - Dashboard 3 - Análises ACES por região, proporção exame HgbA1c

Nas figuras 51, 52, 53 e 54, é possível consultar a dashboard presente nas figuras 49 e 50 com a aplicação de filtros. É possível visualizar apenas os ACES de uma determinada região e ano. Na análise desta dashboard foram escolhidas informações sobre ano 2019 aleatoriamente e foram escolhidas também aleatoriamente as regiões para estudo. As figuras 51 e 52, permitem a consulta da proporção média de utentes inscritos com o exame aos pés realizado por ACES. Na figura 51, é possível consultar a informação relativa à "Região de Saúde do Centro" no ano de 2019. Os ACES da "Região de Saúde do Centro" apresentam alguma discrepância nos valores, podendo significar a falta de recursos em certos ACES da região. O ACES Dão Lafões com 59,42% apresenta mais do dobro da proporção do último ACES Cova da Beira com 23,83%.

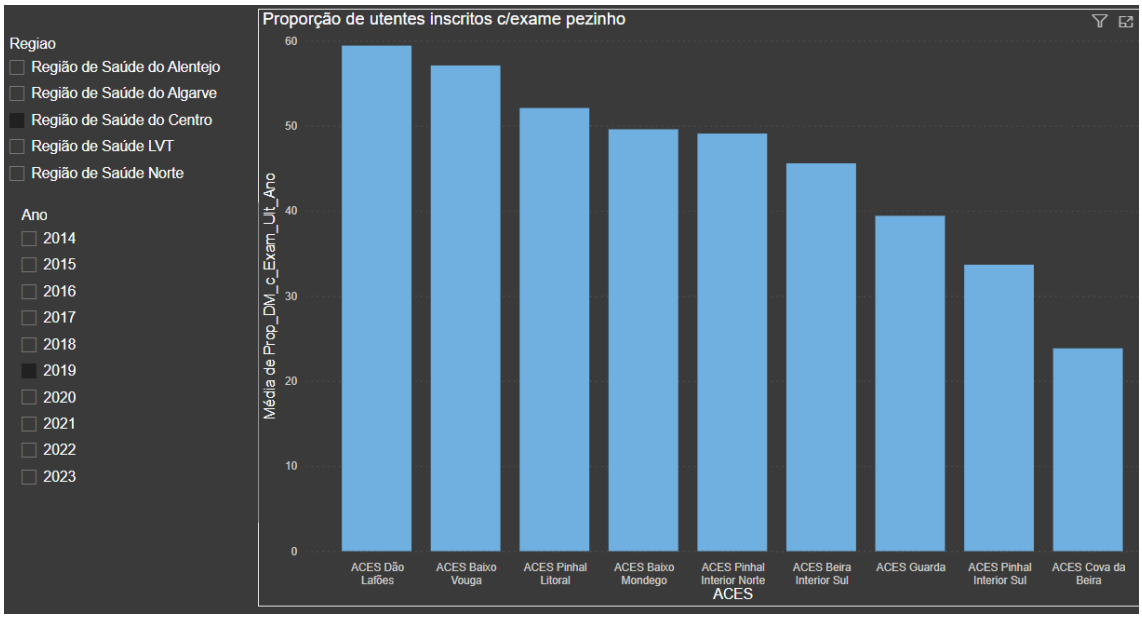

Figura 51 - Dashboard 3 - Análises das ACES na região do centro, proporção exame dos pés

Na figura 52, é possível consultar informação sobre a "Região de Saúde Norte" no ano de 2019. Nos ACES da "Região de Saúde Norte" é possível verificar a tendência negativa presente no gráfico, no entanto, os valores não apresentam uma discrepância igual ao da "Região da Saúde do Centro" cuja variação é de 35,6%. O ACES Entre Douro e Vouga II (75,42%) apresenta a maior proporção de utentes inscritos com o exame realizado no ano de 2019, sendo o ACES do Grande Porto VII é que apresenta uma menor proporção (52,17%).

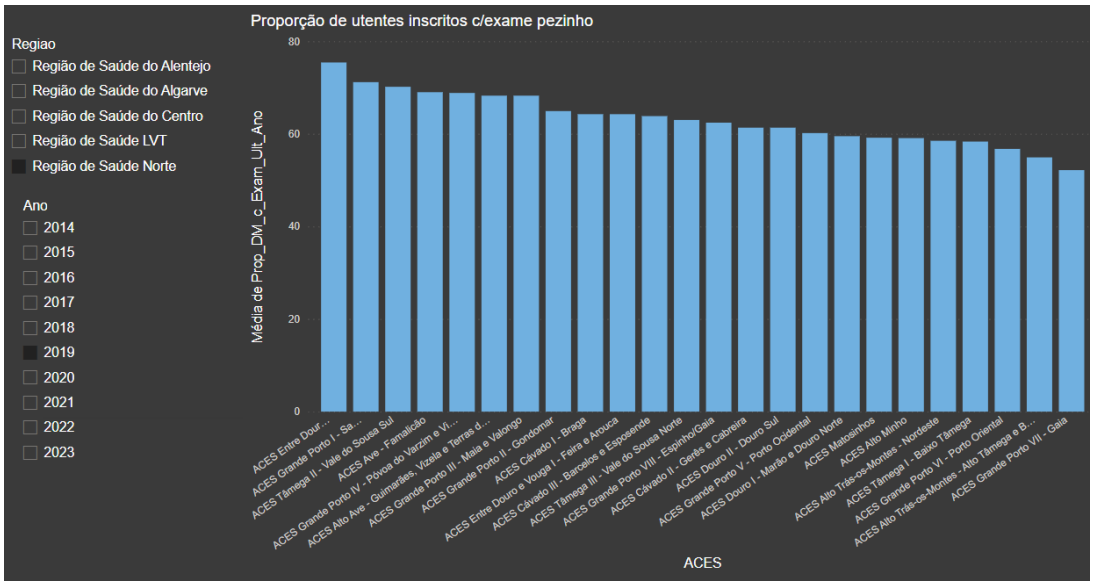

Figura 52 - Dashboard 3 -Análises das ACES na região norte, proporção exame dos pés

Após a análise da proporção de utentes inscritos com exame dos pés visível nesta dashboard, é possível verificar que as Regiões de Saúde Norte e do Alentejo, são as regiões com um controlo da doença semelhante nos diferentes ACES no ano de 2019. No entanto, a "Região de Saúde Norte" apresenta a maior proporção de utentes inscritos em todas os ACES. O ACES Entre Douro e Vouga II é o ACES com maior proporção da "Região de Saúde Norte" (75,42), nas outras regiões o ACES com maior proporção obteve cerca de 60%. Na "Região de Saúde Centro" foi o ACES Dão Lafões como referido, na "Região de Saúde do Alentejo" o ACES Baixo Alentejo, "Região de Saúde do Algarve" o ACES Algarve III e na "Região de Saúde LVT" o ACES Médio Tejo.

As figuras 53 e 54, permitem a consulta da proporção média de utentes inscritos com o exame ao HgbA1c realizado por ACES. Na figura 53, é possível consultar informação relativamente á "Região de Saúde do Centro" no ano de 2019. Os ACES da "Região de Saúde do Centro" apresentam alguma discrepância nos valores, mas não é tão acentuada quanto no exame dos pés. É visível que em relação ao exame aos pés, o exame ao HgbA1c tem uma menor proporção de utentes inscritos.

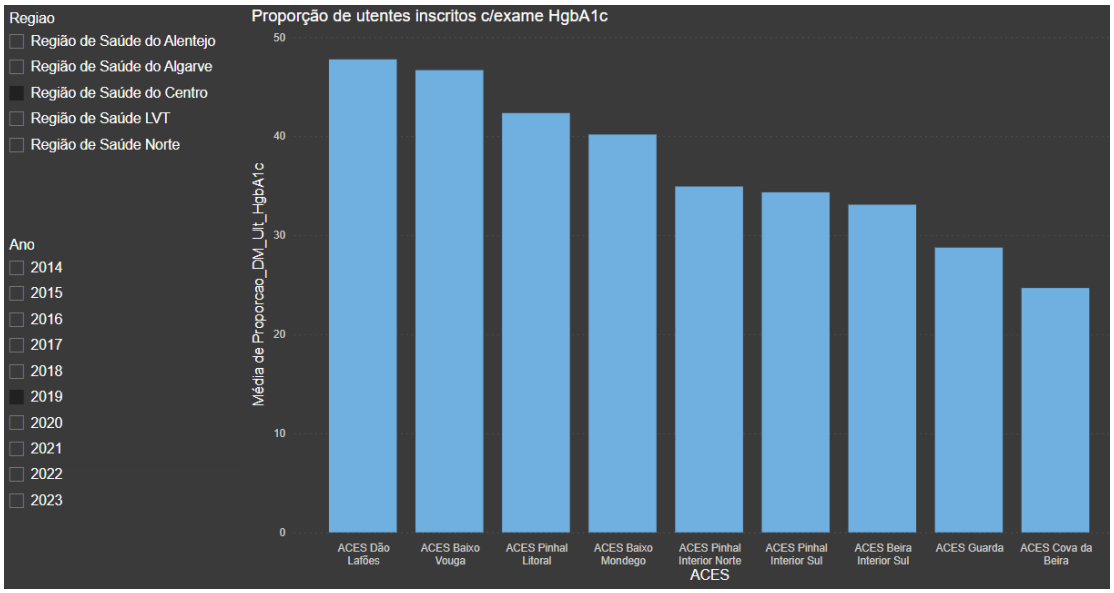

Figura 53 - Dashboard 3 - Análises das ACES na região centro, proporção exame HgbA1c

Na figura 54, é possível consultar informação sobre a "Região de Saúde Norte" no ano de 2019. Nos ACES da "Região de Saúde Norte" os valores não apresentam uma discrepância igual ao da "Região da Saúde do Centro". O ACES Cávado I (48,33%) apresenta a maior proporção de utentes inscritos com o exame realizado ao HgbA1c no ano de 2019, o ACES do Grande Porto V é o que apresenta uma menor proporção (36,25%).

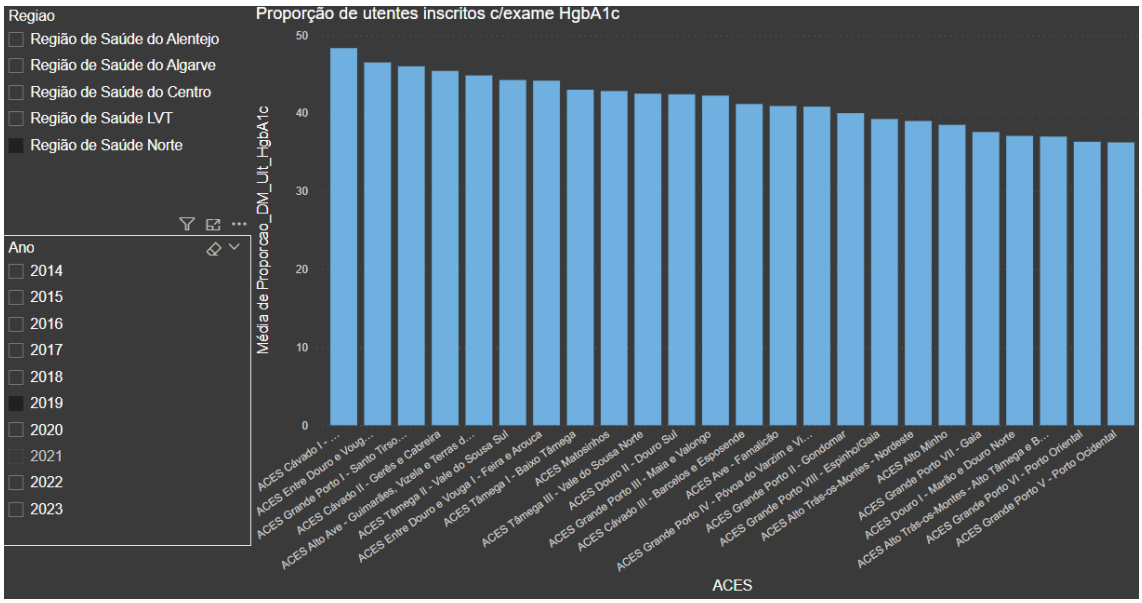

Figura 54 - Dashboard 3 - Análises das ACES na região norte, proporção exame HgbA1c

Em conformidade com a primeira parte desta dashboard a "Região de Saúde Norte" e a "Região de Saúde do Alentejo" apresentam os valores das proporções semelhantes entre os ACES da região. Os valores da proporção de utentes inscritos com o exame ao HgbA1c são significativamente mais baixos relativamente aos valores da proporção de utentes inscritos com exame dos pés em todas as regiões e ACES. O ACES Cávado I é o ACES em 2019 com a maior proporção (48,33%) na "Região de Saúde Norte", na "Região de Saúde do Centro" foi o ACES Dão Lafões (47,75%), na "Região de Saúde do Alentejo" foi o ACES Baixo Alentejo (39,36%), na "Região de Saúde do Algarve" foi o ACES Algarve III (45,17%) e na "Região de Saúde LVT" foi o ACES Lezíria (39,35%).

5.4.1.4 Análises das ACES

Nesta secção, é apresentada e analisada a dashboard com informações relativas à média do número de utentes inscritos com o exame do pezinho e o exame à HgbA1c em cada ano para cada ACES, bem como a média da proporção de utentes inscritos com o exame dos pés e a HgbA1c por ACES em cada região. O objetivo principal desta dashboard é a visualização do aumento da DMe ao longos dos anos. Na figura 55, é possível consultar quatro gráficos já apresentados na secção anterior e um novo sistema de filtros, onde é possível filtrar informação por ACES. A figura permite a consulta sem filtros dos quatros gráficos com a média dos valores dos dois exames e das mesmas proporções por ano.

Analisando os quatro gráficos representados na figura 55, é possível verificar uma tendência e um padrão em todos eles. A tendência presente nos quatro gráficos é de aumento, tanto de utentes inscritos com DMe com os dois exames, bem como da proporção de utentes inscritos com os dois exames realizados. O padrão apresentado nos quatros gráficos está relacionado com a pandemia Covid-19, onde é possível visualizar uma quebra nos números anteriormente registados, nos anos de 2020 e 2021. No ano de 2022, com o fim da pandemia os valores voltaram a regularizar. Através da figura 55, é possível visualizar o continuo aumento.

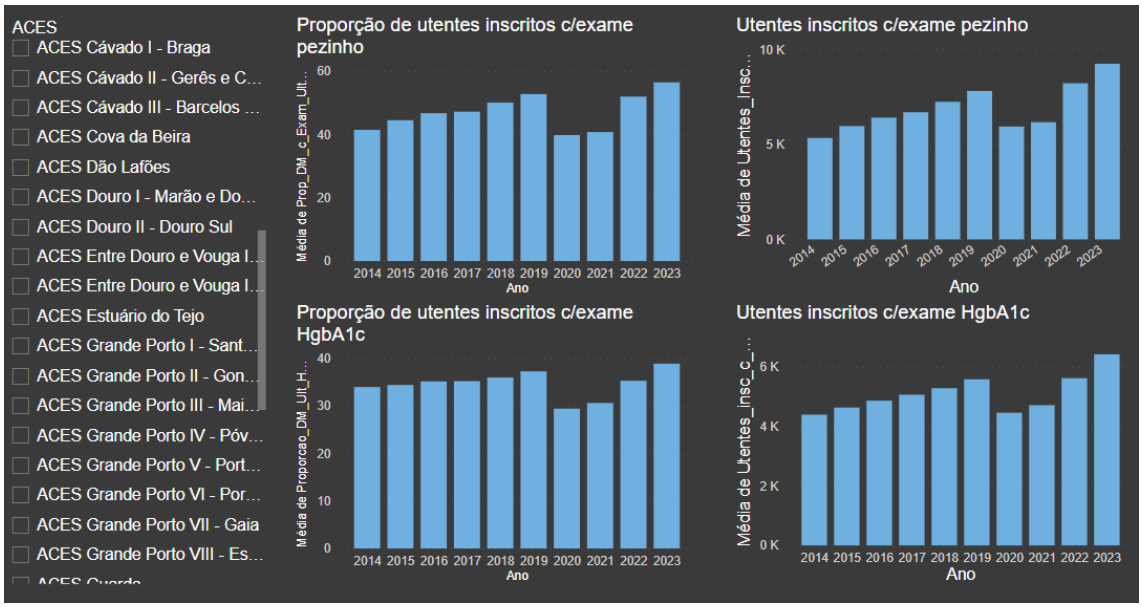

Figura 55 - Dashboard 4 - Análise das ACES

Nas figuras 56 e 57, é possível consultar a dashboard com filtros relativamente aos ACES. Os ACES selecionadas no estudo foram escolhidas aleatoriamente, tendo em consideração não selecionar dois ACES da mesma região. Na figura 56, é possível visualizar a informação relativa ao ACES do Grande Porto VI. Os valores do ACES enquadram-se na média do modelo geral. É possível visualizar o padrão de quebra dos valores nos anos 2020 e 2021, no entanto, é mais difícil visualizar a tendência inicial de crescimento nos quatro gráficos, significando que este ACES apresenta um nível de controlo e acompanhamento da doença mais prolongado.

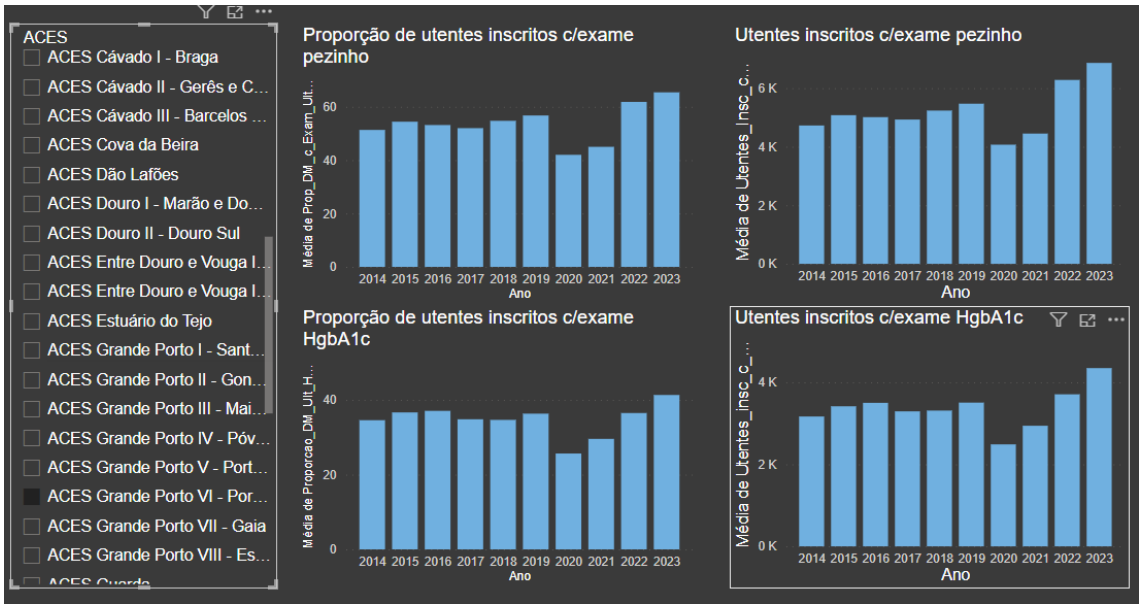

Figura 56 - Dashboard 4 - Análise das ACES, ACES Grande Porto VI

Na figura 57, é possível visualizar informação relativa ao ACES do Algarve II. Os valores deste ACES encontram-se significativamente abaixo da média dos valores gerais, o padrão de quebra nos anos de 2020 e 2021. É possível visualizar a tendência de crescimento desde o ano de 2014, significando um aumento no acompanhamento e controlo da doença.

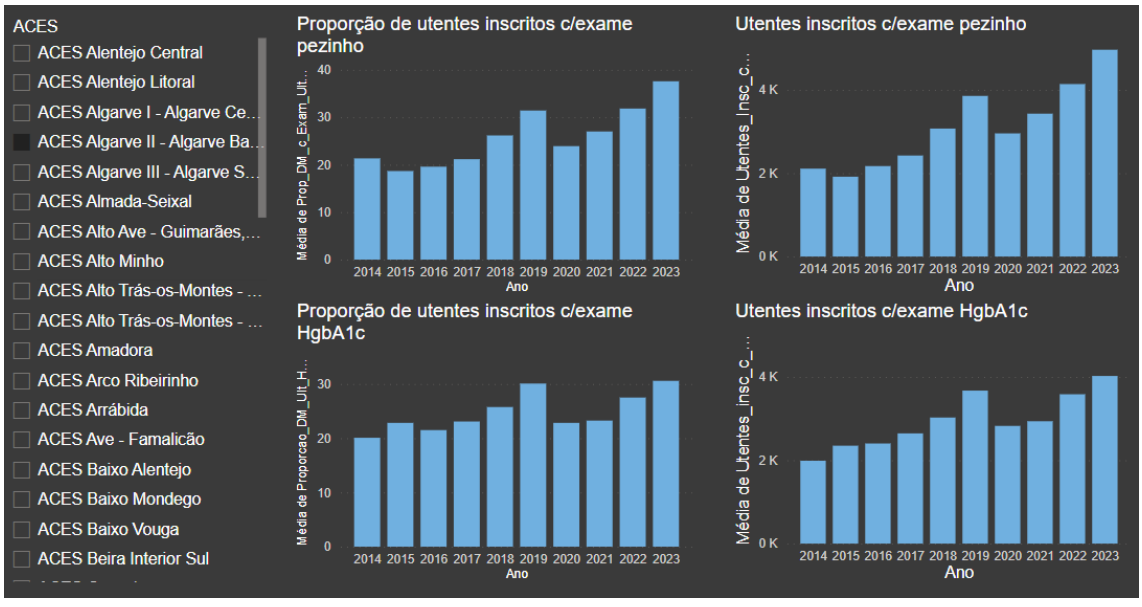

Figura 57 - Dashboard 4 - Análise ACES, ACES Algarve I

5.5 Técnicas de data Mining

Os algoritmos e as técnicas de DM são utilizados para realizar a descrição e/ou a previsão dos dados a partir de um DW. Para a realização desta etapa do desenvolvimento do Data Mining recorreu-se à solução SSAS presente no Microsoft Visual Studio, à imagem do utilizado na criação do cubo OLAP. A ligação a fontes de dados no DW foi realizada através do assistente de mineração de dados, para a aplicação dos algoritmos.

A partir do assistente de mineração de dados dá-se início à definição da estrutura de mineração a partir do DW. Na figura 58, é possível consultar que modelos de dados podem ser usados para a construção da estrutura de mineração. O Microsoft Visual Studio permite a construção baseada na estrutura do DW, ou na estrutura criada no processo do cubo OLAP.

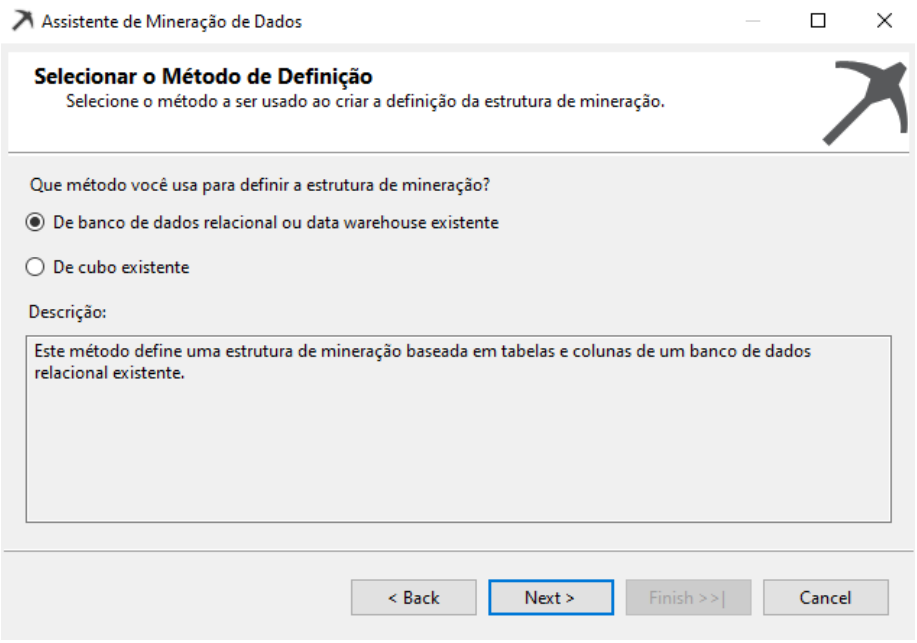

Figura 58 - Seleção da estrutura da fonte de dados

Após a seleção do modelo de dados a utilizar para a criação da estrutura de dados é necessário selecionar que técnica de mineração vai ser utilizada. A figura 59 apresenta os diferentes algoritmos que a ferramenta consegue utilizar para criar estruturas de mineração de dados.

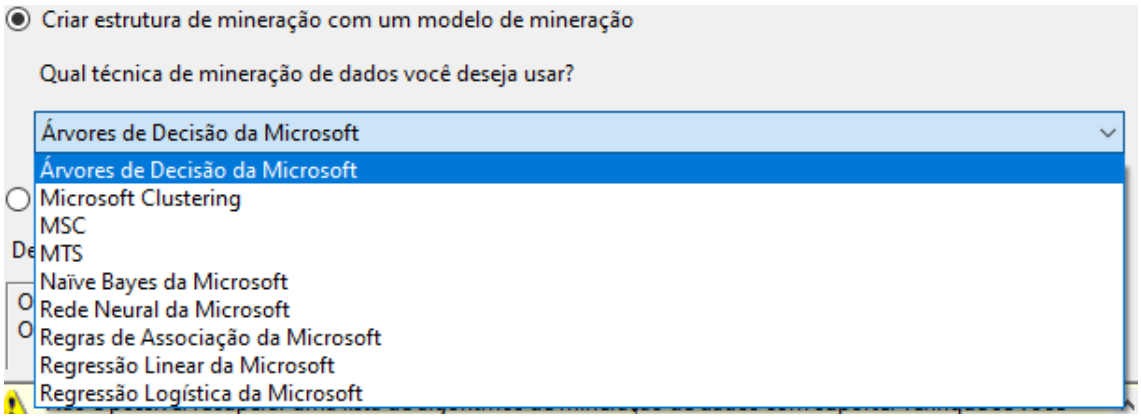

Figura 59 - Algoritmos de DM disponíveis no VS

Os algoritmos apresentam as diferentes técnicas a aplicar na conceção dos modelos de mineração de dados. A utilização do algoritmo mais ajustado, depende do tipo de área de negócio e os resultados que se pretendem alcançar. No caso particular do nosso trabalho, pretende-se aplicar o algoritmo que permita fazer classificações e previsões. Após a realização de alguns estudos foi possível perceber que o algoritmo "Árvore de Decisão" é a tipologia de algoritmo que mais se adequa ao nosso trabalho, ou seja, este algoritmo permite realizar a previsão de valores futuros sobre a DMT2.

Nas figuras 60 e 61, é explicado o processo de definição dos dados para a criação do modelo de DM. A figura 60, apresenta a seleção da tabela que vai ser utilizada para a análise, para realizar a previsão dos dados. A tabela selecionada para a estrutura do modelo de mineração de dados foi a tabela Facto.

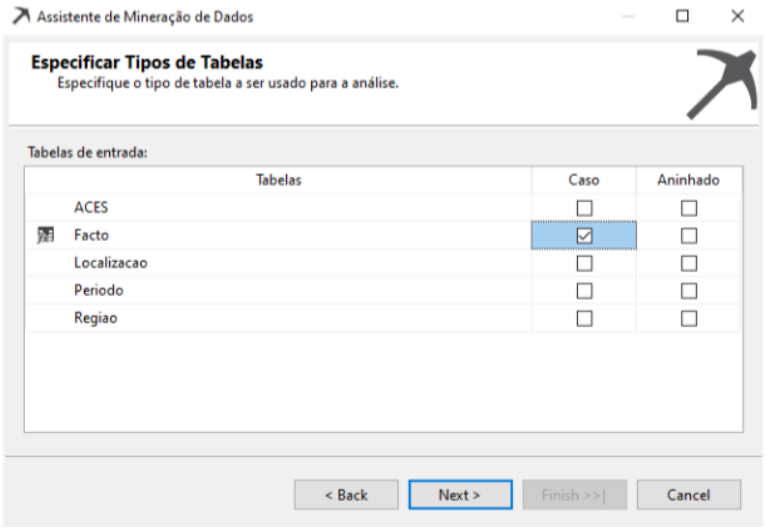

Figura 60 - Seleção da tabela para a mineração de dados

A figura 60, apresenta o processo de seleção dos campos que vão ser valores de entrada e os campos que o algoritmo pretende prever na criação da estrutura de mineração de dados. Após a seleção da tabela de Facto na figura 61, é necessário selecionar os campos de entrada no modelo de dados e os campos a selecionar para realizar a previsão dos valores dos campos, por exemplo, "Proporcao_Dm_Ult_HgbA1c, conforme a figura 61.

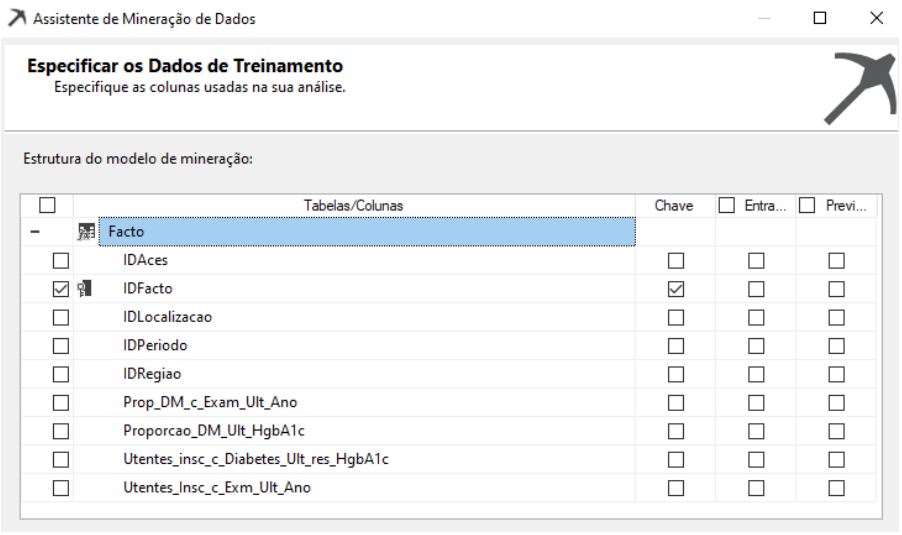

Figura 61 - Processo de seleção dos campos para análise

As figuras 62 e 63, apresentam os dois processos de seleção de variáveis para o processo de mineração de estrutura de dados. A figura 62, apresenta a seleção dos campos de entrada, para a previsão da proporção de utentes inscritos com exame dos pés. A figura 63, apresenta a seleção dos campos que se pretendem prever, ou seja, fazer previsões dos utentes inscritos com exame à HgbA1c.

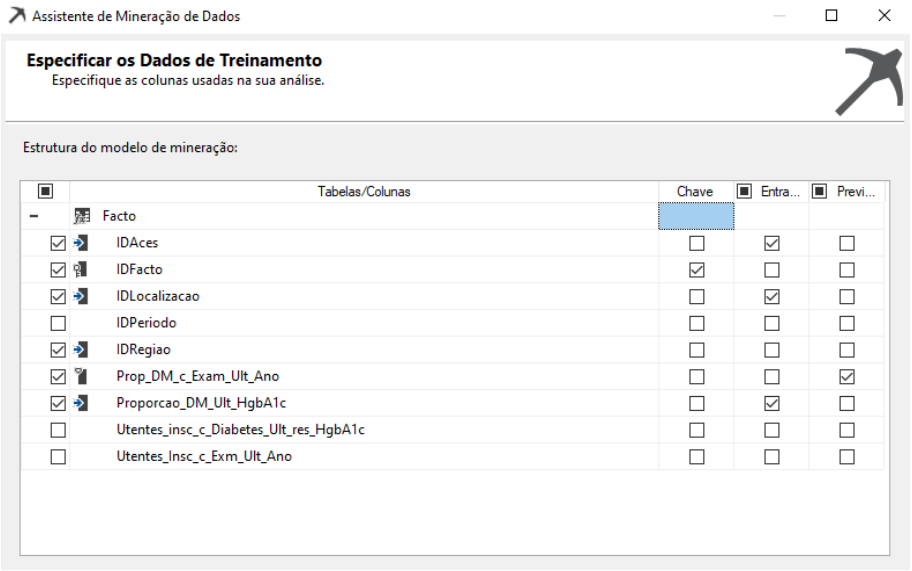

Figura 62 - Seleção dos campos para a previsão dos valores da proporção do exame aos pés

Para a previsão dos utentes inscritos com exame dos pés realizado, foram selecionados como valores de entrada o "IDAces", o "IDLocalizacao" e a "Proporcao_DM_Ult_HgbA1c". Para os valores a prever, foi selecionado o campo "Prop_DM_c_Exam_Ult_Ano".

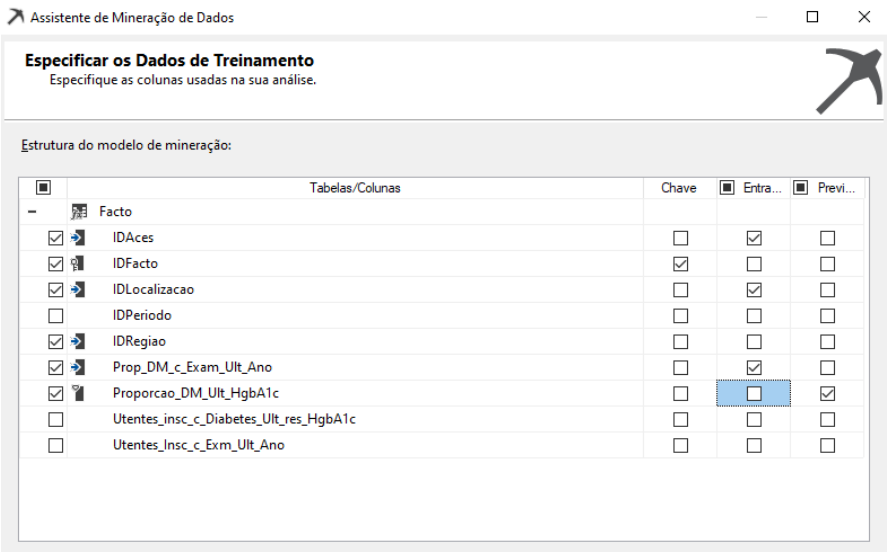

Figura 63 - Seleção dos campos para a previsão dos valores do exame á HgbA1c

Para a previsão dos utentes inscritos com exame à HgbA1c realizado, foram selecionados como valores de entrada o "IDAces", o "IDLocalizacao" e a "Prop_DM_c_Exam_Ult_Ano". Para a obtenção de resultados de previsão, foi selecionado o campo "Proporcao_DM_Ult_HgbA1c".

Após a seleção dos campos, a ferramenta permite fazer uma revisão automática dos campos selecionados e sugere alterações aos dados escolhidos para o estudo. Na figura 64, é possível verificar esta sugestão da ferramenta de DM.

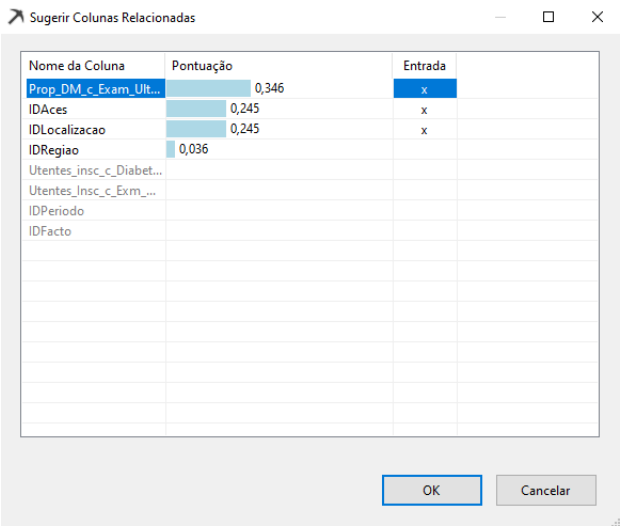

Figura 64 - Exemplo de sugestão dos campos para construção do modelo de previsão

Na próxima fase, o algoritmo executa o treino e teste do modelo em estudo, com a utilização de um conjunto de casos, reservando 30% (figura 65) para testar o modelo e os restantes 70% para treinar o modelo. O teste e treino permitem determinar a precisão do modelo, bem como se o mesmo é ajustável ao modelo em estudo.

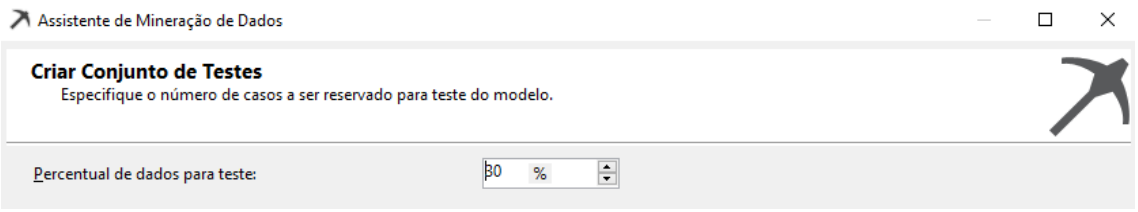

Figura 65 - Exemplo seleção de percentagem dos dados para teste

As figuras 66 e 67, apresentam as estruturas dos modelos de mineração de dados concebidas pela ferramenta, para a previsão da proporção de utentes inscritos com exame dos pés e proporção de utentes inscritos com o exame à HgbA1c.

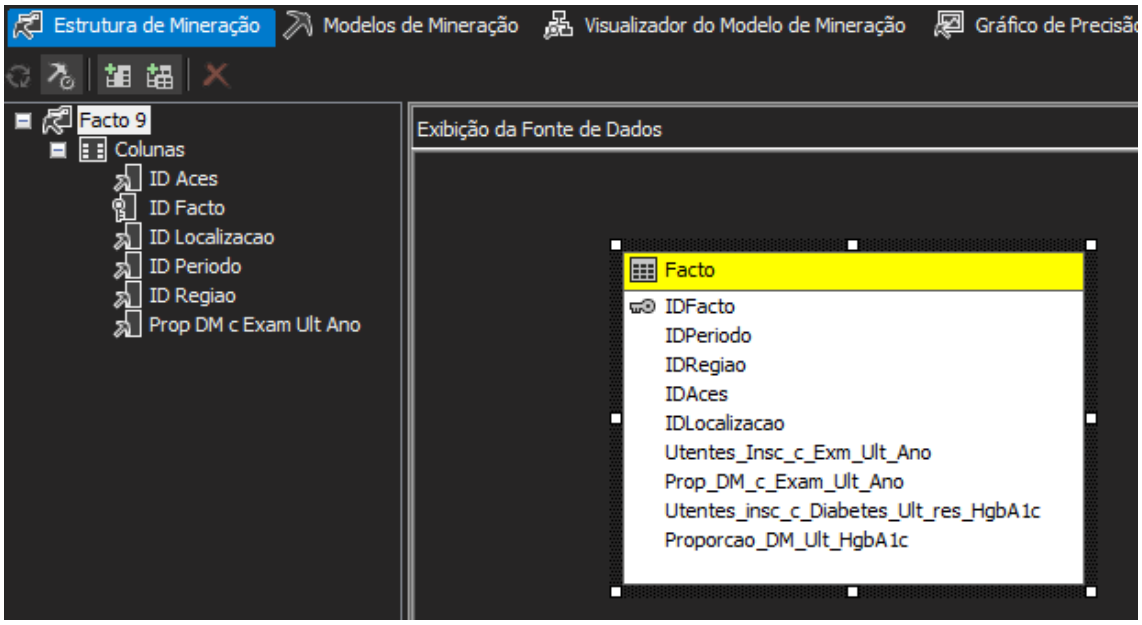

Figura 66 - Estrutura de mineração proporção utentes inscritos com exame aos pés

A figura 66, apresenta a estrutura da fonte de dados e os campos utilizados para o desenvolvimento do modelo de mineração. O modelo implica a aplicação do algoritmo da árvore de decisão, com recurso aos campos apresentados na figura. Como referido, o modelo pretende fazer a previsão dos valores sobre a proporção de utentes inscritos com o exame aos pés realizado.

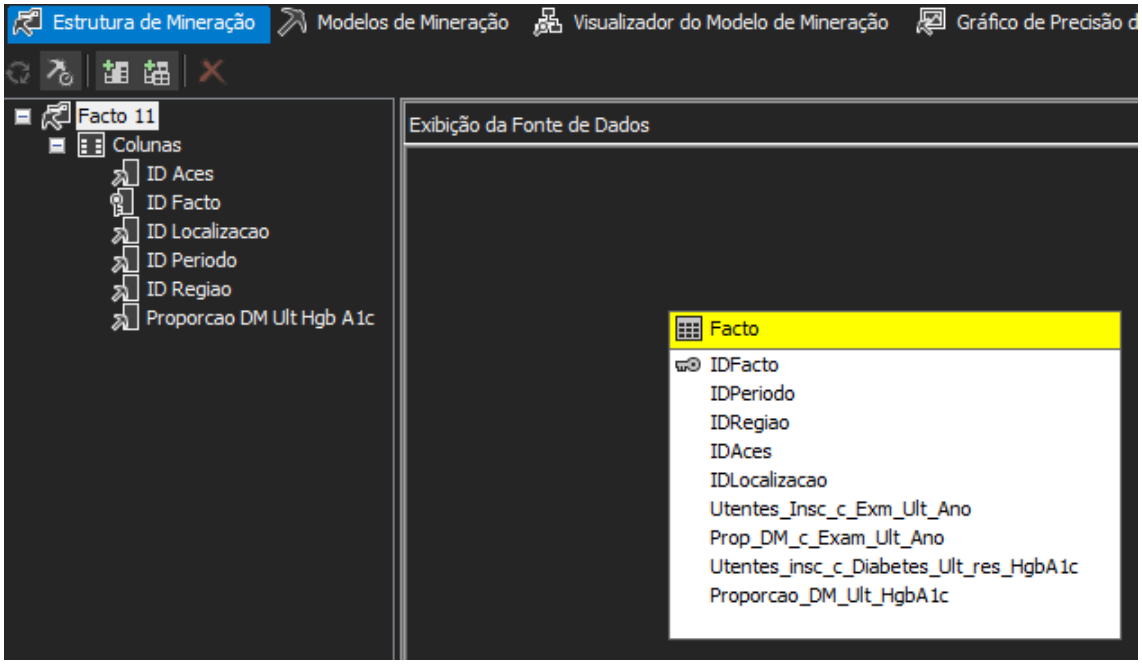

Figura 67 - Estrutura de mineração proporção utentes inscritos com exame á HgbA1c

A figura 67, apresenta a estrutura de mineração de dados, para a previsão do número de utentes inscritos com o exame à HgbA1c. A partir da estrutura, acontece a conceção do algoritmo de "Árvore de Decisão".

Após a aplicação do algoritmo ao modelo de mineração de dados, podemos verificar que os resultados de previsão observados têm por base 70% do treino do modelo. Na figura 68, é possível verificar que apenas 4607 dos 6581 registos foram considerados no algoritmo da árvore de decisão.

Figura 68 - Número de registos nas estruturas de mineração

Na construção do modelo de mineração foi possível constatar que as árvores de decisão estão divididas em diferentes níveis, a árvore de mineração da proporção de utentes inscritos com exame aos pés encontra-se dividida em 124 níveis e a árvore de mineração da proporção de utentes inscritos com exame á HgbA1c encontra-se dividida em 92 níveis.

Na figura 69, é possível consultar a estrutura construída pelo modelo de mineração criado para realizar a previsão dos utentes inscritos com o exame dos pés. O modelo criado dividiu os ACES por região e realizou uma previsão dos valores das proporções para cada ACES. O níveis considerados pelo algoritmo, foram: a região, o ACES e o periodo.

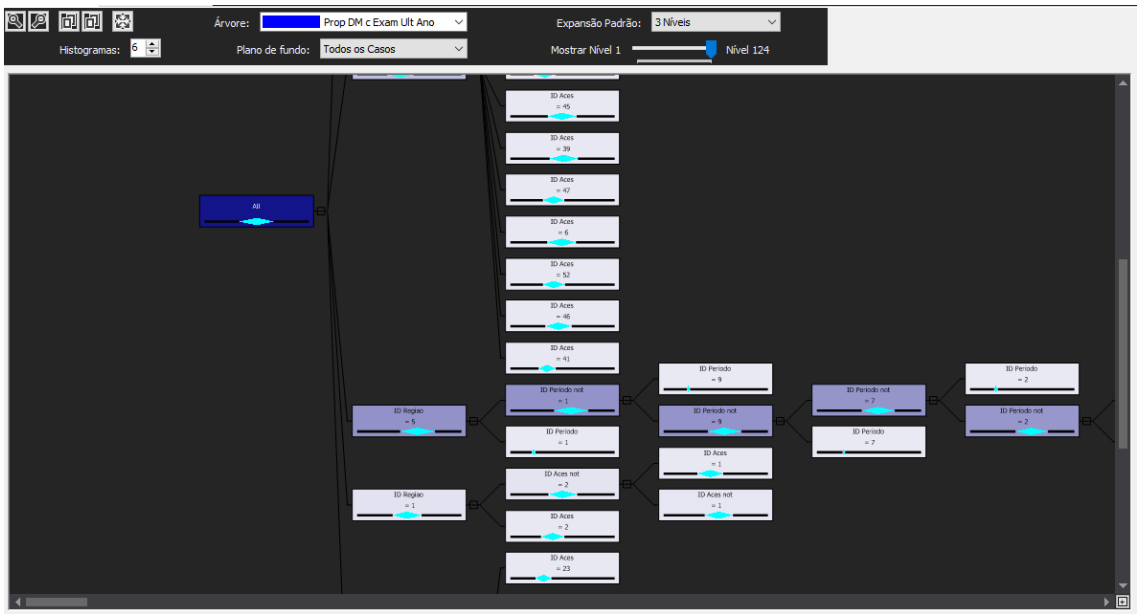

Figura 69 - Estrutura árvore de decisão, proporção utentes inscritos com exame aos pés

A figura 69, apresenta a estrutura da árvore de decisão construída através do modelo de mineração. Este modelo permite prever a proporção de utentes inscritos com o exame dos pés em cada ACES. Também é possível verificar os 124 níveis da árvore de decisão e os valores calculados para cada ACES.

Nas figuras 70 e 71, estão representados os resultados dos casos utilizados pelo algoritmo para efetuar as previsões. Nos 81 casos utilizados, prevê-se que 54,383 ocorram nas ACES utilizadas no estudo (exemplo, ACES ID 24).

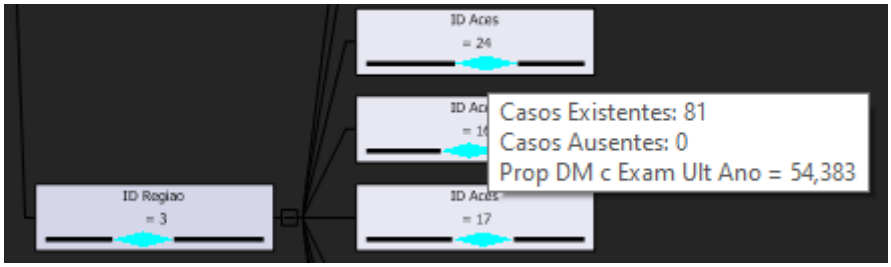

Figura 70 - Valor da proporção de utentes inscritos com exame dos pés ACES 24

Na figura 71, o algoritmo prevê que para a ACES com o ID 13, o numero de casos será de 30,686, podendo significar um aumento proporcional face ao estado atual.

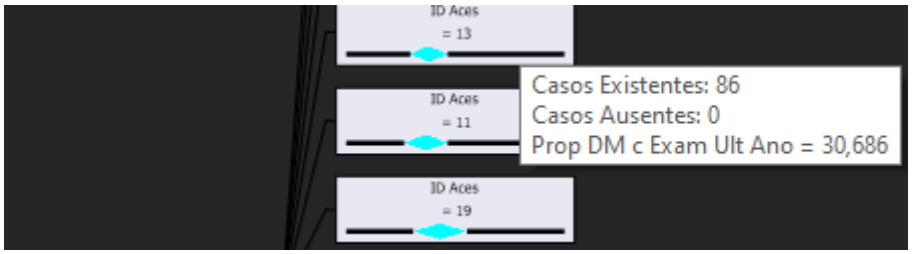

Figura 71 - Valor da proporção de utentes inscritos com exame dos pés ACES 13

Após o desenvolvimento do primeiro modelo de mineração é possível desenvolver o segundo modelo. Este modelo, é responsável por construir uma árvore de decisão para prever a inscrição de utentes para o exame à HgbA1c por ACES.

Na figura 72, é possível consultar a estrutura da árvore de decisão construída pelo modelo de mineração de dados. Como no modelo anterior, a estrututura está dividida por ACES e região.

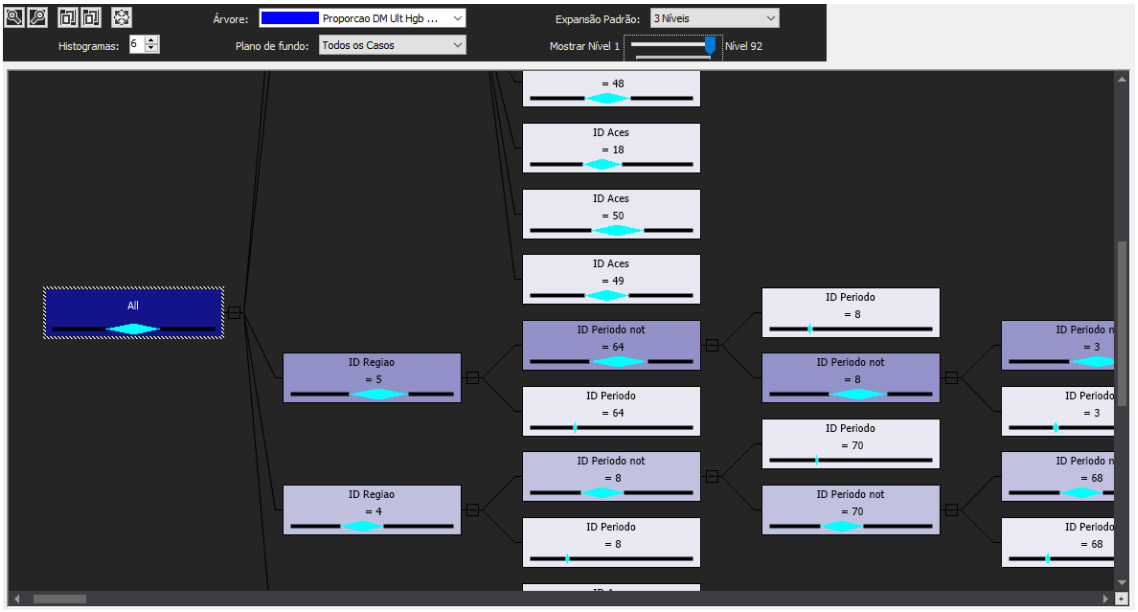

Figura 72 - Estrutura árvore de decisão, proporção utentes inscritos com exame à HgbA1c

A figura 72, apresenta a estrutura da árvore de decisão construída pelo modelo de mineração. Este modelo faz a previsão da proporção de utentes inscritos com o exame à HgbA1c em cada ACES. Podemos verificar a existência dos 92 níveis da árvore de decisão para cada ACES. Nas figuras 73 e 74, estão representados os valores previstos face aos valores históricos, em que para a ACES 24 com uma amostra de 80 casos existentes a previsão é de 44,788 casos a registar na respetiva ACES.

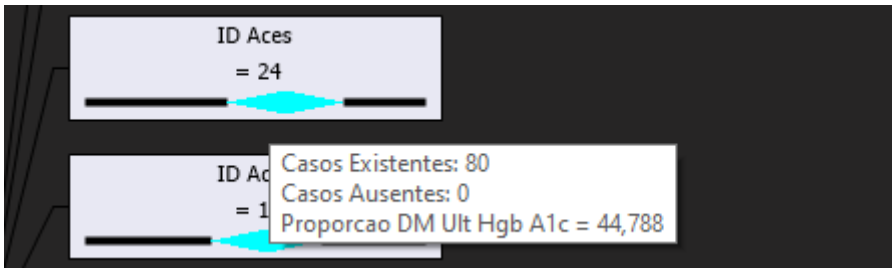

Figura 73 - Previsão das ocorrências de utentes inscritos para o exame à HgbA1c ACES 24

Na figura 74, o algoritmo prevê que para o ID Aces 5 (ACES Algarve III – Algarve Sotavento), o número de casos será de 42,171, podendo significar um aumento proporcional face ao estado atual.

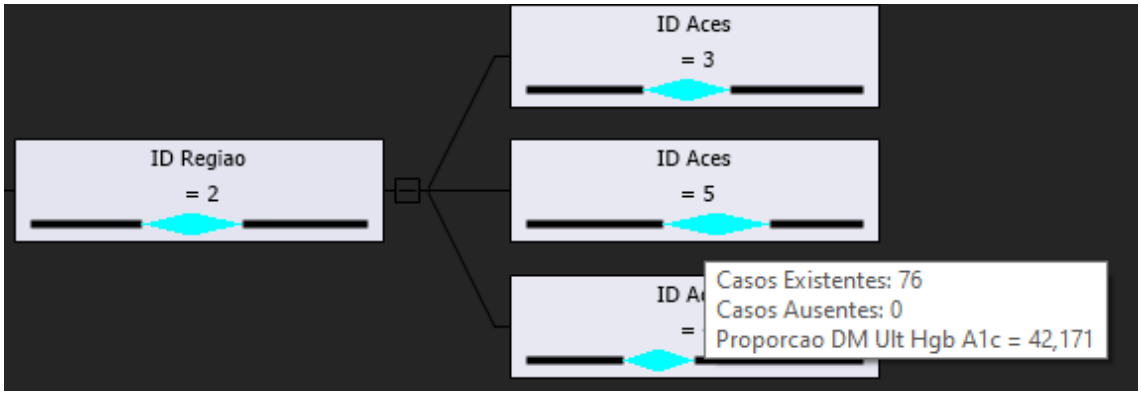

Figura 74 - Valor da proporção de utentes inscritos com exame à HgbA1c ACES 5

Com o desenvolvimento dos dois modelos podemos concluir que se prevê um aumento do número de casos para a realização de exames à HgbA1c nas diferentes ACES, face aos casos existentes, considerando as técnicas de DM com suporte ao algoritmo de árvore de decisão. Os valores previsivos fornecidos pelas técnicas de DM estão completos, estruturados e de fácil compreensão. Também, os valores previsivos estão em consonância com a realidade, face à evolução continua dos exames à HgbA1c.

6 Conclusão

Este documento apresenta o enquadramento do DM como modelo de suporte na tomada de decisão nas organizações. Neste contexto, introduziram-se conceitos como a descoberta de conhecimento em bases de dados e os sistemas de apoio à decisão, para uma melhor compreensão do tema em estudo. Com base na análise destes conceitos é possível verificar a necessidade de estudo sobre técnicas de modelação e tratamento de dados e decidir a que tipo de organização vão ser aplicadas as técnicas de suporte a decisão. Assim, foi necessário estudar sobre o tema de Diabetes Mellitus tipo 2 e sobre técnicas como Processo ETL, Cubo OLAP, Data Mining, Data Science e Business Intelligence.

Depois de uma investigação profunda sobre os temas e as técnicas, foi necessário investigar sobre metodologias de desenvolvimento, sendo que a metodologia de desenvolvimento mais adequada foi a Design Science Research.

Em seguida, na componente prática realizaram-se análises dos dados, sendo possível criar uma definição do dataset para o processo de preparação e tratamento dos dados. Foi analisado o modelo fornecido pelo SNS, sendo percebida a informação disponibilizada e as alterações necessárias para iniciar o desenvolvimento da solução.

Posteriormente, apresentou-se o processo de preparação dos dados, bem como o processo de ETL e o processo de criação do Cubo OLAP. Com os dados devidamente tratados foram realizadas e apresentadas as análises aos dados com a ferramenta Power BI. Através das Dashboards criadas é possível consultar graficamente os dados no dataset, tendo sido possível consultar a evolução do número de exames a serem realizados e a evolução do número de inscritos com diabetes e com os exames realizados. Por último, foram aplicados algoritmos de regressão e classificação.

6.1 Considerações finais

Com o projeto finalizado, é necessário analisar os objetivos e perceber se com o desenvolvimento da investigação foi possível cumprir com os objetivos definidos para o projeto.

Neste sentido com a finalização da investigação é possível verificar que o projeto cumpriu com o objetivo principal. Foi possível aplicar as técnicas de DM para previsão de dados sobre a DMT2. Após verificar o cumprimento do objetivo principal, foi possível verificar o cumprimento dos objetivos secundários.

Após o término da investigação podemos dizer que o resultado obtido para os objetivos foi satisfatório, tendo os mesmos sido cumpridos.

- A revisão da literatura permitiu uma análise e obtenção de conhecimentos relativamente às técnicas de Data Science e Data Mining.
- A preparação dos dados e o processo de ETL permitiu analisar a primeira técnica de DS.
- A criação do cubo OLAP, permitiu analisar outra técnica de DS. A utilização desta ferramenta foi essencial, para o desenvolvimento da solução.
- A criação de Dashboards utilizando a ferramenta Power BI, com o modelo criado pela ferramenta do cubo OLAP, permitiu informar sobre o estado da doença DMT2 em Portugal, e avaliar a utilização de mais uma técnica de DS.
- A utilização de técnicas de DM na criação de modelos de classificação e de regressão permitiu analisar e verificar a integridade destas técnicas para a previsão de dados, contudo era esperado que os resultados fossem diferentes. Devido ao dataset conter dados durante a pandemia do covid-19, estes resultaram numa baixa dos valores esperados.

6.2 Limitações encontradas

No desenvolvimento do projeto foram encontradas algumas dificuldades no processo de desenvolvimento da investigação.

Na revisão de literatura foram encontradas dificuldades em encontrar documentos que relacionassem o tema da DMT2 com técnicas de DM e em encontrar os tópicos que foram objeto de estudo. Esta última dificuldade deveu-se à dificuldade na seleção do dataset, pois como dados da saúde são sensíveis não foi possível encontrar um dataset com histórico maior.

Numa segunda fase, no desenvolvimento da solução foram encontradas duas dificuldades. Na preparação dos dados a fase crucial para o correto desenvolvimento da solução, como referido foi difícil encontrar um dataset que fosse compatível com o objeto do estudo. Esta dificuldade atrasou e complicou o desenvolvimento da solução.

Além das dificuldades referidas, foram também encontradas dificuldades na compatibilidade dos programas com o computador usado para o desenvolvimento, no entanto, as mesmas foram ultrapassadas com certas atualizações manuais. Outra dificuldade foi a conciliação do desenvolvimento deste projeto com as atividades laborais exercidas no momento.

7 Referência Bibliográficas

Ajibade, S. S., & Adediran, A. (2016). An overview of big data visualization techniques in data mining. International Journal of Computer Science and Information Technology Research, 4(3), 105-113.

Algarni, A. (2016). Data mining in education. International Journal of Advanced Computer Science and Applications, 7(6).

Alnoukari, M., & El Sheikh, A. (2012). Knowledge discovery process models: from traditional to agile modeling. In Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications (pp. 72-100). IGI Global.

American Diabetes Association. (2014). Diagnosis and classification of diabetes mellitus. Diabetes Care, 37 Suppl 1, S81-90. https://doi.org/10.2337/dc14-S081

Bielza, C., & Larranaga, P. (2014). Discrete Bayesian network classifiers: A survey. ACM Computing Surveys (CSUR), 47(1), 1-43.

Bonczek, R. H., Holsapple, C. W., & Whinston, A. B. (2014). Foundations of decision support systems. Academic Press.

Burstein, F., W Holsapple, C., Jukic, N., Jukic, B., & Malliaris, M. (2008). Online analytical processing (OLAP) for decision support. In Handbook on Decision Support Systems 1: Basic Themes (pp. 259-276).

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. ACM Sigmod record, 26(1), 65-74.

DeFronzo, R. A., Ferrannini, E., Groop, L., Henry, R. R., Herman, W. H., Holst, & Weiss, R. (2015). Type 2 diabetes mellitus. Nature reviews Disease primers, 1(1), 1-22.

Dey, M., & Rautaray, S. S. (2014). Study and analysis of data mining algorithms for healthcare decision support system. Planning, 5(6).

El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. Journal of King Saud University-Computer and Information Sciences, 23(2), 91-104.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

Ferreira, J., Miranda, M., Abelha, A., & Machado, J. (2010, September). O processo etl em sistemas data warehouse. In INForum (pp. 757-765).

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

Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. Remote sensing of environment, 61(3), 399-409.

Goebel, M., & Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. ACM SIGKDD explorations newsletter, 1(1), 20-33.

Gupta, H., Harinarayan, V., Rajaraman, A., & Ullman, J. D. (1997, April). Index selection for OLAP. In Proceedings 13th International Conference on Data Engineering (pp. 208-219). IEEE.

Han, J. (1998). OLAP mining: An integration of OLAP with data mining. In Data Mining and Reverse Engineering: Searching for semantics. IFIP TC2 WG2. 6 IFIP Seventh Conference on Database Semantics (DS-7) 7–10 October 1997, Leysin, Switzerland (pp. 3-20). Springer US.

Han, J., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques third edition. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS quarterly, 75-105.

Kharroubi AT, Darwish HM. (2015). Diabetes mellitus: The epidemic of the century. World J Diabetes, 6(6), 850-867. https://doi.org/10.4239/wjd.v6.i6.850

Kimball, R., Ross, M., Mundy, J., & Thornthwaite, W. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. John Wiley & Sons.

Khan, R. A., & Quadri, S. M. (2012). Business intelligence: an integrated approach. Business Intelligence Journal, 5(1), 64-70.

Kherdekar, V. A., & Metkewar, P. S. (2016). A technical comprehensive survey of ETL tools. International Journal of Applied Engineering Research, 11(4), 2557-2559.

Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2014). Survey of classification techniques in data mining. International Journal of Computer Sciences and Engineering, 2(9), 65-74.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

Moore, J. H., & Chang, M. G. (1980). Design of decision support systems. ACM SIGOA Newsletter, 1(4-5), 8-14.

Negash, S. (2004). Business intelligence. Communications of the association for information systems, 13(1), 15.

Netto, A. P., Andriolo, A., Fraige Filho, F., Tambascia, M., Gomes, M. D. B., Melo, S. (2009). Atualização sobre hemoglobina glicada (HbA1C) para avaliação do controle glicêmico e para o diagnóstico do diabetes: aspectos clínicos e laboratoriais. Jornal Brasileiro de Patologia e Medicina Laboratorial, 45, 31-48.

Oracle. (2012). Big data for the Enterprise. Redwood Shores, CA: Oracle.

Phyu, T. N. (2009, March). Survey of classification techniques in data mining. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 1, No. 5, pp. 727-731). Citeseer.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of management information systems, 24(3), 45-77.

Pereira, J. (2005). Modelos de Data Mining para multi-previsão: aplicação à medicina intensiva.

Power, D. J. (2002). Decision support systems: concepts and resources for managers. Quorum Books.

Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. International Journal of current engineering and technology, 3(2), 334-337.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. "O'Reilly Media, Inc.".

Rashidi, Maria & Ghodrat, Maryam & Samali, Bijan & Mohammadi, Masoud. (2018). Decision Support Systems. 10.5772/intechopen.79390

Reddy, G. S., Srinivasu, R., Rao, M. P. C., & Rikkula, S. R. (2010). Data Warehousing, Data Mining, OLAP and OLTP Technologies are essential elements to support decision-making process in industries. International Journal on Computer Science and Engineering, 2(9), 2865-2873.

Russom, P. (2011). Big data analytics. TDWI best practices report, fourth quarter, 19(4), 1-34.

Saagari, S., DeviAnusha, P., LakshmiPriyanka, C., & Sailaja, V. S. S. N. (2013). Data Warehousing, Data Mining, OLAP and OLTP Technologies Are Essential Elements to Support Decision-Making Process in Industries. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2(6), 2278-3075.

Santos, M. Y., & Ramos, I. (2009). Introdução [a] Business intelligence: tecnologias da informação na gestão de conhecimento. FCA-Editora de Informática, Lda.

Sethi, M. (2012). Data Warehousing and OLAP Technology. International Journal of Engineering Research and Applications (IJERA), 2(2), 955-960.

Sociedade Brasileira de Diabetes. (2016). Diretrizes da Sociedade Brasileira de Diabetes: 2015-2016 [Internet]. São Paulo: A.C. Farmacêutica.

Sprague Jr, R. H. (1980). A framework for the development of decision support systems. MIS quarterly, 1-26.

Turban E. Sharda R. & Delen D. (2011). Decision support and business intelligence systems (9th ed.). Prentice Hall

Vedder, R. G., Vanecek, M. T., Guynes, C. S., & Cappel, J. J. (1999). CEO and CIO Perspectives on Competitive Intelligence. Communications of the ACM, 42(8), 108–116.

Vercellis, C. (2009). Business Intelligence: Data Mining and Optimization for Decision Making. Politecnico di Milano, Italy: A John Wiley and Sons, Ltd., Publication.

Vercellis, C. (2011). Business intelligence: data mining and optimization for decision making. John Wiley & Sons.

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques, Second Edition. Acm Sigmod Record, 31(1), 76-77.

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100-107.

Younas, M. (2019). Research challenges of big data. Service Oriented Computing and Applications, 13, 105-107.

Zhong, N., Liu, C., Kakemoto, Y., & Ohsuga, S. (1997, August). KDD Process Planning. In KDD (pp. 291-294)