# Abstract

A long, long time ago...

# Resumo

Há muito, muito tempo

# Agradecimentos

Obrigado a todos, obrigado ...

**Dedico a . . .**

# Conteúdo

**Abstract** i

**Resumo** iii

**Agradecimentos** v

**Conteúdo** ix

**Lista de Tabelas** xi

**Lista de Figuras** xiv

**Lista de Blocos de Código** xv

**Acrónimos** xvii

**1 Introdução** 1

1.1 Contexto . . . . . 1

1.2 Motivação . . . . . 2

1.3 Projeto . . . . . 2

1.3.1 Objetivos . . . . . 3

1.3.2 Contribuição . . . . . 4

1.4 Organização . . . . . 4

**2 Fundamentos e Terminologia** 5

2.1 *Diabetes Mellitus* . . . . . 5

2.1.1 Dispositivos para monitorizar a diabetes . . . 7
2.2 *Data Mining* . . . 8
2.2.1 Associação . . . 10
2.2.2 Classificação . . . 11
2.2.3 Validação de modelos . . . 12
2.2.4 Redes *bayesianas* . . . 12
2.2.5 *Data Mining* na diabetes . . . 13

**3 Estado da Arte** **15**
3.1 Medicina personalizada e *data mining* na saúde . . . 15
3.2 Aplicações para *smartphones* Android . . . 17
3.2.1 Diário da Diabetes mySugr . . . 18
3.2.2 Diabetes:M . . . 18
3.2.3 OnTrack Diabetes . . . 19
3.2.4 Diabetes - Diário Glucose . . . 19
3.2.5 Glucose Buddy: Diabetes Log . . . 19

**4 MyDiabetes** **21**
4.1 Objetivo da aplicação . . . 21
4.2 Arquitetura . . . 22
4.3 Variáveis recolhidas . . . 22

**5 Análise de dados** **25**
5.1 Descrição do estudo . . . 25
5.2 Recolha de dados . . . 26
5.2.1 Descrição da experiência e *feedback* dos utilizadores . . . 27
5.2.2 O *data set* . . . 29
5.2.3 Pré-processamento dos dados . . . 30
5.3 Análise estatística básica . . . 34

5.3.1 Média de glicose . . . . . 35
5.3.2 Média de glicose por dia . . . . . 36
5.3.3 Média de glicose por período do dia . . . . . 37
5.3.4 Glicose por hora do dia . . . . . 38
5.3.5 Glicose por hora e por dia . . . . . 44

**6 Resultados** **51**

6.1 Regras de associação . . . . . 51
6.2 Redes *Bayesianas* . . . . . 59
6.3 *Inductive Logic Programming* . . . . . 68

**7 Conclusões** **69**

7.1 Trabalho Futuro . . . . . 70

**Bibliografia** **71**

# Lista de Tabelas

5.1 Tipo das variáveis recolhidas . . . . . 29

5.2 Médias de glicemia dos utilizadores por dia da semana . . . . . 36

5.3 Médias de glicemia dos utilizadores por período do dia . . . . . 37

# Lista de Figuras


4.1 Menu principal da aplicação MyDiabetes . . . . . 23

5.1 Médias de glicemias para os utilizadores . . . . . 35
5.2 Glicemia por horas do utilizador 1 . . . . . 40
5.3 Glicemia por horas do utilizador 2 . . . . . 41
5.4 Glicemia por horas do utilizador 3 . . . . . 42
5.5 Glicemia por horas do utilizador 4 . . . . . 43
5.6 Glicemia por horas do utilizador 5 . . . . . 44
5.7 Glicemia do utilizador 1 por dias da semana . . . . . 45
5.8 Glicemia do utilizador 2 por dias da semana . . . . . 46
5.9 Glicemia do utilizador 3 por dias da semana . . . . . 47
5.10 Glicemia do utilizador 4 por dias da semana . . . . . 48
5.11 Glicemia do utilizador 5 por dias da semana . . . . . 49

6.1 Regras de associação para o utilizador 1 . . . . . 54
6.2 Regras de associação para o utilizador 2 . . . . . 55
6.3 Regras de associação para o utilizador 3 . . . . . 56
6.4 Regras de associação para o utilizador 4 . . . . . 57
6.5 Regras de associação para o utilizador 5 . . . . . 58
6.6 Exemplo de uma rede *bayesiana* . . . . . 60
6.7 Rede *bayesiana* para o utilizador 1 . . . . . 61
6.8 Probabilidade máxima de hiperglicemia para o utilizador 1 . . . . . 62

6.9 Rede *bayesiana* para o utilizador 2 . . . . . 63
6.10 Probabilidade máxima de hiperglicemia para o utilizador 2 . . . . . 63
6.11 Rede *bayesiana* para o utilizador 3 . . . . . 64
6.12 Probabilidade máxima de hiperglicemia para o utilizador 3 . . . . . 65
6.13 Rede *bayesiana* para o utilizador 4 . . . . . 66
6.14 Probabilidade máxima de hiperglicemia para o utilizador 4 . . . . . 66
6.15 Rede *bayesiana* para o utilizador 5 . . . . . 67
6.16 Probabilidade máxima de hiperglicemia para o utilizador 5 . . . . . 68

# Lista de Blocos de Código

# Acrónimos

**IDF** International Diabetes Federation

**OMS** Organização Mundial da Saúde

**NA** Not Available

**HbA1c** Hemoglobina glicada

**WEKA** Waikato Environment for Knowledge Analysis

**SamIAm** Sensivity Analysis, Modeling, Inference And More

# Capítulo 1

# Introdução

## 1.1 Contexto

A diabetes, também conhecida por *diabetes mellitus*, é uma doença crónica bastante comum, conhecida por fazer com que os seus portadores tenham níveis de glicose (açúcar) no sangue mais elevados que o normal. Isto deve-se ao facto de o pâncreas não funcionar da forma devida ou nem sequer funcionar, de todo. Antes de nos aprofundarmos sobre a doença em si, eis alguns factos preocupantes:

Segundo a International Diabetes Federation (IDF) em 2015, cerca de 415 milhões de pessoas tinham diabetes. Em 2040, se continuar ao mesmo ritmo, este número aumentará para 642 milhões [13] e, de acordo com a Organização Mundial da Saúde (OMS), em 2030 a diabetes será a sétima causa de morte no planeta [25].

Como se pode perceber, esta doença afeta muita gente e a tendência é para piorar. Por isso mesmo, torna-se cada vez mais importante conseguir adiar ou prevenir o seu aparecimento, que nem sempre é possível. O problema é que a diabetes não tem cura e portanto é fundamental que um paciente diabético tenha um tratamento adequado, sendo que o objetivo é manter os níveis de glicemia mais ou menos constantes, e dentro de intervalos considerados normais. No entanto, não existe uma forma de tratamento padrão que possa ser aplicada a todos os doentes diabéticos. Além do tratamento médico, como a insulina ou medicamentos, há outros fatores que impactam, de alguma forma, a quantidade de glicose no sangue, como por exemplo o exercício, doenças ou o tipo de alimentos que se ingere. Nem toda a gente tem as mesmas rotinas e, portanto, um tratamento que seja eficaz num paciente pode não ser noutro. É por isso importante que os pacientes diabéticos tenham um tratamento personalizado, de acordo com as suas características e rotinas. Normalmente, o tratamento de um paciente diabético passa por um plano elaborado conjuntamente pelos seus médicos endocrinologista e nutricionista. Este plano será sempre feito tendo em conta o paciente, pelo que é um plano personalizado de acordo com as necessidades e rotinas do mesmo. Isto é a base de um conceito que será abordado no próximo capítulo, medicina personalizada.

Ainda no tratamento da doença, a parte da alimentação e rotinas é bastante importante. A diferença entre fazer sempre as mesmas refeições a horas certas ou não ter qualquer tipo de rotina neste aspeto pode ser a diferença entre valores normais ou descontrolados. Uma medição frequente, para que o paciente vá controlando os seus níveis de glicemia e tomar ações, se necessário, é um fator importante para a estabilização dos valores de glicose. De facto, um controlo apertado dos níveis de glicose pode minimizar ou até prevenir as consequências da diabetes, como vamos ver na próxima secção.

## 1.2 Motivação

Na última secção mencionámos que o controlo dos níveis da glicose, através de medições frequentes, é um fator importante para o aumento da qualidade de vida do doente diabético. Um estudo levado a cabo entre 1983 e 1993 comprova isto mesmo: participaram 1441 voluntários e nesse período de 10 anos tiveram um controlo intensivo da glucose que lhes permitia ter valores próximos dos normais. O controlo intensivo era feito aumentando o número de medições diárias, aumentando o número de injeções de insulina ou com o uso de bomba, ajustando sempre o valor de insulina de acordo com a comida e exercício, seguindo uma dieta e plano de exercícios e fazer visitas mensais ao centro de saúde para avaliar o progesso. O estudo concluiu que um controlo intensivo da glucose levou a uma redução em pelo menos 50% de risco de doenças renais, oculares ou do sistema nervoso [12]. Ou seja, apesar de ser uma doença crónica, é possível aumentar a qualidade de vida dos pacientes diabéticos, desde que tenham os cuidados acima mencionados. Como é possível perceber, a medição e registo da glicose são processos fundamentais para um bom tratamento da doença. No passado, esse registo tinha que ser feito em papel, que tem como inconveniente o facto de ser passível de se perder ou tornar rapidamente confuso e extenso. No entanto, hoje isso já não se verifica. A tecnologia evoluiu de tal forma que foram criados dispositivos com o propósito de medir e registar os níveis de glicemia. Mas os próprios telemóveis, que são cada vez mais baratos e melhores, tornaram-se inteligentes e são hoje ferramentas poderosas que fazem muito mais do que apenas ligar a alguém ou enviar mensagens. Um *smartphone* pode servir para fotografar, jogar ou até ouvir música, mas pode ser usado também como uma ferramenta para o nosso bem-estar, o que se verifica, havendo aplicações destinadas à saúde. A motivação para este trabalho foi a possibilidade de juntar duas áreas diferentes, a saúde e a tecnologia, para desenvolver uma ferramenta que possa ter um impacto positivo na vida dos doentes diabéticos. A próxima secção descreve o projeto com mais detalhe.

## 1.3 Projeto

Esta dissertação integra-se no projeto "Smart Diabetes Self-Management" que conta com uma aplicação para Android chamada "My Diabetes". Esta aplicação visa oferecer aos seus utilizadores uma alternativa para o registo das medições de glicose, que facilita a visualização desses mesmos registos, através de gráficos ou em forma de lista. A aplicação será descrita mais detalhadamente

no capítulo 4.

O trabalho proposto nesta dissertação foi o de desenvolver novas funcionalidades para a aplicação, dando-lhe alguma "inteligência". Foi proposto, então, desenvolver um sistema que, através da análise dos dados inseridos por cada utilizador ao longo do tempo, fosse capaz de aprender as rotinas para que pudesse gerar avisos ou conselhos face a situações anormais ou até mesmo descobrir padrões que levem a resultados indesejados. Ao descobrir uma destas situações e alertar o utilizador para a mesma, estará a contribuir para que este consiga melhorar o seu controlo da glicemia.

Para desenvolver esta nova funcionalidade, foi necessário obter dados de pacientes insulino-dependentes. Deste modo, em parceria com o Hospital de São João do Porto, foi levada a cabo uma sensibilização dos doentes para utilizarem a aplicação de forma voluntária, sendo que, no futuro, serão estes os maiores beneficiados. A utilização voluntária da aplicação por parte dos pacientes tem diversos objetivos: 1) obter *feedback* da aplicação em si, como críticas ou sugestões; 2) poder construir *data sets* de registos glicémicos num espaço temporal, algo escasso na *web*. Esta parte de obtenção e análise dos dados é fundamental uma vez que permite ter mais conhecimento do tipo de dados que vão ser analisados, bem como o tipo de padrões ou regras que podem ser descobertas. Desta forma, será possível saber o que é útil ou não, para que a aplicação apenas mostre o que realmente for importante.

A análise será feita aos dados que os utilizadores inserirem na aplicação e enviarem. Mais informações tais como os dados registados e recolhidos ou o processo de participação no estudo serão abordados com mais detalhe no capítulo 5.

### 1.3.1 Objetivos

O objetivo desta dissertação é analisar dados reais pertencentes a doentes diabéticos e desenvolver um sistema capaz de gerar regras e mostrar avisos ou conselhos a partir dos dados. Este objetivo não é único, sendo que os outros são:

- Obter dados de registos glicémicos através da participação de voluntários diabéticos;
- Fazer diferentes tipos de análises estatísticas sobre esses dados;
- Analisar os dados para reconhecimento de padrões ou anomalias;
- Criar regras a partir da análise de dados;
- Mostar conselhos ou avisos através das regras geradas; (opcional)
- Integrar este sistema na aplicação MyDiabetes. (opcional)

Os dois últimos objetivos são opcionais uma vez que requerem a integração do sistema na aplicação MyDiabetes. Tal pode não ser possível de se fazer no tempo da dissertação.

### 1.3.2 Contribuição

Este trabalho tem algumas contribuições, que são:

- revisão e discussão das tecnologias usadas no controlo da diabetes;
- recolha de dados e criação de *data sets* de registos diabéticos;
- análise estatística de registos diabéticos;
- criação de regras personalizadas para cada utilizador;
- integração de um sistema de aconselhamento para doentes diabéticos numa aplicação para *smartphone*.

## 1.4 Organização

Esta dissertação está organizada da seguinte forma: no Capítulo 2 serão apresentados alguns fundamentos e conceitos relativamente à diabetes e às tecnologias que irão ser utilizadas. No Capítulo 3 faremos uma revisão das tecnologias já aplicadas à saúde e, mais especificamente, à diabetes. Será feita uma comparação entre algumas tecnologias utilizadas. O Capítulo 4 diz respeito à aplicação utilizada neste projeto, a MyDiabetes. Nele, o estado atual da aplicação, como as funcionalidades que disponibiliza e também as variáveis que permite aos utilizadores registar. No capítulo 5 será feita uma análise de dados, que descreve o processo desde a recolha até à análise. Serão também descritos ainda algumas análises estatísticas aplicadas sobre os dados. No capítulo 6 serão descritas as restantes análises, que envolverão o uso de *software* adicional, e que serão análises mais práticas, ao contrário da análise estatística. Finalmente, no capítulo 7 analisaremos o que pode ser concluído deste trabalho, bem como o que pode ainda ser feito no futuro.

# Capítulo 2

# Fundamentos e Terminologia

Este capítulo tem o propósito de explicar, com mais detalhe, conceitos que possam ser relevantes para um melhor entendimento da dissertação e vai ser dividido em duas partes: 1) definição da diabetes e alguns conceitos relacionados e 2) definição de *data mining* e alguns conceitos relacionados. Durante a dissertação será utilizado o termo *data mining*, sem tradução, por ser um termo bastante utilizado na língua portuguesa. Assim sendo, vamos começar por explicar o que é a diabetes, bem como alguns termos associados à doença que possam ser relevantes. Vamos também abordar de forma mais detalhada como pode ser feito o tratamento da doença e quais as ferramentas já existentes que possam auxiliar o mesmo. Tendo uma noção de como funciona, interessa descobrir como é que a informática pode ter algum relevo no tratamento. Para isso vai ser explicado o conceito de *data mining* e alguns conceitos associados a esta área que possam ter algum relevo. Vão ser discutidas diferentes técnicas de *data mining* que poderão ser usadas para diferentes propósitos.

## 2.1 *Diabetes Mellitus*

A diabetes é uma doença que se caracteriza por provocar elevados níveis de glicose (açúcar) no sangue nos seus portadores. A glicose é um dos tipos de hidratos de carbono, que são nutrientes presentes nos alimentos. De forma sucinta, a glicose produz energia que vai ser utilizada pelas células, sendo por isso um dos hidratos de carbono mais importantes.

Numa pessoa sem diabetes, a glicose é regulada através de uma hormona, a insulina, que vai ser libertada pelo pâncreas quando necessário. Depois de cada refeição, a insulina libertada vai ajudar o corpo a usar ou a guardar a glicose. Numa pessoa com diabetes isto não acontece: a glicose em excesso não vai ser usada e portanto a sua concentração no sangue vai aumentar para níveis prejudiciais. Há diferentes razões para que isto aconteça, como por exemplo, o pâncreas deixar de produzir insulina, ou então o corpo ganha resistência à insulina: os músculos, a gordura e as células não conseguem usar a insulina de forma efetiva [16]. Há ainda também causas desconhecidas que fazem com que o corpo deixe de responder à insulina, como alterações

genéticas. Os tipos mais comuns de diabetes são:

- ***Diabetes Mellitus Tipo 1*** - Este tipo de diabetes também é conhecido como diabetes insulino-dependente ou diabetes juvenil, por normalmente aparecer em jovens e representa entre 5% a 10% de todos os casos de diabetes [24]. Neste tipo de diabetes, o pâncreas deixa de produzir insulina pelo que os pacientes têm que tomar doses de insulina diariamente para conseguir regular a glicose.

- ***Diabetes Mellitus Tipo 2*** - Este tipo de diabetes também é conhecido por diabetes não-insulino-dependente e representa cerca de 90% de todos os casos de diabetes [18]. Normalmente está associado a um estilo de vida pouco saudável e por isso mesmo, é frequentemente resultado de excesso de peso ou falta de exercício físico. Neste tipo de diabetes o pâncreas continua a produzir insulina, mas o corpo não a consegue utilizar de forma adequada. É comum os diabéticos de tipo 2 não necessitarem de insulina, sendo a medicação feita através de comprimidos. No entanto, há também diabéticos tipo 2 tratados com insulina, quando a medicação por comprimidos não é suficiente para o controlo. Apesar de a diabetes tipo 2 surgir normalmente em pessoas mais velhas, tem-se vindo a manifestar também em jovens [29].

- **Diabetes gestacional** - Este tipo de diabetes pode aparecer durante a gravidez. Caracteriza-se por ter valores de glicose superiores aos normais mas abaixo dos valores diagnosticados na diabetes. É normalmente descoberto nas consultas de rotina e não por causa dos sintomas. Há também o risco de mulheres que sofram deste tipo de diabetes desenvolverem, no futuro, diabetes do tipo 2.

- **Diabetes LADA** - O nome tem origem no inglês *Latent Autoimune Diabetes in Adults* que significa "Diabetes auto-imune latente em adultos". Este tipo de diabetes é considerado uma variação de diabetes tipo 1, embora com uma evolução mais lenta. Por isso mesmo é às vezes referido como diabetes tipo 1.5 [10]. Muitas vezes este tipo de diabetes é erradamente diagnosticado como diabetes tipo 2. Estima-se que entre 15% e 20% das pessoas diagnosticadas com diabetes tipo 2 tenham na verdade diabetes LADA [10].

O tratamento para qualquer um dos tipos passa por um controlo da glicemia e por um plano de dieta e exercício, em conjunto com a medicação, tal como mencionado no Capítulo 1. A medicação, seja por comprimidos ou por injeção de insulina, também é personalizada para cada doente visto que esta depende do fator de sensibilidade de cada pessoa. O fator de sensibilidade é quanto uma unidade de insulina consegue baixar o valor da glicemia. Portanto, doses iguais podem ter efeitos diferentes sobre a glicemia em pessoas diferentes, pelo que o tratamento através da insulina é personalizado para cada doente.

Uma das formas que o médico tem para saber se o tratamento do seu paciente está a correr da forma adequada é através da Hemoglobina glicada (HbA1c). A hemoglobina é uma proteína existente nos glóbulos vermelhos que se junta com a glicose presente no sangue, tornando-se

glicada. A medição da hemoglobina glicada permite saber a média dos valores de glicemia nas últimas semanas ou meses e o seu valor é dado em percentagem. Quanto maior o valor da HbA1c, maior a probabilidade de se desenvolver complicações relacionadas com a diabetes. Para se ter uma ideia do intervalo de valores, geralmente o objetivo de HbA1c para diabéticos é de 6.5%. Numa pessoa normal o valor é abaixo dos 6% e um valor entre 6.0% e 6.4% indica pré-diabetes [5]. Pré-diabetes significa que o valor não é alto o suficiente para ser considerado diabetes, mas, se não houver intervenção, é provável que a pessoa com pré-diabetes venha a sofrer de diabetes tipo 2 num prazo de 10 anos [8].

Além dos fatores discutidos, existem outros que podem causar alterações nos valores de glicemia, como doenças. Por exemplo, a gripe faz aumentar os valores de glicemia. O exercício também provoca alterações: ao fazer exercício estamos a gastar energia, ou seja, glicose, e portanto naturalmente que os valores de glicemia tendem a baixar depois do exercício. Isto porque alguns órgãos, nomeadamente os músculos, usam a glicose diretamente, sem necessidade de insulina. Por outro lado, uma rotina sedentária não usa a glicose em excesso o que leva a um aumento dos níveis de glicemia. Esta oscilação da quantidade de glicose no sangue por vezes atinge extremos, que não são, de todo, desejáveis. Valores muito baixos de glicemia têm o nome de hipoglicemia e valores muito altos hiperglicemia. Tanto a hipo como a hiperglicemia são estados que podem fazer parte do dia-a-dia dos diabéticos e são ambos perigosos. A hiperglicemia pode provocar complicações a longo prazo, como doenças renais ou cardíacas. Por outro lado, a hipoglicemia é mais perigosa a curto prazo, pois pode levar a um estado de inconsciência. Isto acontece porque o nosso cérebro precisa de açúcar, e, na falta deste, pode haver perda de consciência ou até mesmo lesões cerebrais e morte. Se o paciente diabético não tiver consciência que está em hipoglicemia, pode desmaiar antes de poder ingerir açúcar e, no caso de estar sozinho, pode levar a consequências graves.

Isto vem mais uma vez corroborar aquilo que temos vindo a repetir: o controlo da glicemia é vital. Esta necessidade levou à criação de várias ferramentas que podem ajudar o doente diabético a ter este controlo. Na próxima subsecção vamos abordar algumas destas ferramentas.

### 2.1.1 Dispositivos para monitorizar a diabetes

Há vários dispositivos existentes, alguns mais completos que os outros, mas todos com o mesmo objetivo básico: ajudar o diabético a controlar a glicemia. Alguns dispositivos fazem isto de forma automática, como as bombas infusoras de insulina, que permitem o utilizador escolher uma quantidade de insulina a ser injetada de forma contínua ao longo do dia e também à hora das refeições, enquanto outros fazem-nos de forma indireta, ao alertar o utilizador para que ele possa fazê-lo. Entres estes últimos incluem-se os monitores contínuos de glicose e os glicosímetros. Como já referido anteriormente, os *smartphones* também têm utilidade, ao ter aplicações que permitam o registo de valores de glicemia, que, ao contrário do papel, são facilmente acessíveis e podem ser mostrados ao médico na consulta, caso seja preciso.

**Bombas infusoras de insulina**

Uma bomba infusora de insulina é um pequeno dispositivo que liberta insulina de ação rápida 24 horas por dia. A quantidade de insulina libertada é ajustada de acordo com as necessidades do utilizador. Existem várias marcas e modelos no mercado, e, apesar de todas terem o mesmo objetivo fundamental, têm algumas diferenças nas funcionalidades que oferecem. Um exemplo de bomba é a Accu-Chek Combo: é composta pela bomba e por um monitor de glicemia, que comunicam entre si através de *bluetooth*. O utilizador pode assim escolher a insulina a ser administrada, de acordo com os níveis de glicemia [1]. Note-se que a bomba infusora de insulina não é um pâncreas artificial, uma vez que implica o controlo da glicemia por parte do utilizador. No entanto, o conceito de pâncreas artificial existe e está já a ser desenvolvido [28].

**Glicosímetros**

O glicosímetro é o dispositivo base para qualquer diabético: permite medir os níveis de glicemia a qualquer instante, através de uma pequena quantidade de sangue. São uma importante ferramenta pois permitem ao doente saber qual o seu nível de glicose no sangue a dada altura para que possa assim ajustar a insulina a tomar.

**Monitor contínuo de glicose**

É um pequeno aparelho que o utilizador usa a toda a hora e que está constantemente a medir os níveis de glicemia. Assim, quando estes valores forem demasiado altos ou baixos, emite um aviso para que o utilizador possa tomar a medida mais adequada. Um exemplo de um dispositivo deste tipo é o da Dexcom [2].

## 2.2 *Data Mining*

*Data mining* é uma área de ciência de computadores que permite, através da análise de grandes quantidades de dados, descobrir padrões e regras que uma análise mais simples pode não detetar [21] e é uma etapa de um processo chamado *Knowledge Discovery in Databases* que, como o nome indica, tem o objetivo de extrair informação analisando bases de dados. A área de *data mining* usa diversos métodos de outras áreas tais como matemática, estatística, inteligência artificial e *machine learning* para tratar, explorar e obter conclusões acerca dos dados.

*Machine learning* é outra área de ciência de computadores que tem alguma influência em *data mining*. O seu objetivo é conseguir analisar dados e aprender, de forma automática, para que depois possa também fazer previsões noutros dados. Esta parte de aprendizagem é uma parte importante de *machine learning*, e há dois tipos de aprendizagem:

- **Aprendizagem supervisionada** - Como o nome indica, significa que é feita uma aprendizagem sob os dados com algum tipo de informação. Neste caso, um algoritmo de aprendizagem supervisionada é usado para prever uma variável de classe. O algoritmo é primeiro corrido num conjunto de dados de treino em que o valor da variável de classe é conhecido, ou *labeled*, para que este possa "aprender" como é que as variáveis influenciam

a variável de classe. É então gerado um modelo, que também pode ser chamado de classificador, que será capaz de prever a variável de classe analisando outros dados no mesmo formato. Este método tem o nome de classificação.

- **Aprendizagem não supervisionada** - Neste tipo de aprendizagem não existem variáveis com *labels*, pelo que uma aprendizagem deste tipo não tem o objetivo de classificar mas sim encontrar relações entre as variáveis. Ao analisar os vários valores das diferentes variáveis, o algoritmo vai conseguir aprender alguns padrões que possam ocorrer no conjunto de dados. A isto chama-se associação.

Estes conceitos, **associação** e **classificação**, são algumas das tarefas principais de *data mining* e são relevantes para este trabalho, pelo que serão descritos ainda neste capítulo.

- **deteção de anomalias** - Tem como objetivo a identificação de valores anormais. Esses valores podem ser apenas erros mas também podem ser valores interessantes para uma determinada área. A deteção de anomalias pode ser utilizada para detetar fraude ou invasão de uma rede, por exemplo.
- **associação** - Tem como objetivo encontrar relações entre variáveis e pode quantificar essas relações. Esta tarefa do *data mining* é também chamada de *market basket analysis* uma vez que foi utilizada a primeira vez com o objetivo de negócio.[14] A associação será definida com mais detalhe e com alguns exemplos mais à frente neste capítulo, uma vez que uma parte importante deste trabalho passa por utilizar esta técnica.
- **classificação** -

A área de *data mining* tem-se tornado cada vez mais popular e mais usada em variadas áreas, como economia, educação e saúde. E é fácil perceber o porquê: por exemplo, num supermercado, o conhecimento dos produtos que são mais comprados, ou de quem compra o quê, pode ser usado para maximizar as vendas, ou seja, maximizar o lucro.[14] Por ser um campo da ciência de computadores que permite aumentar o conhecimento sobre tudo o que nos rodeia, pode ser também utilizada na medicina para obter mais informações sobre algumas doenças, como a diabetes, neste caso.

Para este trabalho vão ser utilizadas algumas ferramentas especialmente úteis na área de *data mining*. Uma dessas ferramentas é a linguagem de programação R. R é uma linguagem usada em computação estatística que permite o uso de variadas técnicas, como criação de modelos lineares e não-lineares, análises temporais e associação, que é aquilo que nos interessa, entre outras. É também uma ferramenta que permite a criação e visualização de gráficos com bastante facilidade. O uso da linguagem R com as funcionalidades que já traz de raíz é suficiente para uma primeira fase de análises estatísticas mais básicas, pois estas análises serão feitas recorrendo apenas a médias ou gráficos.

### 2.2.1 Associação

Tem como objetivo encontrar relações entre variáveis e pode quantificar essas relações. Esta tarefa do *data mining* é também chamada de *market basket analysis* uma vez que foi utilizada a primeira vez com o objetivo de negócio.[14] Hoje em dia o seu uso ultrapassa a área de negócios e algoritmos de associação são utilizados em várias outras áreas, tais como diagnóstico médico ou análise de dados científicos [32].

Um exemplo da utilização das regras de associação em negócios é o caso da cadeia de supermercados Walmart que, ao analisar transações passadas, descobriu que nos dias que antecediam um furacão, as compras de lanternas aumentavam bastante, o que faz sentido. No entanto, descobriram um facto curioso: juntamente com as lanternas, as vendas que mais aumentavam eram a de um tipo específico de biscoito de morango. Porquê? Porque este biscoito tinha um grande prazo de validade e não precisava de electricidade ou de outro bem essencial para se consumir. Portanto, sempre que havia previsão de furacões, a cadeia de supermercados enchia as prateleiras com esses biscoitos, que ainda assim esgotavam [26]. Isto é o que as regras de associação podem oferecer: ao analisar grandes quantidades de dados, pode-se descobrir tendências temporais, por exemplo, ou relações entre produtos, e usar essa informação de maneiras úteis.

No campo da associação existem alguns algoritmos populares: Apriori, Eclat e FB-Growth. Todos os algoritmos produzem o mesmo resultado final, sendo que as diferenças entre eles prendem-se com o método utilizado e tempos de computação. [23] Nesta dissertação, optámos por usar o algoritmo Apriori, por ser o algoritmo mais importante e conhecido. O algoritmo Apriori vai produzir regras de associação, que são da forma

$$\{X\} \rightarrow Y$$

sendo que $\{X\}$ é um conjunto de uma ou mais variáveis e Y é apenas uma variável. Também se pode chamar antecedente ao lado esquerdo e consequente ao lado direito. Cada regra de associação tem alguns parâmetros a ela associada, nomeadamente a confiança e o suporte.

A **confiança** é a probabilidade condicional de o consequente ocorrer sabendo que o antecedente ocorre [21]. Por exemplo, uma confiança de 90% numa regra

$$\{X, Z\} \rightarrow Y$$

significa que em 90% das vezes que X e Z ocorrem, Y também ocorre. A confiança é útil para provar a fidedignidade de uma dada regra.

**Suporte** indica a frequência do antecedente em todo o *data set*. Isto é, se $\{X, Z\}$ tiver um suporte de 20%, significa que em 20% das transações, ocorre $\{X, Z\}$ [21]. O suporte serve

para garantir que um dado antecedente pertence a um padrão, ao ocorrer frequentemente. Um antecedente com um suporte muito baixo pode não pertencer a um padrão e ser apenas uma ocorrência pontual.

Um outro parâmetro útil para avaliar uma regra é o *lift*. *Lift* é um quociente que permite averiguar a independência de duas ou mais variáveis e é dado pela fórmula

$$\frac{supp(X \cup Y)}{supp(X) \times supp(Y)}$$

que relaciona o suporte das variáveis. Um *lift* igual a 1 significa que a probabilidade de ocorrência do antecedente e a probabilidade de ocorrência do consequente são independentes e, portanto, não é possível gerar regras. Um *lift* acima de 1 indica que há alguma dependência entre as variáveis, sendo que quanto maior for o *lift* maior é a dependência. Este parâmetro é útil para ajudar a medir a utilidade de regras.

Como mencionado, as funções nativas do R não são suficientes para gerar regras de associação. No entanto, uma das características que ajudou a tornar o R tão popular é o facto de esta linguagem ser facilmente expansível, ou seja, adicionar funções que originalmente não existem. Isto é possível através de *packages*. O CRAN ("Comprehensive R Archive Network") tem atualmente mais de 8000 *packages* disponíveis para *download*, tornando o R altamente personalizável e poderoso [3]. Neste caso em específico em que se pretende usar algoritmos de associação, basta instalar um novo *package* criado propositadamente para esse efeito e passamos a ter uma variedade de funções disponíveis.

### 2.2.2 Classificação

Tem como objetivo estudar conjuntos de dados e gerar modelos com base nesses dados. Depois, ao observar novos dados com padrão similar, vai utilizar o modelo gerado para conseguir classificar corretamente esses dados. Esta categoria pode ser especialmente relevante na saúde. Por exemplo, imaginemos que geramos um modelo de classificação com base num conjunto de dados de pacientes com um tumor na mama, que pode ser maligno ou benigno, e cujo diagnóstico é conhecido. Com esse modelo, será possível prever o diagnóstico em novos dados com uma grande precisão [19].

Existem diferentes algoritmos de classificação, sendo que alguns deles são:

- **Redes neuronais artificais** - É um algoritmo de classificação baseado no sistema nervoso central de um animal. Uma rede neuronal artificial contém nós, e as conexões entre nós têm um valor associado, chamdo de peso. Assim, quando um nó recebe um *input*, esses dados vão ser transformados, por exemplo, multiplicados pelo peso, e o nó envia os dados para outro nó. Quando um determinado valor limite for ultrapassado, é então gerado o *output*. Uma rede neuronal artificial tem a capacidade de fazer aprendizagem ao conseguir

aprender e alterar os valores do peso das conexões. Isto é, se um determinado caminho for melhor que outro terá um peso maior [15].

- **Árvores de decisão** - São árvores em que cada nó de decisão contem um teste num atributo e cada ramo corresponde a um possível valor deste atributo. Cada folha corresponde a uma classe e, consequentemente, cada caminho na árvore, desde a raíz até à folha, corresponde a uma regra de classificação.
- **Regressão logística** - É um método para classificar variáveis binárias. A regressão logística é usada para explicar a relação entre uma variável de classe, ou seja, a variável a ser classificada, e as restantes variáveis.
- **Naive Bayes** -
- **Support Vector Machines** -

### 2.2.3 Validação de modelos

Os algoritmos de classificação vão gerar modelos de previsão para novos dados.

### 2.2.4 Redes *bayesianas*

Uma rede *bayesiana* é um modelo que representa variáveis e as suas relações através de um grafo acíclico dirigido, ou seja, um grafo dirigido que não tem ciclos. [30] Uma rede deste tipo permite, por exemplo, calcular a probabilidade de uma determinada variável ter um determinado valor, tendo em conta as outras variáveis. Por exemplo, se criarmos uma rede *bayesiana* com a variável "Exercício" e outras variáveis como "Insulina", "Hidratos de carbono" ou "Doença", é possível calcular a probabilidade de uma pessoa ter hiperglicemia, sabendo, por exemplo, a quantidade de hidratos de carbono ingerida e o valor de insulina tomado. Também é possível saber, através de uma rede deste tipo, quais as variáveis com mais influência na variável a classificar. Neste trabalho em concreto, uma rede deste tipo pode ser útil para perceber de que forma os vários parâmetros podem ter influência sobre os valores de glicemia.

Para as análises com redes *bayesianas* serão utilizados dois programas: o Waikato Environment for Knowledge Analysis (WEKA) e o Sensivity Analysis, Modeling, Inference And More (SamIAm). O WEKA permite aplicar vários algoritmos de *machine learning* e é possível utilizá-lo para classificação, associação e visualização de dados, entre outros. Embora o WEKA também tenha incorporados algoritmos de associação, nomedamente o *apriori*, não é utilizado com esse intuito uma vez que as funções disponíveis pelo R são suficientes. Neste caso, o WEKA é usado para gerar uma rede *bayesiana* para cada utilizador que vai depois ser analisada por um outro *software*, o SamIAm.

O SamIAm é uma ferramenta desenhada propositadamente para a modelagem e trabalho com redes *bayesianas*. Com o SamIAm é possível fazer diversas análises mas para este trabalho interessa-nos saber como se relacionam as várias variáveis. Isto é, descobrir quais as variáveis que se relacionam mais com o valor de glicemia, por exemplo, e de que forma é que mudar uma determinada variável muda também a glicemia.

Por fim, iremos ainda utilizar uma área de *machine learning*, intitulada *Inductive Logic Programming*. ILP é uma área que permite formular hipóteses (ou regras) a partir da análise de dados.[muggleton.pdf] É composta por três partes:

- *background knowledge*
- *positive evidence*
- *negative evidence*

As hipóteses são formuladas tendo em conta estas três componentes, sendo que cada componente é um programa lógico. Ou seja, uma hipótese é obtida segundo a fórmula

```
Exemplos positivos + exemplos negativos + background knowledge => Hipotese
```

Para esta análise vai ser utilizado um sistema ILP intitulado Aleph (A Learning Engine for Proposing Hypotheses).

Estas análises serão descritas mais pormenorizadamente nos capítulos 5 e 6.

### 2.2.5 *Data Mining* na diabetes

No âmbito desta dissertação, o *data mining* pode ser útil para ajudar a manter os valores da glicose o mais estáveis possível, por exemplo, ao analisar os registos de um paciente durante um mês dos vários parâmetros, como horas das refeições, quantidade de hidratos de carbono a cada refeição, dose de insulina, exercício e doenças. O mais natural será que, algures durante o mês, existam valores demasiado altos e valores demasiado baixos. No entanto, para o paciente isto pode passar despercebido ou, mesmo que não, o paciente pode achar que os valores são isolados e que não têm nenhuma razão específica, e não lhes dar importância. Pode ser esse o caso, e de facto não haver nenhuma razão específica para um valor mais alto, mas também pode haver, e é aqui que o *data mining* pode dar uma ajuda preciosa: perceber o porquê de certos valores altos ou baixos existirem. Por exemplo, se um paciente fizer exercício uma vez por semana ao fim do dia, e depois não se alimentar adequadamente e tiver uma hipoglicemia no dia seguinte. No dia seguinte, ao perceber que está em hipoglicemia, o paciente pode até associar esse valor ao exercício do dia anterior. Mas também é possível que na próxima vez que fizer exercício já não se lembre do que aconteceu, e volte a cometer o mesmo erro. Neste caso, ao analisar os registos do paciente durante um mês, seria possível, através da associação, descobrir um padrão: a grande

maioria das vezes que o paciente faz exercício é seguida por uma hipoglicemia na manhã seguinte. Ao descobrir este padrão, é possível dá-lo a conhecer ao paciente para que este se possa alimentar melhor.

Assim, e imaginando que o paciente utilizaria uma aplicação com um sistema de aconselhamento, uma vez que este padrão fosse aprendido pela aplicação, sempre que o utilizador registasse que iria fazer exercício, ou que já tinha feito, a aplicação mostraria um aviso e aconselharia o paciente a comer mais nessa noite ou a tomar menos insulina. É em casos como estes que aplicar técnicas de *data mining* sobre dados de registos diabéticos pode fornecer uma ajuda importante no controlo da glicemia.

Neste capítulo concluem-se, fundamentalmente, duas coisas: 1) a diabetes, embora sem cura, pode ser controlada e permitir aos doentes levarem uma vida normal, e 2) que a tecnologia, nomeadamente a informática, cada vez mais apresenta ferramentas que possam dar um contributo importante.

# Capítulo 3

# Estado da Arte

Vimos anteriormente que a tecnologia pode ser útil ao serviço da medicina e também que existem dispositivos para medir a glicose e para controlar a glicemia e percebemos também que os *smartphones* podem ser úteis para a diabetes. Neste capítulo pretende-se analisar de que forma é que a tecnologia já está a ser usada para ajudar pacientes diabéticos e vamos abordar duas vertentes: 1) uso da medicina personalizada para controlar a diabetes, através de técnicas de *data mining* e 2) aplicações de registo de glicemias para *smartphones*. Medicina personalizada é a prática de tratar cada doente de forma individualizada, de acordo com as suas características, necessidades e preferências a cada momento, em vez de um tratamento generalizado para todos os pacientes [20].

## 3.1 Medicina personalizada e *data mining* na saúde

Nesta secção pretende-se abordar de que forma a área de *Data Mining* pode ser útil para a saúde. Vamos analisar algum trabalho feito na área da saúde utilizando técnicas de *Data Mining*, de uma forma geral, e também o que já foi feito em específico para a diabetes.

Estas técnicas podem ser utilizadas para fins diferentes: fazer aprendizagem analisando dados já existentes para que se possam criar modelos, que por sua vez irão classificar novos dados; encontrar relações entre variáveis e causas; detetar padrões.

Delen, Walker e Kadam [17] usaram diferentes algoritmos para tentar prever a sobrevivência ao cancro da mama. Neste caso, define-se por sobrevivência o paciente estar vivo pelo menos 5 anos após o diagnóstico do cancro. Foram usados três algoritmos de classificação diferentes: redes neuronais artificiais, árvores de decisão e regressão logística. Os autores usaram um *data set* já existente e, depois de todo o pré-processamento, como limpeza de dados, obtiveram um *data set* com 17 variáveis (16 variáveis de previsão e 1 variável de classe, isto é, a variável a ser prevista). Gerando modelos através dos três algoritmos utilizados, conseguiram classificar, com alta percentagem de precisão, se um dado paciente teria sobrevivido ou não. Além disso, conseguiram também descobrir quais as variáveis mais importantes para a classificação, e, portanto,

atribuir importâncias diferentes a diferentes variáveis. Os diferentes algoritmos conseguiram diferentes precisões: a rede neuronal teve uma precisão de 0.9121; a regressão logística teve uma precisão de 0.8920 e a árvore de decisão teve uma precisão de 0.9362. De notar que estes resultados foram obtidos usando *cross-validation*. *Cross-validation* é um método que divide um *data set* em duas partes: treino e teste. Neste caso, foi usada *10 fold cross-validation* o que significa que o *data set* foi dividido em 10 partes, ou seja, nove partes são usadas para treino e gerar um modelo. Esse modelo vai ser usado na parte restante para classificação e este processo é repetido dez vezes. Em cada repetição, o conjunto de teste é diferente. A precisão obtida nestes testes foram a média das dez repetições.

Palaniappan e Awang [27] criaram uma aplicação *web* para prever o risco de um dado paciente ter doença cardíaca. A partir de um *data set* com 909 registos, com 15 variáveis, usaram três algoritmos diferentes para calcular a probabilidade de um dado paciente ter uma doença cardíaca: Árvores de Decisão, *Naive Bayes* e Redes neuronais. Os registos foram divididos, em igual proporção, num conjunto de treino (455 registos) e conjunto de teste (454 registos). Obtiveram diferentes precisões para os modelos: *Naive Bayes* foi o modelo com maior precisão, 86.12%, seguido da rede neuronal com 85.68% e Árvores de decisão com 80.4%. Neste estudo, os autores conseguiram também encontrar relações entre variáveis. Por exemplo, conseguiram concluir que a variável "Tipo de dor no peito" é a mais influente relativamente a uma doença cardíaca. Conseguiram também obter algumas regras que ajudam a prever, com alta percentagem de correção, se um dado paciente tem doença cardíaca ou não. Uma das regras geradas foi

```
Chest Pain Type = 4 and CA = 0 and Exang = 0 and Trest Blood Pressure >=
    146.362 and < 158.036
```

que diz que 99.61% dos doentes cardíacos cumprem estes requisitos.

Stilou, Bamidis, Maglavers e Pappas [31] aplicaram o algoritmo *apriori* num *data set* com 100 registos de pacientes diabéticos, para tentar gerar regras de associação. Cada registo equivale a um paciente e tem variáveis como idade, regime de insulina, glicose objetivo, glicemia estável ou instável, entre outros. Neste estudo o objetivo era obter conhecimento sobre uma base de dados de pacientes diabéticos e gerar regras com o conhecimento obtido. Uma das regras geradas é

```
IF diabetes mellitus type = 2 AND special condition = no AND target = good AND
    unstable diabetes = no THEN regime = 2
```

Neste caso, regime é a proposta de insulina por dia, sendo que "2" corresponde a duas injeções de insulina mista, com ação curta e intermédia, uma ao pequeno-almoço e uma à tarde.

Han, Luo, Yu, Pan e Chen [22] usaram algoritmos de classificação para gerar um modelo de diagnóstico da diabetes. Neste caso, usam-se SVM's (*support vector machines*) e um *data set* com 56 variáveis que foi dividido em duas partes: 90% para o conjunto de treino e 10% para conjunto de teste. Foi usada *10 fold cross-validation* como método de treino para obter os parâmetros ideais para os modelos. Depois deste processo, a melhor *fold* é escolhida para gerar

conjuntos de regras e para ser usada na classificação do conjunto de teste. Contudo, SVM's têm uma natureza *black-box*, isto é, são capazes de classificar dados mas não são capazes de explicar o porquê dessa mesma classificação. Isto significa que, usando apenas SVM's, não é possível extrair regras. Face a este inconveniente, os autores decidiram combinar SVM's com outros dois algoritmos: *Random Forests* (RF) e C4.5, um algoritmo para árvores de decisão. A combinação de SVM's com outros algoritmos *white-box* já vem sendo utilizada noutros estudos. [**?** **?** ] Neste caso, conseguiram-se gerar regras que ajudam a classificar dados como pertencendo a pacientes diabéticos ou não-diabéticos. Uma das regras geradas é, por exemplo,

```
If HBA1C > 7.15 and HDL > 1.57 and CHOL > 5.9 and AGE>77, then diabetic
```

e outra é

```
If HBA1C > 7.25, then diabetic
```

Os dois algoritmos usados, SVM + RF e SVM + C4.5 conseguiram, respetivamente, 89.6% e 86.3% de precisão.

Após a revisão bibliográfica acerca do uso do *data mining* na medicina, observa-se que a maioria dos trabalhos são para efeitos de classificação. Na pesquisa efetuada sobre o uso de *data mining* só se encontrou um estudo sobre regras de associação para a diabetes, que foi o estudo acima analisado. Esse estudo, apesar de usar um algoritmo de associação para gerar regras, não faz o que é pretendido nesta dissertação. O que se pretende neste trabalho é aplicar esse mesmo algoritmo mas para cada paciente de forma individual, com vários registos ao longo do tempo. Desta forma geram-se regras personalizadas para cada paciente e que, portanto, serão regras específicas para que o paciente possa ter um melhor controlo sobre a sua glicemia. Da pesquisa efetuada não foi encontrado nenhum outro trabalho com uma análise personalizada para cada paciente o que torna este, neste aspeto, diferente do que já foi feito.

## 3.2 Aplicações para *smartphones* Android

Estima-se que em 2016 o número de utilizadores de *smartphones* seja, em todo o mundo, de 2.08 mil milhões [11]. Por outro lado, são ferramentas cada vez mais poderosas e tem havido um crescimento no desenvolvimento de aplicações para saúde e bem-estar. De seguida vamos analisar algumas das aplicações existentes para a diabetes. Para esta análise foram consideradas apenas aplicações para Android, pois é o sistema operativo móvel mais usado no mundo [7] e porque a aplicação na qual este projeto se baseia também é para Android. Foram escolhidas cinco aplicações da *Google Play* com base no número de *downloads* e no número de *ratings*. Cada aplicação foi instalada e testada com o intuito de perceber aquilo que oferece ao utilizador. Alguns dos parâmetros a testar são as variáveis que as aplicações permitem registar e o seu visual. Todas as aplicações escolhidas são grátis.

### 3.2.1 Diário da Diabetes mySugr

Esta aplicação permite ao utilizador adicionar registos e cada registo permite especificar alguns parâmetros, como o nível de glicemia, hidratos de carbono consumidos, tipo de insulina e tipo de refeição. Cada registo pode ser acompanhado para uma foto, caso seja uma refeição, e pode ser também escolhido um tipo para cada registo, como por exemplo "almoço", "jantar", "hipoglicemia", entre outros. Para cada registo é ainda possível escolher um outro tipo que dá mais informação, como "Stressado", "Doente", "Álcool", mas não só. De nota também que é possível especificar o tipo de alimentos caso o registo se trate de uma refeição. Entre os tipos de alimentos existem, entre outros, "Legumes", "Carne", "Peixe", "Ovos", etc.

Esta aplicação permite a sincronização com um glicosímetro, o "iHealth BG5". É ainda possível definir metas como limite para hipo e hiperglicemia, e metas de peso ou exercício. Uma característica interessante da aplicação é ter um sistema de pontos e de desafios. Os desafios são diversos, como por exemplo "Caminhada para a cura", que incentiva o utilizador a registar pelo menos 30 minutos de exercícios em 24 horas. Desafios completos desbloqueiam novos desafios.

Por cada registo efetuado ganha-se uma quantidade de pontos, que é maior quantos mais parâmetros forem preenchidos em cada registo. A aplicação tem um pequeno boneco animado que vai sendo desbloqueado com pontos. Estes dois sistemas são interessantes porque podem funcionar como um incentivo extra para o uso regular da aplicação.

Por fim, a aplicação possibilita a exportação dos registos efetuados para três formatos possíveis: xls, pdf ou csv. Esta característica, no entanto, está disponível apenas na versão paga.

### 3.2.2 Diabetes:M

Esta aplicação permite o registo de glicose, hidratos de carbono consumidos, insulina de efeito rápido e longo, peso, colesterol, pressão arterial, atividade física e hemoglobina glicada. À primeira vista, nota-se logo o ecrã principal que se pode tornar confuso pela grande quantidade de botões que oferece. As funções disponibilizadas são bastante semelhantes às da aplicação anterior. Uma função nova é a de alarme, que ajuda os utilizadores a não se esquecerem de medir a glicose. Em termos de visualização dos dados inseridos, a aplicação mostra os mesmos em forma de gráficos para se poder acompanhar os registos num determinado intervalo de tempo. É possível verificar que se podem usar unidades de medida diferentes para os vários parâmetros. Por exemplo, para a glicemia pode-se usar mg/dL ou mmol/L. Uma vantagem do ecrã principal é mostrar a quantidade de insulina ativa presente num dado momento. Isto é, se um utilizador tomar 5 doses de insulina, a aplicação mostra, ao longo do tempo, um valor denominado "Insulina Ativa", ou seja, a previsão da insulina que "sobra" desde a última toma.

Uma outra característica interessante é a de possibilitar sincronização com aplicações externas, como Dropbox, Google Drive e Google Fit. A aplicação permite ainda fazer *backup* dos dados.

É também possível exportar e importar dados nos formatos csv e xls, bem como importar

dados de glicosímetros de diferentes modelos, tais como OneTouch, Dexcom ou Accu-Chek.

### 3.2.3 OnTrack Diabetes

Esta aplicação permite registar glicose, refeições, exercício, medicação, peso, pressão arterial, pulsação e HbA1c. Tem uma interface bastante simples relativamente às outras aplicações experimentadas. Tem apenas três menus no ecrã principal, que permite ver relatórios, o histórico e alguns gráficos relativamente aos dados inseridos. O ecrã principal mostra também as médias dos níveis de glicose diários, semanais e mensais. Ao explorar a aplicação foi possível verificar que esta oferece vários gráficos. Por exemplo, é possível visualizar, através de gráficos, valores de glicose, média diária de glicose, glicose por hora do dia, exercício, etc.

Ao consultar o menu "Histórico" os dados aparecem na forma de lista e por ordem de refeição, ou seja, para um mesmo dia, os dados relativamente ao pequeno almoço aparecem antes do jantar. Este menu apresenta, portanto, todos os dados registados em cada dia. No menu "Relatórios", podemos observar médias de glicose, que são diárias, semanais, mensais ou trimestrais. Existe uma outra opção chamada "glicose por categoria", que mostra os valores médios da glicose registados em cada tipo de refeição. Uma outra funcionalidade, "Logbook", permite a visualização dos dados através de gráficos, permitindo ver qualquer parâmetro registado e partilhar esses mesmos gráficos por *e-mail*.

É possível exportar os dados para csv, xml ou html. É também possível criar *backup* ou apagar todos os dados num determinado intervalo de tempo.

### 3.2.4 Diabetes - Diário Glucose

De todas as aplicações analisadas, esta é a mais simples. É a que menos funções oferece, permitindo registar apenas o peso e a glicose, que é feito no ecrã principal. A aplicação é composta por outros três separadores que permitem visualizar os níveis de glicose em lista e em gráfico. É possível exportar os dados registados para um ficheiro pdf ou partilhar por *e-mail*.

### 3.2.5 Glucose Buddy: Diabetes Log

Esta aplicação permite registar o tipo de diabetes, peso, altura, pressão arterial, glicose, HbA1c, exercício, refeições e a atividade do registo (refeição, antes de exercício, depois de exercício, etc.).

Pode-se observar os registos de glicose em forma de lista, utilizando o menu "Logs" ou em forma de gráfico usando o menu "Graphs". No gráfico pode-se visualizar apenas o parâmetro da glicose bem como a média de todos os valores registados por dia.

A aplicação oferece ainda um alarme que pode ser ativado para uma determinada hora ou então pode ser coordenado com um evento. Por exemplo, o utilizador pode definir um alarme

para 30 minutos depois do almoço, sendo que quando fizer um registo com o tipo de refeição "almoço", ativará o alarme para o tempo definido.

É possível exportar os registos selecionando intervalos pré-estabelecidos pela aplicação e enviar para o *e-mail*.

Como se pode perceber, as aplicações não diferem muito entre si e todas elas oferecem praticamente as mesmas funcionalidades, que são de registo e visualização de dados. Desta forma, pode-se concluir que um sistema de aconselhamento numa aplicação para registo de glicemias será um aspeto inovador. Para este trabalho vamos utilizar a aplicação MyDiabetes, que será também a aplicação onde o sistema desenvolvido será integrado. Uma vez que para este projeto poder ser feito, será necessário recolher dados de pacientes diabéticos, a aplicação MyDiabetes será também a plataforma para a recolha desses dados, através da utilização da aplicação por pacientes diabéticos. Os motivos para a necessidade de recolha dos dados serão explicados no capítulo 5. O próximo capítulo descreve de forma mais detalhada a aplicação MyDiabetes.

# Capítulo 4

# MyDiabetes

Neste capítulo vamos analisar com mais detalhe a aplicação utilizada neste projeto de dissertação. A aplicação chama-se *MyDiabetes* e foi desenvolvida no âmbito do mesmo projeto em que esta dissertação se insere.

Um dos objetivos opcionais é a integração de um sistema de aconselhamento personalizado numa aplicação para registo de glicemias. Assim, e embora a aplicação, de momento, ofereça apenas a possibilidade de registo e visualização de dados, no futuro vai ter um sistema de aconselhamento. A aplicação permite também que os utilizadores enviem os seus dados, para que estes possam ser analisados. Esta é uma ferramenta importante visto que houve a necessidade de criar o nosso próprio *data set*. Essa necessidade é explicada no capítulo 5.

## 4.1 Objetivo da aplicação

O objetivo desta aplicação já foi mencionado anteriormente: ajudar o doente diabético, ao oferecer uma ferramenta alternativa que permita registar e visualizar todos os parâmetros importantes, como glicose, insulina e hidratos de carbono. O *smartphone* é um dispositivo bastante interessante para ter aplicações como esta: a maioria das pessoas tem um e portanto, se tiver uma aplicação pode registar a glicemia a qualquer hora e em qualquer lugar. Além da função de registo, a aplicação permite também a visualização dos registos efetuados em forma de gráfico. Assim torna-se mais fácil detetar hiperglicemias, por exemplo. Ou visualizar registos de meses anteriores ou até mesmo mostrar ao médico, durante a consulta, os valores de glicemia no período entre as consultas. Tudo isto contribui para o objetivo principal: tornar mais simples e eficaz o controlo da glicemia.

Para esta dissertação, o objetivo da aplicação foi que os utilizadores a pudessem utilizar para que se familiarizassem com uma aplicação deste tipo e também para enviar os dados necessários para análise posterior.

## 4.2 Arquitetura

A aplicação é bastante simples e facilmente um utilizador se habitua às suas funcionalidades. Sempre que um utilizador a use pela primeira vez é necessário preencher os seus dados. Todos os dados são obrigatórios mas não relevantes para esta dissertação, como nome, data de nascimento ou altura, por exemplo. Desta forma, qualquer utilizador que quisesse enviar os seus registos para o projeto de forma anónima poderia fazê-lo, bastando para isso meter dados falsos. Por outro lado há dados que são relevantes para o cálculo da insulina, como o fator de sensibilidade. Outro tipo de dados como os limites para hipo e hiperglicemia são úteis para as futuras funcionalidades que a aplicação possa vir a ter, como a amostragem de avisos, já que alguns avisos são baseados nestes valores. De notar que estes dados são registados aquando da primeira utilização mas podem ser alterados a qualquer momento a partir do menu das definições. A aplicação em si é constituída por um ecrã principal que tem os submenus de registo, tais como refeições, exercício, insulina, entre outros. Tem também o submenu "Logbook" que permite visualizar registos anteriores, através da escolha de um intervalo de tempo. Estes dados são mostrados na forma de gráfico e de lista.

A aplicação é bastante intuitiva: tem menus de registo, como registo de refeições, insulinas, exercício ou doenças. Esses registos são guardados no *smartphone*, numa base de dados *sqlite*. Tem também um menu para visualização dos registos efetuados, o "Logbook". No menu de definições é possível aceder a outras funcionalidades da aplicação. É possível alterar os dados, como já mencionado, mas também enviar os registos para o projeto ou gerar um relatório. Este relatório pode ser especialmente útil para mostrar ao médico na consulta pois é possível imprimi-lo e mostrar os valores de qualquer parâmetro que o utilizador queira, como glicemias por exemplo. É possível escolher, por exemplo, todas as glicemias registadas entre consultas, exportar para pdf e depois imprimir, sendo um registo limpo e de fácil leitura.

## 4.3 Variáveis recolhidas

Como já explicado anteriormente, a aplicação permite recolher qualquer tipo de dados que possa ser relevante para um bom controlo da glicemia. Além dos valores da glicose, hidratos de carbono e insulina, a aplicação permite também a recolha de outros dados, como doença e exercício, que são eventos que têm um impacto direto nos valores de glicemia. Existe também a opção de registar a hemoglobina glicada. É possível ainda registar outros dados que, apesar de terem menos ou nenhum impacto na diabetes, podem ser indicadores do estado de saúde do utilizador e portanto permite também ter um maior controlo sobre eles, como pressão arterial, colesterol ou peso. Sempre que é efetuado um registo, seja de que parâmetro for, a aplicação regista também o dia e hora desse mesmo registo. Saber a hora e dia de cada registo será especialmente importante para o objetivo de detetar padrões ou anomalias, ou também para análises estatísticas, uma vez que possibilitará dar uma ideia ao utilizador dos dias ou períodos do dia em que a glicemia é mais elevada. De notar que nem todos os dados registados pelos utilizadores foram recolhidos para

esta dissertação. No entanto, todos os que foram recolhidos foram usados única e exclusivamente para a análise.

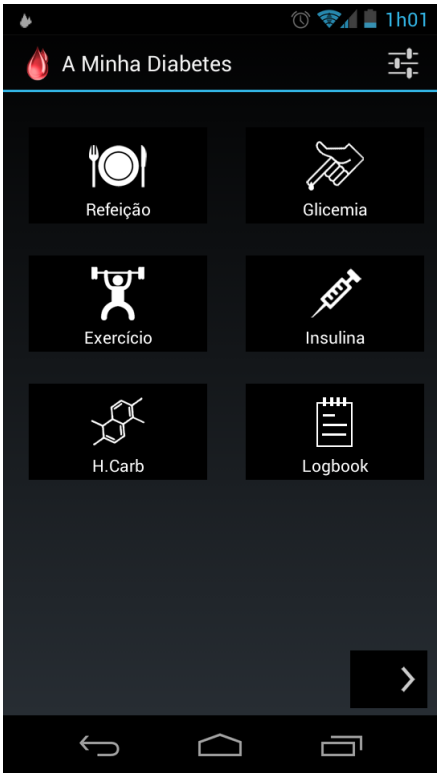


Figura 4.1: Menu principal da aplicação MyDiabetes

# Capítulo 5

# Análise de dados

Neste capítulo será feita uma descrição do trabalho: no que consiste, quais os requisitos para um trabalho deste tipo e as metodologias abordadas. Serão feito dois tipos de análises neste trabalho: uma análise estatística básica que servirá de introdução para um outro tipo de análise, envolvendo ferramentas direcionadas para o *data mining*, ainda que este último tipo de análise seja feita apenas no próximo capítulo.

## 5.1 Descrição do estudo

Para que este estudo pudesse ser feito, o primeiro passo era ter um conjunto de dados de pacientes diabéticos de forma a ser analisado. Este conjunto de dados tinha que ter algumas características específicas para que pudesse ser utilizado da forma pretendida: tinha que ser um conjunto de registos para cada paciente ao longo de algum tempo, sendo que idealmente, no mínimo, cada paciente teria registos correspondentes a quatro semanas. Depois de uma pesquisa, foi possível concluir que não existia, na *web*, qualquer *data set* com estas características. Para que este estudo pudesse ser feito, era necessário ter um *data set* com os diferentes parâmetros que pudessem ser relevantes, como glicemia, insulina, hidratos de carbono e exercício. Como já mencionado, nenhum *data set* existente *on-line* tinha as características desejadas pois só possuíam um registo por pessoa e/ou o seu propósito era a classificação de um paciente como diabético ou não. O único *data set* encontrado que tinha vários registos por pessoa ao longo do tempo tinha o inconveniente de ter apenas dados relativos a glicemias, pelo que não é suficiente [4].

Face à falta de dados disponíveis, a solução encontrada foi recolher dados para a criação do *data set* desejável. Ao podermos construir o nosso próprio *data set* tínhamos a vantagem de podermos recolher as variáveis que queríamos, e portanto, ter um conjunto de dados que torne este trabalho mais eficiente. Por outro lado, visto que a recolha dos dados seria feita através da aplicação MyDiabetes, havia a desvantagem desta ser ainda fechada ao público e, portanto, não ter doentes diabéticos a utilizá-la. Uma outra desvantagem era o tempo que a criação de um *data set* com a variedade e tamanho desejáveis poderia levar. Uma maior variedade de dados de

diferentes de pessoas seria útil para também haver a possibilidade de uma maior variedade de rotinas diferentes e, portanto, regras diferentes. Por outro lado, a quantidade relativa a cada utilizador seria também importante para que os padrões encontrados para cada utilizador sejam mais fidedignos, isto é, um padrão encontrado em registos relativos a algumas semanas tem mais confiança que um padrão encontrado em registos de apenas alguns dias. Posto isto, o objetivo era recolher dados referentes ao máximo de utilizadores possível e durante o máximo de tempo possível, desde o início até ao fim deste trabalho, para que a variedade e quantidade disponíveis fossem maiores.

Feitas as contas, apesar das desvantagens enunciadas, a solução seria mesmo recolher os dados através da aplicação, sendo que o primeiro passo seria arranjar voluntários. O processo de recolha de dados é descrito na próxima secção.

## 5.2 Recolha de dados

A recolha de dados foi feita em parceria com o Hospital de S. João, através do seu serviço de endocrinologia. Para tal, foi pedido à comissão de ética autorização para falar com os pacientes diabéticos do hospital e para que estes utilizassem a aplicação, enviando os respetivos registos, pedido esse que foi aceite. Qualquer paciente seria elegível para o projeto desde que tivesse mais de 18 anos e fosse insulino-dependente.

O Dr. Celestino Neves, médico endocrinologista daquele serviço, serviu de ponte entre a faculdade e o hospital. Nesta primeira fase, o Dr. Celestino dava uma pequena explicação do projeto ao paciente antes, durante ou depois da consulta, sendo que depois re-encaminhava o paciente para nós, investigadores. Esta primeira abordagem do Dr. Celestino era bastante importante uma vez que tornava os pacientes mais recetivos à participação. No nosso encontro com os pacientes, explicávamos o que era o projeto, mostrando no que a aplicação consistia e como funcionava. Depois desta parte, dávamos a conhecer ao paciente os passos seguintes, nomeadamente a integração de um sistema de conselhos baseados no *input* do utilizador. Durante este processo, explicávamos ao paciente a importância de termos dados reais de utilizadores e portanto a importância do envio dos registos, realçando que a longo prazo, os utilizadores seriam os maiores beneficiados, pois poderiam melhorar o seu controlo da glicemia. Os pacientes eram ainda informados que os registos enviados seriam utilizados apenas para fins de investigação e que, caso aceitassem participar no projeto, teriam de assinar um consentimento informado que explicava isso mesmo.

Se os pacientes aceitassem fazer parte do projeto, a aplicação era então instalada nos seus *smartphones* e seria-lhes também dado acesso à aplicação na *Google Play*. Assim, a aplicação passaria a estar sempre disponível para esse utilizador, mesmo que este formatasse ou trocasse o telemóvel, a partir da loja. Outra vantagem deste acesso seria os utilizadores poderem usufruir das atualizações entretanto feitas: sempre que houvesse uma atualização, apareceria o aviso e o utilizador poderia atualizar sem perder os registos já feitos.

Além da aplicação, eram também dadas aos voluntários as credenciais para terem acesso ao *website* do projeto, para onde poderiam exportar os registos efetuados e visualizá-los através de gráficos. As credenciais garantiam assim que apenas os voluntários tinham acesso à parte privada da página, que tem tutoriais e a parte de visualização dos dados.

### 5.2.1 Descrição da experiência e *feedback* dos utilizadores

O processo em cima referido, era feito duas vezes por semana, que correspondia aos dias de consultas da diabetes, e decorreu durante aproximadamente três meses. Durante este tempo falámos com várias dezenas de pacientes, sendo que foi possível obter, logo à partida, uma conclusão: a esmagadora maioria dos pacientes com quem falámos estavam bastante recetivos à ideia e concordavam que podia ser uma mais-valia na melhoria da sua qualidade de vida. Muito poucos pacientes tinham a opinião de que a aplicação não teria utilidade, principalmente devido ao facto de estes utilizarem bomba, que torna o processo de controlo da glicemia muito mais reduzido e simples. Além da ideia em si, que foi bastante bem recebida, os utilizadores também deram *feedback* positivo relativamente ao *design* e funcionalidades da aplicação.

Apesar de todos os pacientes acharem boa ideia ter uma aplicação como esta, nem todos se tornaram voluntários devido a diferentes razões: 1) nem todos os utilizadores tinham *smartphone* e cerca de metade dos pacientes tinha outro sistema operativo que não Android e 2) alguns pacientes com idade mais avançada ou com outros problemas de saúde simplesmente não tinham tempo ou disponibilidade para participar como voluntário. No final deste período tínhamos conseguido a participação de 31 voluntários. Este número dava-nos alguma garantia de variedade bem como quantidade de dados mas, como mais tarde viríamos a concluir, nem todos os voluntários enviaram, de facto, os registos.

Dos 31 voluntários apenas 8 enviaram registos pelo menos uma vez e apenas 5 destes enviaram registos regularmente e relativamente a algumas semanas. Este foi um dos principais problemas encontrados: para integrar um sistema de aconselhamento baseado nos dados introduzidos numa aplicação para *smartphone* é fundamental saber que tipo de avisos ou conselhos se deve ou não mostrar. Por exemplo, um conselho como "Ajuste a insulina à quantidade de hidratos de carbono ingerida" é algo óbvio e que qualquer paciente diabético sabe, e portanto não seria uma grande mais valia para a aplicação. Por outro lado, um aviso como "Hoje é segunda-feira e geralmente à segunda-feira tem valores de glicemia mais elevados" pode ser útil porque possivelmente é um padrão que o paciente não detetou. Neste caso, este conselho poderia fazer o utilizador controlar mais frequentemente a sua glicemia às segundas-feiras e, portanto, normalizar os valores das segundas-feiras a partir daí.

Esta distinção entre avisos ou conselhos realmente úteis ou descartáveis é importante: se a aplicação mostrar demasiados conselhos intuitivos ou regras "banais" das quais os utilizadores já tenham conhecimento, estes podem acabar por achar que a aplicação não traz benefícios. Pelo contrário, se a aplicação mostrar conselhos baseados em comportamentos não percetíveis pelos utilizadores, estes podem ser alertados.

Por isso mesmo, e tal como explicado anteriormente, torna-se importante obter a maior quantidade e variedade possível de dados, para que tenhamos uma melhor distinção entre regras realmente importantes ou não. Imaginemos que temos apenas um *data set* de um paciente com registos de uma semana. Esta quantidade de dados não permite concluir nada sobre eventuais regras descobertas nem sequer permite descobrir vários tipos de regras; tanto a variedade de pacientes como a quantidade de registos serão demasiado reduzidas para tal.

Embora a quantidade de utilizadores que usaram regularmente a aplicação tenha ficado um pouco áquem das expetativas, estes utilizadores fizeram uma quantidade de registos aceitável o que permitiu mostrar que, mesmo com esta quantidade reduzida de dados, é possível extrair diferentes padrões para diferentes utilizadores.

O objetivo da aplicação é gerar os tais conselhos de forma automática e, pensado desta forma, esta parte inicial pode parecer inútil: para quê estar a gerar regras se elas não vão ser utilizadas na aplicação em si?

Esta questão prende-se, uma vez mais, com a filtragem de regras úteis ou não. Os registos de um único paciente podem gerar, por exemplo, centenas ou até milhares de regras. Obviamente que em milhares de regras muitas vão ser bastante parecidas e, possivelmente, a maioria vão ser regras que não têm relevância. É possível, por exemplo, que em alguns milhares de regras só se consiga aproveitar 5 ou 6, ou até 0. A importância de ter registos de vários pacientes prende-se com a variedade das regras geradas: dois pacientes podem gerar poucas regras cada um mas essas regras serem diferentes, ou seja, relativas a rotinas ou anomalias com efeitos ou causas diferentes. Esta análise dos dados à procura de regras servirá, então, para construir um conjunto de regras passíveis de ocorrer no dia-a-dia de um diabético que serão então integradas nesta aplicação. Assim, serão apenas detetados padrões e situações que possam ter alguma relevância e conselhos como "Adeque a insulina aos hidratos de carbono" não serão mostrados. Isto porque, apesar de ser um bom conselho, também é óbvio e portanto provavelmente será tido como um conselho inútil por parte dos utilizadores.

Ao ter esta filtragem para mostrar apenas conselhos realmente importantes, não só diminuiremos a quantidade de vezes que algum conselho é mostrado, tornando assim a aplicação menos "chata" para os utilizadores, como também tornam os conselhos mais especiais, ou seja, se um conselho ou aviso é mostrado é porque o utilizador tem de facto algum comportamento que pode ser corrigido.

Neste sentido, a reduzida quantidade de voluntários que enviaram registos, embora útil para análise, não é suficiente para o desenvolvimento de um sistema de aconselhamento, pelo menos tão fidedigno como desejável. Assim, esta foi a primeira limitação encontrada, pelo que o objetivo de desenvolver o sistema de aconselhamento, que era opcional, foi posto de lado. Serão feitos então diferentes tipos de análises com os dados recolhidos, que vão ser descritas neste capítulo.

Já com os dados dos utilizadores, o passo seguinte era o seu tratamento para depois serem utilizados.

### 5.2.2 O *data set*

Os registos dos utilizadores são armazenados no telemóvel numa base de dados *sqlite* que é depois convertida para um ficheiro csv através de um *script* [9]. O *script* converte cada uma das tabelas num ficheiro csv e depois, os ficheiros csv que interessa manter, são copiados para um único ficheiro através do comando da *shell*, *copy*.

Depois de selecionar as variáveis pretendidas e as juntar no mesmo ficheiro, o *data set* fica com a estrutura apresentada na tabela 5.1

| **Variável** | **Tipo** |
|---|---|
| DateTime | Integer |
| Value_Carbs | Integer |
| Value_Glucose | Integer |
| Target_BG | Integer |
| Value_Insulin | Double |
| Exercise | Integer |

Tabela 5.1: Tipo das variáveis recolhidas

A variável **DateTime** diz respeito ao dia e hora exatos de cada registo; **Value_Carbs** é a quantidade de hidratos de carbono; **Value_Glucose** é o nível de glicemia à hora do registo; **Target_BG** é o objetivo de glicemia para a hora do registo, pois o utilizador pode ter diferentes objetivos de glicemia para diferentes períodos do dia. Esta variável é importante para o cálculo da insulina a administrar. **Value_Insulin** é a quantidade de insulina tomada pelo utilizador à hora do registo. **Exercise** corresponde ao exercício feito à hora do registo.

Este *data set* não representa, no entanto, todos os dados enviados pelos utilizadores, pois há alguns dados que não são relevantes para a análise e por isso são descartados. Alguns dos dados descartados são os dados pessoais do utilizador, que incluem, entre outros, nome, idade, peso ou altura. A forma como a aplicação funciona faz com que, a cada registo, sejam registados não só os valores dos parâmetros como outras informações: em cada registo são guardados outros dados como o ID do registo ou a "Tag", que é uma variável opcional que pode associar um registo a um evento. Por exemplo, ao registar uma refeição é possível escolher várias "Tags" como "pequeno-almoço" ou "almoço"; é também guardada uma variável "Note" que permite ao utilizador adicionar uma nota a uma medição. Por exemplo, se o utilizador fizer um registo de exercício, pode adicionar uma nota com "Corrida". Portanto, o registo de cada parâmetro faz com que estas variáveis também sejam guardadas, mesmo que o utilizador não as preencha. O facto de descartarmos estas variáveis não significa que não tenham importância: a variável "Note", por exemplo, pode ser importante se o utilizador quiser adicionar alguma informação extra que ache relevante para depois se lembrar ou até mostrar ao médico. Apenas não são relevantes para este tipo de análise: uma nota sobre uma refeição não pode ser utilizada diretamente por um

algoritmo. O tratamento deste tipo de dados, textual, está fora do âmbito desta dissertação, pelo que esses dados não foram considerados.

Neste ponto, temos as variáveis com as quais iremos trabalhar, mas que ainda precisam de ser pré-processadas. O pré-processamento do *data set* será descrito de seguida.

### 5.2.3 Pré-processamento dos dados

Uma parte crucial de *data mining*, ainda antes de aplicar quaisquer técnicas, é a parte de pré-processamento dos dados. Esta parte é composta por várias etapas:

- Limpeza dos dados
- Redução dos dados
- Transformação dos dados

O pré-processamento dos dados é necessário para tratar inconsistências que possam existir no *data set*. Estas inconsistências podem ser valores em falta ou valores errados. Por exemplo, num *data set* com um campo "Idade", um valor negativo neste campo é um valor errado. Então, para chegar a um estado em que o *data set* esteja pronto a ser utilizado, é preciso percorrer um ou mais dos passos acima mencionados.

**Limpeza dos dados** - é o primeiro passo a fazer num conjunto de dados. Neste caso específico, o *data set* vinha com alguns valores em falta ou valores errados. Apesar destes casos serem uma percentagem pequena do total, é importante que sejam resolvidos. Para valores em falta há duas opções: ou remover o registo em que um dos valores falte, ou tentar prever o falor em falta e preenchê-lo. A segunda opção pode ser feita obtendo, por exemplo, a média dessa variável e, em todos os registos com valor em falta, colocar a média. No entanto, para o caso específico da diabetes esta alternativa não parecia a mais viável. Por exemplo, se num dado registo faltar o valor da glicemia, não faz sentido encontrar a média da glicemia para todos os registos e colocar no registo em falta. A glicemia pode oscilar bastante e portanto, estar a preencher um valor em falta com um valor médio da glicemia pode estar a comprometer a veracidade da análise posterior: um registo que até podia ter um valor de hiperglicemia estaria a ser substituído com um valor de glicemia mais normal e portanto, esse registo já não seria uma exceção. Repetindo isto para todos os valores de glicemia em falta, estaria-se a normalizar situações que podiam não ser normais.

Posto isto, a alternativa escolhida foi a de deixar todos os valores em falta assim, sendo que no R serão tratados como NA e, portanto, não terão efeito sobre a descoberta de regras. Contudo, para as médias dos valores de glicemia, valores NA não podem existir e portanto, apenas para as análises feitas utilizando algum valor média, estes valores foram removidos. Para todas as outras análises valores NA são mantidos por não provocarem qualquer efeito. Quanto

a valores impossíveis, estes foram removidos. Por exemplo, alguns registos tinham valores de glicemia “0” ou “7”, que não são valores possíveis para glicemia, tratando-se por isso de um erro de registo. Assim garante-se que todos os registos tenham valores e que esses valores sejam valores realistas. Esta medida foi tomada apenas para as variáveis necessárias para análise, isto é, para uma análise apenas com a glicemia, as outras variáveis não são tidas em conta e portanto não são limpas. Já para uma análise de procura de regras, como todas as variáveis são utilizadas, todas as variáveis são limpas.

Isto garante também que as regras geradas não são enviesadas: se quisermos descobrir regras que relacionem os valores de insulina e hidratos de carbono com os valores de glicemia, mas eliminarmos registos que eram 0, não estaríamos a usar tantos registos como os possíveis. Por exemplo, para uma dada linha, uma das variáveis pode não ter valor mas, se as outras tiverem, então essa linha pode ter a sua contribuição, sendo que as variáveis com valores poderão ser úteis. Se apagássemos as linhas que têm uma variável sem valor, estaríamos a eliminar registos, que por sua vez tornariam os padrões menos frequentes.

**Redução dos dados** - Num *data set* nem todos os dados têm a mesma relevância: uns são importantes e outros são menos importantes ou até mesmo irrelevantes. É importante perceber quais os dados que não interessa ter, pois assim vamos reduzir a quantidade de regras irrelevantes que são geradas. Por isso mesmo, o segundo passo seria analisar todos os dados recolhidos e perceber quais os que valiam a pena manter e os que se podia remover. Um exemplo que ajuda a perceber melhor este passo é olharmos para os valores de glicose, insulina e hidratos de carbono: na aplicação, cada um destes parâmetros permite registar o valor em si mas também adicionar uma nota para acompanhar o registo. Essa nota poderá ser uma breve descrição feita pelo utilizador aquando de um registo. Existe também um outro atributo chamado “Tag” que permite, por exemplo, associar uma refeição a um período do dia. Se a “Tag” 1 for “pequeno-almoço”, então sempre que o utilizador regista o pequeno-almoço, pode escolher essa fase do dia, e esse registo ficará com a “Tag” 1. Como se pode perceber, atributos como “Note” ou “Tag” podem ajudar a perceber o porquê de alguma alteração nos valores de glicose mas não têm qualquer uso para a descoberta de padrões e consequente geração de regras. Portanto, atributos como estes podem ser retirados do *data set* de forma a reduzir a quantidade de informação não relevante. Existem outros atributos que foram retirados por não apresentarem qualquer utilidade para o resultado pretendido, tais como, “Idade”, “Nome”, “Altura” ou “Sexo”. No final deste passo, o conjunto de dados é consideravelmente mais pequeno, em termos de quantidade de variáveis, e portanto permite fazer uma análise mais objetiva, com menos informação desinteressante.

**Transformação dos dados** - Por fim, este passo serve para transformar os valores existentes em valores que possam ser utilizados da forma mais conveniente. Ter uma variável "DateTime"no formato “AAAA-MM-DD HH:MM” não é tão útil visto que o dia e hora, que até podem ser variáveis interessantes, estão numa só variável e torna o seu uso mais difícil. Neste caso, seria mais vantajoso separar as duas variáveis e portanto, transformar “DateTime” em “Day” e “Hour”. Ainda assim, ter um dia do ano e uma hora específica do dia não é exatamente o formato mais útil, pois são variáveis demasiado específicas e por isso pouco repetidas, ou até nunca repetidas,

pelo que não irão contribuir para a descoberta de padrões. Para melhorar este detalhe, podia ainda fazer-se outra alteração: transformar o dia do ano em dias da semana e transformar a hora do dia em período do dia, como "manhã", "noite", ou "tarde". Assim, transformámos a variável "DateTime" em duas variáveis, "Day" e "Period".

Sempre que um utilizador regista uma refeição, com base na glicemia, na glicemia que pretende atingir e nos hidratos de carbono que ingere, é calculada a insulina a tomar. Contudo, o utilizador pode optar por não seguir a dosagem recomendada e tomar mais ou menos. Tendo isto em conta, pode ser importante ver os efeitos que isto provoca e portanto criou-se uma nova variável "Insulin_Difference". A variável "Insulin_Difference" é calculada com base na fórmula

```
Insulin_Difference = Value_Carbs / RH + ((Value_Glucose - OG)/FS)
```

que "RH" é o rácio de hidratos de carbono, "OG" é o objetivo de glicemia e "FS" é o fator de sensibilidade. Estes valores são diferentes para cada utilizador e cada utilizador pode ter vários objetivos de glicemia por dia. Com esta nova variável, se o utilizador optar por tomar uma quantidade de insulina diferente da sugerida com frequência, e isto tiver efeito negativo na glicemia, este padrão será detetado.

Neste momento o *data set* é composto por 7 variáveis: "Day", "Period", "Value_Carbs", "Value_Glucose", "Value_Insulin", "Insulin_Difference" e "Blood_TG".

Relembrando que "Day" e "Period" foram transformadas através da variável original, "DateTime". A variável "Day" foi obtida aplicando a função *weekdays* do *package "base"* do R. A variável "Period" tem como objetivo discretizar a hora do registo: "Period" não diz respeito a uma hora mas sim a um intervalo. Assim sendo, "Period" tem três valores possíveis:

- **1** - Manhã (06:00 - 11:59)
- **2** - Tarde (12:00 - 19:59)
- **3** - Noite (20:00 - 05:59)

o que garante que um dado registo vá pertencer a um dos períodos existentes. Se em vez de usar um intervalo de horas, fossem usadas as horas certas, isso faria com que fosse muito mais difícil encontrar padrões: por exemplo, um registo às 08:30 e outro às 09:30 pertencem ambos ao período 1 mas a horas diferentes. Para que haja um padrão, é necessário que haja repetição de valores. Neste exemplo, se fossem usadas as horas certas os valores seriam diferentes mas usando intervalos de tempo, ambos os registos pertencem ao mesmo período, o da "Manhã". Se pensarmos que um dia tem 24 horas e cada hora tem 60 minutos, usando uma variável "Hora" teríamos 1440 valores possíveis. Por outro lado, usando "Period" temos apenas 3 valores diferentes.

Quanto às variáveis "Value_Glucose", "Value_Insulin" e "Value_Carbs", também tiveram que ser discretizadas, uma vez que se tratavam de variáveis contínuas, ainda que em diferentes intervalos. A glicemia de um paciente diabético pode oscilar entre 50 mg/dL e 300 mg/dL, por exemplo. Já a insulina e os hidratos também podem variar mas os intervalos são mais pequenos, principalmente na insulina. Posto isto, a solução foi discretizar estras três variáveis em intervalos, tal como no período do dia. Uma solução seria dividir em três partes para cada um dos atributos mas visto que se tratam de variáveis importantes, o melhor foi dividir em mais intervalos, nomeadamente 5. Tome-se o exemplo da glicemia: para pessoas com diabetes, os valores recomendados de glicemia antes das refeições estão entre 70 mg/dL e 130 mg/dL e depois das refeições são entre 90 mg/dL e 160 mg/dL [6]. Uma vez que esta doença costuma provocar oscilações na glicemia, é comum que os valores estejam no intervalo recomendado mas também estejam acima ou abaixo desse intervalo. Por vezes podem estar muito acima, tratando-se de uma hiperglicemia, ou muito abaixo, no caso de uma hipoglicemia. Sabendo apenas que um valor está acima do recomendado não dá muita informação. Depois de uma refeição, tanto 170 mg/dL como 300 mg/dL são valores acima do intervalo acima mostrado. A diferença é que o primeiro valor é um pouco acima e não é preocupante, enquanto que o segundo valor é muito mais preocupante e requer ação imediata. Isto para mostrar que, ao discretizarmos os valores de glicemia, é importante que o façamos com vários níveis: dizer que um valor está acima do recomendado não chega, é preciso diferenciar o quão acima está. Deste modo, tanto para a glicemia, como insulina e hidratos de carbono, decidiu-se discretizar os valores em cinco níveis:

- **1** - Valor muito abaixo do normal e hipoglicemia;
- **2** - Valor um pouco abaixo do normal;
- **3** - Valor normal;
- **4** - Valor um pouco acima do normal;
- **5** - Valor muito acima do normal e hiperglicemia;

Estes intervalos não são fixos, variando para cada utilizador. Os extremos são definidos pelo próprio utilizador, ao escolher na aplicação os limites para hipo e hiperglicemia. O valor 3 é definido tendo em conta a média de todas as glicemias do utilizador e os valores 2 e 4 são definidos através do 1º e 3º quartil de todos os valores, respetivamente. Além de ter isto em conta, é preciso também ter em conta os intervalos recomendados em cima definidos. Isto é, se a média dos valores de glicemia de um utilizador for acima do intervalo recomendado, então não fará tanto sentido definir o valor 3 como essa média. No entanto, sempre que a média das glicemias estiver dentro de um intervalo considerado normal, o valor 3 será definido como essa média.

O processo de discretização para os hidratos de carbono e insulina é o mesmo: um valor 5 para hidratos de carbono mostra que o utilizador ingere uma quantidade bastante maior que a recomendada tal como um valor 5 para insulina mostra que o utilizador toma uma dose de insulina muito maior que a recomendada. Naturalmente que um extremo (1 ou 5) em qualquer

variável é sempre algo indesejável: consumir demasiados hidratos de carbono pode levar a uma hiperglicemia assim como tomar demasiada insulina pode provocar uma hipoglicemia.

Quanto ao valor da insulina calculada para cada registo, "Insulin_Difference", o processo é ligeiramente diferente. A insulina é calculada com base na fórmula apresentada na figura 5.1 para calcular a insulina recomendada para cada registo, tendo em conta os valores de glicemia e hidratos de carbono desse mesmo registo. A insulina calculada é então subtraída à insulina tomada pelo utilizador e essa diferença será o valor da variável "Insulin_Difference". Depois, tal como nas outras variáveis, esta é também dividida em 5 intervalos, que são:

- **1** - O valor de insulina tomado é muito menor que o valor calculado
- **2** - O valor de insulina tomado é ligeiramente menor que o valor calculado
- **3** - O valor de insulina tomado é o calculado
- **4** - O valor de insulina tomado é ligeiramente maior que o valor calculado
- **5** - O valor de insulina tomado é muito maior que o valor calculado

Desta forma será possível encontrar relações, se existirem, entre mudanças no valor da insulina a tomar que possam levar a valores de glicemia indesejados.

Depois deste processo, o *data set* está num estado em que já pode ser utilizado para algumas análises. Na próxima secção serão mostradas algumas análises básicas de estatística envolvendo os dados referentes a alguns utilizadores. Na última secção serão enumeradas outras análises efetuadas, sendo que a descrição será feita no próximo capítulo. Algumas dessas análises requerem novas mudanças nos dados, principalmente questões técnicas associadas com algumas funções do R. Essas alterações serão descritas sempre que necessário. De notar as variáveis discretizadas serão utilizadas apenas na parte de associação, pelo que na próxima secção serão feitas análises ainda com os valores numéricos de glicemia. Também é importante referir que a varíavel "Insulin _Difference" é calculada tendo em conta os valores numéricos originais de glicemia e hidratos de carbono, pelo que essa variável é calculada antes da discretização das outras variáveis.

## 5.3 Análise estatística básica

Com o *data set* pré-processado, estávamos em condições de começar a utilizá-lo para análise. Antes de começar a aplicar técnicas de *data mining* começou-se por fazer algumas estatísticas com os valores de glicose dos utilizadores. Nas próximas subsecções vamos descrever algumas estatísticas feitas para alguns utilizadores que enviaram registos. Para fazer uma análise minimamente fidedigna é necessário ter uma quantidade considerável de dados pelo que vamos analisar os dados de utilizadores que enviaram registos referentes a quatro ou mais semanas. Começaremos com algumas análises mais simples, fazendo algumas estatísticas que possam permitir observar

algumas anormalias nos valores de glicemia. Serão então feitas as diferentes análises para cinco utilizadores de forma totalmente anónima.

### 5.3.1 Média de glicose

Uma primeira estatística poderia ser simplesmente a média de glicose para um determinado utilizador. Relembrando que o HbA1c é um parâmetro importante para verificar o controlo da diabetes num paciente visto que é possível determinar a média de glicose de algumas semanas ou meses. Portanto, de uma maneira mais ou menos semelhante, a média de glicose de um determinado paciente dá para ter uma ideia do quão bem esse paciente controla a glicemia. Apresentam-se de seguida as médias de glicose para os cinco utilizadores:

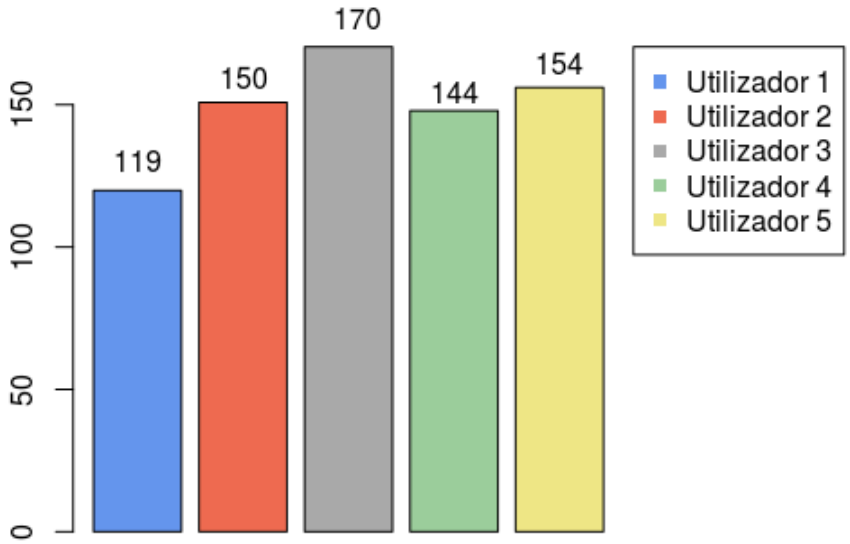


Figura 5.1: Médias de glicemias para os utilizadores

Como mencionado anteriormente, estas médias pertencem a diferentes utilizadores e dizem respeito a registos durante pelo menos quatro semanas, e portanto, são um indicador geral do controlo da glicemia por parte de cada utilizador. Pode perceber-se que alguns utilizadores têm uma média com um valor mais normal que outros o que pode ser um indicador de um bom controlo. No entanto, isto não é necessariamente verdade. Um bom controlo da glicemia passa não só por manter os valores em intervalos normais mas também em prevenir grandes oscilações. Neste caso, uma média de glicemia de 119 mg/dL pode parecer melhor que uma média de glicemia de 170 mg/dL mas pode não ser: o primeiro utilizador pode ter 2 registos em que um seja 60 mg/dL e outro seja 180 mg/dL, o que dá uma média de 120 mg/dL; por outro lado, o terceiro utilizador pode ter ambos os valores a 170 mg/dL. Ou seja, embora a média de glicemia do primeiro utilizador seja mais baixa, não significa que os valores sejam estáveis, tal como no terceiro utilizador não significa que os valores não sejam estáveis. Medir apenas a média da glicemia é uma análise demasiado vaga: pode ser um indicativo de um bom ou mau controlo

da glicemia mas não permite ter certezas. Torna-se portanto necessário fazer uma análise mais aprofundada.

### 5.3.2 Média de glicose por dia

Uma forma mais detalhada de tentar perceber alguns hábitos dos utilizadores é verificar como é que a sua glicemia oscila durante a semana. Tal pode ser feito obtendo a média de glicose para cada dia. Assim, é possível saber, por exemplo, quais os dias em que a glicemia é mais elevada ou mais baixa, sem esquecer que se trata de uma média, e que por isso não é possível ter conhecimento sobre eventuais hipo ou hiperglicemias. Ainda assim, ao saber a média de glicemia por cada dia, será possível avisar o utilizador de quais os dias a ter em atenção, isto é, se a segunda-feira for um dia com uma média elevada de glicemia, então o utilizador pode ser avisado deste facto para tomar as medidas necessárias ou, pelo menos, aumentar o controlo da glicemia neste dia. Portanto, embora este tipo de estatística não permita descobrir situações em particular, pode ainda assim ser útil. Eis os valores médios de glicemia para os utilizadores:

| **Utilizador/Dia** | **Domingo** | **Segunda** | **Terça** | **Quarta** | **Quinta** | **Sexta** | **Sábado** |
|---|---|---|---|---|---|---|---|
| **Utilizador 1** | 132 | 117 | 119 | 127 | 110 | 110 | 115 |
| **Utilizador 2** | 161 | 131 | 154 | 156 | 149 | 148 | 165 |
| **Utilizador 3** | 182 | 162 | 189 | 166 | 181 | 163 | 146 |
| **Utilizador 4** | 149 | 170 | 165 | 114 | 142 | 131 | 136 |
| **Utilizador 5** | 182 | 153 | 154 | 150 | 170 | 140 | NR |

Tabela 5.2: Médias de glicemia dos utilizadores por dia da semana

Com uma análise deste tipo, embora não se possam tirar conclusões sobre a forma como a glicemia varia, podemos começar a ter algum conhecimento sobre o que podem ser rotinas dos utilizadores. Por exemplo, para 4 dos 5 utilizadores acima mostrados, o dia em que o valor médio de glicemia é mais elevado é aos Domingos. Normalmente ao fim-de-semana as pessoas tendem a ter um estilo de vida mais sedentário e relaxado, o que por sua vez pode levar a um controlo menos apertado da glicemia; mesmo que o controlo não seja afetado ao fim-de-semana, pode haver quem não tenha tanto cuidado com a alimentação, cometendo alguns excessos; pode até apenas ser o facto de as pessoas estarem menos *stressadas* ao fim-de-semana. Mais importante de saber o porquê de estes valores serem mais elevados ao Domingo, o importante é o conhecimento de que isto acontece. É impossível adivinhar o porquê de os valores serem tendencialmente maiores a um determinado dia, apenas sabendo a média dos valores, mas o simples facto de descobrir esse padrão e de alertar um utilizador para o mesmo, pode fazer com que ele próprio se aperceba o que faz de diferente e que provoque a alteração, podendo assim corrigi-la.

No contexto da aplicação, uma simples estatística como esta acima mostrada, podia fazer aparecer um aviso, alertando o utilizador que a um determinado dia da semana, os valores de glicemia são, em geral, mais altos. Este processo seria automático: a aplicação calcularia o

valor da média para cada dia e mostraria um aviso sempre que este valor fosse mais alto que o desejado. Imaginemos que o utilizador 3 definiu o limite de hiperglicemia como 180. A aplicação ao calcular a média de glicemia para cada dia verificava que em três dias diferentes da semana os valores costumam ser acima desse limite e portanto lançaria um aviso.

Pode concluir-se que este tipo de estatística é mais vantajoso que o primeiro, porque pode ajudar a identificar dias em que os valores são mais anormais. Uma vez mais, a média de glicemia não permite saber cada valor exato de cada medição, pelo que ainda não é possível verificar se os valores são estáveis ou oscilatórios. Ainda assim, se a média de um determinado dia é próxima ou até superior ao limite de hiperglicemia definido pelo utilizador, permite saber que há uma grande probabilidade de nesse dia o utilizador ter tido valores acima do limite.

### 5.3.3 Média de glicose por período do dia

Como já vimos, pode ser útil saber os valores médios de glicemia para cada dia da semana, que nos permite saber que, por exemplo, o Domingo é um dia com valores tendencialmente mais elevados. Mas apenas isso não nos dá qualquer tipo de informação sobre a variação dos valores ao longo do dia, o que também seria útil. Se um utilizador tiver o conhecimento de que tem valores mais elevados à tarde do que no resto do dia, se calhar percebe que o tipo de almoço que normalmente faz pode não ser o mais indicado. Obviamente que isto é só um exemplo, mas o facto é que saber como varia a glicemia durante o dia pode ajudar os utilizadores a perceber eventuais hábitos errados. Tendo isto em conta, para a próxima estatística dividiu-se o dia em três partes, como já explicado:

- Manhã: 06:00 - 11:59;
- Tarde: 12:00 - 19:59;
- Noite: 20:00 - 05:59;

Os valores médios para cada utilizador em cada fase do dia são:

| **Utilizador/Período** | **Manhã** | **Tarde** | **Noite** |
|---|---|---|---|
| **Utilizador 1** | 112 mg/dL | 117 mg/dL | 128 mg/dL |
| **Utilizador 2** | 160 mg/dL | 151 mg/dL | 138 mg/dL |
| **Utilizador 3** | 162 mg/dL | 141 mg/dL | 199 mg/dL |
| **Utilizador 4** | 158 mg/dL | 134 mg/dL | 146 mg/dL |
| **Utilizador 5** | 154 mg/dL | 157 mg/dL | 153 mg/dL |

Tabela 5.3: Médias de glicemia dos utilizadores por período do dia

Este tipo de análise já permite ter uma ideia mais concreta das oscilações da glicemia para cada utilizador ao longo do dia. É possível verificar que alguns utilizadores conseguem manter os níveis de glicemia mais ou menos estáveis, como os utilizadores 1 e 5. É possível observar algumas oscilações mais acentuadas em alguns utilizadores: o utilizador 1 por exemplo, embora consiga manter sempre os níveis de glicemia num intervalo considerado normal, tem tendência para ter valores ligeiramente mais elevados à noite. Isto pode ser devido a menos atividade durante a noite, que normalmente é usada para relaxar. Pode ser também por não adequar a insulina em relação ao jantar, por exemplo.

Já no utilizador 2 acontece precisamente o contrário: tem valores mais elevados pela manhã, que vão diminuindo ao longo do dia. Os valores altos pela manhã podem indicar que o utilizador ingere alimentos antes de ir para a cama ou que não adequa a insulina depois do jantar, por exemplo. Uma vez mais, neste tipo de estatística é mais importante detetar comportamentos anormais do que os explicar.

Quanto ao utilizador 3 é aquele que, pelo que se pode observar apenas com estes dados, apresenta mais oscilações. Começa o dia com valores ligeiramente elevados, pelo que os consegue baixar um bocado durante o dia até que voltam a subir bastante durante a noite. Isto pode não acontecer todos os dias mas geralmente é o que acontece. O aviso nesta situação seria importante para que o utilizador tentasse evitar hiperglicemias durante a noite.

É natural que os valores de glicemia oscilem durante o dia por vários fatores: o *stress* do trabalho pode causar aumento no nível de glicose no sangue. Por outro lado, o esforço físico durante o dia, quando existente, pode provocar o contrário. Mesmo sem fatores externos, por vezes há situações que levam a glicemia a variar de forma natural. Uma dessas situações é conhecida por *dawn phenomenon*, que causa o aumento da glicose no sangue durante a noite, devido a hormonas que o corpo produz nesse período. Isso faz com que o utilizador tenha valores mais altos de manhã.[dawn] Este pode até ser o motivo para que três utilizadores tenham valores mais elevados de manhã do que nos outros períodos do dia, tal como os motivos anteriores. Independentemente do motivo, a consciência de que isto acontece é o primeiro passo para corrigir a situação.

Na próxima subsecção vamos fazer uma análise ainda mais detalhada, mostrando os níveis de glicemia ao longo do dia e por cada dia da semana. Como já mencionado, apesar de a média de glicemia ser um bom indicativo do controlo da mesma, é mais útil ter valores exatos que permitem ver se, de facto, há oscilações e o quão grandes são essas variações.

### 5.3.4 Glicose por hora do dia

Os valores de glicose numéricos originais, ao contrário da média, permitem perceber se há variações ou estabilidade na glicemia. Numa primeira análise mostraremos os valores de glicose por hora para todos os dias e posteriormente para cada dia da semana. Desta forma será possível

saber a que horas do dia, e em que dias, é que o utilizador geralmente tem mais hiper ou hipoglicemias.

**Utilizador 1**

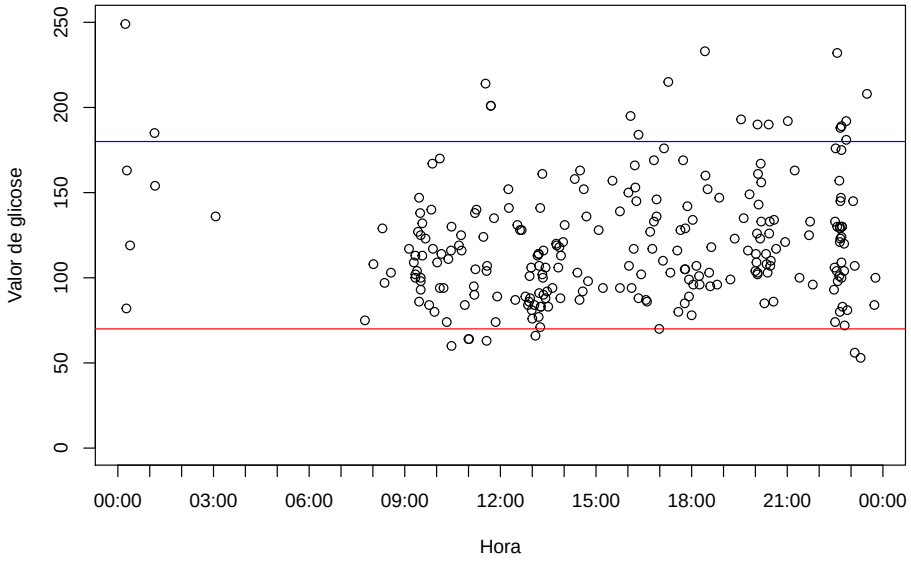


Figura 5.2: Glicemia por horas do utilizador 1

Como se pode perceber pela figura 5.2, a maioria dos registos é feita entre as 08:00 e as 24:00, sendo que neste intervalo, grande parte dos registos estão divididos em três momentos, provavelmente na hora das refeições. A linha vermelha representa o limite mínimo desejável e a linha azul o máximo desejável. Estas linhas correspondem a hipo e hiperglicemia, respetivamente, e foram definidas pelo utilizador, sendo que o limite para hipoglicemia é 70 e para hiperglicemia é 180. Ao analisar o gráfico percebe-se imediatamente que quase todos os valores se encontram entre estes dois valores, o que é desejável, visto que é o intervalo definido como "normal" pelo utilizador. Percebe-se também que há algumas hiperglicemias, sendo que todas elas ocorrem entre o início da tarde e o início da madrugada. Obviamente que isto não significa que durante a madrugada não haja valores elevados, apenas não são registados.

**Utilizador 2**

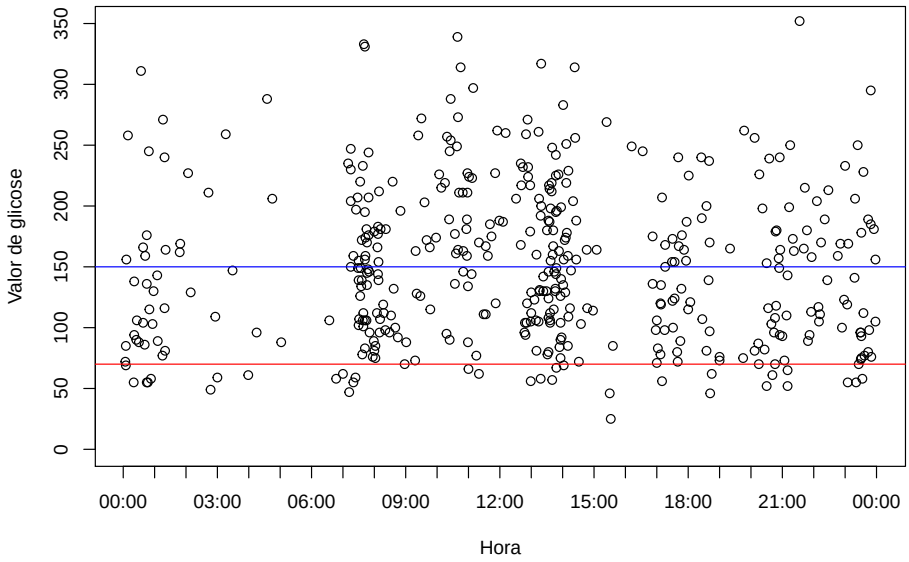


Figura 5.3: Glicemia por horas do utilizador 2

Ao analisar a figura 5.3 percebe-se um dos problemas anteriormente mencionados: as médias podem ser um bom indicador geral mas também podem não o ser. E neste caso não é: a média de glicemia do utilizador 2 é de 150 mg/dL o que poderia levar a pensar que este utilizador tinha valores mais ou menos estáveis. No entanto, observando o gráfico, verifica-se que não é o que acontece: os valores oscilam entre extremos muito separados, havendo valores acima de 300 mg/dL e outros abaixo de 50. Neste caso os limites definidos pelo utilizador foram de 70 mg/dL para hipoglicemia e de 150 mg/dL para hiperglicemia. Também é possível observar que cerca de metade dos registos efetuados encontram-se fora dos intervalos estabelecidos, o que leva a pensar que estes limites escolhidos não foram os mais corretos e talvez tenham de ser adaptados. De qualquer das formas, as medições são feitas mais frequentemente de manhã cedo e ao início da tarde, presumivelmente à hora de pequeno-almoço e almoço.

**Utilizador 3**

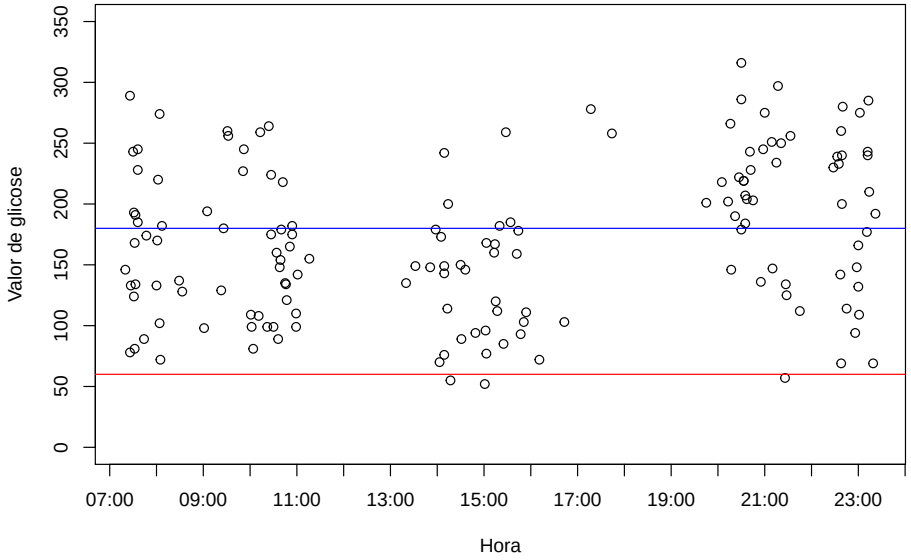


Figura 5.4: Glicemia por horas do utilizador 3

Ao observar a figura 5.4, percebe-se que os valores oscilam bastante, havendo alguns valores próximos de 50 mg/dL e outros próximos de 300 mg/dL. Os limites definidos pelo utilizador foram de 60 mg/dL para hipoglicemia e 180 mg/dL para hiperglicemia. No gráfico do utilizador anterior, havia vários períodos do dia com hiperglicemia mas estes eram, aproximadamente, em igual número à quantidade de valores normais, isto é, pelo gráfico do utilizador 2, observa-se que no geral, para um determinado período, a quantidade de valores normais ou demasiado elevados são mais ou menos iguais. Por outro lado, neste gráfico observa-se o contrário: há períodos que notoriamente têm mais hiperglicemias que valores normais e outros períodos com mais valores normais que hiperglicemias. Por exemplo, entre as 07:00 e as 10:00, que será o período do pequeno-almoço ou imediatamente depois, verifica-se que a quantidade de valores altos e valores normais é parecida. Já no período imediatamente a seguir à hora de almoço, há poucas hiperglicemias em relação à quantidade de registos feitos. Por último, à hora de jantar ou logo a seguir, a quantidade de hiperglicemias é bastante alta em relação à quantidade das medições. Isto pode ser uma consequência da rotina do utilizador. Por exemplo, se o utilizador jantar e logo a seguir descansar, pode causar esta subida. Por outro lado, e olhando para o gráfico, uma solução poderia ser algum tipo de exercício leve depois do jantar, como uma caminhada, para conseguir baixar um pouco os valores. Outra alternativa seria ajustar a insulina depois do jantar. De qualquer das formas, com este tipo de análise, não se consegue descobrir qual a causa para estes valores anormais, pelo que sem um tipo de análise mais profunda, cabe ao utilizador perceber o porquê de isto acontecer.

**Utilizador 4**

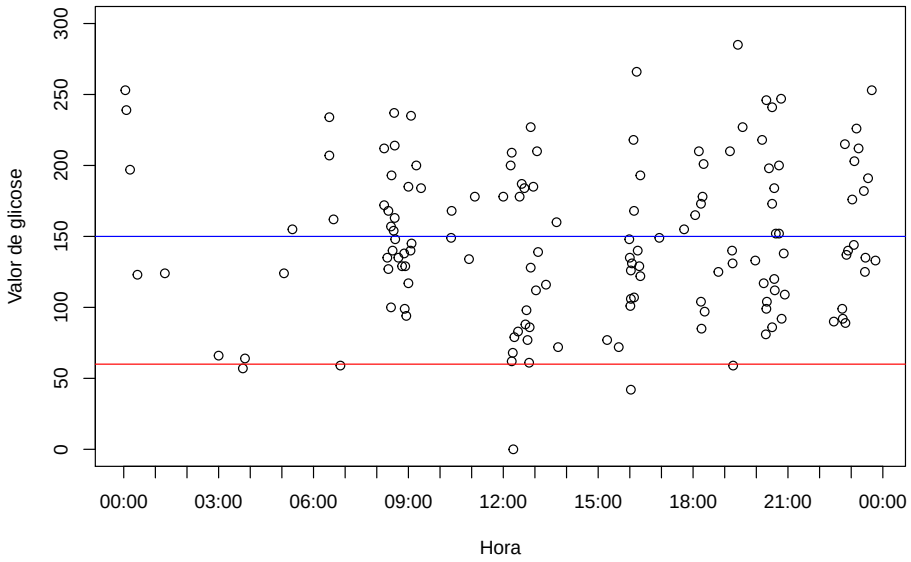


Figura 5.5: Glicemia por horas do utilizador 4

O utilizador 4 definiu como 60 mg/dL o limite para hipoglicemia e 150 mg/dL o limite para hiperglicemia. Tal como já aconteceu com outros utilizadores, os valores de hiperglicemia são abundantes o que pode levar a crer que o limite para hiperglicemia é demasiado baixo e poderia ser ajustado. Este utilizador tem geralmente hiperglicemias depois das três refeições o que significa que talvez a insulina devesse ser ajustada. Verifica-se também a presença de alguns valores demasiado altos por volta da meia noite seguidos por valores demasiados baixos durante a madrugada.

**Utilizador 5**

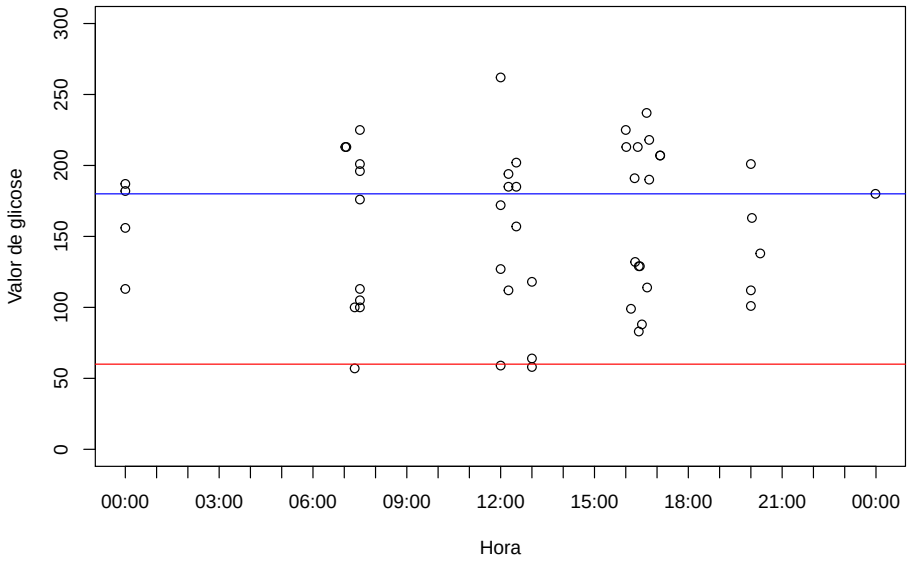


Figura 5.6: Glicemia por horas do utilizador 5

Tal como no utilizador 4, neste utilizador, apenas observando o gráfico não se consegue obter conclusões, isto porque não há nenhum período do dia que se destaque dos outros. O utilizador definiu um limite para hipoglicemia de 70 mg/dL e não definiu um limite para hiperglicemia, pelo que o limite utilizado foi 180 mg/dL. Conseguem-se observar alguns períodos com tendência para hiperglicemia, nomeadamente de manhã, ao início da tarde e a meio da tarde. Neste utilizador não é possível concluir nada apenas com esta análise.

Pelos gráficos observados de todos os utilizadores, verifica-se que em todos eles há hiperglicemias, embora nuns notoriamente mais que noutros. Embora não se consigam descobrir padrões ou detetar anomalias apenas com uma análise deste tipo, fica visível a necessidade de um melhor controlo da glicemia por parte de quase todos os utilizadores, que pode ser conseguida através de um uso de uma aplicação de registo, tal como a MyDiabetes. Embora na análise feita nesta subsecção se consigam descobrir mais detalhes do que na subsecção anterior, ainda não é descoberto o suficiente. Desta forma vamos aprofundar ainda mais e fazer uma análise por dia e por hora para cada utilizador.

### 5.3.5 Glicose por hora e por dia

Nas últimas subsecções fizemos uma análise que nos permite saber que, por exemplo, um determinado utilizador tem tendencialmente valores de glicemia mais elevados à terça-feira ou que um outro utilizador normalmente tem valores de glicose no sangue mais elevados à noite. Mas estas duas conclusões são independentes: saber que os valores são mais altos à terça-feira não

nos permite saber como é que os valores variam durante a terça-feira ou saber que os valores são mais altos à noite não nos permite saber em que noites da semana é que os valores são de facto mais elevados. Face a este problema, fez-se uma análise ainda mais detalhada que as anteriores e relacionámos estas duas variáveis, dia da semana e período do dia, para tentar perceber de que forma é que a glicose varia durante o dia, para cada dia da semana. De seguida apresentaremos os gráficos para os cinco utilizadores.

**Utilizador 1**

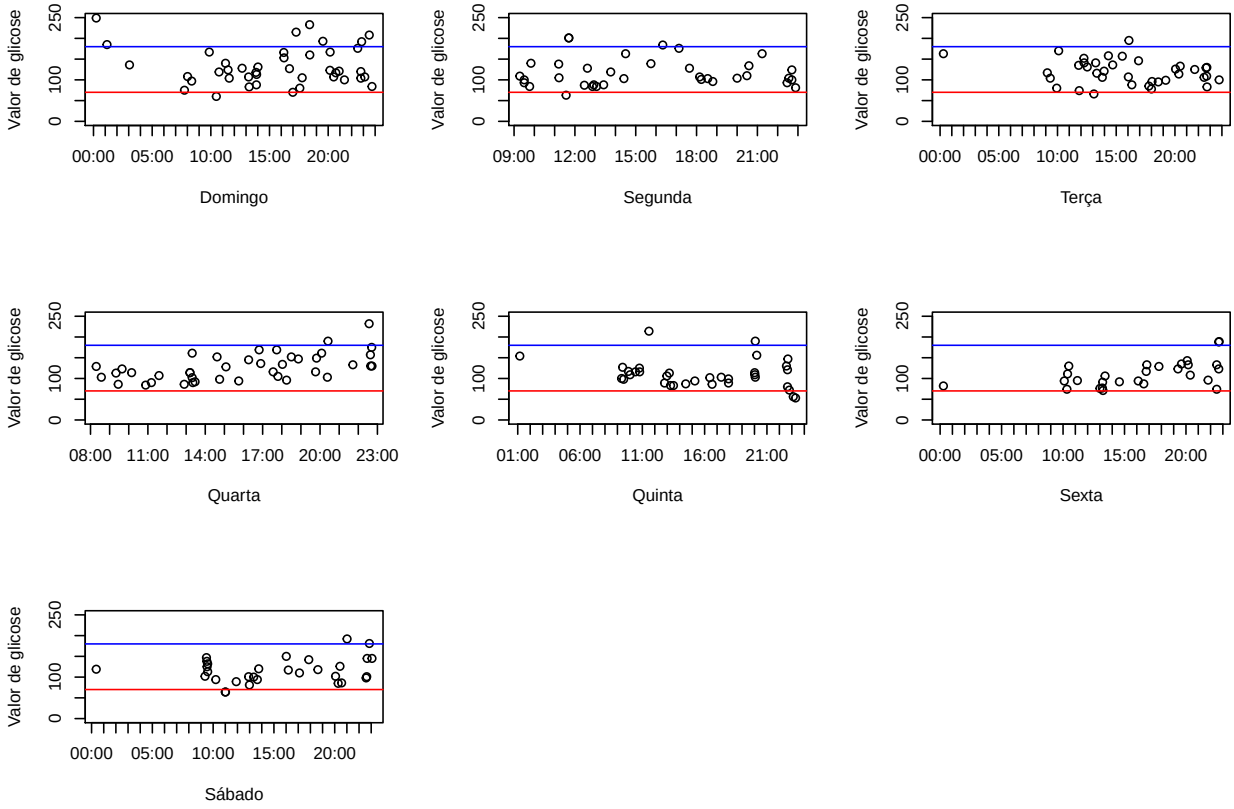


Figura 5.7: Glicemia do utilizador 1 por dias da semana

Analisando as figuras **??**, nota-se que o utilizador mantém, grande parte das vezes, os valores de glicemia dentro dos intervalos normais todos os dias. No entanto, nota-se claramente que o Domingo tem mais hiperglicemias que os outros dias todos e que estas ocorrem ao final da tarde ou início da noite. Embora as estatísticas anteriores sobre este utilizador mostrem que de facto o Domingo é o dia com a média de glicemia mais elevada e que a noite é o período com média de glicemia mais elevada, tal não significa necessariamente que estas duas médias fossem verdade sempre. Ou seja, nada garantia que ao Domingo a glicemia também fosse mais elevada à noite. Contudo, com esta análise pode perceber-se que de facto isso acontece. Nos outros dias verifica-se que há um ou outro caso de hiperglicemias ou hipoglicemias mas em muito pouca quantidade. Uma análise que utilize os outros parâmetros como insulina ou hidratos de carbono

talvez ajude a perceber o porquê de o Domingo ser um dia com valores de glicemia mais altos.

**Utilizador 2**

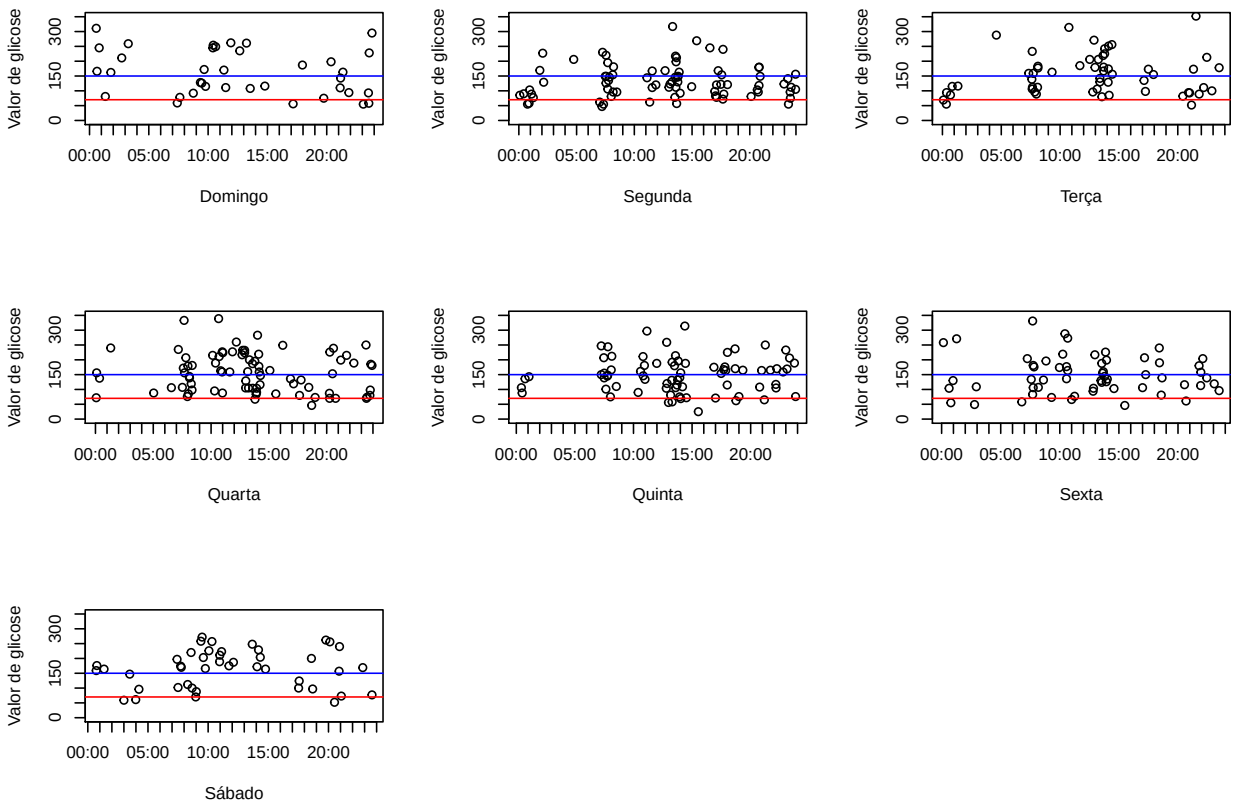


Figura 5.8: Glicemia do utilizador 2 por dias da semana

Observando a figura 5.8 verifica-se que o utilizador apresenta hiperglicemias todos os dias, maioritariamente durante a tarde. Ao Domingo apresenta menos hiperglicemias porque também tem menos registos. No resto dos dias apresenta uma distribuição semelhante entre valores normais e valores acima do limite de hiperglicemia, isto é, há mais ou menos quantidade de valores normais e valores demasiado altos. Contudo, ao Sábado o número de hiperglicemias é bastante maior em relação ao número de valores no intervalo normal.

**Utilizador 3**

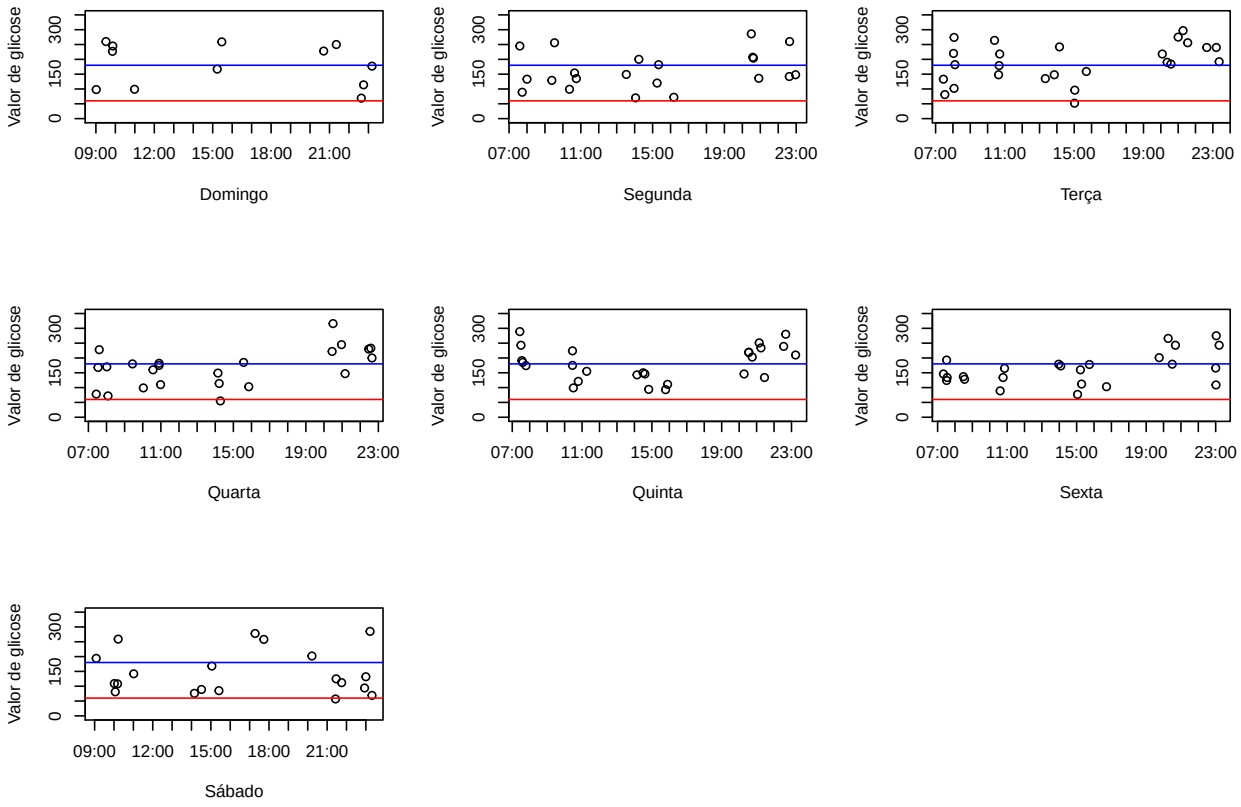


Figura 5.9: Glicemia do utilizador 3 por dias da semana

Neste utilizador é possível ver um padrão: tem tendência para ter hiperglicemias de manhã e ao fim da tarde ou início da noite. Uma vez mais, a causa pode estar relacionada com o trabalho, por exemplo, que pode ser mais desgastante à tarde, daí fazer com que os valores não subam tanto. Esta teoria seria ainda suportada pelo gráfico de Sábado, que mostra algumas hiperglicemias também a meio da tarde, o que faria sentido se de facto a causa dos valores durante a semana fosse o trabalho.

**Utilizador 4**

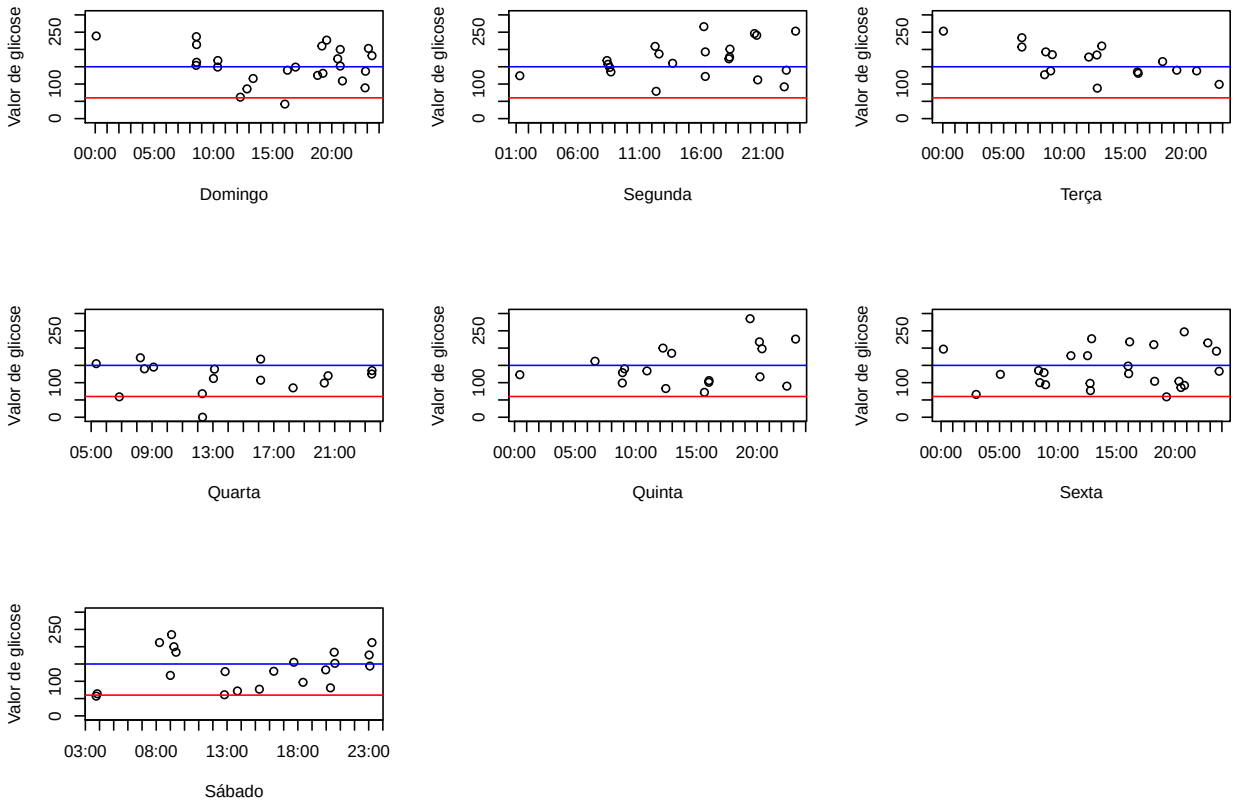


Figura 5.10: Glicemia do utilizador 4 por dias da semana

Observando a figura refdias4 em cima, é possível observar que em todos os dias ocorrem hiperglicemias mas há três dias que se destacam: Domingo, Segunda e Terça, pois têm mais hiperglicemias em comparação com valores normais que o resto dos dias. Verifica-se também que os valores mais altos ocorrem de manhã ou ao início da noite, sendo que apenas em dois dias há mais que uma hiperglicemia a meio da tarde, o que pode significar que o utilizador possa fazer algo de diferente nestes dois dias, embora sem qualquer tipo de certeza. Em vários dias da semana notam-se alguns valores muito próximos ou até abaixo do limite de hipoglicemia sendo notórios dois períodos em que eles são mais frequentes: durante a madrugada e perto da hora de almoço. Isto pode significar que o utilizador não come nada antes de ir dormir, para o primeiro caso, e para o segundo caso pode significar que o utilizador não come nada entre o pequeno-almoço e o almoço, e que talvez devesse comer alguma coisa.

**Utilizador 5**

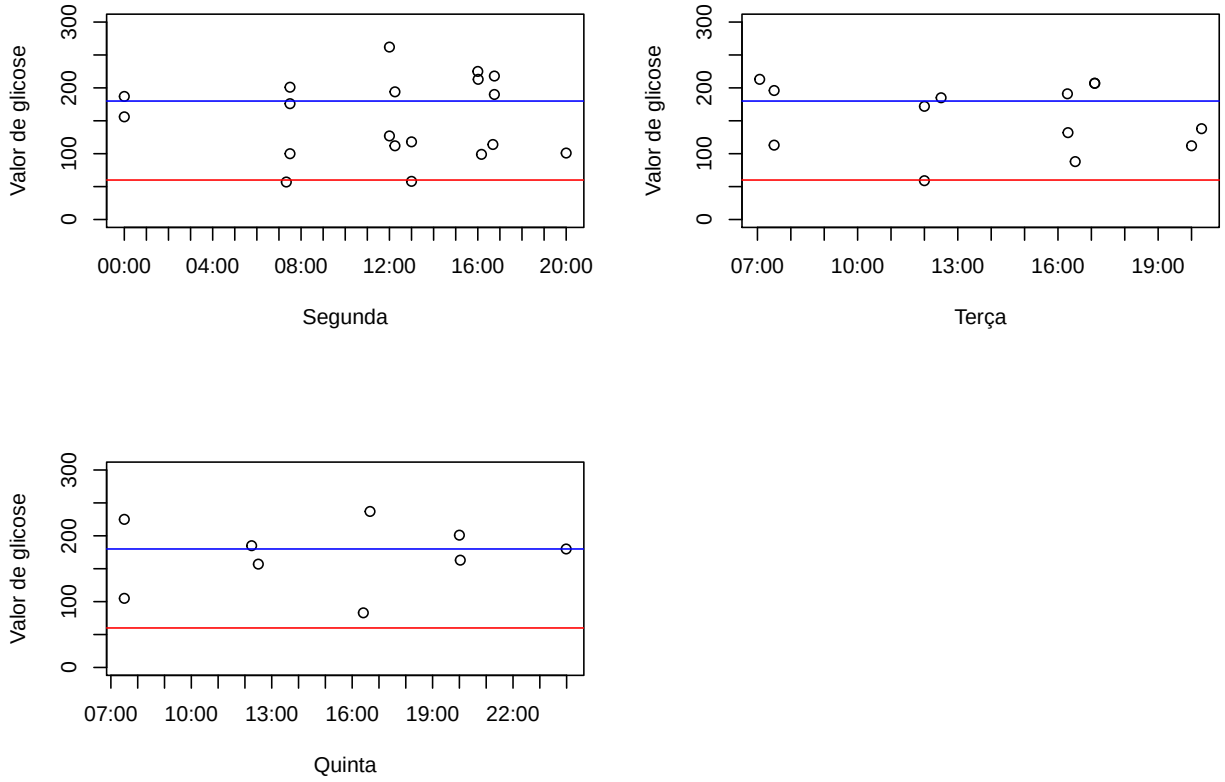


Figura 5.11: Glicemia do utilizador 5 por dias da semana

Para este utilizador mostramos apenas gráficos referentes a três dias já que os outros não têm registos suficientes. É notório, contudo, que à Segunda-feira o utilizador tende a ter hiperglicemias ao fim da tarde/início da noite.

# Capítulo 6

# Resultados

Até agora utilizaram-se os dados para fazer algumas análises estatísticas que permitem visualizar alguns padrões existentes. Neste capítulo pretende-se explorar os dados de uma forma avançada, para ver o que daí se retira. Vamos fazer três análises diferentes, envolvendo diferentes ferramentas. As análises são:

- Regras de associação
- Redes *bayesianas*
- ILP

Ao fazermos análises diferentes, vamos aumentar a probabilidade de conseguirmos obter algumas regras. Enquanto que no capítulo anterior fizemos uma análise utilizando apenas duas variáveis, dia/hora e glicemia, neste capítulo utilizamos métodos que relacionem todas as variáveis à procura de padrões.

## 6.1 Regras de associação

O R não tem funções para regras de associação, de origem, pelo que foi necessário instalar um *package* adicional, *arules*, que oferece uma vasta quantidade de funções direcionadas para regras de associação. O processo de procura de regras de associação é de fácil compreensão: o algoritmo vai percorrer o *data set* e analisar cada "transação", que corresponde a cada linha do ficheiro csv. Dependendo da confiança e suporte escolhidos, vão ser geradas regras, se existirem, que respeitem os limites escolhidos. As regras geradas podem ser de qualquer tipo, isto é, o consequente pode ser qualquer uma das variáveis analisadas, sendo que neste caso interessa-nos descobrir regras com a variável "Value_Glucose" como consequente, ou seja, no lado direito. Contudo, o *data set* ainda não está estruturado da melhor forma para a criação das regras. Da forma que o *data set* está feito, cada linha corresponde ao mesmo momento, ou seja, se uma linha tiver um valor de glicemia, hidratos de carbono e insulina, tudo isso corresponde a um registo feito à mesma hora.

O mesmo acontece para o exercício. O tipo de conclusões que se pretende obter é de que forma estes parâmetros causam algum impacto no valor de glicemia, como por exemplo, descobrir de que forma o exercício vai alterar a quantidade de glucose no sangue ou de que forma a insulina tomada ou os hidratos de carbono vão fazer alterar este valor. Se todos esses parâmetros forem registados à mesma hora, não se consegue concluir nada: uma vez que o algoritmo Apriori vai analisar linha a linha de forma independente, no formato atual, o valor de glicemia só vai ser comparado com o valor de insulina e de hidratos de carbono registados ao mesmo tempo. No entanto, o importante é saber como é que a quantidade de hidratos de carbono ingerida e a insulina tomada vão afetar a glicemia no espaço de tempo a seguir, e não no mesmo espaço de tempo. Ou seja, como é que os valores de hidratos e de insulina num registo, vão afetar a glicemia no registo seguinte. A questão é que o valor de glicemia do registo seguinte vai pertencer a outra linha do csv e portanto não vai ser tido em conta para os valores de outra linha. Posto isto, a solução foi alterar a estrutura do ficheiro para que cada linha passe a ter o valor de glicemia seguinte, e não o atual. Assim, imaginando que a variável "Value_Glucose" passe a chamar-se "Next_Glucose", é possível ter uma regra como

```
Se Value_Carbs==5 entao Next_Glucose==5
```

ou seja, descobrir um padrão em que quando o utilizador ingere demasiados hidratos de carbono, então o valor seguinte de glicemia será demasiado elevado, mais precisamente uma hiperglicemia. Se esta alteração não fosse feita, as regras descobertas apenas dariam relações entre valores registados à mesma hora o que não tem qualquer utilidade. Embora este não seja o único tipo de regras a descobrir, certamente que é um dos tipos de conselhos que podem levar um utilizador a conseguir controlar a diabetes de forma mais eficiente. Convém relembrar que se uma regra é descoberta, é porque essa ação ocorre várias vezes, ou seja, trata-se de um padrão. Ao informar o utilizador de que este padrão existe, e que tem uma influência negativa nos seus valores de glicemia, o utilizador pode tomar a ação que achar mais apropriada de forma a evitar que aconteça. Na presença de uma regra como esta, a aplicação poderia mostrar um aviso quando o utilizador fosse fazer um registo de refeição e inserisse um valor elevado de hidratos de carbono, em que o aviso poderia ser, por exemplo, "Inseriu um valor de hidratos de carbono elevado e, quando faz isso, normalmente tende a ter uma eventual hiperglicemia algum tempo depois da refeição".

Uma vez alterados os *data sets* de cada utilizador, os dados estavam em condições de ser utilizados pelo algoritmo apriori. De seguida, apresentaremos, para cada utilizador, algumas das regras mais relevantes que foram descobertas. Como mencionado no capítulo 2, o algoritmo apriori tem alguns parâmetros que permitam avaliar a utilidade de uma regra. Neste trabalho, definimos os valores de confiança a 0.6 e suporte a 0.01. Na prática, este valor de suporte significa que as regras geradas dizem respeito a transações que correspondem a pelo menos 1% das transações totais. Uma confiança de 0.6 mostra que, para um dado antecedente, o consequente aparece pelo menos 60% das vezes.

Interessa-nos descobrir, para cada utilizador, comportamentos que levem a valores anormais

de glicemia. Na variável "Next_Glucose" os valores anormais correspondem a 1 e 2, para valores baixos, e 4 e 5, para valores altos. Isto significa que o tipo de regras que procuramos tem como consequente "Next_Glucose=1" ou "Next_Glucose=5". Procuramos apenas regras com valores extremos porque são valores que correspondem a situações mais preocupantes, isto é, hipoglicemia e hiperglicemia. Embora os valores 2 e 4 também sejam valores fora do intervalo considerado normal, são valores que ainda não ultrapassam os limites definidos pelos utilizadores, pelo que estas situações não são prioritárias.

A forma de obter apenas regras com o consequente pretendido passa por fazer uma filtragem do conjunto de todas as regras encontradas. O R permite gerar um sub-conjunto de regras com a variável pretendida como consequente, com os valores pretendidos.

No R, o comando para a procura de regras através do algoritmo *apriori* é

```
rules <- apriori(dataset, parameter=list(confidence=0.6, support=0.01))
```

em que "dataset" é o ficheiro a ser analisado com os parâmetros confiança e suporte definidos com os valores já mencionados. Para mostrar apenas as regras com o valor de glicose pretendido, como por exemplo 5, o comando é

```
rules.sub <- subset(rules, subset = rhs %in% "Next_Glucose=5")
```

que gerará então um subconjunto de regras em que o consequente será "Next_Glucose=5".

É comum um conjunto de regras ter alguns milhares de regras, pelo que a análise individual de cada regra se torna impossível. A criação de um sub-conjunto de regras com o consequente pretendido ajuda nesta questão, mas ainda assim não é suficiente: um sub-conjunto pode, ainda assim, ter algumas dezenas ou centenas de regras. É, portanto, importante utilizar ferramentas que possam dar uma ajuda extra na seleção de regras úteis. Nesta parte utilizamos um *package* do R, o *arulesViz*, que permite visualizar as regras geradas, através de gráficos. É possível, por exemplo, representar as regras num gráfico, ordenadas por suporte, confiança ou *lift*. Assim torna-se mais fácil filtrar as regras potencialmente mais útil, sendo que este processo é interativo: ao selecionar uma região no gráfico, é possível mostrar todas as regras presentes nessa região. Apresentamos, de seguida, algumas das regras geradas para cada utilizador.

**Utilizador 1**

Para este utilizador foram geradas 1239 regras, sendo que nenhuma delas tem como consequente o valor de glicose 0 ou 5. Neste caso, procuramos então regras com o valor de glicose 2 ou 4. Verifica-se que há 2 regras para valores baixos de glicemia e 23 para valores altos, pelo que se torna necessário usar o package arulesViz para as regras para valores altos. O comando é

```
plot(rules.sub, method=NULL, measure="support", shading = "lift", interactive =
    TRUE, data = NULL, control = NULL)
```

que gera o gráfico

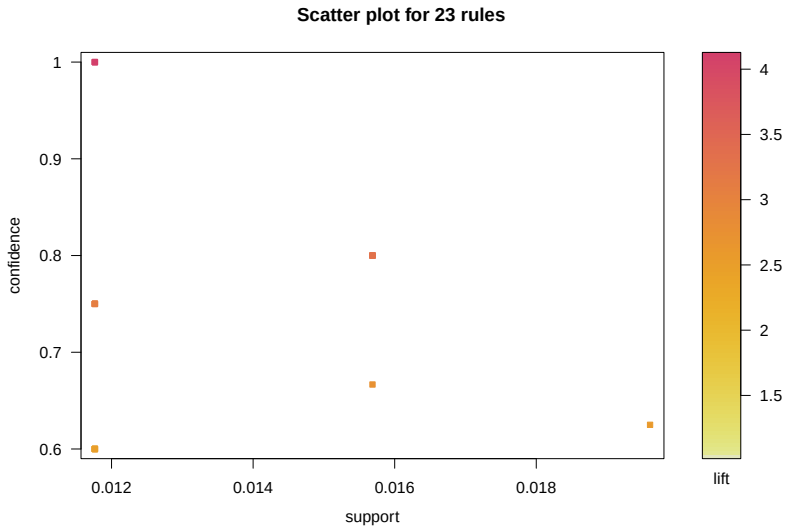


Figura 6.1: Regras de associação para o utilizador 1

Ao escolher a região no canto inferior esquerdo, que tem o suporte mais elevado, obtém-se a seguinte regra:

```
{Day=Quarta,Period=2,Value_Carbs=4} => {Next_Glucose=4}
```

Note-se que cada quadrado no gráfico não corresponde necessariamente a uma só regra, mas sim a um grupo de regras, embora neste caso exista apenas uma. Esta regra permite perceber que o utilizador tende a ter valores de glicemia mais elevados à quarta-feira à tarde, quando consome uma grande quantidade de hidratos de carnobo.

Para valores de glicemia baixos, foi encontrada uma regra:

```
{Day=Sexta,Period=1,Value_Carbs=3} => {Next_Glucose=2}
```

que mostra que à sexta-feira de manhã o utilizador tende a ter valores mais baixos que nos outros dias. Ao contrário das análises anteriores, este tipo de análise permite saber com mais detalhe o porquê destes valores ocorrerem. Neste caso, isto acontece quando o utilizador ingere uma quantidade normal de hidratos de carbono. Como o valor da glicose é ligeiramente baixo, isto pode indicar que o utilizador não fez uma contagem certa de hidratos de carbono e portanto a insulina tomada não foi a adequada.

### Utilizador 2

Para este utilizador foram descobertas consideravelmente menos regras, 498. No entanto, foram descobertas 42 regras para hiperglicemia. Uma vez mais, é importante o uso do *package arulesViz* para selecionar as regras mais úteis. Foi gerado o gráfico:

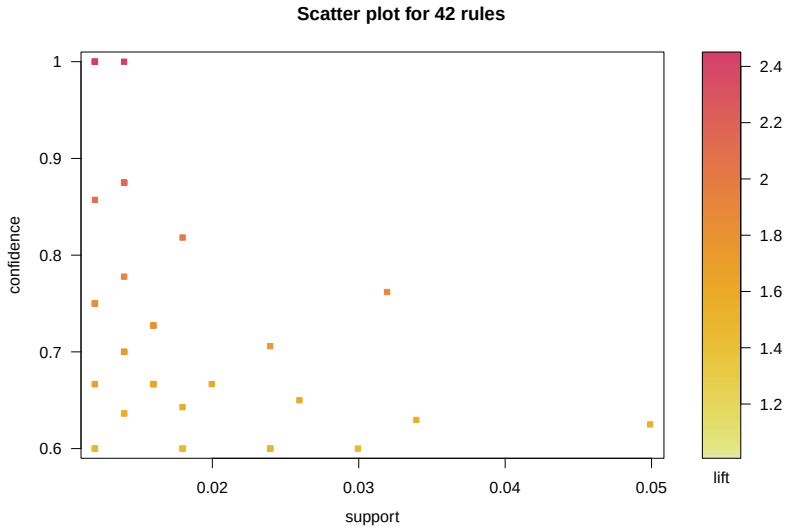


Figura 6.2: Regras de associação para o utilizador 2

e escolhido o grupo de regras com maior suporte, ou seja, o quadrado mais à direita. Esse grupo continha duas regras:

```
{Day=Quinta,Period=2}     => {Next_Glucose=5} 0.0499002 0.625      1.527439
{Day=Sexta,Period=2} => {Next_Glucose=5}
```

Neste utilizador, ao contrário do primeiro, foram descobertas regras para valores de glicose 5. As regras são bastante parecidas: em dois dias da semana o utilizador tende a ter o valor de glicose elevado algures durante a tarde. Tendo esta informação, o utilizador pode tentar perceber o que faz nestes dias que possa provocar estas alterações. O simples facto de o utilizador perceber que isto acontece pode ser o suficiente para corrigir e, consequentemente, estabilizar mais os valores de glicemia.

**Utilizador 3**

Para este utilizador, foram geradas 150 regras com o consequente "Next_Glucose". Uma vez mais, recorrendo a uma ajuda visual, obtemos o gráfico do sub-conjunto de regras:

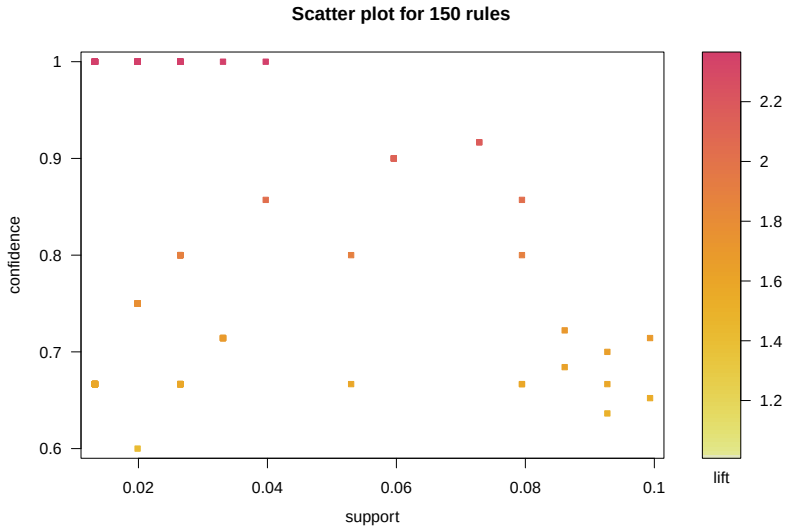


Figura 6.3: Regras de associação para o utilizador 3

A região escolhida para inspecionar foi a região à direita de 0.08, ou seja, com suporte maior que 0.08. Embora o *lift* seja também um parâmetro importante para a avaliação da utilidade das regras, o suporte permite ter uma ideia da frequência, e garante também que uma dada regra faz parte do padrão e não seja um caso pontual. Tendo em conta que os *data set* para este trabalho não são muito grandes, vamos utilizar sempre o suporte como escolha para análise de regras. Como já mencionado, uma regra com suporte muito pequeno pode dizer respeito a um acontecimento único. Algumas das regras na região selecionada são:

```
{Period=2, Value_Carbs=4}       => {Next_Glucose=5} 0.07284768  0.9166667
    2.162760
{Period=3, Value_Insulin=3}     => {Next_Glucose=5} 0.09933775  0.6521739
    1.538723
{Value_Carbs=4} => {Next_Glucose=5} 0.07947020  0.8000000 1.887500
```

Como se percebe, as regras fazem sentido: quando o utilizador ingere uma grande quantidade de hidratos de carbono, tem valores de glicemia demasiado elevados. Mais frequente e específico, isto acontece algures durante a tarde. À noite o utilizador tem tendência a ter hiperglicemia embora tome um valor de insulina adequado. Isto pode acontecer por a noite ser um período mais calmo e parado, e portanto a glicose não é utilizada tão rapidamente, pelo que se acumula em maior quantidade.

Será interessante também ver uma regra com maior *lift*. Essa regra é:

```
{Day=Terca,Period=2} => {Next_Glucose=5} 0.0397351 1          2.359375
```

que é a regra mais à direita da linha de regras na parte superior do gráfico. Embora esta regra seja do mesmo tipo de regras dos utilizadores anteriores, é diferente dessas mesmas regras, pois o

dia é outro.

Para valores baixos de glicemia não foram geradas quaisquer regras.

**Utilizador 4**

Para este utilizador foram geradas 41 regras para valores de hiperglicemia, isto é, "Next_Glucose=5". O gráfico obtido com as regras geradas foi:

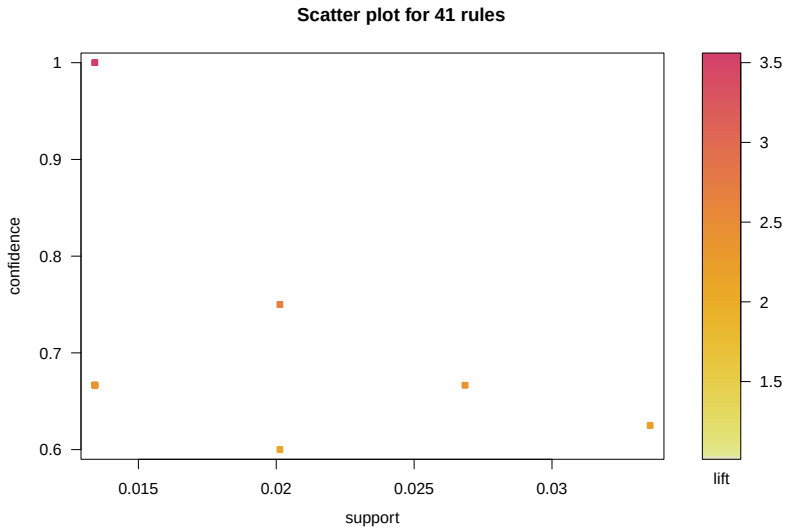


Figura 6.4: Regras de associação para o utilizador 4

Foi escolhido o grupo de regras com suporte maior que 0.03. Uma das regras nesse grupo é:

```
{Day=Sexta, Value_Insulin=3, Insulin_Difference=4} => {Next_Glucose=5}
    0.03355705      0.625 2.217262
```

Como se pode verificar, a regra é parecido com as regras até aqui mostradas, mas com uma diferença: para este utilizador já foi encontrado um padrão em que existem diferenças entre a insulina tomada e a calculada. Nomeadamente, um valor 4 de "Insulin_Difference" significa que o utilizador tomou mais insulina que a insulina calculada. Esta informação parece contraditória, porque insulina a mais causaria um valor baixo de glicemia e não o contrário, mas de facto é o que acontece, tendo em conta que foi descoberto numa regra. Isto pode significar que a insulina calculada tem um valor baixo por erro nas contas de contagem de hidratos e, portanto, a insulina tomada na realidade não seria maior que a insulina calculada.

Algumas das regras com maior *lift*, embora com um suporte consideravelmente mais pequeno, são:

```
{Day=Terca, Value_Carbs=5, Insulin_Difference=2} => {Next_Glucose=5} 0.01342282
            1 3.547619
{Day=Sabado, Period=3, Value_Carbs=3}    => {Next_Glucose=5} 0.01342282
```

```
              1 3.547619
```

Na primeira regra aparece outra vez a variável "Insulin_Difference" mas desta vez com o valor 2, que significa que o utilizador tomou menos insulina que o calculado. Aliando isso à grande quantidade de hidratos de carbono consumida, faz com que às terças-feiras o utilizador tenha tendência para ter hiperglicemia. Ao sábado o mesmo acontece, mas no período da noite.

**Utilizador 5**

Por fim, para o utilizador 5 foram descobertas 136 regras para hiperglicemia e o seguinte gráfico foi obtido:

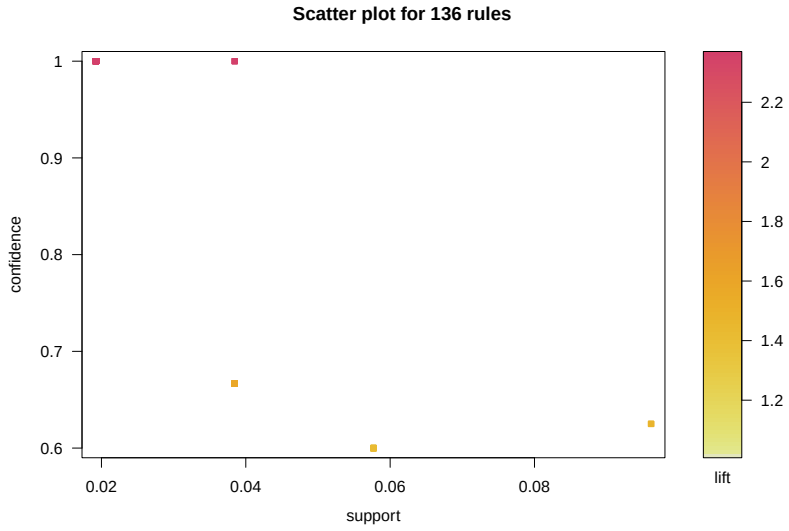


Figura 6.5: Regras de associação para o utilizador 5

Foi escolhido o grupo com maior suporte, ou seja, o quadrado mais à direita, que contém as regras

```
{Period=2,Value_Insulin=3} => {Next_Glucose=5} 0.09615385 0.625      1.477273 3
{Period=2,Value_Carbs=3} => {Next_Glucose=5} 0.09615385 0.625      1.477273 3
```

Isto mostra que, com alguma frequência, o utilizador tem hiperglicemias durante a tarde, mesmo ingerindo uma quantidade de normal de hidratos de carbono ou tomando uma dose de insulina normal. Isto pode indicar que há algum erro na contagem de hidratos de carbono e portanto, ou a insulina ou os hidratos de carbono podem estar com valores errados, refletindo um aumento na glicemia. Note-se também que isto ocorre à tarde, o que pode significar que estes erros possam acontecer ao almoço, por exemplo.

Com as regras de associação deu para perceber que diferentes utilizadores têm diferentes rotinas e padrões e, portanto, diferentes motivos para hipo ou hiperglicemias, que neste caso se

traduz em diferentes regras. Isto mostra que, para estes utilizadores, as soluções para estabilizar os valores de glicemia seriam diferentes, pelo que fica evidente a necessidade de um tratamento personalizado e individualizado. Fica também a noção que uma aplicação capaz de, em tempo real, aprender o que é normal e anormal e dar *feedback*, através de avisos/alertas, teria um papel importante no controlo da glicemia. Um simples aviso como "Hoje é segunda-feira e tendencialmente tem hiperglicemia" pode deixar o utilizador em alerta e, quanto mais não seja, ter um controlo mais apertado nesse dia. De notar que a variável "Exercise" não apareceu em nenhuma situação nesta análise por ser tão pouco frequente.

## 6.2 Redes *Bayesianas*

Nesta secção vamos fazer uma análise um pouco diferente da feita na secção 6.1. As regras de associação exploram os dados e tentam encontrar relações frequentes com esses dados, para gerar regras do tipo "Às terças-feiras normalmente tens hiperglicemia". Esta é uma abordagem interessante e importante, mas a abordagem inversa também pode ser útil: "Hoje é terça-feira, qual será a probabilidade de teres hiperglicemia?". Para tal, criou-se uma rede *bayesiana* para cada utilizador, de forma a perceber de que forma os vários parâmetros influenciam a glicemia, em vez de sabermos apenas que influenciam. Isto será possível porque uma rede *bayesiana* consegue calcular a probabilidade de cada variável tomar um determinado valor, com base no *data set*, isto é, se em todo o *data set* o período mais frequente for "Tarde" então este será o valor mais provável para o período. Assim, para cada utilizador será possível conjugar os vários parâmetros e ver com que valores é que a probabilidade de hipo ou hiperglicemia aumentam.

Nesta fase, o primeiro passo é, então, criar uma rede *bayesiana*. Para tal, recorremos ao *software* Waikato Environment for Knowledge Analysis (WEKA). O WEKA cria a rede em função de uma variável de classe, que neste caso é "Next_Glucose" sendo por isso necessário colocar esta variável como última coluna no *data set*. Note-se que o WEKA é utilizado apenas para criar a rede, sendo que a análise será feita num outro programa, o Sensivity Analysis, Modeling, Inference And More (SamIAm).

Para relembrar, uma rede *baysesiana* é um grafo dirigido. Um exemplo de uma rede está representado na figura 6.6

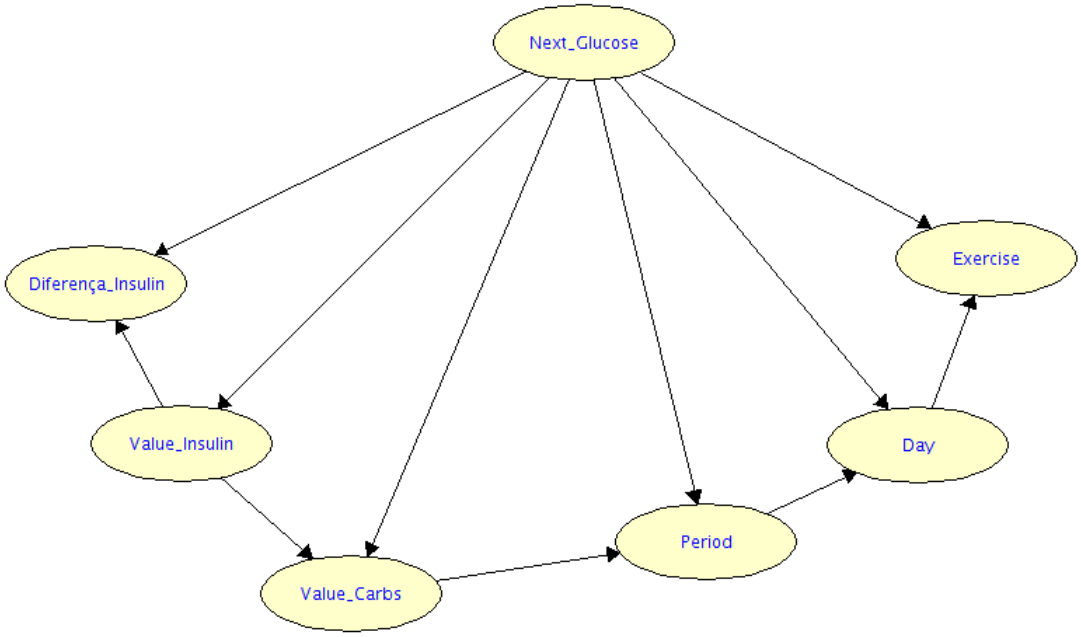


Figura 6.6: Exemplo de uma rede *bayesiana*

O SamIAm tem dois modos diferentes: modo *edit* e modo *query*. O modo *edit* serve para editar a estrutura da rede, tal como adicionar nós ou arestas, mas esse não é o objetivo visto que a rede é criada previamente. O modo *query* permite perceber de que forma as variáveis se relacionam umas com as outras e perceber os efeitos que os valores de uma variável provocam nos valores de outras variáveis, sendo que é este o objetivo. Assim, vamos utilizar o programa SamIAm apenas no modo *query*.

Este modo permite ver a probabilidade de cada variável tomar cada valor, de acordo com o *data set* e permite também mudar esses valores e ver de que forma as outras variáveis mudam também. Por exemplo, uma regra possível seria "Um utilizador tem tendência para ter valores de glicemia altos à quarta-feira à tarde". É interessante e útil para o utilizador conhecer esta regra, por si só. Mas também pode ser interessante o processo inverso, ou seja, "Qual a probabilidade de ter hiperglicemia, sendo que hoje é quarta-feira à tarde?". É isso que o SamIAm permite fazer no *query mode*, entre outras coisas, e que pode dar mais informação útil ao utilizador. A figura seguinte mostra a mesma rede mas em *query mode*.

A rede *bayesiana* é feita com base nos dados de cada utlizador, sendo que será criada uma rede para cada utilizador e cada utilizador pode ter uma rede diferente. O passo seguinte é criar uma rede para cada utilizador e ver que informação adicional se consegue obter.

É possível ver as tabelas de probabilidade para cada variável, sendo que cada nó tem por predefinição o valor mais frequente, como por exemplo, "Next_Glucose=3". Note-se que os valores Not Available (NA) correspondem a registos que não usaram essas variáveis, isso é, um utilizador pode registar uma glicemia e não registar insulina ou hidratos d carbono. Estas tabelas também permitem perceber a distribuição de algumas variáveis: por exemplo, é possível

verificar que, neste caso, os dias têm probabilidades muito parecidas, o que significa que têm um número parecido de registos. No entanto, tal como já mencionado, o objetivo nesta parte é tentar perceber de que forma estas variáveis afetam o valor da glicemia e também ver se estas alterações estão em concordância com as regras mostradas na secção anterior ou até ver se novos padrões surgem. Vamos, por isso, mostrar a rede *bayesiana* para cada utilizador.

**Utilizador 1**

A rede criada para o utilizador 1 está representada na figura **??**

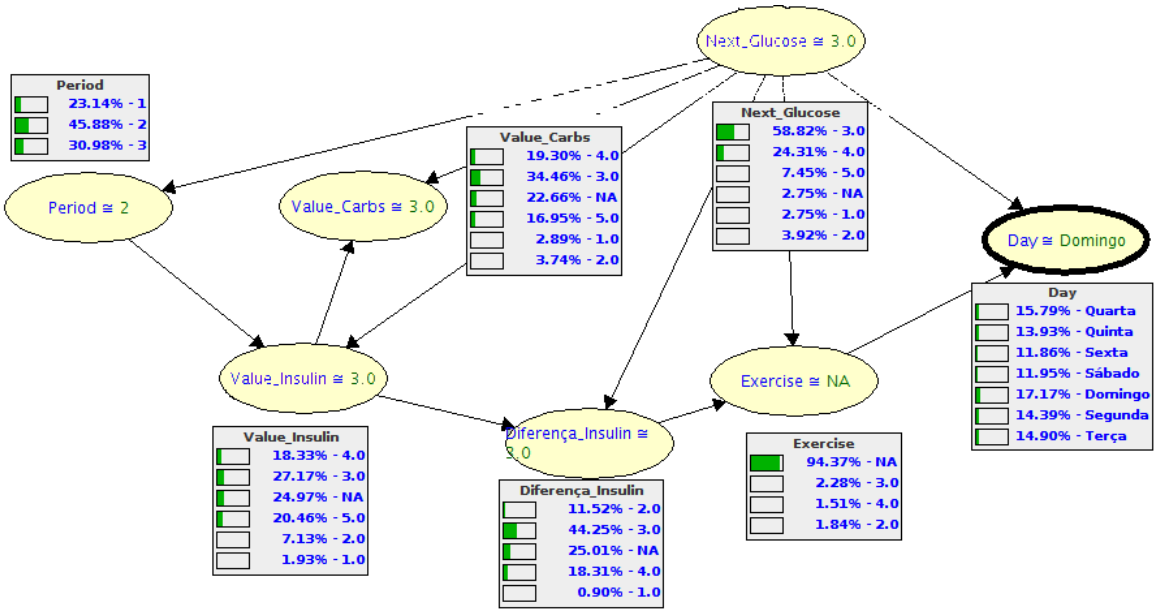


Figura 6.7: Rede *bayesiana* para o utilizador 1

Para cada variável é possível ver as tabelas de probabilidade para os seus possíveis valores: por exemplo, probabilidades parecidas para os dias da semana indicam que os dias têm uma quantidade de registos parecida. Por sua vez, o valor 3 para glicose tem maior probabilidade que os outros, mostrando que a maioria dos registos de glicemia está dentro dos valores normais.

Podemos ter duas abordagens diferentes: alterar os parâmetros de acordo com as regras de associação geradas, para comprovar a concordância entre os dois programas ou tentar alterar os parâmetros para tentar descobrir padrões que provoquem valores alterados e que não tenham sido descobertos pelas regras. Para o primeiro caso, na secção anterior, uma das regras do utilizador 1 mostrava que, às quartas-feiras à tarde, este ingere valores altos de hidratos de carbono pelo que a glicemia tem também valores elevados. No SamIAm, ao selecionar esses três parâmetros com os valores definidos na regra, verifica-se que a probabilidade de "Next_Value=4" sobe de 24.32% para 62.68%, tal como seria de esperar.

Por outro lado, se colocarmos as variáveis de acordo com a regra para hipoglicemia neste utilizador, verificamos que a probabilidade de "Next_Glucose=2" sobe para 34.74%.

No entanto, é possível descobrir outros padrões: por exemplo, se selecionarmos apenas o dia "Domingo" verificamos que a probabilidade de uma hiperglicemia, ou seja um valor de glicemia 5, sobe para 18.67%.

Na figura 6.8 pode-se ver quais os parâmetros que provocam uma probabilidade máxima de uma hiperglicemia.

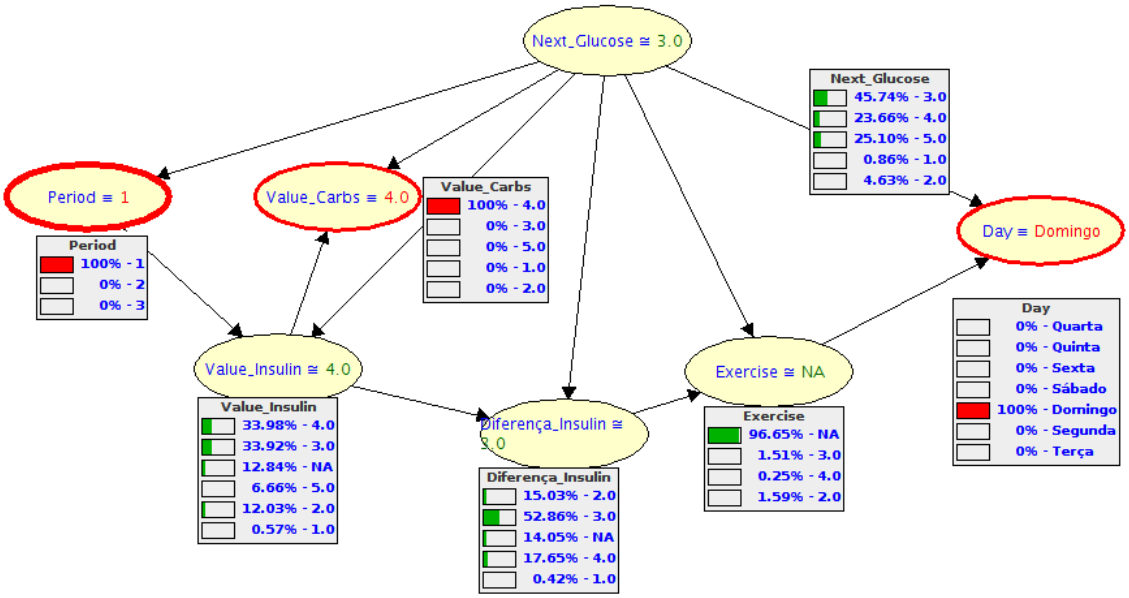


Figura 6.8: Probabilidade máxima de hiperglicemia para o utilizador 1

Ou seja, uma nova regra é descoberta: ao Domingo de manhã, quando o utilizador ingere uma quantidade grande de hidratos de carbono, tem maior probabilidade de ter uma hiperglicemia.

**Utilizador 2**

A rede para o utilizador 2 é mostrada na figura 6.9

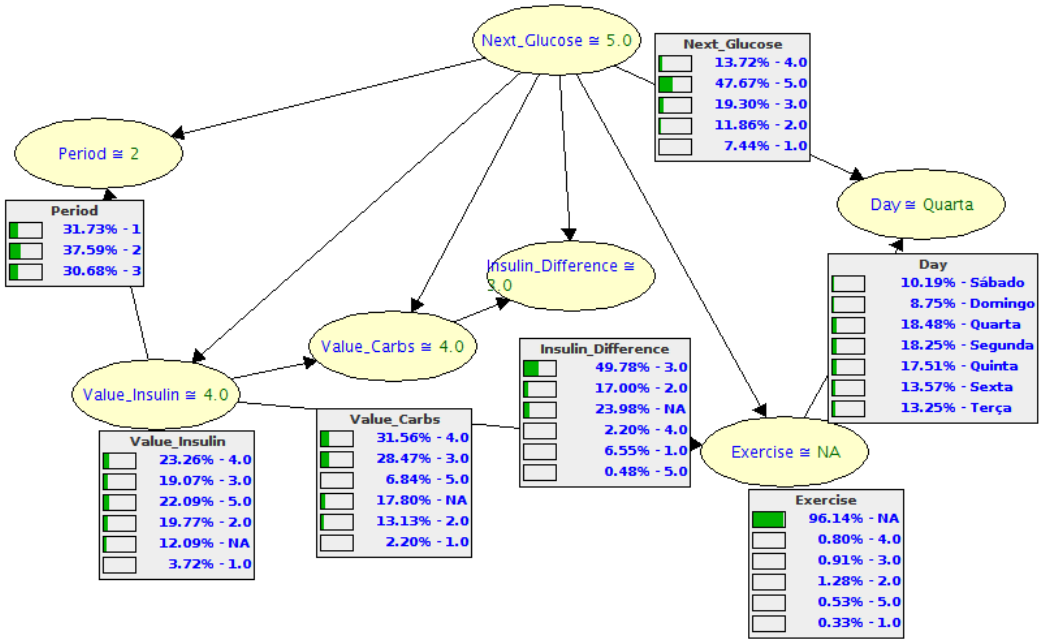


Figura 6.9: Rede *bayesiana* para o utilizador 2

que, como se verifica, é ligeiramente diferente da rede gerada para o utilizador 1. Uma vez mais, é interessante ver quais os valores dos vários parâmetros que maximizam o valor de glicemia.

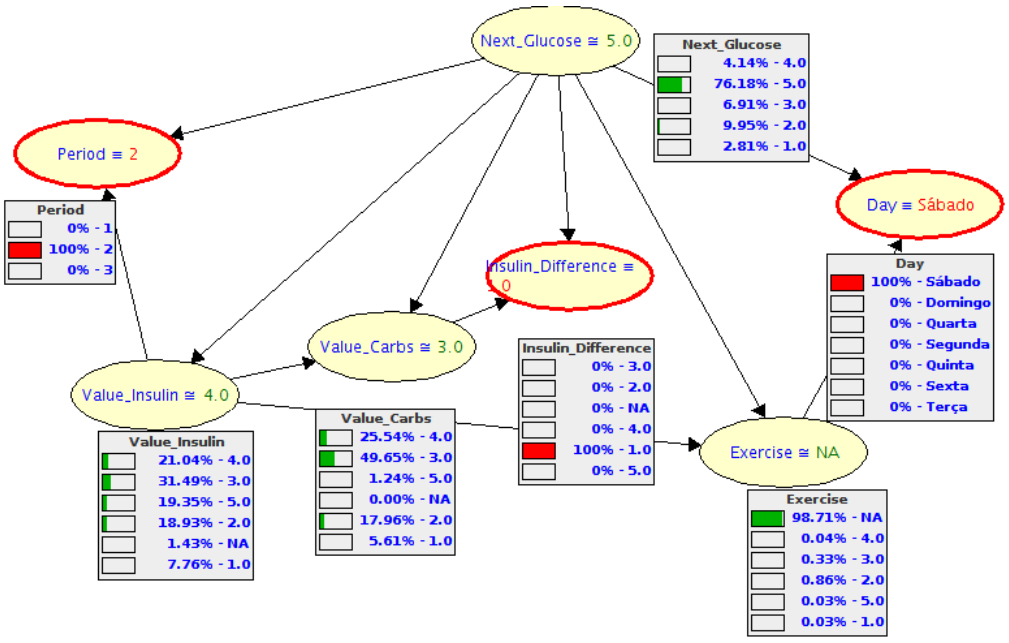


Figura 6.10: Probabilidade máxima de hiperglicemia para o utilizador 2

Como se pode observar, a probabilidade de hiperglicemia aumenta bastante aos sábados à tarde, quando o utilizador toma menos insulina que a recomendada. Esta regra não foi gerada pelo algoritmo *apriori*, possivelmente por ter um valor de suporte abaixo do mínimo definido, mas

ainda assim é uma regra interessante e que mostra uma utilidade desta análise. É interessante saber quais as condições em que o risco de hiperglicemia é maior, tendo em conta os dados já registados.

### Utilizador 3

A rede criada para o utilizador 3 está representada na figura 6.11.

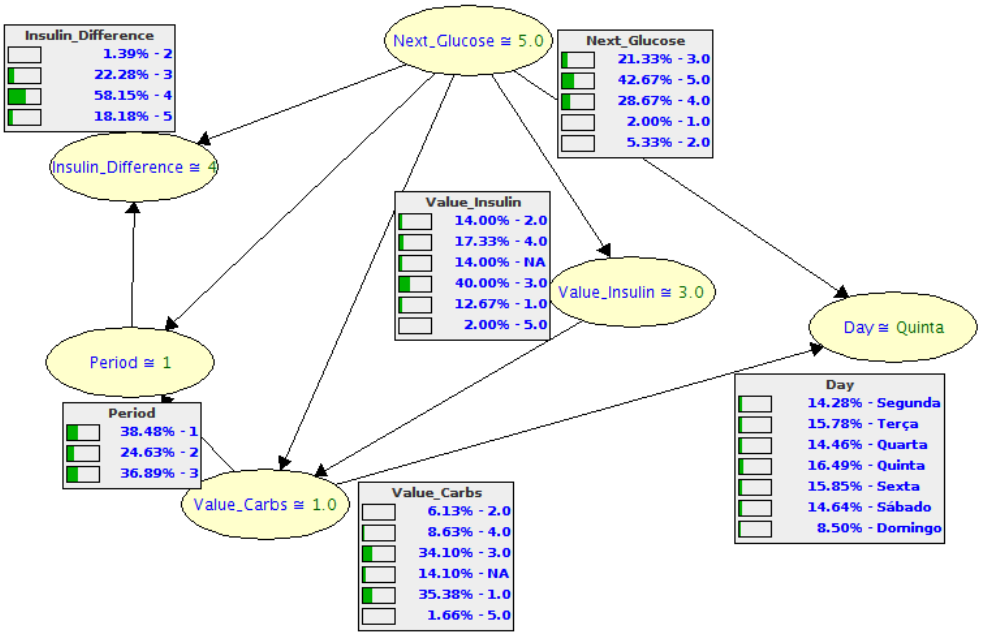


Figura 6.11: Rede *bayesiana* para o utilizador 3

Esta rede é diferente das redes anteriormente apresentadas. A principal diferença que se nota logo é a falta do nó "Exercise" que se deve pelo facto de este utilizador não ter qualquer registo de exercício e portanto, a variável não foi tida em conta. A rede já com as probabilidades é:

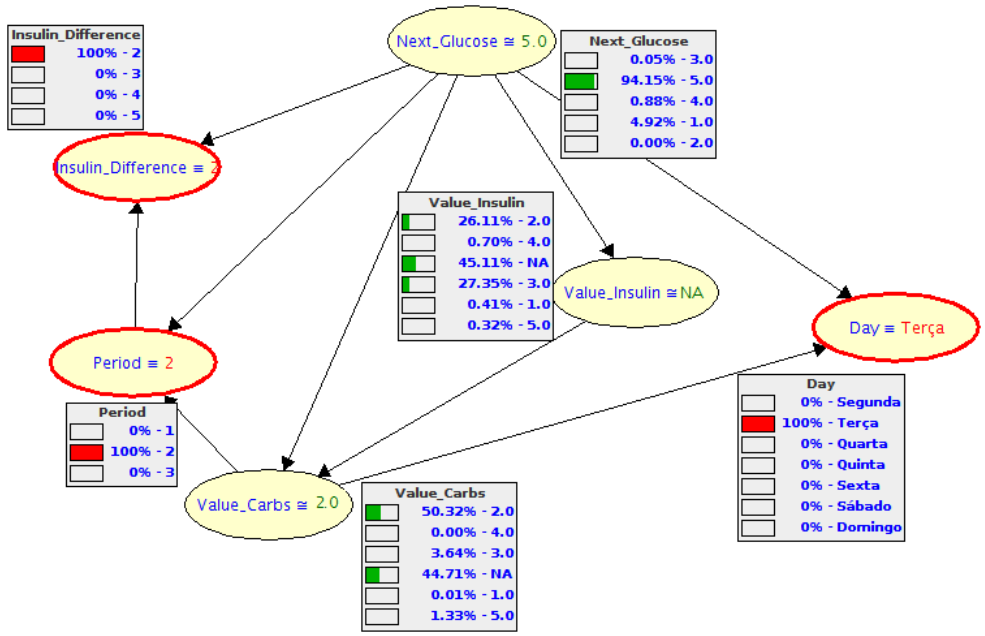


Figura 6.12: Probabilidade máxima de hiperglicemia para o utilizador 3

Como se pode perceber pela figura 6.12, este utilizador tem uma probabilidade enorme de ter uma hiperglicemia à terça-feira à tarde, especialmente se tomar menos insulina do que a calculada.

**Utilizador 4**

A rede *bayesiana* para o utilizador 4 está representada na figura 6.13, sendo que também é diferente das outras.

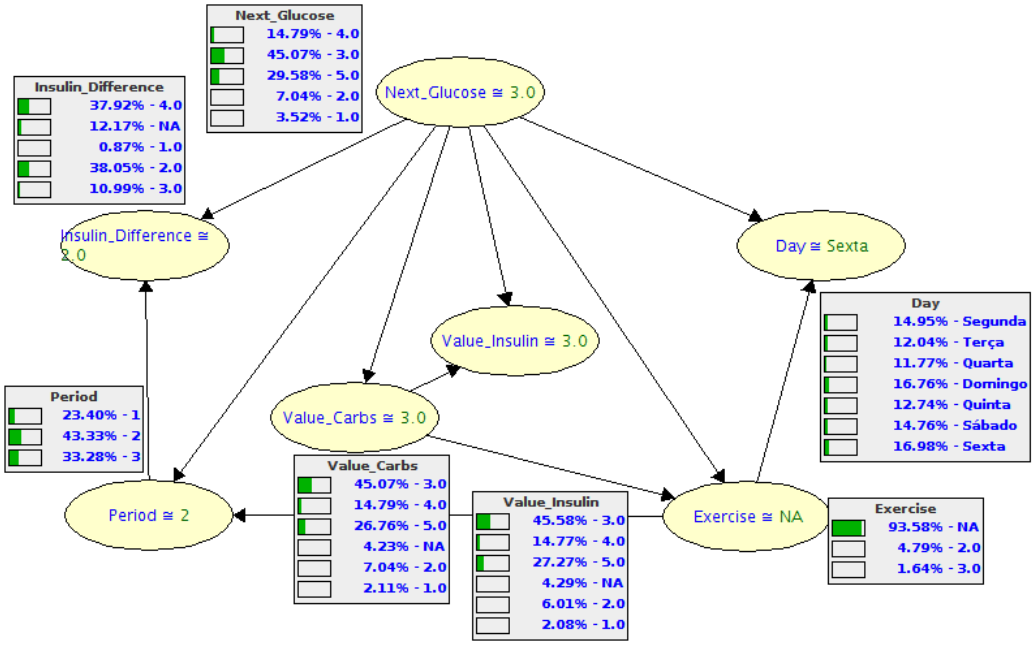


Figura 6.13: Rede *bayesiana* para o utilizador 4

Desde logo é possível notar que para a variável "Insulin_Difference" os valores com probabilidades maiores são o 2 e 4, o que mostra que este utilizador, na maior parte das vezes, toma uma quantidade de insulina diferente da quantidade calculada, o que pode causar consequências nos valores da glicemia. De resto também é possível ver que o exercício é algo raro, tal como nos outros utilizadores.

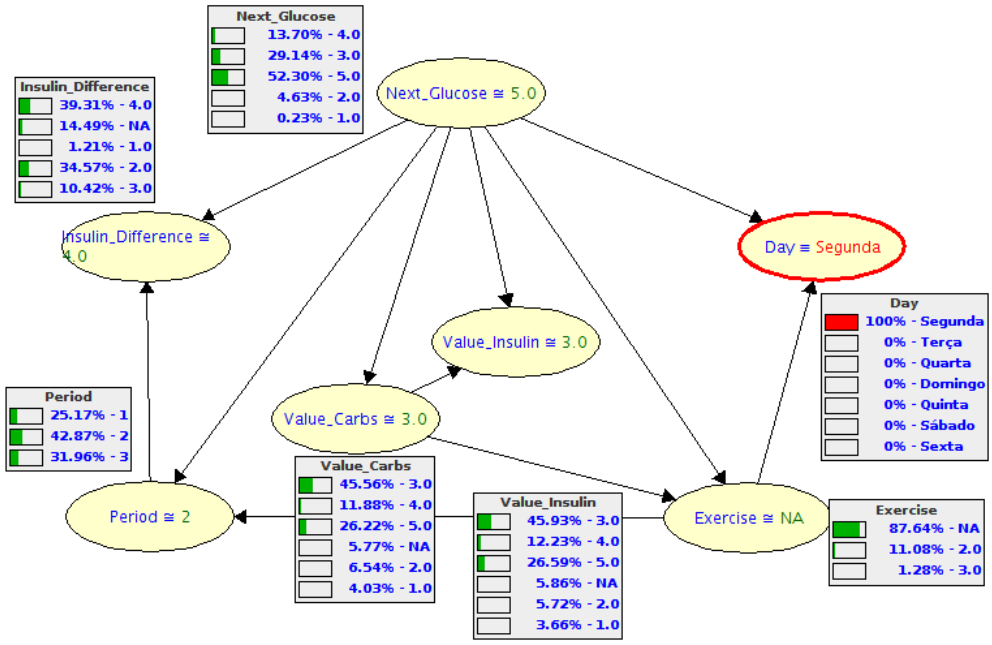


Figura 6.14: Probabilidade máxima de hiperglicemia para o utilizador 4

Na figura 6.14 é possível observar que à segunda-feira a probabilidade de ocorrer uma hiperglicemia aumenta consideravelmente, sendo que, neste caso, os outros parâmetros não influenciam muito a probabilidade de hiperglicemia, pelo menos para este dia escolhido. Por exemplo, à segunda-feira a probabilidade de hiperglicemia é parecida para qualquer período.

Por outro lado, verificamos que o utilizador tem uma maior probabilidade de ter um valor de hipoglicemia, isto é, "Next_Glucose=1" à sexta-feira à tarde, cuja probabilidade aumenta para os 10.19%.

**Utilizador 5**

Para o último utilizador, a rede gerada está representada na figura 6.15.

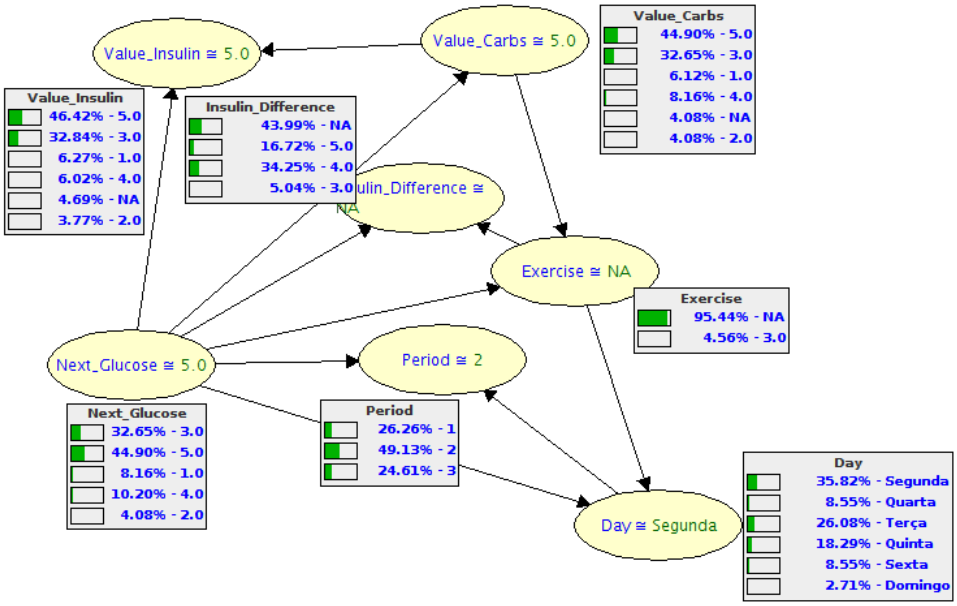


Figura 6.15: Rede *bayesiana* para o utilizador 5

Uma vez mais, um exemplo de uma combinação de parâmetros que aumentam a probabilidade de hiperglicemia é representado na figura 6.16.

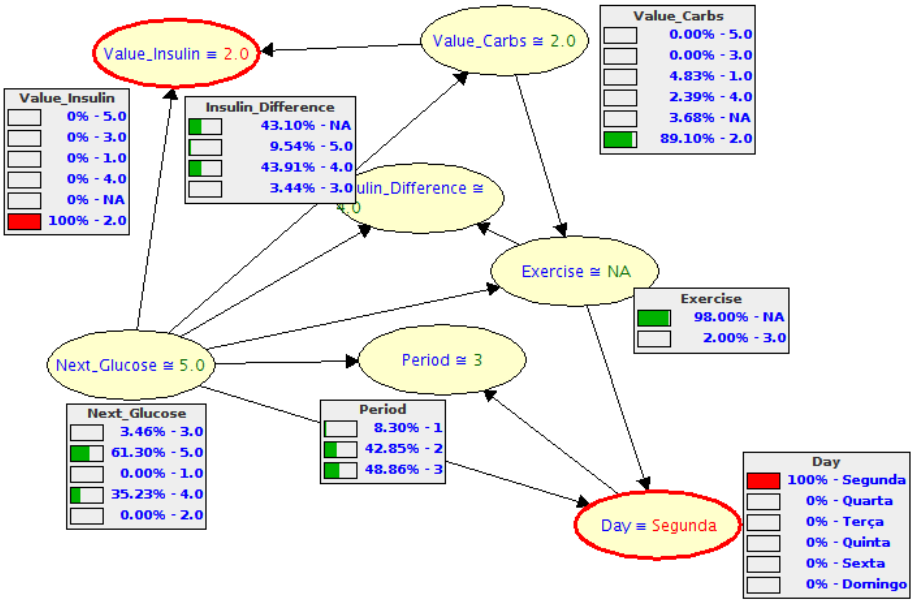


Figura 6.16: Probabilidade máxima de hiperglicemia para o utilizador 5

É possível notar que à segunda-feira, e quanto o utilizador toma pouca insulina, o valor de glicose no sangue aumenta, ocorrendo uma hiperglicemia. Nesta situação, um simples alerta para aumentar a dose de insulina poderia ser suficiente para corrigir esta situação e acabar com este padrão.

Quanto a hipoglicemias, para este utilizador, a rede mostra que a probabilidade de "Next_Glucose=1" aumenta nas manhãs em que o utilizador toma mais insulina do que aquela recomendada, sendo que neste caso a probabilidade é de 20.09%.

## 6.3 *Inductive Logic Programming*

Para este tipo de análise vamos usar o já mencionado sistema Aleph. Para tal, são necessários três ficheiros: um para exemplos positivos, um para exemplos negativos e outro para o *background knowledge*, sendo que estes três ficheiros teriam de ser em prolog.

# Capítulo 7

# Conclusões

Ao longo desta dissertação fomos enfatizando duas coisas acerca da diabetes: é uma doença sem cura e, com tratamento adequado, os doentes diabéticos podem levar uma vida normal. Vimos também que não há um tratamento generalizado e que este tem de ser personalizado tendo em conta as características de cada doente. Assim sendo, propusemos fazer diversos tipos de análise para um conjunto de utilizadores. Estas análises foram personalizadas para cada utilizador para tentar perceber que tipo de comportamentos é que os doentes diabéticos têm que provocam alterações na glicemia.

Numa primeira análise estatística apenas com os valores de glicemia, foi possível observar alguns padrões, nomeadamente perceber que há dias em que a glicemia é mais elevada do que outros. Na parte de regras de associação, por sua vez, foi possível aprofundar esse conhecimento e passar a ter razões para essas oscilações nos valores de glicemia. Este conhecimento é importante porque assim permite saber qual pode ser a solução. Verificou-se que cada utilizador tem a sua própria rotina o que faz com que cada utilizador tenha também diferentes regras. Note-se que as regras podem ser do mesmo tipo, como serem regras envolvendo períodos do dia, mas são diferentes, pois têm valores diferentes. Esta diferença entre regras para cada utilizador é importante, uma vez que torna esta análise personalizada. Finalmente, na parte das redes *bayesianas* foi possível tentar uma abordagem diferente, ao conseguir saber quais as condições que, para cada utilizador, provocam valores mais baixos ou mais altos de glicemia. Este tipo de análise também é importante uma vez que, permite saber, por exemplo, qual é a probabilidade de uma hiperglicemia a cada momento. Se um dado momento tiver alta probabilidade de hiperglicemia então o utilizador pode ser alertado para esse facto, tomando as precauções necessárias.

Como se pode perceber, esta personalização é fundamental: a análise é feita para cada utilizador, com base nos dados que cada utilizador introduz e, consequentemente, com base nos seus hábitos de vida, pelo que qualquer regra gerada pode ser considerada fidedigna. A consciencialização de que um comportamento pode levar a valores de glicemia indesejáveis é o primeiro passo para que isso possa ser corrigido.

Por fim, conclui-se que é viável usar *data mining* em registos de pacientes diabéticos a fim

de melhor o seu tratamento: foi possível concluir alguns factos diferentes sobre um pequeno conjunto de utilizadores o que dá a ideia de que, para uma maior quantidade de utilizadores, se consiga obter ainda mais variedade de regras. Esta variedade de regras será fundamental para poder, eventualmente, integrar um sistema destes numa aplicação.

## 7.1 Trabalho Futuro

No futuro, seria interessante integrar um sistema de aconselhamento personalizado numa aplicação de registo de diabetes, nomeadamente na aplicação MyDiabetes. O facto de uma pessoa diabética poder ter, em tempo real, aconselhamento sobre o que pode estar a fazer de errado contribuirá para um melhor tratamento e, por isso, uma melhor qualidade de vida. Também para trabalho futuro seria interessante fazer um outro tipo de análise: em vez de analisar cada utilizador individualmente, pode ser interessante utilizar todos os utilizadores de forma conjunta. Este tipo de análise mais geral pode permitir descobrir grupos de pessoas com características ou rotinas idênticas, criando grupos de pessoas diferentes. Por exemplo, poderia chegar-se à conclusão que mulheres e homens têm rotinas parecidas, dentro de cada um dos dois grupos. Esse tipo de análise permitia identificar a que grupo pertenceria cada utilizador e portanto, apresentar regras mais específicas para esse grupo.

De qualquer das formas, para que tanto a integração como uma análise mais geral possam ser feitas, é necessária uma maior quantidade e variedade de dados. Para isto, talvez um passo a curto prazo fosse alargar o grupo de utilizadores da aplicação, através do alargamento a outros hospitais ou até mesmo à disponibilização pública da aplicação. Isto permitirá recolher mais dados e fazer o que já foi mencionado.

# Bibliografia


[1] Accu-Check Combo. https://www.accu-chek.com.br/br/produtos/sic/index.html. [Online; acedido a 17 de Junho de 2016].

[2] Continuous Glucose Monitoring | What is CGM? https://www.dexcom.com/continuous-glucose-monitoring. [Online; acedido a 17 de Junho de 2016].

[3] The Comprehensive R Archive Network. https://cran.r-project.org/. [Online; acedido a 17 de Junho de 2016].

[4] Diabetes data set. https://archive.ics.uci.edu/ml/datasets/Diabetes. [Online; acedido a 17 de Junho de 2016].

[5] What is HbA1c? http://www.diabetes.co.uk/what-is-hba1c.html. [Online; acedido a 17 de Junho de 2016].

[6] Blood sugar level ranges. http://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html. [Online; acedido a 17 de Junho de 2016].

[7] Operating system market share. https://www.netmarketshare.com/operating-system-market-share.aspx?qprid=8&qpcustomd=1. [Online; acedido a 17 de Junho de 2016].

[8] Prediabetes. http://www.mayoclinic.org/diseases-conditions/prediabetes/basics/definition/con-20024420. [Online; acedido a 17 de Junho de 2016].

[9] sqlite2csv. https://gist.github.com/Heart1010/3e2b2c528257e18e9b98. [Online; acedido a 17 de Junho de 2016].

[10] Type 1.5 diabetes. http://www.diabetes.co.uk/type15-diabetes.html. [Online; acedido a 17 de Junho de 2016].

[11] Number of smartphone users worldwide from 2014 to 2019 (in millions). http://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/. [Online; acedido a 17 de Junho de 2016].

[12] James W Albers, William H Herman, Rodica Pop-Busui, Eva L Feldman, Catherine L Martin, Patricia A Cleary, Barbara H Waberski, John M Lachin, DCCT/EDIC Research Group, et al. Effect of prior intensive insulin treatment during the diabetes control and complications

trial (dcct) on peripheral neuropathy in type 1 diabetes during the epidemiology of diabetes interventions and complications (edic) study. *Diabetes Care*, 33(5):1090–1096, 2010.

[13] IDF Diabetes Atlas. International diabetes federation, brussels, 2015. 2015.

[14] Michael J Berry and Gordon Linoff. *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc., 1997.

[15] Judith E Dayhoff and James M DeLeo. Artificial neural networks. *Cancer*, 91(S8):1615–1635, 2001.

[16] Sociedade Portuguesa de Diabetologia. Definição, diagnóstico e classificação da diabetes mellitus. [Online; acedido a 25 de Junho de 2016].

[17] Dursun Delen, Glenn Walker, and Amit Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–127, 2005.

[18] Diabetes.co.uk. Type 2 diabetes. http://www.diabetes.co.uk/type2-diabetes.html. [Online; acedido a 17 de Junho de 2016].

[19] Pedro Ferreira, Nuno A Fonseca, Inês Dutra, Ryan Woods, and Elizabeth Burnside. Predicting malignancy from mammography findings and image–guided core biopsies. *International journal of data mining and bioinformatics*, 11(3):257–276, 2015.

[20] Food and Drug Administration. *Paving the Way for Personalized Medicine: FDA's Role in a New Era of Medical Product Development*, pages 5–11. CreateSpace Independent Publishing Platform, 2013.

[21] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[22] Longfei Han, Senlin Luo, Jianmin Yu, Limin Pan, and Songjing Chen. Rule Extraction From Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes. *IEEE journal of biomedical and health informatics*, 19(2):728–734, 2015.

[23] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter*, 2(1): 58–64, 2000.

[24] David M Maahs, Nancy A West, Jean M. Lawrence, and Elizabeth J Mayer-Davis. Epidemiology of type 1 diabetes. *Endocrinology and metabolism clinics of North America*, 39(3):481–497, Setembro 2010.

[25] Colin D Mathers and Dejan Loncar. Projections of global mortality and burden of disease from 2002 to 2030.

[26] Viktor Mayer-Schonberger and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Company, 2013.

[27] Sellappan Palaniappan and Rafiah Awang. Intelligent Heart Disease Prediction System Using Data Mining Techniques . In *International Conference on Computer Systems and Applications*, pages 108–115, 2008.

[28] Jerry Radziuk. The artificial pancreas. *Diabetes*, 61(9):2221–2224, 2012.

[29] Arlan L Rosenbloom, Jennie R Joe, Robert S Young, and William E Winter. Emerging epidemic of type 2 diabetes in youth. *Diabetes care*, 22(2):345–354, 1999.

[30] Fabrizio Ruggeri, Ron S. Kennet, and Frederick W. Faltin. *Encyclopedia of Statistics in Quality and Reliability*. Wiley, 2007.

[31] S. Stilou, P.D. Bamidis, N. Maglaveras, and C. Pappas. Mining Association Rules from Clinical Databases: An Intelligent Diagnostic Process in Healthcare. *Studies in health technology and informatics*, 2001.

[32] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2008.