

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	xiii
Lista de Blocos de Código	xv
1 Introdução	1
1.1 Contexto	1
1.2 Motivação	2
1.3 Projeto	2
1.3.1 Objetivos	3
1.3.2 Contribuição	3
1.4 Organização	4
2 Fundamentos e Terminologia	5
2.1 <i>Diabetes Mellitus</i>	5
2.1.1 Dispositivos para monitorizar a diabetes	7

2.1.1.1	Bombas infusoras de insulina	7
2.1.1.2	Glicosímetros	8
2.1.1.3	Monitor contínuo de glicose	8
2.2	<i>Data Mining</i>	8
2.2.1	<i>Data Mining</i> na diabetes	9
3	Estado da Arte	11
3.1	Aplicações para <i>smartphones</i> Android	11
3.1.1	Diário da Diabetes mySugr	11
3.1.2	Diabetes:M	12
3.1.3	OnTrack Diabetes	13
3.1.4	Diabetes - Diário Glucose	13
3.1.5	Glucose Buddy: Diabetes Log	13
3.2	Medicina personalizada e data mining na saúde	14
4	MyDiabetes	17
4.1	Objetivo da aplicação	17
4.2	Arquitetura	18
4.3	Variáveis recolhidas	18
5	Análise de dados	21
5.1	Descrição do estudo	21
5.2	Recolha de dados	22
5.2.1	Números e <i>feedback</i>	23
5.2.2	O <i>data set</i>	24
5.2.3	Pré-processamento dos dados	25
5.3	Análise estatística básica	27
5.3.1	Média de glicose	27
6	Conclusões	29

6.1 Trabalho Futuro	29
Bibliografia	31
A Acrónimos	31

Lista de Tabelas

5.1	<i>Data set</i> original	25
-----	------------------------------------	----

Lista de Figuras

Lista de Blocos de Código

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 2014, cerca de 387 milhões de pessoas tinham diabetes. Em 2035, este número aumentará para 592 milhões e, de acordo com a Organização Mundial da Saúde, em 2030 a diabetes será a sétima causa de morte no planeta.

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 [controlo.pdf] 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 progresso. 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. 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 final desta dissertação é desenvolver um sistema capaz de gerar regras e mostrar avisos ou conselhos a partir dos dados inseridos, em tempo real e integrá-lo numa aplicação já existente. No entanto há mais objetivos:

- 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;
- Integrar este sistema na aplicação MyDiabetes.

1.3.2 Contribuição

A contribuição principal deste trabalho vai ser um sistema de aconselhamento inteligente para diabéticos, integrado numa aplicação para Android. Serão também considerados como

contribuições os seguintes pontos:

- 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.

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 os diferentes tipos de análises efetuadas bem como os diferentes *softwares* utilizados.

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. 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 coisas.

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 na comida. 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 e 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, sendo que todas elas provocam diabetes mas de tipos diferentes. 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. 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. 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, apesar de haver também diabéticos tipo 2 insulino-dependentes, e a medicação é feita através de comprimidos. Apesar de a diabetes tipo 2 surgir normalmente em pessoas mais velhas, tem-se vindo a manifestar também em jovens.
- *Diabetes gestacional* Este tipo de diabetes pode aparecer durante a gravidez. Caracteriza-se por ter valores de glicose superiores aos normais mas, ainda assim, 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 Autoimmune 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. [1.5] 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.[1.5]

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 a 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. Um outro parâmetro importante é o rácio de hidratos de carbono. [acabar]

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 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. 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. [prediabetes]

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. 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 chamam-se de 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 uma hipoglicemia 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 uma consequência grave.

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. De seguida 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: medir a glicemia e colocá-la a valores normais, se necessário. Alguns dispositivos fazem isto de forma automática, como as bombas infusoras de insulina, outros fazem-nos de forma indireta, ao alertar o utilizador para que ele possa fazê-lo. Entre 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. No entanto, a análise a aplicações para ajuda na diabetes será feita apenas no Capítulo 3.

2.1.1.1 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* para que a insulina injetada seja de acordo com os níveis de glicemia. [akku]

2.1.1.2 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.

2.1.1.3 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. [continuous]

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.[oracle] A área de *data mining* usa diversos métodos de outras áreas tais como matemática, inteligência artificial e *machine learning* para tratar, explorar e obter conclusões acerca dos dados. Esta área é utilizada para diversos fins, sendo que alguns são deteção de anomalias, associação e classificação.

- 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. Por exemplo, ao analisar os dados relativos a compras num supermercado, pode-se concluir que quem compra cerveja e pão, tem tendência a comprar leite. Esta categoria do *data mining* é especialmente interessante em negócios.
- classificação Tem como objetivo estudar conjuntos de dados e gerar modelos com base nesses dados. Depois, ao observar novos dados dentro com igual formato, 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.

A área de *data mining* tem-se tornado cada vez mais popular e mais usada em variadas áreas, como economia, educação e saúde. 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.

2.2.1 *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 seja que, algures durante o mês, hajam 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 ter 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 voltar 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. Basta descobrir este padrão e dá-lo a conhecer ao paciente para que ele se alimente melhor, e acaba-se com alguns valores hiperglicémicos.

Assim, e imaginando que o paciente utilizaria a aplicação MyDiabetes, 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.

Capítulo 3

Estado da Arte

Vimos anteriormente que a tecnologia pode ser útil ao serviço da medicina. Vimos que existem dispositivos para medir a glicose e para controlar a glicemia e percebemos também que os *smartphones* também 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) aplicações para *smartphones* e 2) uso de medicina personalizada para controlar a diabetes, através de técnicas de *data mining*. 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. [PersMed.pdf]

3.1 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.[smartUsers] 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 [mercadomovel] 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.

3.1.1 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 glicómetro, o "iHealth BG5". [ref] 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.

3.1.2 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 glicosilada (também conhecida por hemoglobina glicada ou HbA1c). 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 glicómetros de diferentes modelos, tais como OneTouch, Dexcom ou Accu-Chek.

3.1.3 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.

3.1.4 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*.

3.1.5 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 seleccionando 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.

3.2 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.

Em [tiago.pdf], os autores 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.

Em [palaniappan2008.pdf], os autores 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). Obteram 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.

Em [associationDiabetes.pdf], os autores 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.

Finalmente, em [SVMDiabetes.pdf], os autores usaram algoritmos de classificação para gerar um modelo de diagnóstico da diabetes. Neste caso, usaram-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*, 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. [svm1.pdf][svm2.pdf] 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". 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.

Uma parte importante deste trabalho é a aplicação MyDiabetes, será a aplicação onde o sistema desenvolvido será integrado e que, até lá, serve de plataforma para os clientes registarem e enviarem os dados a serem utilizados no projeto. Assim, o próximo capítulo aborda de forma mais detalhada a aplicação.

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. Por não haver *data sets* existentes com o formato e tipo de dados pretendidos, tornou-se necessário recolhê-los. Para tal, disponibilizámos a aplicação para ser utilizada de forma voluntário por alguns utilizadores diabéticos. Este processo será explicado no próximo capítulo. Esta aplicação foi a escolhida para disponibilizar aos voluntários não só por ter sido desenvolvida neste mesmo projeto mas também pelo maior controlo que permite ter ao permitir registar todos os dados que possam ser relevantes para esta dissertação. Outra característica que a torna relevante é o facto de ter um sistema que permite o envio dos dados dos utilizadores para o projeto, para que estes possam ser utilizados. Sendo então esta a aplicação utilizada, vamos descrever alguns aspetos importantes.

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. No futuro, a aplicação será mais interativa com o utilizador ao mostrar diferentes avisos ou conselhos. 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 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 e o rácio de hidratos de carbono. 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.

Capítulo 5

Análise de dados

Neste capítulo será discutido o que foi feito nesta dissertação. Serão abordados requisitos para realizar o trabalho, as decisões tomadas e respetiva explicação e as dificuldades encontradas, que levaram a eventuais alterações aos objetivos inicialmente propostos. Serão ainda mostrados os vários tipos de análises efetuadas aos dados recolhidos, bem como alguns resultados daí obtidos.

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.

Face à falta de dados disponíveis, a solução encontrada foi recolher, de raíz, os 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. Por variedade e tamanho desejáveis entenda-se dados relativos a cerca de 20 pessoas durante algumas semanas.

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 Números e *feedback*

Este processo era feito duas vezes por semana, que correspondia aos dias de consultas da diabetes, 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 não disponibilidade para participar como voluntário. No final deste período tínhamos conseguido a participação de 31 voluntários. Apesar de este número ser suficiente para o que se pretendia, pois inicialmente tinha sido definido como objetivo pelo menos 20 participantes, 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 relativamente a algumas semanas. Isto 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 errados para cada utilizador, estes podem aperceber-se de que realmente têm esses comportamentos errados e usufruir verdadeiramente deste sistema de aconselhamento.

Para fazer esta distinção torna-se importante obter a maior quantidade e variedade possível de dados. 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.

Da mesma forma, os registos de 5 pacientes, embora seja obviamente mais vantajoso do que o registo de apenas 1 paciente, não é suficiente para esta parte. Uma vez mais, 5 pacientes não oferecem muita quantidade nem variedade em relação a rotinas ou comportamentos errados, que se traduz na variedade de regras produzidas.

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 a esmagadora 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. 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 diminuíremos 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 errado e que pode ser corrigido.

Neste sentido, a pouca quantidade de voluntários que enviaram registos foi a primeira dificuldade encontrada e que nos fez alterar a estratégia: uma vez que os dados existentes não seriam suficientes para fazer o sistema de aconselhamento de uma forma tão fidedigna quanto desejável, esse objetivo foi posto de lado. Face a estas novas circunstâncias, o objetivo passou a ser analisar os dados existentes das mais variadas formas para ver que tipo de conclusões se consegue obter. Os diferentes tipos de análises feitas vão ser descritos ainda neste capítulo.

Apesar de os dados serem poucos, a parte positiva é que estavam no formato pretendido e portanto o passo seguinte era começar a tratá-los.

5.2.2 O *data set*

Os registos dos utilizadores são armazenados no telemóvel numa base de dados *sqlite*[*sqlite*] que é depois convertida para um ficheiro csv através de um *script*. Os dados recebidos contêm alguns dados que são descartados, como os dados pessoais do utilizador, registo de peso ou medicação,

que não são tão relevantes para a análise. Mesmo relativamente a algumas variáveis, como a glicemia por exemplo, só nos interessa saber o valor e a hora, e portanto todos os outros dados sobre a glicemia são apagados, como o "ID", "Tag" que é um valor opcional que pode associar a medição a um evento, como "pequeno-almoço" e "Note" que associa uma nota à medição. O mesmo é feito para as outras variáveis. Um *data set* no estado inicial em formato csv tem a seguinte estrutura:

Nome	DateTime	Value__Carbs	Value__Glucose	Value__Insulin	Exercise
Tipo	integer	integer	integer	double	double

Tabela 5.1: *Data set* original

O **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; **Value__Insulin** é a quantidade de insulina tomada pelo utilizador à hora do registo; **Exercise** corresponde ao exercício feito à hora do registo.

No entanto, os dados não estavam ainda prontos a ser utilizados e precisavam de ser pré-processados.

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 valor 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

registro faltar o valor da glicemia, não faz sentido encontrar a média da glicemia para todos os registros e colocar no registro 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 registro que até podia ter um valor de hiperglicemia estaria a ser substituído com um valor de glicemia mais normal e portanto, esse registro 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 tomada foi a de apagar todos os registros que tivessem valores de glicemia em falta ou valores anormais. Por exemplo, alguns registros tinham valores "0" ou "7", que não são valores realistas para glicemia e portanto trata-se um erro de registro. Assim garante-se que todos os registros 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 utilizarmos registros que tenham alguns valores de insulina ou hidratos de carbono a zero, então esses registros vão contribuir tanto como os outros para o resultado final, pelo que esse resultado poderá estar adulterado. Limpando um registro inteiro, qualquer que seja o valor da variável em falta ou com um valor errado, garantimos que qualquer resultado obtido numa análise será realista.

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 valem 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 registro. Essa nota poderá ser uma breve descrição feita pelo utilizador aquando de um registro. 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 registro 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". 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* era composto por 6 variáveis: "Day", "Period", "Value_Carbs", "Value_Glucose", "Value_Insulin" e "Insulin_Difference".

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 um determinado utilizador e algumas possíveis conclusões a retirar. O utilizador escolhido foi um utilizador com uma quantidade considerável de dados, de forma a tornar a análise mais fidedigna.

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. Posto isto, escolheu-se um paciente com uma grande quantidade de registos e calculou-se a média da glicose.

Média de glicose: 116

Esta média é referente a vários registos ao longo de 5 semanas e portanto, indica que o utilizador tem um bom controlo geral da glicemia. No entanto isto não é suficiente: a média dá um indicador demasiado geral sobre os valores de glicemia e não permite detetar anomalias. Isto é, apesar de 116 ser um ótimo valor para média, pode não significar necessariamente que os valores de glicemia são estáveis. Se tivermos 2 registos em que um valor é 40 e o outro 200, a média será de 120 e no entanto ambos os valores são perigosos: o primeiro representa uma hipoglicemia e o segundo uma hiperglicemia. No entanto, olhando apenas para a média diríamos que os valores eram ótimos e o utilizador teria um bom controlo da glicemia. Torna-se portanto fazer uma análise mais aprofundada.

Capítulo 6

Conclusões

Teste de citações e sua ordem [? ? ?]

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

6.1 Trabalho Futuro

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget,

consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Apêndice A

Acrónimos

BS Base Station

BSN Body Sensor Network

HTTP Hypertext Transfer Protocol

TCP Transmission Control Protocol

UDP User Datagram Protocol