Abstract

A long, long time ago...

Resumo

Há muito, muito tempo

Agradecimentos

Obrigado a todos, obrigado ...

Dedico a . . .

Conteúdo

Lista de Tabelas

Lista de Figuras

Lista de Blocos de Código

Capítulo 1

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 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.

1.1 *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: 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, embora 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.[15] 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 nao-insulino-dependente e representa cerca de 90% de todos os casos de diabetes. [12] 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 ingês *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. [9] 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. [9]

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.

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. [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. 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.

1.1.1 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. Entres estes últimos incluem-se os monitores contínuos de glicose e os glicosímetros. Como já referido anteriormente, os *smartphones* também têm utilidade, ao ter aplicaçes 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.

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. [1]

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. [3]

1.2 *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. [**?**] 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. 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. 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** - Tem como objetivo estudar conjuntos de dados e gerar modelos com base nesses dados. Depois, ao observar novos dados 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. 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. 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 classificação, 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.

No entanto, as funções pré-existentes no R não são suficientes para uma análise mais avançada, como associação, que vai ser utilizada neste trabalho. O objetivo da associação é descobrir regras escondidas nos dados de um grande *data set*, através da relação de variáveis. Embora a associação tenha sido criada com o objetivo de usar em negócios, hoje em dia é utilizada em várias outras áreas, tais como diagnóstico médico ou análise de dados científicos. [20] 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. [16] 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. 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. [2] 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 $\{X, Z\}$. [2] 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.

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*. Um *package* é um conjunto de funções e/ou dados, que pode ser feito por qualquer pessoa e pode ser adicionado ao R. [7] Ao adicionar um novo *package* ao R, todas as suas funções passam a ficar disponíveis para utilização, bastando para isso carregar a extensão em cada sessão. Existe uma rede de servidores *web* e ftp que guardam versões atualizadas de código e documentação para o R, chamada CRAN. O CRAN ("Comprehensive R Archive Network") tem atualmente mais de 8000 *packages* disponíveis para *download*, tornando o R altamente personalizável e poderoso. [4] 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.

Depois da associação de regras, iremos fazer ainda um outro tipo de análise que envolve *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. [18] Não ter ciclos significa que é impossível ter um vértice com um caminho que começa e acaba nesse vértice. 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 "Cancro do Pulmão" e outras variáveis como "Fumador", "Amianto" ou "Arsénio", é possível calcular a probabilidade de uma pessoa ter cancro do pulmão sabendo que é fumadora ou que esteve continuamente exposta a amianto. Também é possível saber, através de uma rede deste tipo, qual 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.

Esta e outras eventuais análises requerem ferramentas que não estão disponíveis no R, nem sequer com *packages* e por isso será necessário usar *software* adicional. No capítulo 5 esses programas utilizados serão enumerados em cada secção da respetiva análise e serão também explicadas as razões que levaram à sua escolha, e não a outro qualquer.

1.2.1 *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, 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, pode 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 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. 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.

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 2

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* 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. [13]

2.1 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.

Em [11], 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 [17], 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 varíaveis, 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.

Em [19], 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 [14], os autores 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. [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
```

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.

2.2 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. [10] 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 [6] 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.

2.2.1 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 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.

2.2.2 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 glicómetros de diferentes modelos, tais como OneTouch, Dexcom ou Accu-Chek.

2.2.3 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.

2.2.4 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*.

2.2.5 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.

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] Apriori. https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/algo_apriori.htm. [Online; acedido a 17 de Junho de 2016].

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

[4] The Comprehensive R Archive Network. https://cran.r-project.org/. [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] 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].

[7] R Packages. http://www.statmethods.net/interface/packages.html. [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] Type 1.5 diabetes. http://www.diabetes.co.uk/type15-diabetes.html. [Online; acedido a 17 de Junho de 2016].

[10] 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].

[11] Dursun Delen, Glenn Walker, and Amit Kadam. Predicting breast cancer survivability: a comparison of three data mining methods, 2004.

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

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

[14] 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. In *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19*, 2015.

[15] David M Maahs, Nancy A West, Jean M. Lawrence, and Elizabeth J Mayer-Davis. Chapter 1: Epidemiology of Type 1 Diabetes, Setembro 2011.

[16] Viktor Mayer-Schonberger and Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think. chapter 4, page 118. 2013.

[17] 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.

[18] Fabrizio Ruggeri, Ron S. Kennet, and Frederick W. Faltin. Encyclopedia of Statistics in Quality and Reliability. page 490. 2007.

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

[20] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining. chapter 6, page 328. 2008.

Apêndice A

Acrónimos

BS Base Station

BSN Body Sensor Network

HTTP Hypertext Transfer Protocol

TCP Transmission Control Protocol

UDP User Datagram Protocol