

# Algoritmos de Mineração de Dados em Sistema de Monitoramento de Diabetes

Robson Cezário<sup>1</sup>, Alexandre de O. Zamberlan<sup>1</sup>

<sup>1</sup>Curso de Sistemas de Informação – Universidade Franciscana (UFN)  
Santa Maria – RS

<sup>2</sup>Laboratório de Práticas – Universidade Franciscana (UFN)  
Santa Maria – RS

{robson.cezario, alexz}@ufn.edu.br

**Resumo.** *Este trabalho está no contexto de extração e descoberta de conhecimento em bases de dados referentes às pessoas com diabetes. A base contém dados de monitoramento de data, hora e quantidade de insulina utilizada, quantidade de calorias e carboidratos ingeridos naquela data, qualidade do sono (escala Likert) e tempo em minutos de atividades físicas. Contudo, uma base com muitos dados, há dificuldade de se encontrar padrões de forma visual, mesmo que por relatórios. Portanto, esta pesquisa busca estudar, aplicar, avaliar e apontar métodos de mineração de dados adequados para esse contexto. A pesquisa é baseada em revisão bibliográfica e estudo de caso, em que o trabalho de [Rath et al. 2014] é estendido e integrado com uma base de três (3) anos de dados coletados em tabela de planilha eletrônica. E por meio dessa nova base, os algoritmos de mineração devem ser avaliados.*

**Abstract.** *This work is in the context of discovering knowledge in databases of people with diabetes. The database contains data on the date, time and amount of insulin used, amount of calories and carbohydrates ingested on that date, sleep quality (Likert scale) and time in minutes of physical activities. However, in a database with a lot of data, it is difficult to find patterns visually, even through reports. Therefore, this research intends to study, apply, evaluate and point out suitable data mining methods in this context. The research is based on literature review and case study, in which the work of [Rath et al. 2014] will be extended and integrated with a 3-year database of data collected in a spreadsheet table. And through this new data base, mining algorithms must be evaluated.*

## 1. Introdução

A área da Saúde produz muitos e variados dados, sejam eles na Farmácia, Medicina, Fisioterapia, Enfermagem, entre outros. Tratamento de pacientes com diabetes, por exemplo, também acabam gerando uma quantidade significativa e diversificada de dados, principalmente, quando pacientes, médicos, farmacêuticos e nutricionistas monitoram a evolução (ou controle) da doença. Esse controle pode ser via monitoramento de valores diários de unidades de insulina aplicada, valores diários de glicemia em jejum, quantidade de calorias e carboidratos consumidos, tempo e quantidade de exercícios realizados, entre outros. Entretanto, muitos e diferentes dados dificultam a análise e o reconhecimento de padrões

que possam estar embutidos (até mesmo ocultos) numa base de dados. Assim, médicos, nutricionistas e até educadores físicos podem elaborar estratégias falhas, ou incompletas, pois a análise acaba sendo superficial.

### **1.1. Justificativa**

Em 2014, um aluno do curso de Sistemas de Informação da UFN [Rath et al. 2014] projetou e implementou um Sistema de Recomendação para Diabetes, construído via a linguagem PHP e banco de dados MySQL. O projeto disponibilizou um sistema web para que pacientes diabéticos, seus médicos e nutricionistas pudessem registrar dados de insulina, glicemia, calorias, carboidratos, exercícios, qualidade de sono, etc. Porém, o projeto aplicou parcamente técnicas de mineração para descoberta de padrões. O projeto trabalhou com mineração, mas não era o foco principal, ou seja, o foco era o sistema web para registro e acompanhamento de dados de pessoas diabéticas. A mineração foi prototipada mais em um sentido de criar algo funcional com o mínimo possível para ser analisado e testado. Dessa forma, este trabalho pretende estender a pesquisa realizada em [Rath et al. 2014], principalmente na análise e aplicação de melhores técnicas de mineração de dados na base projetada para diabetes. Registra-se que há uma base de dados criada e populada entre os anos 2012 e 2014 (independente do sistema projetado em [Rath et al. 2014]), com 722 dias monitorados. E essa base será utilizada para o estudo, aplicação e avaliação das diferentes técnicas de mineração.

### **1.2. Objetivos**

O objetivo inicial é estudar, comparar, aplicar e avaliar algoritmos presentes no ambiente WEKA e em pacotes do universo Python, por exemplo, PyBrain para reconhecer precisamente padrões presentes na base de dados do estudo.

Objetivos específicos são: i) entender e aplicar as categorias de algoritmos de mineração de dados aos diferentes contextos [Tan et al. 2009], [Marques et al. 2008]; ii) estudar e avaliar algoritmos presentes no ambiente WEKA [Group 2021] e em pacotes do universo Python [Schaul et al. 2010]; iii) realizar adequação da base de dados do sistema projetado por [Rath et al. 2014]; iv) mapear e compilar trabalhos relacionados que usaram técnicas de mineração em bases de dados [Marques et al. 2008], [Vieira 2016]; v) definir, aplicar, analisar um estudo de caso para validar a proposta.

## **2. Revisão Bibliográfica**

Nesta seção, são apresentados e discutidos conceitos que fundamentam a pesquisa e fornecem entendimento à proposta do trabalho.

### **2.1. Descoberta de Conhecimento e Mineração de Dados**

As áreas de conhecimento produzem, diariamente, uma quantidade de dados e de informação muito grande. Esses dados e informações precisam ser coletados, tratados e armazenados, produzindo assim uma base de dados e/ou conhecimento específico.

Por exemplo, no campo da Medicina, dados coletados não são o suficiente para uma tomada de decisão precisa e eficiente. Assim, para se ter uma decisão mais assertiva, são necessárias ferramentas que facilitam e auxiliam a verificação/análise desses dados. Logo, conforme Prieto et. al. 2004 *apud* [Rath et al. 2014], as boas tomadas de decisão

em relação aos quadros clínicos de pacientes diabéticos, por exemplo, influenciam diretamente no tratamento e na prevenção da doença.

Conforme Prieto et. al. 2004 *apud* [Rath et al. 2014], a análise de dados nas áreas da Saúde tem sido realizada por meio estatístico, que usa um processo matemático estabelecido com suporte teórico que permite interpretações. No entanto, há um método alternativo que auxilia em análises e interpretações de grandes quantidades de dados, conhecido como Mineração de Dados (*Data Mining* - DM). Esse método, por meio do uso ou não da estatística, busca encontrar modelos ou padrões ocultos dentro de uma base, que dificilmente seriam detectados com técnicas estatísticas.

De acordo com o trabalho de [Rath et al. 2014], alguns autores consideram os termos Mineração de Dados e Descoberta de Conhecimento em banco de dados (*Knowledge-Discovery in Databases* - KDD) como processos diferentes. De fato, mineração faz parte da linha de descoberta de conhecimento. Entretanto, neste trabalho, assume-se que os termos têm o mesmo significado: extrair conhecimento de dados.

A definição aceita e citada por vários autores sobre mineração, segundo [Rezende 2003], é: “extração de conhecimento de base de dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados”.

Há variações na quantidade de etapas que compõe o processo de mineração, mas todas as etapas tratam de algumas características, como [Rezende 2003]:

- Dados: elementos armazenados em um repositório;
- Padrões: um subconjunto de dados em alguma linguagem descritiva que aponta para um tema comum;
- Processo: qualquer uma das etapas referentes à extração de conhecimento, como preparação de dados, busca por padrões e avaliação do conhecimento;
- Válidos: são os padrões encontrados obedecendo regras e/ou princípios e que sejam admissíveis;
- Novos: um padrão com um conjunto de informações definindo um novo padrão;
- Úteis: algum padrão passível de uso ou aproveitamento;
- Compreensíveis: qualquer padrão descoberto e registrado em alguma linguagem que possa ser entendida por usuários, possibilitando uma análise mais adequada dos dados;
- Conhecimento: é definido de acordo com o seu escopo de aplicação, utilidade, originalidade e compreensão, ou seja, é um conjunto de informações que fornecem algum tipo de contexto para processos de raciocínio ou inferência.

No trabalho de [Rath et al. 2014], a mineração foi dividida em 3 etapas: coleta de dados, processamento e identificação de padrões, e pós processamento.

### **3. Algoritmos clássicos para mineração de dados**

De acordo com [Larose 2005], mineração de dados é classificada conforme as tarefas realizadas, sendo as mais comuns:

- Descrição: tarefa que descreve os padrões descobertos e disponibiliza uma possível interpretação dos resultados, via técnicas exploratórias de dados;

- Classificação: tem como objetivo detectar a qual classe uma determinada informação pertence, como um processo de categorização, geralmente realizada por processo de aprendizado supervisionado;
- Regressão: semelhante à classificação, mas é utilizada quando a informação é identificada por um valor numérico e não categórico;
- Predição: também semelhante à classificação e regressão, no entanto busca descobrir o valor futuro de um determinado atributo;
- Agrupamento: identifica e aproxima informações semelhantes. Um agrupamento é um conjunto de informações similares entre si, mas diferentes de outras informações nos demais agrupamentos;
- Associação: identifica a relação entre atributos de um conjunto de informações.

Segundo o trabalho realizado em [Furlan and de Souza Poletto 2018], a mineração de dados possui várias implementações distintas por meio de diversos algoritmos. Esses algoritmos são segmentados, novamente, pelas tarefas. A Figura 1 mostra uma relação dos principais algoritmos, suas descrições, tarefas e exemplos.

Técnica	Descrição	Tarefas	Exemplos
Árvore de Decisão	Baseada em estágios de decisão (nós) e na separação de classes e subconjuntos, organiza os dados de forma hierárquica.	- Classificação - Predição	CART, CHAID, C5.0, ID-3
Redes Neurais	Modelos inspirados na fisiologia do cérebro, nos quais o conhecimento é fruto do mapa de conexões neuronais e dos pesos dessas conexões.	- Classificação - Agrupamento - Predição	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB.
Raciocínio Baseado em Casos	Baseado no método do vizinho mais próximo combina e compara atributos para estabelecer hierarquia de semelhança.	- Classificação - Agrupamento	BIRCH, CLARANS CLIQUE
Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, em que a cada nova geração, soluções melhores têm mais chance de ter "descendentes".	- Classificação - Agrupamento	Algoritmo Genético Simples, Genitor, GA-Nuggets, GAPVMINER
Conjuntos Fuzzy	Oferece uma grande vantagem para classificar dados com um alto nível de abstração.	- Classificação - Agrupamento	K-means, FCMdd
Regras de Indução	Processo para obter uma hipótese a partir de dados e fatos já existentes.	- Classificação - Predição	CART, CHAID
Regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados.	- Associação	Apriori, AprioriTid, AprioriHybrid, AIS, SETM

**Figura 1. Técnicas e Tarefas empregadas na Mineração de Dados [Goldschmidt 2005].**

### 3.0.1. Ambientes ou Ferramentas para Mineração de Dados

Um dos ambientes mais conhecidos e utilizados em instituições de ensino e de pesquisa, no contexto de descoberta de conhecimento e mineração de dados, é o *Waikato Environment for Knowledge Analysis* (WEKA). É uma ferramenta de código aberto, com licença *General Public Licence* (GPL), multiplataforma (Windows, MAC OS e Linux), contendo inúmeras funções, como por exemplo, a importação de banco de dados em arquivo para mineração (.arff). Além disso, possui diversos algoritmos de mineração que podem ser aplicados, testados e analisados em diferentes bases importadas.

O Weka, enquanto ferramenta, possui ambientes de importação, de visualização de resultados e de algoritmos de mineração, como algoritmos de redes neurais para aprendizado de máquina (*machine learning*), entre outros.

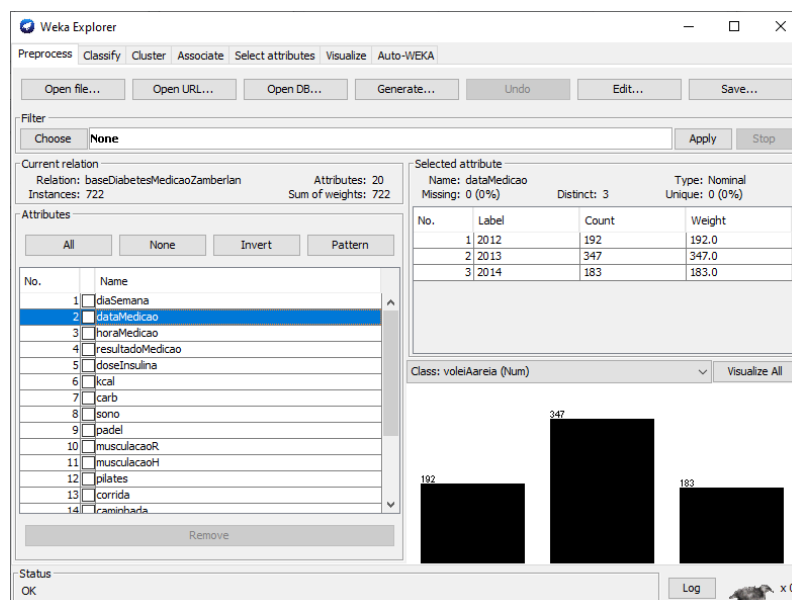
No ambiente de algoritmos para mineração, há sub-ferramentas de pré-processamento, oferecendo suporte para todo processo de mineração, incluindo a preparação dos dados de entrada, avaliação estatística de aprendizagem, visualização dos dados de entrada e seus resultados.

A ferramenta WEKA foi desenvolvida na linguagem JAVA, em que é possível tanto usar uma interface gráfica específica dela, quanto importar pacotes WEKA, contendo os algoritmos de mineração, em aplicações JAVA.

WEKA possui funcionalidades divididas em ambientes, que são chamadas em botões na interface principal, como:

- *Explorer*: possui aplicações de pré-processamento, análise e visualização dos resultados;
- *Experimenter*: aqui, o usuário pode realizar testes estatísticos entre as estruturas de aprendizagem da ferramenta, onde se pode utilizar vários algoritmos concomitantes e comparar os resultados, escolhendo assim o algoritmo mais adequado para a base;
- *Knowledge Flow*: com funções similares ao *Explorer*, diferencia-se pela representação gráfica dos resultados;
- *Workbench*: ambiente que combina todas as interfaces gráficas em uma única interface;
- *Simple CLI*: ambiente para inserção de comandos em forma de *shell* ou terminal.

Por exemplo, após o banco de dados ser carregado é possível realizar a edição para uma eventual correção e filtrar pelos atributos conforme ilustrado na Figura 2.



**Figura 2. Preprocess - WEKA Explorer com a base de dados carregada.**

Por outro lado, o universo da linguagem Python, via bibliotecas, tem apresentado muitos recursos para os processos de descoberta de conhecimento e mineração de dados. As bibliotecas que mais tiveram destaque são:

- Pandas: Fornece ferramentas de análise e manipulação de dados (de forma rápida e flexível) em código aberto. Pandas é adequado para diferentes tipos de dados, tais como uma tabela do padrão SQL (banco de dados relacional), planilha eletrônica, dados temporais, dados de matriz e qualquer forma de conjuntos de dados estatísticos [The pandas development team 2021];
- PyTables/HDF5: quando se tem uma quantidade muito grande de dados é utilizado o pacote Pytables (Desenvolvido com base na biblioteca HDF5) que gerencia de forma eficaz conjuntos de dados hierárquicos [PyTables Developers Team 2021];
- Theano: utilizado para determinar, incrementar e avaliar expressões matemáticas em CPUs/GPUs que envolvem matrizes multidimensionais [Laboratório LISA, Universidade de Montreal 2021];
- SeaBorn: prepara uma interface de alto nível que mostra gráficos estatísticos, interativos e informativos [Waskom 2021];
- Airflow: plataforma para criar, agendar e monitorar fluxos de trabalho [Apache Airflow 2021];
- Pybrain: é uma biblioteca de aprendizado de máquina exclusiva para Python, com uma proposta de interface amigável, com algoritmos e diferentes ambientes de teste desses algoritmos. Destaca-se, que PyBrain é uma biblioteca de aprendizagem por reforço via Redes Neurais [Schaul et al. 2010];
- Scikit-learn: possui algoritmos de classificação, regressão e agrupamento [Pedregosa et al. 2011];
- Keras: tem um interface simples para resolver problemas de aprendizado de máquina e auxilia no aproveitamento máximo de escalabilidade com redes neurais [Grupo de Interesse Especial Keras (Keras SIG) 2021];
- Dask: Utilizado para computação paralela e composto por coleções de "Big Data" e Agendamento de tarefas dinâmico [Rocklin 2015].

### 3.1. Trabalhos Relacionados

Neste seção, buscou-se discutir trabalhos que abordaram ou mineração de dados e/ou mineração de dados aplicada à base de dados na área da Saúde.

O primeiro trabalho é a base para o estudo de caso desta pesquisa, onde [Rath et al. 2014] desenvolveu um sistema Web, na linguagem PHP, para cadastro e controle diário de dados do diabético, que são armazenados em banco de dados MySQL. O trabalho utilizou algoritmos presentes no ambiente WEKA de mineração.

No trabalho realizado em [Marques et al. 2008], foram aplicadas técnicas de mineração de dados para um sistema de apoio à tomada de decisão no contexto de monitoramento de jogadas em partidas de Voleibol. O sistema é conhecido como *scout*, com apelo estatístico, que possui informações muitas vezes desnecessárias, que na mineração de dados são descartadas. Dessa forma, o processo de mineração no *scout* de Voleibol foi útil para destacar informações confiáveis e algumas até desconhecidas, assim ajudando de maneira mais eficaz as comissões técnicas. O sistema foi construído na linguagem de programação Java e com o apoio do ambiente WEKA.

Segundo o que foi apresentado em [Furlan and de Souza Poletto 2018], um estudo foi realizado para aplicar conceitos de descoberta de conhecimento e de mineração em banco de dados. Alguns algoritmos do ambiente WEKA para visualização do processo foram testados.

### 3.1.1. Análise dos Trabalhos Relacionados

No trabalho apresentado em [Rath et al. 2014], havia o objetivo de aplicar técnicas de mineração de dados para encontrar padrões na sua base, para poder recomendar alimentos e atividades físicas aos diabéticos. No trabalho, foi projetado e implementado um sistema Web para isso. Porém, essa recomendação não foi totalmente finalizada, uma vez que a base de dados não tinha um conjunto de dados significativo para que algoritmos de mineração fossem utilizados. Dessa forma, a principal justificativa deste trabalho. No trabalho de [Marques et al. 2008], destaca-se o processo de limpeza da base, a revisão bibliográfica das categorias dos diferentes algoritmos de mineração, facilitando a escolha de algoritmos para esta pesquisa. Já no trabalho de [Furlan and de Souza Poletto 2018], foi mostrada a análise de técnicas de algoritmos de mineração de dados presente no ambiente WEKA. Também compilou informações sobre mineração de dados, descoberta de conhecimento. Dessa forma, ficou evidente o uso da ferramenta WEKA de forma simplificada.

### 3.2. Monitoramento de Diabetes

De acordo com a Sociedade Brasileira de Diabetes [SBD 2021], a diabetes *mellitus* é uma doença crônica em que o corpo não é capaz de produzir ou absorver corretamente a insulina gerada pelo pâncreas. A insulina é o hormônio responsável pelo controle da quantidade de glicose no sangue que um ser humano recebe na alimentação como fonte de energia para o organismo [SBD 2021].

Novamente conforme a SBD, quando o ser humano tem diabetes e caso não for controlada de forma adequada por longos períodos, podem aparecer complicações, como doenças renais, infarto do miocárdio, acidente vascular cerebral, pé diabético, glaucoma, catarata, entre outros [SBD 2021]. De acordo com o portal da secretaria de atenção primária a saúde estima-se que no Brasil 9 milhões de pessoas com diabetes (<https://aps.saude.gov.br/noticia/10336>).

Cabe ressaltar que o tratamento com insulina, busca, de forma artificial e externa ao corpo, a regulação desse hormônio, para que cumpra seu papel no organismo. Sendo assim, todos portadores de diabetes Tipo 1 e Tipo 2 podem necessitar de insulina para controlar a glicose no sangue. Para essa finalidade, existem vários tipos de insulina disponíveis para o tratamento. E tal diferença se dá pelo tempo que ficam ativas no corpo, pelo tempo que levam para agir e em qual situação do dia são mais eficientes.

Independente do tipo de diabetes ou o tipo de terapia ou tratamento, o monitoramento é essencial para o controle, para o tratamento, para o acompanhamento dos profissionais Médicos, Nutricionistas e Educadores Físicos. Há profissionais da área da Saúde que monitoram seus pacientes com exames de sangue tradicionais, realizados de tempos em tempos, como glicemia em jejum, glicemia glicada, insulina, triglicerídeos, colesterol, etc. Porém, há outros tipos de acompanhamento, ditos diários: glicemia, unidades aplicadas de insulina, qualidade do sono, quantidade de calorias e carboidratos ingeridos, tempo e tipo de atividade física.

## 4. Proposta de trabalho

A partir do exposto na seção anterior, esta pesquisa busca estender o trabalho [Rath et al. 2014] (via estudo de caso) e aplicar algoritmos de mineração em base com

um conjunto significativo de dados para que os algoritmos possam retornar resultados reais e mais fidedignos.

#### 4.1. Materiais e métodos

Este trabalho é baseado em pesquisa exploratória com revisão bibliográfica amparado com estudo de caso. O estudo de caso é aplicação de mineração na base de dados criada entre 2012 e 2014. Já no projeto e desenvolvimento da solução, são utilizados a metodologia *Scrum* [Wykowski and Wykowska 2019] com a técnica *Kanban*.

As ferramentas a serem utilizadas são: Trello - técnica kanban; GitHub - controle de versão; Astah - diagramação UML; Ambiente WEKA - simulador para testes de algoritmos de mineração de dados; Pacotes do universo Python para mineração em banco de dados.

#### 4.2. Modelagem do sistema

Em termos funcionais da proposta do estudo de caso, que é um sistema Web adaptado do trabalho de [Rath et al. 2014], na Figura 3 é possível entender todos os atores que podem ter relação com o sistema, suas principais funcionalidades e a funcionalidade de aplicação de algoritmos de descoberta de conhecimento.

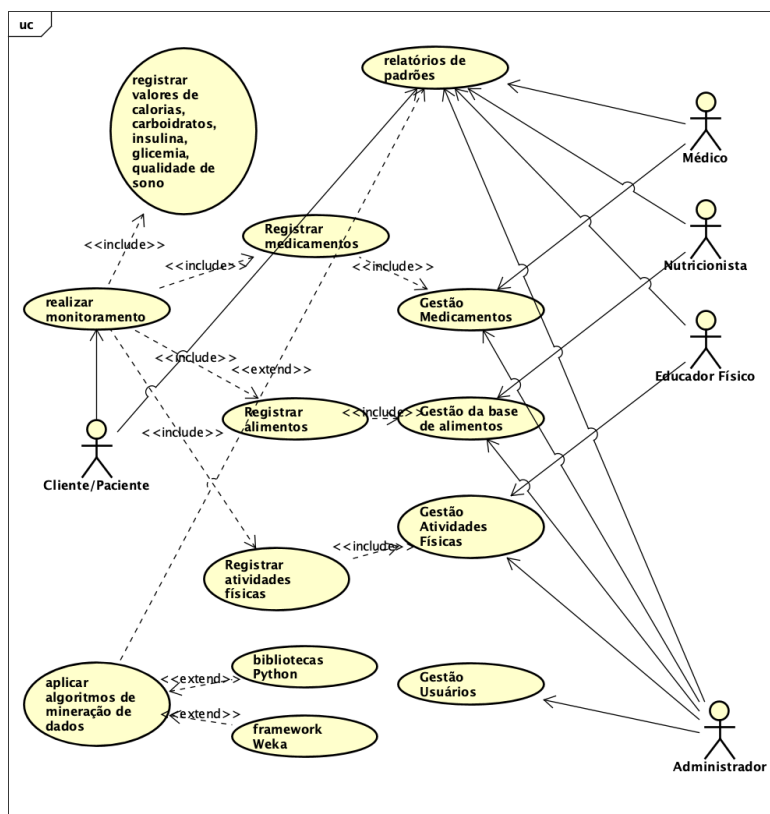
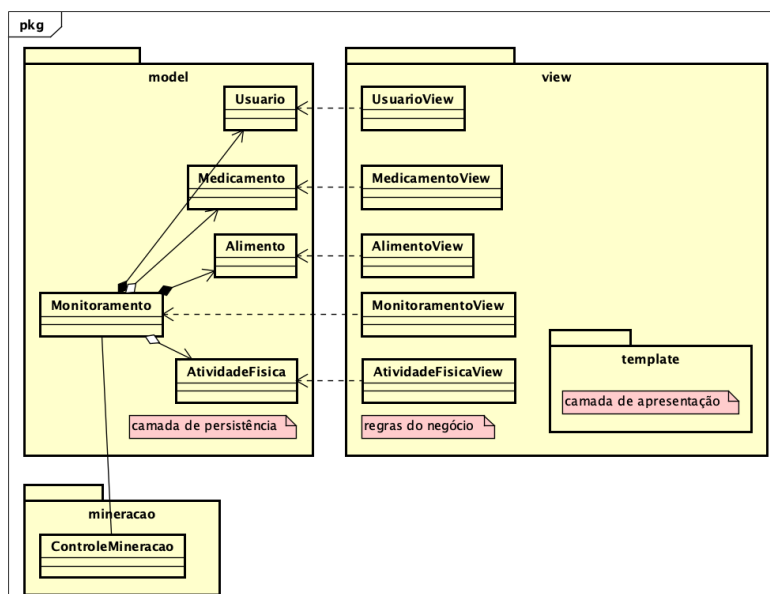


Figura 3. Diagrama de Casos de Uso para o estudo de caso adaptado do sistema de [Rath et al. 2014].

Já em relação à organização do protótipo do estudo de caso, inicialmente, pensou-se orientado a objetos tendo como referência o padrão arquitetural *Model-View-Template*



(MVT) utilizado pelo *framework* Django. Na Figura 4, é possível visualizar os pacotes e as classes que fazem a persistência no banco de dados na camada *model*.



**Figura 4. Diagrama de pacotes ou de domínio do protótipo.**

Registra-se que o *framework* Django utiliza de forma automática o Mapeamento Objeto-Relacional (MOR), ou seja, a geração do banco de dados é realizada a partir das classes existentes na camada *model* do sistema. Assim, a Figura 4 mostra a estrutura de pacotes/classes e, de forma equivalente, o modelo Entidade-Relacionamento.

### 4.3. Preparando a base para a mineração

A Figura 5 mostra como os dados monitorados (dias da semana, ano, glicemia, quantidade de insulina, quantidade calorias, quantidade de carboidratos, qualidade do sono, tempo de atividades físicas) foram coletados. A ressalva é que a média de glicemia era analisada sempre de 3 em 3 dias, pois essa média sugeria alteração nas unidades de insulina aplicada. Ou seja, se a média fosse maior que 100 mg/dL, a quantidade de unidades de insulina deveria ser aumentada em duas. Caso fosse menor que 80 mg/dL deveria diminuir em 2 unidades.

Destaca-se que a plataforma WEKA precisa receber uma base tratada, preferencialmente com valores categóricos, assim foi preciso melhorar alguns campos (colunas ou atributos). Dessa forma, foram realizadas conversões de valores numéricos para valores categóricos, principalmente nos campos quantidade de exercícios realizados, quantidade de carboidratos, quantidade de calorias ingeridos em um dia e resultado da medição da glicemia.

Em relação à quantidade de exercícios, optou-se em utilizar a escala (ou tabela) de Borg (de 0 a 20), que trabalha com uma percepção subjetiva de quanto esforço o próprio atleta realizou, sendo que 1 é quase nenhum esforço e 20 o esforço máximo possível [Borg and Borg 2001]. De acordo com [], os valores entre 6 e 20 são baseados na Frequência Cardíaca de 60 a 200 batimentos por minuto (bpms), sendo que o valor

12 corresponde aproximadamente 55% e o 16 a 85% da Frequência Cardíaca Máxima. Dessa forma, a escala corresponderia ao esforço da seguinte forma:

- 7 - muito fácil;
- 9 - fácil;
- 11 - relativamente fácil;
- 13 - ligeiramente cansativo;
- 15 - cansativo;
- 17 - muito cansativo;
- 19 - exaustivo.

Dia Semana	Data	Antes Comer / Depois Comer	Resultado	Dose Insulina	kcal	carb	noite de sono 1-5	padel	musculação R	musculação H	pilates	corrida	caminhada	tênis	sauna	bike	natação	elíptico	volei de areia
Quinta	2013	ac	99	6	Recomendado	Recomendado	3	11	0	0	0	0	0	0	0	0	0	0	0
Sexta	2013	ac	97	6	Recomendado	Acima	3	0	0	0	0	0	0	0	0	0	9	0	0
Sábado	2013	ac	98	6	Recomendado	Acima	3	0	0	0	0	0	0	0	0	0	0	0	0
Domingo	2013	ac	98	6	Recomendado	Acima	3	0	0	0	0	0	0	0	15	0	0	0	0
Segunda	2013	ac	92	6	Recomendado	Acima	4	15	0	0	0	0	0	0	0	0	0	0	0
Terça	2013	ac	106	6	Abaixo	Recomendado	3	0	0	0	0	0	0	0	0	0	0	0	0
Quarta	2013	ac	100	6	Recomendado	Acima	4	17	11	0	0	0	0	0	0	0	0	0	0
Quinta	2013	ac	91	6	Recomendado	Recomendado	4	0	0	0	0	0	0	0	0	0	0	0	0
Sexta	2013	ac	89	6	Recomendado	Acima	3	11	0	0	13	0	0	0	0	0	0	0	0
Sábado	2013	ac	100	6	Recomendado	Recomendado	4	0	0	0	0	0	0	0	15	0	0	0	0
Domingo	2013	ac	92	6	Recomendado	Acima	4	0	0	0	0	0	0	0	15	0	0	0	0
Segunda	2013	ac	95	6	Recomendado	Recomendado	4	15	0	0	0	0	0	0	0	0	0	0	0

**Figura 5. Planilha eletrônica com melhorias nas colunas e nos dados, para facilitar a conversão ao arquivo .arff (arquivo de mineração no WEKA).**

Em relação ao consumo de carboidratos e calorias, também foram necessárias modificações na tabela original, transformando dados numéricos em dados categóricos, como "abaixo", "recomendado" e "acima" do recomendado. Dessa forma, se uma pessoa tem como orientação (referência), na sua dieta, ingerir no máximo 200 gramas carboidratos e 2600 calorias, o abaixo do recomendado é quando essa pessoa ingere menos de 80% desses valores de referência, recomendado quando ingere entre 80% a 120% dos valores e acima do recomendado quando consome mais de 120% dos valores sugeridos na dieta. Essas orientações são estratégias básicas no contexto da nutrição []. Por fim, em relação à resultados de medição de glicemia há as categorias "abaixo", "recomendado" e "acima" do recomendado. Porém, há dois momentos de medição: em jejum e pós-prandial (após 2 horas de uma refeição). Dessa forma, para medições em jejum, o abaixo do recomendado é um valor menor que 80 mg/dL (miligramas por decilitro), recomendado é um valor entre 80 mg/dL e 100 mg/dL e acima do recomendado um valor acima de 100 mg/dL. Para medições pós-prandial, o recomendado é um valor abaixo de 140 mg/dL e acima, um valor maior que 140 mg/dL [].

Uma vez melhorada a tabela, já é possível no ambiente WEKA importar a base da planilha eletrônica. A importação (automática) converte a tabela em um arquivo com extensão .arff, com os campos (colunas) para sua manipulação interna, conforme mostra a Figura 6.

Ressalta-se que @attribute representa cada coluna da planilha eletrônica. Alguns atributos precisam ser valorados, como por exemplo, o atributo diaSemana que contém os dias da semana. Finalmente, a partir da linha 25, há os dados da planilha convertidos para a plataforma WEKA. Por exemplo, na linha 25 há uma medição referente a uma quinta-feira, do ano de 2012, com uma medição antes de comer, com 90 mg/dL de glicemia, 6 unidades de insulina aplicada, 2037 calorias, 246 gramas de carboidratos, 4 pontos na qualidade do sono e nenhuma atividade física realizada.

```

1 @attribute dia {Quinta,Sexta,Sabado,Domingo,Segunda,Terca,Quarta}
2 @attribute ano numeric
3 @attribute 'hora comida' {ac}
4 @attribute glicemia numeric
5 @attribute 'insulina' numeric
6 @attribute kcal {Abaixo,Recomendado,Acima}
7 @attribute 'carb' {Acima,Recomendado,Abaixo}
8 @attribute sono numeric
9 @attribute padel numeric
10 @attribute musculacaoR numeric
11 @attribute musculacaoH numeric
12 @attribute pilates numeric
13 @attribute corrida numeric
14 @attribute caminhada numeric
15 @attribute tenis numeric
16 @attribute sauna numeric
17 @attribute bike numeric
18 @attribute natacao numeric
19 @attribute eliptico numeric
20 @attribute voleiareaia numeric|
21 @data
22 Quinta,2012,ac,90,6,Abaixo,Acima,4,0,0,0,0,0,0,0,0,0,0,0,0
23 Sexta,2012,ac,96,6,Recomendado,Acima,4,0,13,0,0,11,0,0,0,0,0,0,0
24 Sabado,2012,ac,90,6,Recomendado,Acima,4,17,0,0,0,0,0,0,0,0,0,0,0
25 Domingo,2012,ac,105,6,Abaixo,Acima,5,0,0,0,0,0,0,0,0,0,0,0,0
26 Segunda,2012,ac,86,6,Recomendado,Acima,5,17,0,0,0,0,0,0,0,0,0,0,0
27 Terca,2012,ac,86,6,Recomendado,Acima,4,0,0,0,0,0,0,0,0,0,0,0,0
28 Quarta,2012,ac,97,6,Recomendado,Acima,5,0,0,0,0,0,0,13,0,0,0,0,0

```

**Figura 6. Dados tratador pela plataforma WEKA e armazenados em arquivo .arff.**

#### 4.4. Aplicando os algoritmos de mineração via WEKA

Uma vez que a base foi preparada para ser trabalhada no WEKA, testaram-se todos os algoritmos presentes no ambiente que ficaram disponíveis para a base cadastrada. Ou seja, alguns algoritmos não foram habilitados à execução dessa base.

Dois algoritmos apresentaram respostas significativas:

- Agrupamento ou clusterização pelo algoritmo *Expectation Maximisation* (EM): agrupou a base em 3 grupos;
- Classificação pelo algoritmo *Random Tree* (árvore de decisão): XXXXXX

Os resultados parciais obtidos pelo algoritmo EM podem ser visualizados nas Figuras 7 e 8.

E com isso, foi possível concluir que:

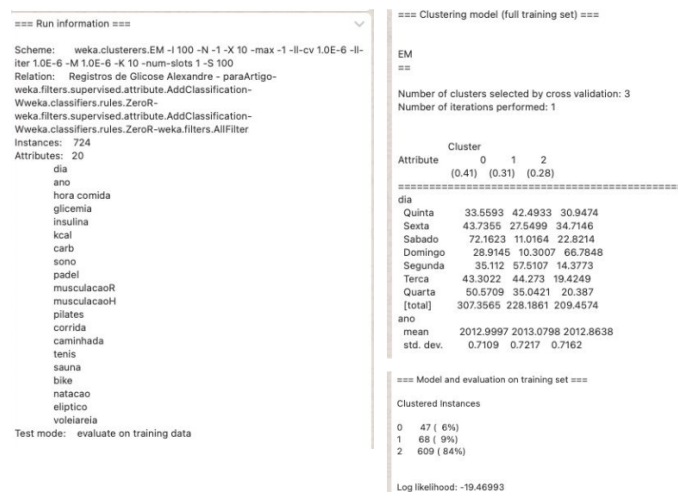
- exercícios com carga muscular elevada ajudam mais no controle da glicemia, logo na aplicação de insulina;
- a variável sono não tem influência na quantidade de glicemia;
- a quantidade elevada consumida de calorias influenciou diretamente na quantidade elevada consumida de carboidratos;

Já os resultados obtidos pelo algoritmo Random Tree foram:

IMAGENS DE RESULTADO VIA RANDOM TREE:

E com isso, foi possível concluir:

- .....



**Figura 7. Aplicação do algoritmo EM e o agrupamento da base em 3 grupos.**

insulina					
mean	8.6701	8.8879	7.8633		
std. dev.	2.1696	2.1817	2.008		
kcal					
Abaixo	23.6208	125.8603	18.5189		
Recomendado	247.7823	88.6168	179.601		
Acima	31.9535	9.709	7.3375		
[total]	303.3565	224.1861	205.4574		
carb					
Acima	278.2237	65.4183	73.3579		
Recomendado	19.3826	136.4682	124.1492		
Abaixo	5.7501	22.2996	7.9503		
[total]	303.3565	224.1861	205.4574		
sono					
mean	3.9058	3.9301	3.9342		
std. dev.	0.8322	0.794	0.7268		
padel					
mean	4.7396	3.5937	3.7665		
std. dev.	7.4001	6.7721	6.7933		
musculacaoR					
mean	0.7257	0.4707	0.2168		
std. dev.	2.8742	2.3787	1.6401		
musculacaoH					
mean	1.1779	1.5921	0.0447		
std. dev.	3.7317	4.2618	0.7609		
pilates					
mean	0.355	0.9523	0.6508		
std. dev.	2.107	3.3746	2.8065		
corrida					
mean	0.5627	0	0		
std. dev.	2.6654	1.7402	0.0001		
caminhada					
mean	1.1118	0.1077	0.2332		
std. dev.	3.0507	0.982	1.4899		
tenis					
mean	1.2897	0.5757	3.0392		
std. dev.	4.002	2.8146	7.0092		
sauna					
mean	0.3864	0.176	0		
std. dev.	2.3401	1.5458	0.0001		
bike					
mean	1.7515	0.1572	0.5787		
std. dev.	4.5145	1.1813	2.5137		

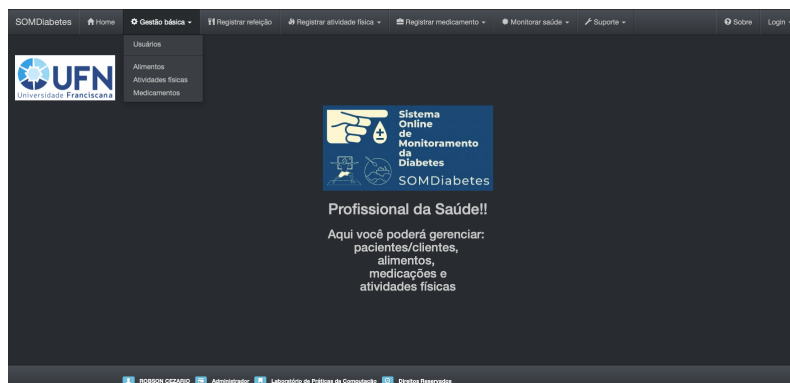
**Figura 8. Aplicação do algoritmo EM e o agrupamento da base em 3 grupos: dados insulina, carboidrato, calorias, atividades físicas.**

#### 4.5. SOMDiabetes - Sistema Online de Monitoramento da Diabetes

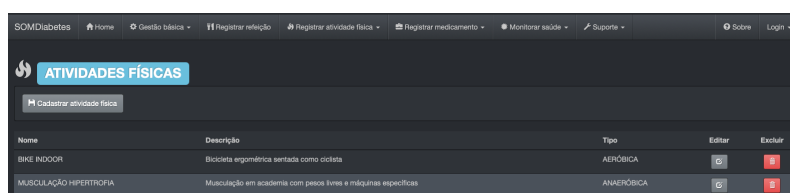
O protótipo do estudo de caso, já possui um conjunto de funcionalidades básicas de gestão de usuários (pacientes, nutricionistas, educadores físicos, médicos), de alimentos e de atividades físicas. Além disso, há toda a gestão de registro de refeições por paciente em um determinado dia, com contagem de calorias e carboidratos. Também, há a gestão de registro de atividades físicas realizadas em um período. Dessa forma, as Figuras 9, 10, 11 e 12 mostram essas funcionalidades em operação.

## 5. Conclusões

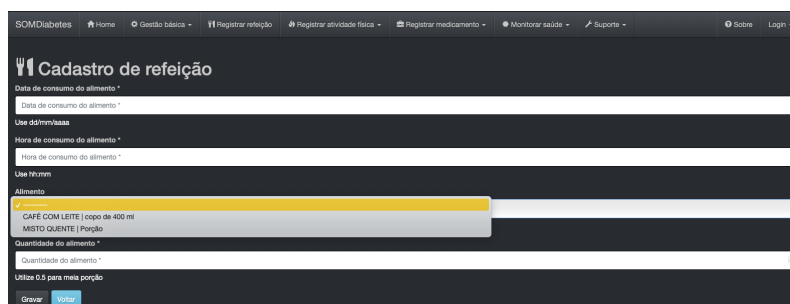
Ao longo do texto, foram apresentados e discutidos conceitos referentes à diabetes (como uma doença que exige acompanhamento para um tratamento mais preciso), descoberta de conhecimento e a reconhecimento de padrões por meio de mineração de dados. Também



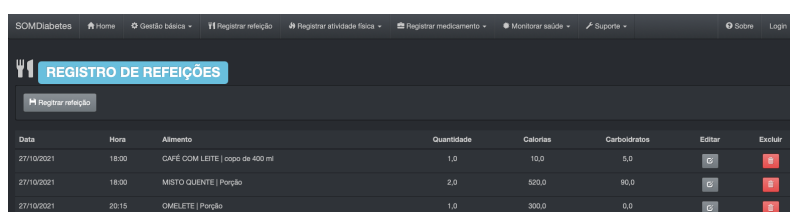
**Figura 9. Interface de visão do Administrador.**



**Figura 10. Interface gestão de atividades físicas.**



**Figura 11. Interface de registro de refeições em um dia e hora.**



**Figura 12. Interface da lista de refeições registradas.**

foram analisados trabalhos relacionados que fornecem suporte a esta pesquisa, principalmente, o trabalho de Rodrigo Rath [Rath et al. 2014].

Os resultados atingidos, até este momento, foram a remodelagem de alguns aspectos estruturais e funcionais existentes em [Rath et al. 2014], a adequação da planilha eletrônica (com dados de monitoramento de 3 anos de um doente em diabetes) ao formato do ambiente WEKA.

## Referências

- Apache Airflow (2021). Apache Airflow. <http://airflow.apache.org/>. Acessado em Março de 2021.
- Borg, G. and Borg, E. (2001). A new generation of scaling methods: Level-anchored ratio scaling. *Psychologica*, 28:15–45.
- Furlan, M. B. and de Souza Poletto, A. S. R. (2018). Algoritmos e técnicas para mineração de dados. *Fundação Educacional do Município de Assis/SP*.
- Goldschmidt, E. P. R. (2005). *Data Mining: Um Guia Prático*. Editora Elsevier.
- Group, M. L. (2021). Weka: The workbench for machine learning. <https://www.cs.waikato.ac.nz/ml/weka/>. Acessado em Março de 2021.
- Grupo de Interesse Especial Keras (Keras SIG) (2021). Keras SIG. <https://keras.io/>. Acessado em Maio de 2021.
- Laboratório LISA, Universidade de Montreal (2021). Biblioteca Python: Teano. <https://pypi.org/project/Theano/>. Acessado em Maio de 2021.
- Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. Editora John Wiley and Sons.
- Marques, É. B., Zamberlam, A. d. O., de Oliveira, R. F., Raimann, L. H., and de Oliveira, L. V. (2008). Projeto de módulo de data mining para scout voleibol. *Seminário de Informática-RS (SEMINFO RS 2008), Torres (RS)*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- PyTables Developers Team (2021). PyTables: Hierarchical Datasets in Python. <http://www.pytables.org/>. Acessado em Maio de 2021.
- Rath, R., Zamberlan, A., and Vieira, S. (2014). Sistema de recomendação para controle da diabetes. In *7o Congresso Sul Brasileiro de Computação*, Criciúma. SULCOMP, UNESC.
- Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.
- Rocklin, M. (2015). Dask: Parallel computation with blocked algorithms and task scheduling. In Huff, K. and Bergstra, J., editors, *Proceedings of the 14th Python in Science Conference*, pages 130 – 136.
- SBD (2021). Sociedade Brasileira de Diabetes. <https://www.diabetes.org.br/>. Acessado em Março de 2021.
- Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., Rückstieß, T., and Schmidhuber, J. (2010). PyBrain. *Journal of Machine Learning Research*, 11:743–746.
- Tan, P.-N., Steinbach, M., and Kumar, V. (2009). *Introdução ao datamining: mineração de dados*. Ciência Moderna, São Paulo.

- The pandas development team (2021). pandas-dev/pandas: Pandas. <https://pandas.pydata.org/>. Acessado em Maio de 2021.
- Vieira, S. A. G. (2016). *Identificação de padrões de expressão em doenças genéticas usando uma rede de integração de vias de manutenção do Genoma, Angiogênese, Hipóxia e Vigilância Imunológica*. PhD thesis, Centro Universitário Franciscano, Santa Maria.
- Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.
- Wykowski, T. and Wykowska, J. (2019). Lessons learned: Using scrum in non-technical teams. <https://www.agilealliance.org/resources/experience-reports/lessons-learned-using-scrum-in-non-technical-teams/>. Acessado em Março de 2021.