UNIVERSIDADE DE TAUBATÉ

Vanessa Rovida Loschi

USO DA MINERAÇÃO DE DADOS PARA AUXILIAR NA PREVENÇÃO DA DOENÇA RENAL CRÔNICA

TAUBATÉ

2019

Vanessa Rovida Loschi

USO DA MINERAÇÃO DE DADOS PARA AUXILIAR NA PREVENÇÃO DA DOENÇA RENAL CRÔNICA

Trabalho de Monografia apresentado para a obtenção do Certificado de Especialização pelo Curso de Gestão de Projetos em BI, do Departamento de Informática, da Universidade de Taubaté.

Área de Concentração: Mineração de Dados

Orientador: Prof. Dr. Luis Fernando de Almeida

TAUBATÉ

2019

Ficha catalográfica elaborada pelo
SIBi – Sistema Integrado de Bibliotecas / UNITAU

L879u Loshi, Vanessa Rovida
Uso da mineração de dados para auxiliar na prevenção da doença renal crônica / Vanessa Rovida Loshi. - 2019.
45f.:il.

Monografia (especialização) - Universidade de Taubaté, Departamento de Informática, 2019.
Orientação: Prof. Dr. Luis Fernando de Almeida, Departamento de Informática.

1. Mineração de dados (Computação). 2. Algoritmo de associação. 3. Algoritmo de classificação. 4. Rins – Doenças. I. Universidade de Taubaté. II. Título

CDD 006.3

Ficha catalográfica elaborada pela Bibliotecária Regina Márcia Cuba – CRB 8ª/7416

VANESSA ROVIDA LOSCHI

USO DA MINERAÇÃO DE DADOS PARA AUXILIAR NA PREVENÇÃO DA DOENÇA RENAL CRÔNICA

Trabalho de Monografia apresentado como requisito parcial para a obtenção do Certificado de Especialização pelo Curso de Gestão de Projetos em Business Intelligence, do Departamento de Informática, da Universidade de Taubaté.

Área de Concentração: Mineração de Dados

Data:

Resultado: Aprovado

BANCA EXAMINADORA

Prof .Dr.Luis Fernando de Almeida Universidade de Taubaté

Assinatura_____

Prof. Me. Dawilmar Guimarães Araújo Universidade de Taubaté

Assinatura_____

Prof. Esp. Fabio Rosindo Daher de Barros Universidade de Taubaté

Assinatura_____

Este trabalho é dedicado a todos os profissionais da área de saúde que trabalham com a prevenção de doenças.

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus por tudo.

Um agradecimento especial ao meu orientador Luis Fernando de Almeida pela paciência e ajuda.

Nestes dois anos de curso, grandes pessoas passaram por mim e me fizeram crescer como pessoa e como profissional. Gostaria de agradecer a todos os amigos de classe e professores, sempre tão dedicados e comprometidos. Em especial às professoras Rita Rigotti e Ana Clara Mota por todo o apoio.

Agradeço ao meu ex-chefe Alisson Valério por me dar a oportunidade de migrar para a área de Business Intelligence, sem isso, esse curso não teria acontecido.

Quero agradecer também à amiga Liliane Moreira pelas conversas sobre a área de prevenção de doenças, e por me ajudar a entender um pouquinho deste universo.

Enfim, agradeço à minha família pelo apoio e compreensão. Amo vocês.

“Descobrir consiste em olhar para o que todo mundo está vendo e pensar uma coisa diferente.”

Roger Von Oech

RESUMO

Devido à dificuldade do diagnóstico nas fases iniciais da doença renal crônica, ao sofrimento que ela causa aos pacientes e aos custos elevados no seu tratamento, este trabalho, de natureza exploratória, visa encontrar, por meio da mineração de dados, um padrão associativo e modelos preditivos que consigam identificar pacientes nas fases iniciais da doença renal crônica. Para a execução dos algoritmos, foram utilizados dados históricos de atendimentos médicos de pacientes já dependentes de hemodiálise. No desenvolvimento, foram aplicados o algoritmo de associação APRIORI e os algoritmos de classificação Random Forest e J48. Apesar do algoritmo de associação não ter conseguido resultados satisfatórios nesta pesquisa, bons modelos preditivos foram criados com os algoritmos de classificação, que faz com que o objetivo principal deste trabalho tenha sido alcançado.

Palavras-chave: Mineração de Dados, Algoritmo de Associação, Algoritmo de Classificação, Doença Renal Crônica.

ABSTRACT

Due to the difficulty of diagnosis in the early stages of chronic kidney disease, the suffering it causes to patients and the high costs of its treatment, this exploratory work aims to find, through data mining, an associative pattern and models predictive that can identify patients in the early stages of chronic kidney disease. For the execution of the algorithms, we used historical data of medical care of patients already dependent on hemodialysis. In development, the association algorithm APRIORI and the classification algorithms Random Forest and J48 were applied. Although the association algorithm did not reach satisfactory results in this research, good predictive models were created with the classification algorithms, which has achieved the main objective of this work.

Keywords: Data Mining, Association Algorithm, Classification Algorithm, Chronic Kidney Disease.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

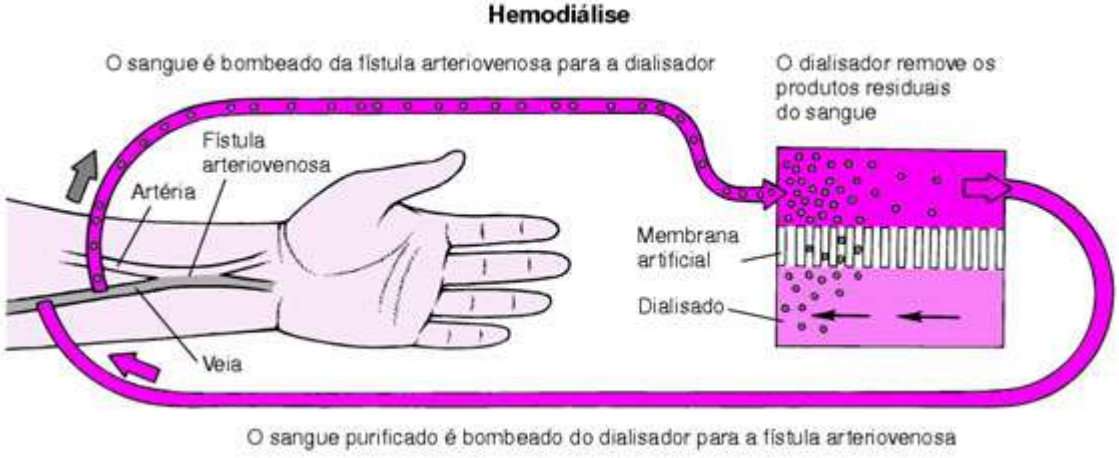
1 INTRODUÇÃO

Segundo a SOCIEDADE BRASILEIRA DE NEFROLOGIA (2019), a insuficiência ou doença renal é a perda das funções dos rins, ela pode ser aguda ou crônica. A insuficiência renal aguda é a perda súbita da função dos rins, que normalmente é reversível. Comum em pacientes internados, ela pode ser causada por desidratação, medicamentos, intoxicações, traumatismos e outras doenças. Já a insuficiência renal crônica é a perda lenta e progressiva das funções renais, e ela é irreversível.

A doença renal crônica possui uma dificuldade do diagnóstico nas fases iniciais da doença. Seu tratamento nas fases mais avançadas é a hemodiálise, que causa sofrimento e custos elevados (MINISTÉRIO DA SAÚDE, 2019). Esta particularidade apresenta-se como um motivador para um estudo mais aprofundado sobre ela.

Segundo a SOCIEDADE BRASILEIRA DE NEFROLOGIA (2019), a hemodiálise é um procedimento onde uma máquina é responsável por limpar e filtrar o sangue, fazendo parte do trabalho que o rim doente não pode fazer. Na hemodiálise (Figura 1), a máquina recebe o sangue do paciente por um acesso vascular, o sangue é levado até o filtro de hemodiálise, onde são retirados os líquidos e as toxinas em excesso e, o sangue limpo volta para o paciente pelo acesso vascular.

Figura 1: Hemodiálise.



Fonte: Hechanova, 2017.

O tempo que o paciente fica ligado à máquina, varia conforme o seu estado clínico, mas geralmente é de quatro horas, e de três a quatro vezes por semana (SOCIEDADE BRASILEIRA DE NEFROLOGIA, 2019).

Mas, se um paciente for identificado nas fases iniciais da doença, o tratamento pode ser feito com medicamentos, dieta e a adoção de condutas terapêuticas para retardar a progressão da doença.

Segundo o MINISTÉRIO DA SAÚDE (2019), a evolução da doença renal crônica é assintomática, causando um diagnóstico tardio e levando os pacientes diretamente à dependência da hemodiálise.

A prevenção de doenças, neste caso o retardamento da evolução da doença renal crônica, é uma aliada, postergando o sofrimento dos pacientes e diminuindo os gastos relacionados com o tratamento.

Este trabalho pesquisa a mineração de dados como uma alternativa a ser utilizada na prevenção da doença renal crônica. A mineração de dados é formada por um conjunto de ferramentas e técnicas, que utilizam algoritmos de aprendizagem ou classificação para descobrir padrões, tendências e mecanismos de regras, auxiliando na descoberta de conhecimento.

Todo o desenvolvimento deste trabalho parte da seguinte pergunta: A mineração de dados poderia auxiliar na prevenção, identificando pessoas na fase inicial da Doença Renal Crônica (DRC)?

1.1 OBJETIVOS

Os objetivos deste trabalho se dividem em Objetivo Geral e Objetivos Específicos.

1.1.1 **ObjetivoGeral**

Encontrar um padrão associativo e modelos preditivos nos dados históricos da assistência médica de pacientes dependentes de hemodiálise, a fim de identificar pessoas que estejam na fase inicial da DRC (Doença Renal Crônica).

1.1.2 **ObjetivosEspecíficos**

Os objetivos principais deste trabalho são:

- Estruturar os dados históricos de pacientes dependentes de hemodiálise para aplicação dos algoritmos de Mineração de Dados;
- Aplicar o algoritmo de associação APRIORI e os algoritmos de classificaçãoJ48 e Random Forest a partir da linguagem de programação R;
- Analisar os resultados obtidos e apresentá-los.

1.2 RELEVÂNCIA

A doença renal crônica consiste na perda progressiva e irreversível da função dos rins. Na sua fase avançada, os rins não conseguem mais manter o meio interno da pessoa. Se a doença renal for detectada precocemente, serão utilizadas, em seu tratamento, condutas terapêuticas para retardar a sua progressão, reduzindo o sofrimento dos pacientes e diminuindo os custos financeiros associados à doença. (ROMÃO JUNIOR, 2004)

1.4 APRESENTAÇÃO DO TRABALHO

O trabalho presente está estruturado em cinco capítulos:

- O presente capítulo apresenta a ideia geral do trabalho, o problema a que se dispõe a resolver, seus objetivos geral e específicos e ainda qual a relevância do mesmo.

- No Capítulo 2 são descritos os conceitos relacionados à pesquisa apresentada. Encontra-se aqui conceitos sobre a doença renal crônica, o processo KDD, a mineração de dados e a linguagem de programação utilizada no desenvolvimento deste trabalho.
- O Capítulo 3 descreve quais os métodos e as ferramentas utilizadas para atingir o objetivo proposto,detalha todo o desenvolvimento da pesquisa e descreve todos os passos executados a fim de se alcançar o objetivo deste trabalho. Aqui, encontra-se a descrição da base de dados inicial e de todo o processo de limpeza e transformação dos dados, o processo de mineração utilizado e a descrição dos parâmetros escolhidos.
- O Capítulo 4 apresenta a conclusão do trabalho, com o resultado encontrado e as considerações finais e propostas para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo são apresentados os conceitos sobre o objeto de pesquisa deste trabalho. A doença renal crônica será descrita a fim de contextualizar o leitor das dificuldades de sua prevenção. Também serão apresentados, os conceitos do processo *Knowledge Discovery in Databases (KDD*), da mineração de dados, dos algoritmos de associação e classificação, e da linguagem de programação R.

2.1 A DOENÇA RENAL CRÔNICA - DRC

“O equilíbrio da química interna de nossos corpos se deve em grande parte ao trabalho dos rins. Nossa sobrevivência depende do funcionamento normal destes órgãos vitais.”(SOCIEDADE BRASILEIRA DE NEFROLOGIA, 2019).

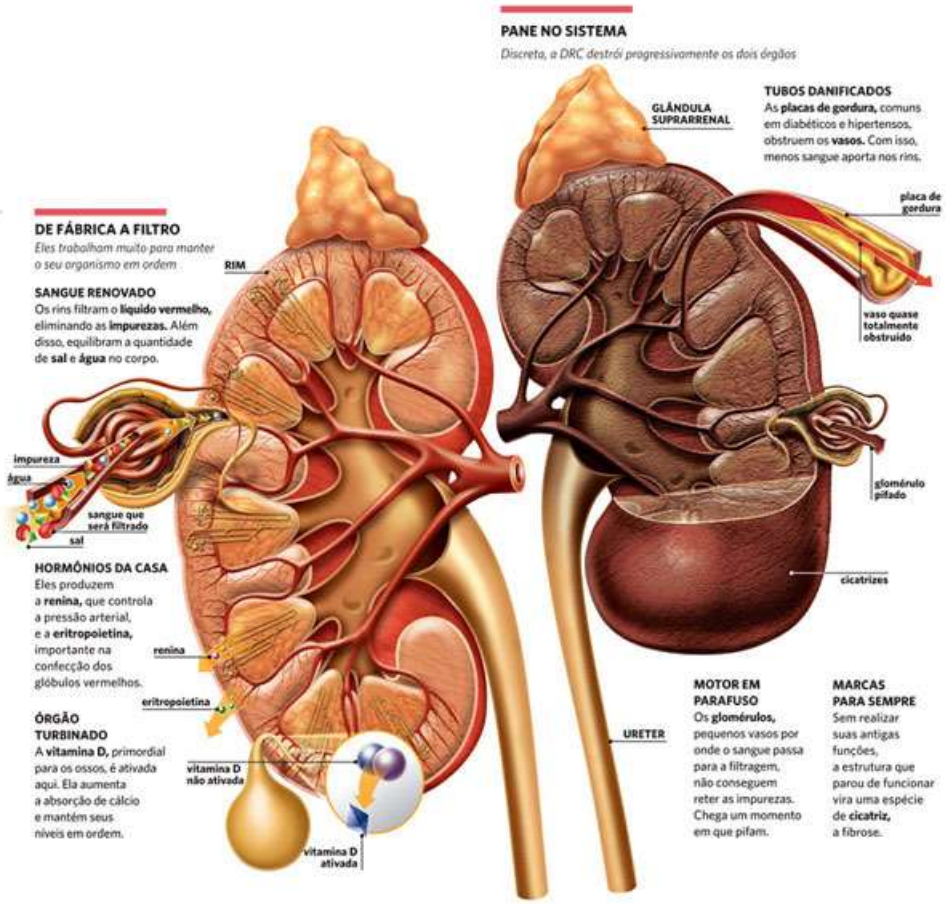
Segundo o MINISTÉRIO DA SAÚDE (2019), a principal função do rim é retirar os resíduos e o excesso de água do organismo. Os rins executam funções fundamentais para o nosso organismo se manter vivo e funcionando. As principais funções renais são a excreção de produtos finais de diversos metabolismos, a produção de hormônios, o controle do equilíbrio hidroeletrolítico, o controle do metabolismo ácido-básico e o controle da pressão arterial.

Os rins recebem, aproximadamente, 1,2 litros de sangue por minuto, que é cerca de um quarto do sangue bombeado pelo coração, então, podemos dizer que os rins filtram todo o sangue de uma pessoa quase doze vezes por hora (SOCIEDADE BRASILEIRA DE NEFROLOGIA, 2019).

A perda lenta e gradual das funções renais, se não for tratada, pode levar à paralisação dos rins (PAVÃO, 2012). Segundo a SOCIEDADE BRASILEIRA DE NEFROLOGIA (2019), o mau funcionamento dos rins pode acarretar em vários problemas além da hipertensão. Por produzirem um hormônio chamado de eritropoetina, que ajuda na maturação dos glóbulos vermelhos do sangue e da medula óssea, alterações em seu funcionamento podem causar anemia. Por serem responsáveis pela função de filtragem das toxinas, problemas nos rins podem acumular toxinas e fazer o “filtro” reduzir a ponto de não passar nem água, situação que impede a pessoa de urinare causa um acúmulo mais rápido das toxinas, essa

situação é chamada de uremia. Os sintomas da uremia incluem náuseas, debilidade, fadiga, desorientação, dispneia e edema nos braços e pernas. A Figura 2 apresenta um esquemático da descrição dos rins.

Figura 2: Descrição dos Rins.



Fonte: IHS, 2013.

Como, normalmente, a doença renal crônica evolui lentamente, o organismo acaba por se adaptar à diminuição da função renal, o que dificulta a identificação da doença em fases iniciais.

Segundo a SOCIEDADE BRASILEIRA DE NEFROLOGIA (2019), nas fases iniciais da DRC, o tratamento pode ser feito com medicamentos e dieta. Chamado de tratamento conservador, ele utiliza medidas clínicas como remédios, modificações na dieta e estilo de vida para retardar a evolução da função renal, diminuir os sintomas e evitar as complicações associadas à DRC. Mesmo com estas medidas, a doença renal crônica é progressiva e irreversível. O tratamento conservador consegue apenas reduzir a

velocidade da progressão ou estabilizar a DRC. Quanto mais cedo começar o tratamento conservador, maiores as chances da função dos rins serem preservados por mais tempo.

Mas, com o avanço da doença, a hemodiálise se faz necessária. Segundo a SOCIEDADE BRASILEIRA DE NEFROLOGIA (2019), na hemodiálise, uma máquina é responsável por receber o sangue do paciente através de um acesso vascular (cateter ou fístula arteriovenosa), este sangue é impulsionado por uma bomba até o filtro de diálise, onde é exposto à solução de diálise por uma membrana semipermeável responsável por retirar o líquido e as toxinas em excesso. Após o processo, o sangue limpo volta para o paciente pelo acesso vascular. O paciente é submetido a este processo de hemodiálise de três a quatro vezes por semana com uma duração média de quatro horas cada sessão.

A hemodiálise é um tratamento que deve ser feito até o fim da vida ou até que o paciente seja submetido a um transplante renal. Segundo a SOCIEDADE BRASILEIRA DE NEFROLOGIA (2019) e PAVÃO (2012), o Brasil tem o maior programa de transplante renal público do mundo. Em 2011, foram realizados 4.957 transplantes dos rins e o número cresceu para 5.402 em 2012. A doença renal crônica é considerada hoje um importante problema médico e de saúde pública. O número de pacientes em programas de hemodiálise mais que dobrou entre os anos de 1994 e 2004, passando de 24.000 para 59.153 pacientes em 2004. Este índice cresce aproximadamente 8% ao ano.

2.1.1 **Causas**

A doença renal crônica está associada a duas doenças: hipertensão arterial e diabetes. Segundo PAVÃO (2012), a hipertensão pode causar a disfunção renal, já que, como o controle da pressão arterial é função dos rins, mudanças nos níveis da pressão podem sobrecarregá-los. Assim, o controle da hipertensão é fundamental para a prevenção da doença. Em 2011, trinta e cinco por cento dos pacientes que precisaram de hemodiálise, já haviam sido diagnosticados com hipertensão arterial.

A diabetes pode causar dano nos vasos sanguíneos dos rins, dificultando a correta filtragem do sangue. Normalmente vinte e cinco por cento de portadores de

diabetes tipo I e, de cinco a dez por cento de portadores de diabetes tipo II desenvolvem a doença renal.
Outras doenças causadoras da insuficiência renal são:nefrite, cistos hereditários, infecções urinárias frequentes e doenças congênitas.

2.1.2 **Sintomas**

Apesar da dificuldade de identificar a doença nas fases iniciais, alguns sintomas podem ser observados (SOCIEDADE BRASILEIRA DE NEFROLOGIA, 2019):

- Pressão Alta;
- Inchaço ao redor dos olhos e nas pernas;
- Fraqueza constante;
- Náuseas e vômitos frequentes;
- Dificuldade de urinar;
- Queimação ou dor quando urina;
- Urinar muitas vezes, principalmente à noite;
- Urina com aspecto sanguinolento;
- Urina com muita espuma;
- Dor lombar, que não piora com movimentos;
- História de pedras nos rins.

O diagnóstico pode ser feito, através de exames laboratoriais simples, como o exame de urina que apresenta a presença da proteína albumina e o exame de sangue que verifica o nível da creatinina. Quando os rins não estão funcionando corretamente, eles causam desequilíbrio no organismo, eliminando ou absorvendo substâncias desordenadamente.

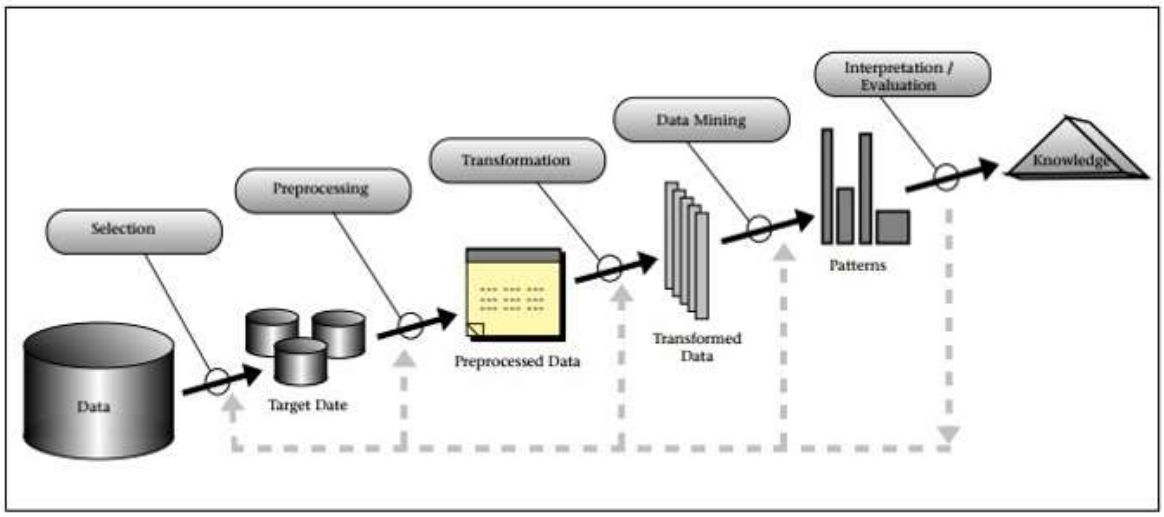
Segundo o MINISTÉRIO DA SAÚDE (2019), existem diversas formas de aferir as funções renais, como exame de urina e exames detalhados dos rins. No entanto, a função excretora tem maior correlação com os desfechos clínicos. A função excretora pode ser medida pela Taxa de Filtração Glomerular (TFG), que é calculada com base na idade, gênero e valor da creatinina.

2.2 PROCESSO KDD

KDD (Figura 3) é uma abreviatura que vem do Inglês "*Knowledge Discovery in Databases*" que significa Descoberta do Conhecimento em Bases de Dados. O KDD é um processo contendo tarefas para transformar dados brutos em conhecimento. Segundo Fayyad (Apud AZEVEDO, 2018), este conjunto de tarefas é sequencial e organizado como segue:

1. Seleção dos dados;
2. Pré-processamento dos dados;
3. Transformação dos dados;
4. Data Mining;
5. Interpretação e avaliação dos resultados.

Figura 3: Processo KDD.



Fonte: FAYYAD et al., 1996 Apud AZEVEDO, 2018

Resumidamente, suas etapas podem ser descritas como segue:

- Na etapa de seleção dos dados, são selecionados e extraídos os dados pertinentes ao objetivo do processo de KDD.
- O pré-processamento é a etapa de limpeza dos dados, do preenchimento de dados faltantes e da exclusão de *outliers*.
- Na transformação, os dados são estruturados de acordo com o tipo de mineração que será submetido. Nesta etapa, ocorre a redução dos dados, que pode ser por agregação, onde os objetos são combinados criando

uma visão alto nível, por redução da dimensionalidade, onde são retirados atributos irrelevantes e, também, por seleção de subconjunto, onde são criados novos atributos da combinação de atributos antigos. Ainda na fase de transformação, os dados podem sofrer normalização, discretização e binarização.

- Em seguida vem a etapa da mineração de dados, onde são aplicados algoritmos de mineração que podem ser de agrupamento, classificação, associação, entre outros.
- Na última etapa, de pós-processamento, o modelo gerado é interpretado e seus resultados são apresentados através de gráficos e planilhas.

2.3 MINERAÇÃO DE DADOS

Dados são gerados continuamente, a quantidade de dados armazenados aumenta a cada dia. Todas as escolhas, compras, acessos, pesquisas, viagens, enfim, tudo que se faz pode gerar um registro em um repositório de dados.

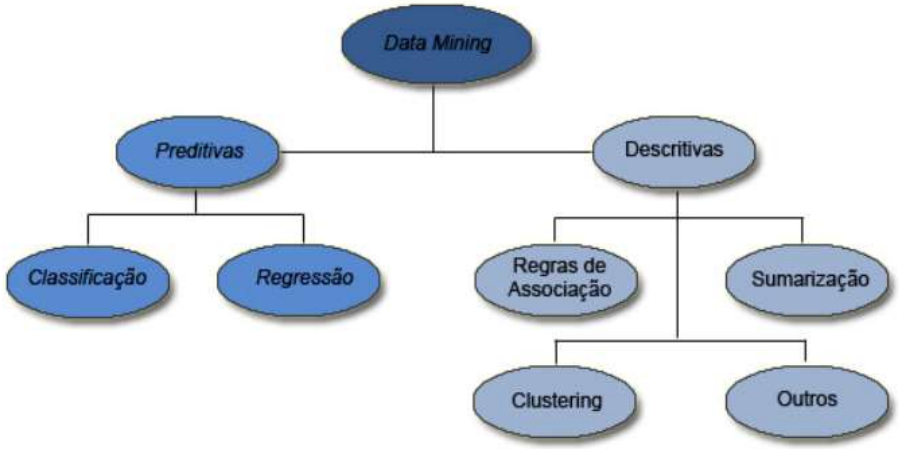
O aumento da capacidade de armazenamento dos hardwares e a diminuição no custo deste recurso causaram um tsunami de dados armazenados, criando infinitas possibilidades à mineração de dados.

> A mineração de dados é definida como o processo de descoberta de padrões nos dados. As pessoas vêm buscando padrões nos dados desde que a vida humana começou. Os caçadores buscam padrões no comportamento de migração de animais, os agricultores buscam padrões no crescimento das plantações, os políticos buscam padrões na opinião dos eleitores e os amantes buscam padrões nas respostas de seus parceiros. O trabalho de um cientista é dar sentido aos dados, descobrir os padrões que governam o funcionamento do mundo físico e encapsulá-los em teorias que podem ser usadas para prever o que acontecerá em novas situações. (WITTEN, 2011)

“Dados analisados de forma inteligente são um recurso valioso. Pode levar a novos insights e, em ambientes comerciais, a vantagens competitivas.” (WITTEN, 2011).

A mineração de dados, em sua busca por informações úteis, utiliza algoritmos específicos a fim de descobrir padrões, tendências e mecanismos de regras. Aqui, são escolhidos, configurados e executados apenas um ou vários algoritmos de mineração. A escolha da tarefa depende dos objetivos a serem alcançados. E, ainda, as tarefas podem ser agrupadas em atividades preditivas ou descritivas, conforme Figura 4.

Figura 4: Tarefas de Mineração de Dados.



Fonte: Rezende, 2003. p.318

2.3.1 **Algoritmos deAssociação**

Os algoritmos de associação procuram relações entre itens contidos em grande quantidade de transações, e, onde seria impossível sua descoberta por observação.

Segundo GONÇALVES (2012), as Regras de Associação tem grande aplicabilidade e são consideradas como um importante tipo de conhecimento minerado. Este tipo de algoritmo apresenta padrões de relacionamento dos itens de uma base de dados. É muito utilizado em análise de transações de compra.

A análise de transações de compra consiste na procura por padrões para determinar produtos que normalmente são adquiridos juntos em uma mesma compra.

A Tabela 1apresenta um exemplo do funcionamento do algoritmo de associação APRIORI.

Tabela 1: Cesto de Compra.

Transação	Itens Comprados
1	Produto 1, Produto 2
2	Produto 1, Produto 2
3	Produto 1, Produto 2, Produto 3
4	Produto 1, Produto 2, Produto 3
5	Produto 1, Produto 2, Produto 3
6	Produto 1, Produto 2, Produto 3
7	Produto 1, Produto 3
8	Produto 1, Produto 2, Produto 3
9	Produto 1, Produto 2
10	Produto 2, Produto 3
11	Produto 2, Produto 3
12	Produto 1, Produto 2, Produto 3

Fonte: O autor.

Como os dados precisam estar estruturados para a utilização do algoritmo,na Tabela 2 pode-se visualizar os dados no formato necessário.

Tabela 2: Dados de Compra Estruturados.

Transação	Produto 1	Produto 2	Produto 3
1	1	1	0
2	1	1	0
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	0	1
8	1	1	1
9	1	1	0
10	0	1	1
11	0	1	1
12	1	1	1

Fonte: O autor.

Segundo AGRAWALET al. apud GONÇALVES (2012), são necessárias duas fases para encontrar as regras de associação.

A primeira é a descoberta de cestos frequentes, onde são levantados todos os possíveis conjuntos de itens comprados. Para cada conjunto sua frequência é verificada em relação ao total de transações. São considerados somente os conjuntos que a frequência esteja acima do parâmetro de suporte mínimo.

Ainda, utilizando o exemplo da cesta de compras da Tabela 1, a frequência dos conjuntos de itens foi calculada e, dado um suporte mínimo de 0,6, foram realçadas de verde as linhas cuja frequência foi maior ou igual ao suporte mínimo. A Tabela 3 apresenta a informação resultante.

Tabela 3: Cestos Frequentes – Suporte Mínimo 0,6.

Conjunto de Itens	Suporte
Produto 1	0,83
Produto 2	0,92
Produto 3	0,75
Produto 1, Produto 2	0,75
Produto 1, Produto 3	0,58
Produto 2, Produto 3	0,67
Produto 1, Produto 2, Produto 3	0,5

Fonte: O autor.

E, a segunda fase é a descoberta de regras interessantes, onde os cestos frequentes são utilizados para gerar regras de análise combinatória. Cada regra com um conjunto antecedente gerando um item consequente é validada pela frequência deste item consequente em relação ao conjunto antecedente. São consideradas somente as regras que a frequência esteja acima da confiança mínima.

Prosseguindo com o mesmo exemplo, foram realçadas de verde todas as linhas onde um conjunto de itens antecedente gera um item consequente com um fator de confiança acima de 0,8. A Tabela 4 ilustra estes dados.

Tabela 4: Regras Interessantes – Confiança Mínima 0,8.

Conjunto de Itens	Suporte	Confiança
Produto 1 -> Produto 2	0,75	0,90
Produto 2 -> Produto 1	0,75	0,82
Produto 1 -> Produto 3	0,58	0,70
Produto 3 -> Produto 1	0,58	0,78
Produto 2 -> Produto 3	0,67	0,73
Produto 3 -> Produto 2	0,67	0,89
{Produto 1, Produto 2} -> Produto 3	0,50	0,67
{Produto 1, Produto 3} -> Produto 2	0,50	0,86
{Produto 2, Produto 3} -> Produto 1	0,50	0,75

Fonte: O autor.

Assim, o algoritmo APRIORI faz a mineração dos dados baseado na geração e poda. Gera todos os conjuntos de itens possíveis dos dados e faz a poda, retirando todos os conjuntos de itens com a frequência inferior ao valor mínimo de suporte. E também faz a poda das regras associativas, retirando aquelas que não alcançaram o valor mínimo de confiança.

2.3.2 **Algoritmos de Classificação**

A classificação é uma tarefa preditiva de aprendizagem supervisionada. Os dados que irão gerar o modelo preditivo são conhecidos desde o início. Segundo PACHECO (2016), a classificação de dados pode ser aplicada facilmente no dia a dia. Reconhecer padrões em imagens, diferenciar espécies de plantas, classificar tumores benignos e malignos, entre outros, são aplicações deste tipo de algoritmo. A classificação consiste em atribuirum rótulo a algum objeto, com base em um conjunto de atributos pertencentes ao objeto. Para isso, é necessário criar um modelo a partir de um conjunto de objetos onde os rótulos são conhecidos.

Os dados são separados em duas partes, uma parte maior para treino, onde o modelo é construído, e uma pequena parte para teste, onde o modelo será avaliado.

No treino, os dados, com a classe definida, são lidos pelo algoritmo de classificação, que criará o seu modelo para alcançar os resultados esperados. Já no

teste, o algoritmo irá testar o modelo criado com outros dados, e apresentará a precisão do modelo construído.

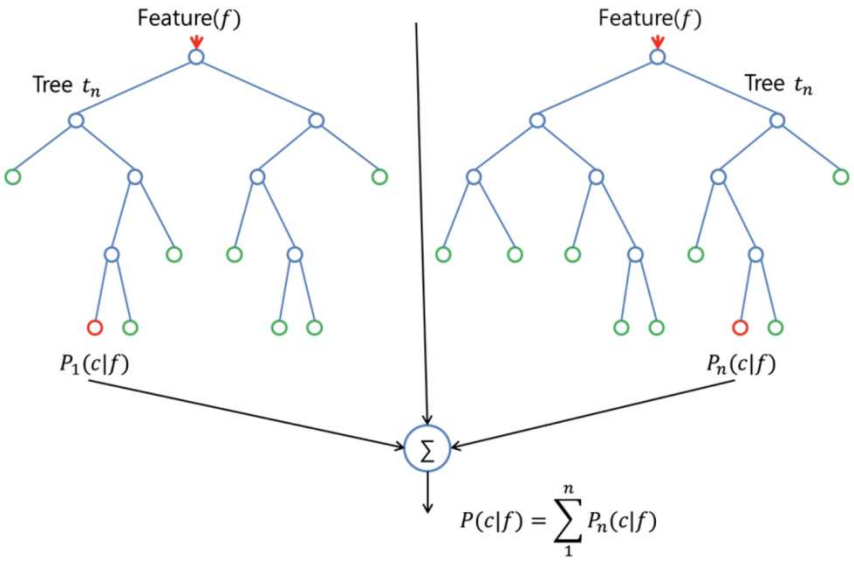
Neste trabalho, serão utilizados algoritmos de classificação baseados em árvores de decisão.

Uma árvore de decisão começa com um único nó, chamado nó raiz, que se divide em possíveis resultados, os nós filhos. Cada um desses filhos leva a outros nós filhos adicionais, que se ramificam em outras possibilidades até chegar ao nó terminal chamado de nó folha. Assim, cria-se uma forma de árvore.

Cada nó contém um teste de um atributo, cada ramo se refere a um possível valor deste atributo, cada folha é associada a uma classe e, cada percurso na árvore representa uma regra de classificação.

Um exemplo de algoritmo de classificação é o Random Forest (Figura 5). O algoritmo de florestas aleatórias cria várias árvores de decisão e as combina para conseguir uma predição com melhor resultado e maior estabilidade.

Figura 5: Random Forest.



Fonte: Medium, 2019.

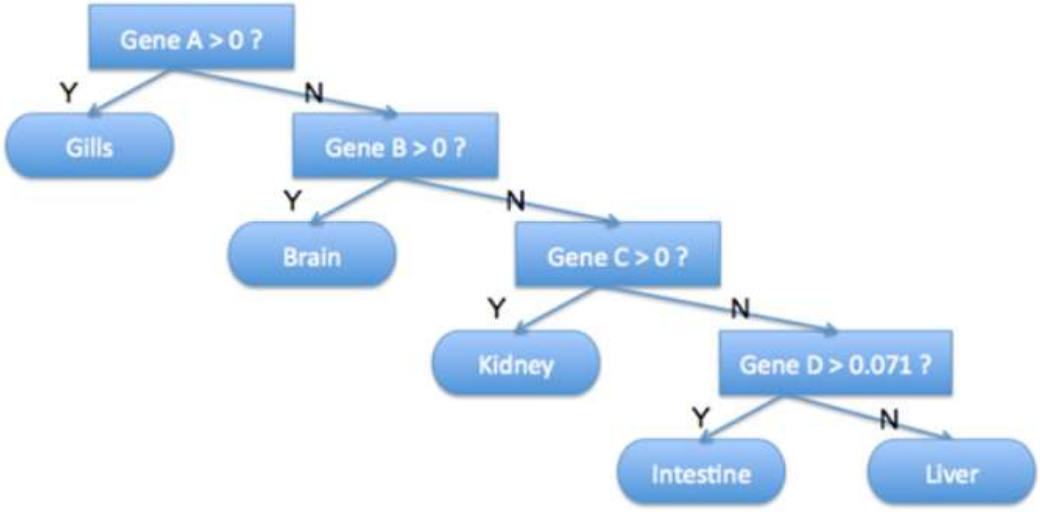
Segundo SILVA (2018), o algoritmo de floresta aleatória aumenta a aleatoriedade do modelo, enquantocria as árvores. No momento que faz a partição de nodos, o algoritmo busca a melhor característica em um subconjunto aleatório das características, criando uma grande variedade e gerando modelos melhores.

O algoritmo Random Forest combina um grupo de modelos fracos para formar um modelo mais forte.

Outro algoritmo de classificação é o J48, que é a implementação na linguagem JAVA de um algoritmo muito conhecido de indução de árvore de decisão, o C4.5, criado na década de 90 por Ross Quinlan. É muito utilizado,pois aceita tanto atributos categóricos (ordinais ou não-ordinais) como também atributos contínuos. (VIEIRAet al., 2018)

Para trabalhar com atributos contínuos, o algoritmo J48 (Figura 6) define um limiar e divide os dados binariamente, como maiores e menores ou iguais ao limiar definido. Apesar de aceitar valores faltantes, eles não são utilizados nos cálculos de ganho e entropia. O J48 gera árvores com maior precisão e menor complexidade, por fazer a divisão dos dados com o atributo selecionado pela medida de razão de ganho. Faz, também, a pós-poda das árvores geradas, buscando de baixo para cima ramos sem ganho significativo e transformando-os em nós folha.

Figura 6: Árvore podada J48.



Fonte:Techyv, 2019.

2.4 LINGUAGEM R

O R é um ambiente para computação e uma linguagem de programação orientada a objetos que permite a manipulação de dados, realização de cálculos e confecção de gráficos (PETERNELLI E MELLO, 2013). R é uma linguagem de programação amplamente usada por estatísticos e mineradores de dados para manipulação de dados estatísticos e gráficos.

Segundo Matos (2015) o R é uma implementação da linguagem de programação S. Foi criado por Ross Ihaka e Robert Gentleman na Universidade de Auckland, Nova Zelândia. É escrito principalmente em C, Fortran e R. Está disponível gratuitamente sob a licença GNU General Public e em vários sistemas operacionais. Pode ser usado através de uma interface de linha de comando ou por meio de *front-ends* gráficos como o RStudio.

Na linguagem R existem vários atalhos e expressões úteis, onde o usuário pode executar operações muito complicadas de forma sucinta. Muitos destes atalhos se tornam naturais quando se está familiarizado com a linguagem. Em alguns casos, existem várias maneiras de se executar uma mesma tarefa (R CORE TEAM, 2019).

Apesar de ter sido desenvolvido para fornecer suporte e inovações em computação estatística, com gráficos de alta qualidade e com a possibilidade da inclusão de fórmulas e símbolos matemáticos, a linguagem R não é somente um software estatístico, pois dispõe de funcionalidades de operações matemáticas, banco de dados, manipulação de vetores e matrizes, entre outras. É uma ferramenta madura e expansível com o uso de pacotes com funções específicas que podem ser encontradas na rede de distribuição do R, conforme ilustrado na Figura 7.

Figura 7: R Packages - *Leaderboard*.

Fonte: Datacamp, 2019

3 METODOLOGIA

Neste capítulo, encontra-se a descrição dos métodos utilizados a fim de alcançar os objetivos idealizados neste trabalho. Será apresentada a natureza da pesquisa, a fonte de dados utilizada, como serão apresentados os resultados e quais as ferramentas utilizadas no desenvolvimento.

A natureza da pesquisa apresentada é exploratória, e visa, baseada no histórico de utilização da assistência médica de pacientes dependentes de hemodiálise, encontrar um padrão de utilização através do algoritmo de associação APRIORI e encontrar modelos preditivos com a aplicação dos algoritmos de classificação Random Forest e J48.

3.1 DEFINIÇÃO DA FONTE DE PESQUISA

A fonte, de origem primária, é composta por dados históricos de utilização da assistência médicade pacientes dependentes de hemodiálise. Os dados selecionados para a pesquisa são referentes a dois anos de utilização antes da primeira ocorrência do procedimento de hemodiálise.

Se algum padrão for encontrado neste período de dois anos, o algoritmo poderá ser utilizado para alertar pessoas de um possível problema renal e solicitar que façam exames específicos.

A fonte utilizada não contém dados pessoais dos pacientes. O código de paciente foi gerado aleatoriamente, e serve apenas para distinguir os atendimentos.

3.2 APRESENTAÇÃO DOS RESULTADOS

Os resultados serão apresentados quantitativamente. Através de valores calculados e Gráficos gerados na Linguagem de programação R. Para as regras associativas, um gráfico será gerado pela função PLOT no RStudio para demonstrar o suporte e a confiança das regras criadas pelo algoritmo APRIORI.

Para o algoritmo de classificação Random Forest, o resultado, calculado a partir da matriz de confusão, será apresentada em forma de texto e também será apresentado um gráfico gerado pela função PLOT tanto para o treino quanto para o teste.

E, o resultado obtido no algoritmo de classificação J48 será apresentado apenas em forma de texto com base no cálculo da matriz de confusão gerada.

3.3 FERRAMENTAS UTILIZADAS

Para o execução deste trabalho, foi utilizada a linguagem de programação R através da *Integrated Development Environment* (IDE) RStudio. Nela, foi desenvolvida toda a programação utilizada na limpeza e transformação dos dados, na aplicação do algoritmo de mineração e na apresentação dos resultados obtidos.

O algoritmo de associação APRIORI foi utilizado para a mineração das regras associativas. E, os algoritmos de classificação Random Forest e J48 foram utilizados para a criação dos modelos preditivos.

3.4 DESENVOLVIMENTO

No desenvolvimento, estão descritos todos os passos executados na pesquisa. Aqui, será apresentada a fonte de dados, quais os processos de limpeza e transformação que foram aplicados na base inicial e a execução dos algoritmos de associação e classificação.

3.4.1 **Fonte de dados**

A fonte principal de dados utilizada no trabalho é baseada nos dados históricos da assistência médica de pacientes dependentes de hemodiálise.

Foram extraídos os dados de dois anos de utilização antes da primeira ocorrência da hemodiálise.

Foram também utilizados, para os algoritmos de classificação, dados históricos de pacientes que não são portadores da doença renal crônica. Estes pacientes foram escolhidos por pertencerem à mesma faixa etária dos pacientes renais.

As fontes utilizadas não possuem dados pessoais dos pacientes. O código de paciente foi gerado aleatoriamente, e serve apenas para distinguir os atendimentos .A fonte, inicial, contém os seguintes dados:

- Código do Paciente (Aleatório);
- Tipo do Atendimento (Consulta, Internação, Ambulatório e Diagnóstico / Terapia);
- Código da doença;
- Especialidade do atendimento;
- Código TUSS procedimento realizado;
- Data de realização;
- Data da primeira ocorrência de hemodiálise.

Também foram utilizados dados complementares como as descrições dos grupos das doenças e as descrições dos grupos de procedimentos. Estes dados complementares foram extraídos do portal DataSus do Governo Federal através dos endereços “http://www.datasus.gov.br/cid10/V2008/cid10.htm” para os grupos de doenças e, “http://sigtap.datasus.gov.br/tabela-unificada/app/sec/inicio.jsp” para os grupos de procedimentos. (Tabela 5)

Tabela 5: Exemplo dos Dados Extraídos.

Descrição do Campo	**Dado**
Código do Paciente	136
Tipo de Atendimento	INTERNACAO
Código Doença	I770
Descrição do Grupo da Doença	IX. Doenças do aparelho circulatório
Especialidade do Atendimento	CIRURGIA VASCULAR
Código TUSS	30908094
Descrição Grupo do Procedimento	FÍSTULAS ARTERIOVENOSAS CONGÊNITAS OU ADQUIRIDAS
Data Realização	29/12/2015
Data 1a Hemodiálise	01/02/2016

Fonte: O autor.

3.4.2 **Limpeza dos dados**

Esta pesquisa visa aplicar as técnicas de mineração de dados em datasets contendo dois anos de utilização da assistência médica de pacientes renais antes da primeira ocorrência da hemodiálise, então, durante o processo de limpeza foram excluídos todos os atendimentos com a data de realização maior que a data da primeira ocorrência de hemodiálise e, também, todos os atendimentos com a data de realização menor que vinte e quatro meses antecedentes à primeira ocorrência da hemodiálise.

Foram desconsiderados os registros de internação sem a informação de um Código Internacional de Doenças (CID) válido e as consultas que não possuíam uma especialidade médica válida.

3.4.3 **Transformação dos dados**

A transformação é a parte mais onerosa da pesquisa. Transformar atendimentos de saúde em uma estrutura de cesta de compras custou grande parte do tempo do trabalho.

Foram incluídos os grupos do código de doenças e os grupos dos procedimentos, já que nos dois casos, existe um leque grande de variações de doenças e procedimentos que iriam dificultar o processo de criação de regras associativas.

Os dados foram agrupados em:

- Consultas por especialidade;
- Ocorrência de atendimentos ambulatoriais;
- Internações por grupo do código de doença;
- E, procedimentos de diagnose e terapia por grupo de procedimentos médicos.

O dataset criado foi transposto horizontalmente. Assim, cada código de paciente resultou em apenas uma linha e nela foram incluídas colunas,

apresentando a ocorrência das consultas, atendimentos ambulatoriais, internações e procedimentos de diagnose e terapia, cada uma com sua respectiva especialidade ou grupo. Estas ocorrências também foram separadas por semestres, os dados foram organizados em semestres anteriores ao evento da hemodiálise, constando o primeiro, segundo, terceiro e quarto semestres antecedentes.

Devido a necessidade de melhora no desempenho do algoritmo, os dados faltantes ficaram em branco para não acarretar na criação de regras para a não ocorrência dos eventos, que demandaria recurso de hardware.

Após a criação das linhas, o código do paciente foi excluído do DataSet, restando apenas as ocorrências dos atendimentos.

Um dataset contendo apenas dados de pacientes dependentes de hemodiálise foi separado para a aplicação do algoritmo de associação.

Já, para a utilização do dataset nos algoritmos de classificação, os dados de pacientes sem problemas renais (Classe “-1”) foram mesclados com os dados de pacientes renais (Classe “1”) e, ainda, os dados foram divididos em um dataset para treino com oitenta por cento dos registros e outro dataset contendo vinte por cento dos dados para teste.

A Tabela 6 exemplifica os dados do paciente da Tabela 5 após o processo de transformação. Os dados estão organizados com o tipo de atendimento, seguido dos dados de agrupamento, distintos entre os tipos de atendimento, e complementados com o semestre anterior à primeira hemodiálise. O número 1 na segunda coluna indica a ocorrência do evento.

Tabela 6: Dados após o processo de transformação.

Tipo Atendimento "." Agrupador "." Semestre	Ocorrência
INTERNACAO . IX. Doenças do aparelho circulatório .1	1
AMBULATORIO .AMBULATORIO .4	1
AMBULATORIO .AMBULATORIO .3	1
AMBULATORIO .AMBULATORIO .2	1
SADT .RETINA .4	1
SADT .RETINA .3	1
SADT .RETINA .2	1
SADT . ENDOCRINOLOGIA LABORATORIAL .2	1
SADT . ENDOCRINOLOGIA LABORATORIAL .1	1
SADT .PROCEDIMENTOS .4	1
SADT .PROCEDIMENTOS .3	1
SADT .PROCEDIMENTOS .2	1
SADT .PROCEDIMENTOS .1	1
SADT . ECG - TE .2	1
SADT .URINÁLISE .2	1
SADT . HEMATOLOGIA LABORATORIAL .2	1
SADT . HEMATOLOGIA LABORATORIAL .1	1
SADT . BIOQUÍMICA (SANGUE, URINA E OUTROS MATERIAIS) .2	1
SADT . BIOQUÍMICA (SANGUE, URINA E OUTROS MATERIAIS) .1	1
SADT .IMUNOLOGIA .2	1
SADT .IMUNOLOGIA .1	1
SADT . CONSULTAS, VISITAS HOSPITALARES OU ACOMPANHAMENTO DE PACIENTES .4	1
SADT . CONSULTAS, VISITAS HOSPITALARES OU ACOMPANHAMENTO DE PACIENTES .3	1
SADT . CONSULTAS, VISITAS HOSPITALARES OU ACOMPANHAMENTO DE PACIENTES .2	1
SADT . ULTRASSONOGRAFIA DIAGNÓSTICA .2	1
SADT . ULTRASSONOGRAFIA DIAGNÓSTICA .1	1
CONSULTA . CIRURGIA VASCULAR .1	1
CONSULTA .NEFROLOGIA .2	1
CONSULTA .NEFROLOGIA .1	1
CONSULTA . CLINICA MEDICA .1	1
CONSULTA .OFTALMOLOGIA .3	1
CONSULTA . CIRURGIA GERAL .4	1

Fonte: O autor.

3.4.4 **Aplicação dos algoritmos**

Foram aplicados nos dados de pesquisa o algoritmo de associação APRIORI e os algoritmos de classificação Random Forest e J48. A seguir serão descritos os processos executados na aplicação destes algoritmos e quais foram os resultados encontrados.

3.4.4.1Algoritmo de Associação APRIORI

O algoritmo de associação APRIORI foi aplicado no Dataset através da linguagem R no RStudio utilizando o pacote “arules”. Os parâmetros especificados na execução estão descritos na Figura 8.

Figura 8: Parâmetros APRIORI.

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.6    0.1    1 none FALSE            TRUE       5     0.6      2     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE
```

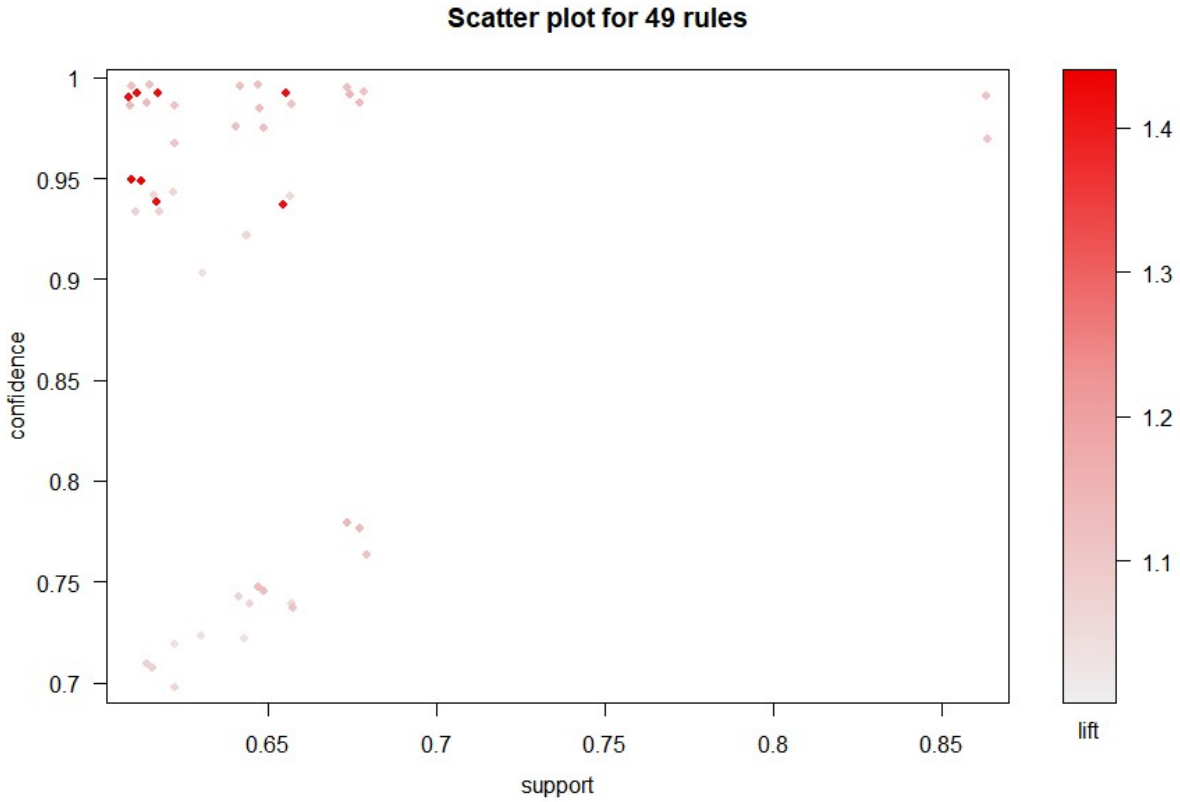
Fonte: A autora.

As regras de associação encontradas foram, em sua grande maioria, referente a exames, já que a variedade solicitada de exames gera uma ocorrência quantitativa maior que consultas, ambulatório e internações. Os grupos de maior ocorrência foram:

- BIOQUÍMICA (SANGUE, URINA E OUTROS MATERIAIS);
- HEMATOLOGIA LABORATORIAL;
- URINÁLISE;
- IMUNOLOGIA.

Foram encontradas algumas ocorrências de ambulatório também. As regras criadas, com o suporte acima de 0.6 (Figura 9), trouxeram apenas exames referentes aos doze meses antecedentes à primeira hemodiálise, com sua maioria referente ao primeiro semestre.

Figura 9: Regras com suporte 0.6.

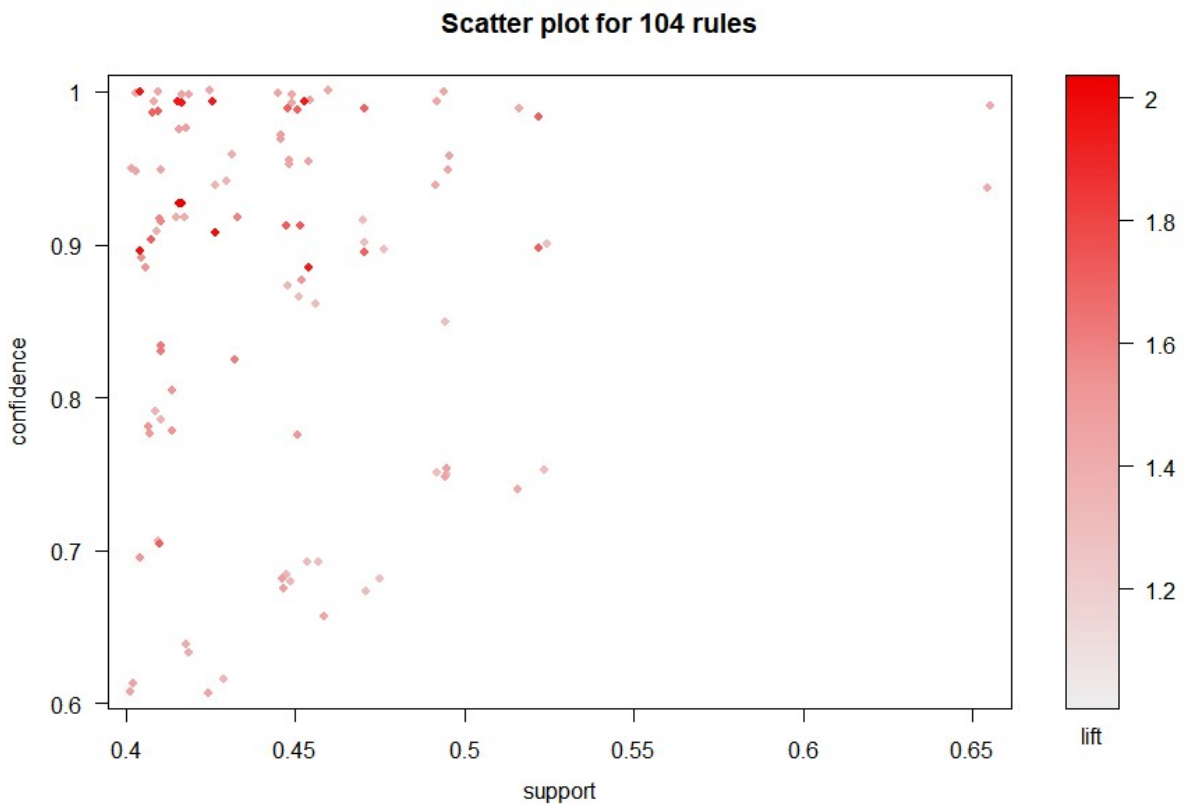


Fonte: A autora.

Assim,o algoritmo foi aplicado novamente sem considerar os dados do primeiro semestre imediatamente anterior à hemodiálise e com os parâmetros de suporte e confiança superior a 0.4.

Nas novas regras encontradas, os grupos se mantiveram os mesmos, mas agora apresentando ocorrência em todos os semestres com exceção do grupo Urinálise, que, nesta nova execução, continuou apresentando ocorrência até o segundo semestre anterior à hemodiálise. Neste grupo de exames, podemos encontrar o exame de Microalbuminúria utilizado na detecção de problemas renais, conforme ilustrado na Figura 10.

Figura 10: Regras com suporte 0.4.



Fonte: A autora.

O algoritmo de associação foi aplicado também para um dataset contendo apenas consultas e internações, mas o suporte encontrado foi muito baixo.

3.4.4.2Algoritmo de Classificação Random Forest

Para a aplicação do algoritmo de classificação Random Forest, foram utilizados datasets de treino com oitenta por cento dos dados e de teste com vinte por cento dos dados. As classes foram definidas como “-1” para pacientes sem problemas renais e “1” para pacientes dependentes de hemodiálise.

O algoritmo de classificação Random Forest foi aplicado nos Datasets através da linguagem R no RStudio utilizando o pacote “randomForest”. No treino, o algoritmo classificou corretamente 99,1% dos pacientes sem problemas renais e 63,8% dos pacientes dependentes de hemodiálise. No teste, o modelo gerado no algoritmo Random Forest classificou corretamente 93% dos pacientes sem

problemas renais e 95,7% dos pacientes renais. As Figuras 11-13 ilustram os resultados obtidos.

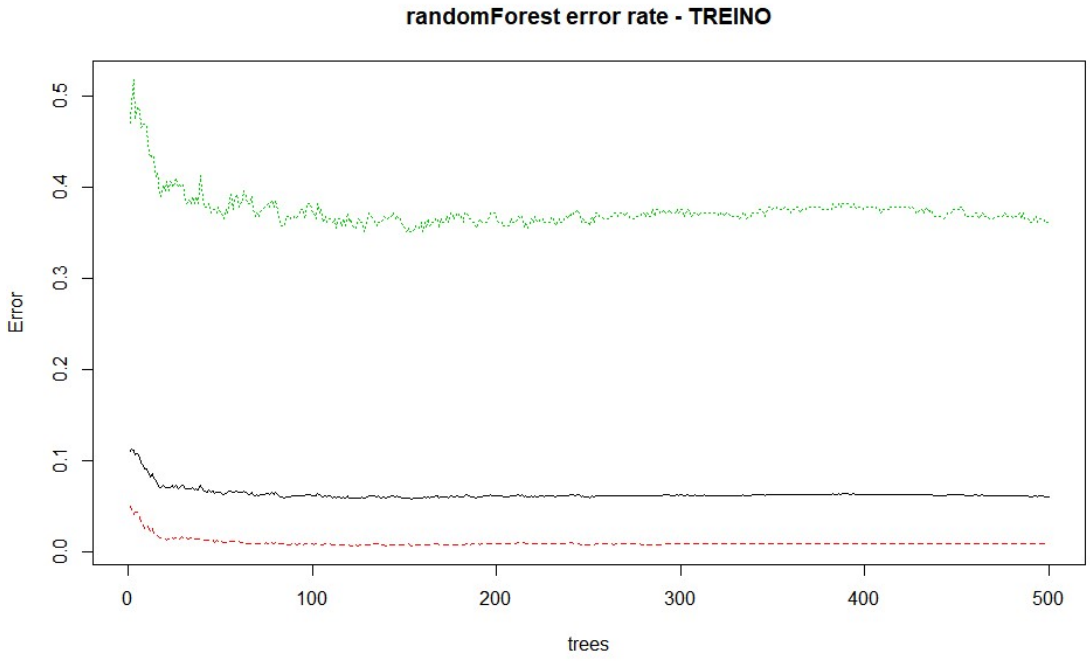
Figura 11: Treino do Algoritmo Random Forest.

```
Call:
 randomForest(formula = Classe ~ ., data = trainData, importance = TRUE,      proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 23

        OOB estimate of  error rate: 6.07%
Confusion matrix:
     -1   1 class.error
-1 1672  15 0.008891523
1   105 185 0.362068966
```

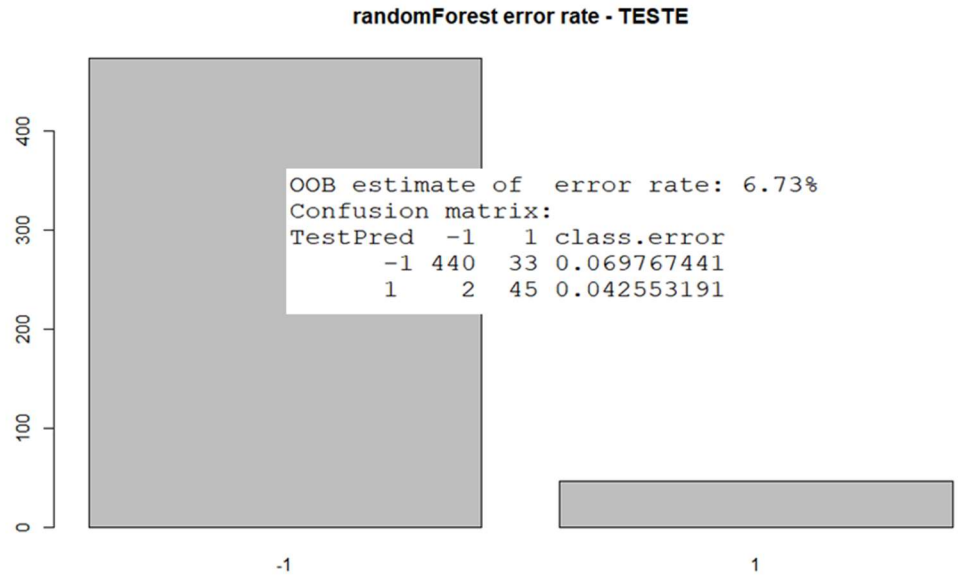
Fonte: A autora.

Figura 12: Taxa erro do treino -Random Forest.



Fonte: A autora.

Figura 13: Teste do modelo criado -Random Forest.



Fonte: A autora.

O modelo preditivo criado pelo algoritmo Random Forest conseguiu um percentual muito bom na classificação de ambos os tipos pacientes.

3.4.4.3Algoritmo de Classificação J48

Para a aplicação do algoritmo de classificação J48, também foram utilizados datasets de treino com oitenta por cento dos dados e de teste com vinte por cento dos dados. As classes foram definidas como “-1” para pacientes sem problemas renais e “1” para pacientes dependentes de hemodiálise.

O algoritmo de classificação J48 foi aplicado nos Datasets através da linguagem R no RStudio utilizando o pacote “RWeka”. No treino, o algoritmo J48 classificou corretamente 97,66% dos pacientes sem problemas renais e 93,98% dos pacientes dependentes de hemodiálise, com 2,83% de erro total. No teste, o modelo gerado no algoritmo J48 classificou corretamente 94,13% dos pacientes sem problemas renais e 85% dos pacientes renais com 6,92% de erro total. As Figuras 14 e 15 ilustram o resultado deste algoritmo.

Figura 14: Treino do Algoritmo J48.

```
=== Summary ===

Correctly Classified Instances        1921               97.1674 %
Incorrectly Classified Instances        56                2.8326 %
Kappa statistic                          0.8828
Mean absolute error                      0.0502
Root mean squared error                  0.1585
Relative absolute error                 20.0417 %
Root relative squared error             44.7904 %
Total Number of Instances             1977

=== Confusion Matrix ===

    a    b   <-- classified as
 1671   16 |    a = -1
   40  250 |    b = 1
```

Fonte: A autora.

Figura 15: Teste do modelo criado - J48.

```
> testPred <- predict(Principalj48, newdata = testData)
> table(testPred, testData$Classe)

testPred  -1   1
      -1 433  27
       1   9  51
```

Fonte: A autora.

O modelo preditivo criado pelo algoritmo J48 também conseguiu um percentual muito bom na classificação de ambos os tipos pacientes.

4 CONCLUSÃO

O algoritmo de associação não apresentou um resultado satisfatório neste trabalho de pesquisa. A maneira como o dataset foi modelado e os dados escolhidos não foram suficientes para encontrar regras associativas que ajudassem na seleção de pacientes para a prevenção da doença renal crônica.

Em contrapartida, os algoritmos de Classificação Random Foreste J48, foram ambos capazes de criar bons modelos preditivos para detecção de pacientes que estariam desenvolvendo a doença renal crônica. No teste dos modelos, o algoritmo Random Forest conseguiu classificar corretamente 95,7%, enquanto o algoritmo J48 conseguiu classificar corretamente 85% dos pacientes renais.

Então, a resposta para a pergunta do problema, apresentado no Capítulo 1 deste trabalho, se a mineração de dados poderia auxiliar na prevenção, identificando pessoas na fase inicial da DRC (Doença Renal Crônica) é sim. A mineração seria uma forma identificar pacientes que podem estar desenvolvendo a doença renal crônica e auxiliar as áreas, responsáveis por captar pacientes gerenciáveis, no processo de prevenção da doença.

Para trabalhos futuros, propõem-se novos estudos com regras de associação, utilizar o algoritmo APRIORI com outra modelagem de dados ou ainda utilizar outro algoritmo associativo que considere a quantidade de eventos. Também, é proposto continuar a pesquisa com outras doenças onde, a detecção prévia, ajudaria a minimizar complicações.

REFERÊNCIAS

AZEVEDO, R. O. S. **utilização de mineração de dados na descoberta de padrões de relacionamento entre fatores associados ao câncer oral**: um estudo sobre dados coletados na zona rural da Região do Seridó Potiguar. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação). Universidade Federal do Rio Grande do Norte, Caicó, 2018.

ALVAREZ, A. R Packages: A Beginner'sGuide. **Portal Datacamp**, 2019 Disponível em:.<https://www.datacamp.com/community/tutorials/r-packages-guide> Acesso em: 12/10/2019.

GONÇALVES, E. C.. **Data Mining de Regras de Associação** – Parte 1. 2012. https://www.devmedia.com.br/data-mining-de-regras-de-associacao-parte-1/6533

HECHANOVA, L. A. Diálise. **Portal Manual MSD**. Disponível em: <http://www.manualmerck.net/?id=149&cn=2106> Acesso em: 12/07/2019.

IHS. Saiba como cuidar bem dos rins. **Portal do Instituto de Hemodiálise de Sorocaba**, 2013. Disponível em: <http://www.ihs.med.br/noticias/saiba-como-cuidar-bem-dos-rins/20130507140107_A_402> Acesso em 15/08/2019.

MATOS, D. Uma Breve Introdução ao R. **Portal Ciência e Dados**. 2015. Disponível em: <http://www.cienciaedados.com/uma-breve-introducao-ao-r/> Acesso em: 12/10/2019.

MEDIUM. O Algoritmo floresta aleatória. **Portal Medium**. Disponível em: <https://medium.com/machina-sapiens/o-algoritmo-da-floresta-aleatória-3545f6babdf8> Acesso em: 15/15/2019.

MINISTÉRIO DA SAÚDE. Doenças de A a Z: Doenças Renais. **Portal do Governo Brasileiro. Brasil**. Disponível em: <http://www.saude.gov.br/saude-de-a-z/doencas-renais> Acesso em: 13/07/2019.

PACHECO, A. Classificação de dados. **PortalComputação Inteligente**. 2016. Disponível em: <http://computacaointeligente.com.br/artigos/classificacao-de-dados> Acesso em: 12/10/2019.

PAVÃO, O. Doença Renal Crônica. **Portal do Albert Einstein Sociedade Beneficente Israelita Brasileira**. 2012. Disponível em: <https://www.einstein.br/doencas-sintomas/doenca-renal-cronica> Acesso em: 12/08/2019.

PETERNELLI, L. A.; MELLO, M. P. **Conhecendo o R: uma visão mais que Estatística**, Editora UFV: Viçosa, 2013.

REZENDE, S. O. **Mineração de Dados**. 2005. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/enia/2005/0102.pdf>. Acesso em: 20/09/2019.

R CORE TEAM. **R LanguageDefinition**, 3.6.1, 2019. Disponível em: <https://cran.r-project.org/doc/manuals/r-release/R-lang.html> Acesso em: 12/10/2019.

ROMÃO JUNIOR; J. E. Doença renal crônica. **Jornal Brasileiro Nefrologia**, v. XXVI, n. 3, S. 1, 2004.

SOCIEDADE BRASILEIRA DE NEFROLOGIA. Doenças Comuns: Insuficiência Renal. **Portal da Sociedade Brasileira de Nefrologia**. Disponível em: <https://sbn.org.br> Acesso em: 15/04/2019.

SILVA, J. C. Aprendendo em uma Floresta Aleatória: Veja a Floresta e não as Árvores.**Portal Medium**, 2018. Disponível em: <https://medium.com/machina-sapiens/o-algoritmo-da-floresta-aleatória-3545f6babdf8> Acesso em: 15/10/2019.

TECHYV.Top10dataminingalgorithmsthatyoucaneasilyimplementonweka. **Portal Techvy**. Disponível em: <https://www.techyv.com/article/top-10-data-mining-algorithms-that-you-can-easily-implement-on-wekatanagra/> Acesso em: 15/10/2019.

VIEIRA, E. M. de A.; NEVES, N. T. de A. T.; OLIVEIRA, A. C. C. de; MORAES, R. M. de; NASCIMENTO, J. A. do. Avaliação da performance do algoritmo J48 para construção de modelos baseados em árvores de decisão. **Revista Brasileira de Computação Aplicada**, Vol. 10, N 2, pp. 80–90, JUL/2018.

WITTEN, I. H. **Data mining**: practicalmachinelearning tools andtechniques.3rd edition. Morgan KaufmannPublishers. 2011.

Vanessa RovidaLoschi

Taubaté, dezembro de 2019.