CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS
Recredenciado pela Portaria Ministerial nº 3.607, de 17/10/05, D.O.U. nº 202, de 20/10/2005
ASSOCIAÇÃO EDUCACIONAL LUTERANA DO BRASIL

Silas Gonçalves dos Reis

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS:
Um estudo de caso utilizando tarefas de predição aplicado a base de dados de germinação de sementes

Palmas – TO
2016

Silas Gonçalves dos Reis

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS:

Um estudo de caso utilizando tarefas de predição aplicado a base de dados de germinação de sementes

Trabalho de Conclusão de Curso (TCC) II, elaborado e apresentado como requisito parcial para obtenção do título de bacharel em Sistemas de Informação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).
Orientador: Prof. M.Sc. Fernando Luiz de Oliveira.
Co-orientadora: Prof. Dra. Conceição A. Previero.

Palmas – TO

2016

Silas Gonçalves dos Reis

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS:

Um estudo de caso utilizando tarefas de predição aplicado a base de dados de germinação de sementes

Trabalho de Conclusão de Curso (TCC) II, elaborado e apresentado como requisito parcial para obtenção do título de bacharel em Sistemas de Informação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. M.Sc. Fernando Luiz de Oliveira.
Co-orientadora: Prof. Dra. Conceição A. Previero.

Aprovado em: _____/_____/_____

BANCA EXAMINADORA

_____

Prof. M.Sc. Fernando Luiz de Oliveira

Orientador

Centro Universitário Luterano de Palmas – CEULP

_____

Prof. M.Sc. Cristina D'Ornellas Filipakis

Centro Universitário Luterano de Palmas – CEULP

_____

Prof. M.Sc. Madianita Bogo

Centro Universitário Luterano de Palmas – CEULP

Palmas – TO

2016

RESUMO

O Data Mining é uma etapa essencial do processo KDD, sendo o responsável por realizar o objetivo principal de todo processo, que é a descoberta do conhecimento em base de dados. Devido sua ampla aplicabilidade, busca-se neste trabalho aplicar as tarefas e técnicas de predição, presentes no Data Mining, em uma base de dados sobre germinação de sementes, cedidas pelo Laboratório de Pós-colheita de Produtos Agrícolas do Centro Universitário Luterano de Palmas (CEULP/ULBRA), a fim de predizer a taxa de germinação das sementes em novas situações. Para execução do projeto foi utilizada a metodologia do CRISP-DM, munida de suas seis fases que buscam a execução com sucesso de um projeto de mineração de dados para descoberta do conhecimento almejado.

LISTA DE ILUSTRAÇÕES

Figura 1 - Etapas do Processo KDD. 10
Figura 2 - Categorias das tarefas de Data Mining 14
Figura 3 - Árvore de Decisão Jogar Tênis 17
Figura 4 - DAG de uma Rede Bayesiana 18
Figura 5 - Estrutura de um Classificador Bayesiano em estrela 19
Figura 6 - Modelo de neurônio 20
Figura 7 - Fases do CRISP-DM 21
Figura 8 - Interface RapidMiner Studio 23
Figura 9 - Exemplo de operadores RapidMiner Studio 23
Figura 10 - Imagem da tabela de dados 27
Figura 11 - Imagem da nova tabela de dados 28
Figura 12 - Carregando dados 30
Figura 13 - Formato das colunas 31
Figura 14 - Dados prontos para mineração 32
Figura 15 - Modelo Preditivo 33
Figura 16 - Cross Validation 34
Figura 17 - Parâmetros Set Role 34
Figura 18 - Edição de Parâmetros 35
Figura 19 - Sub-Processo Cross Validation 35
Figura 20 - Tabela de performance 36
Figura 21 - Tabela usada para treinamento 37
Figura 22 - Descrição da Árvore de Decisão 38
Figura 23 - Árvore de Decisão 39
Figura 24 - Parâmetros Decision Tree 39

LISTA DE ABREVIATURAS E SIGLAS

KDD	*Knowledge Discovery in Databases*
BI	*Business Intelligence*
CEULP	Centro Universitário Luterano de Palmas
ULBRA	Universidade Luterana do Brasil
DM	*Data Mining*
CRIP-DM	*Cross Industry Standard Process for Data Mining*
RNA	Rede Neural Artificial
LBDES	Laboratório de Banco de Dados e Engenharia de Software
DAG	Gráfico Acíclico Direcionado
TCC I	Trabalho de Conclusão de Curso I

SUMÁRIO

INTRODUÇÃO **6**
REFERENCIAL TEÓRICO **9**
1.1 DADOS 9
1.2 PROCESSO KDD 10
1.3 DATA MINING 12
1.4 TAREFAS PREDITIVAS 14
1.4.1 Classificação **15**
1.4.2 Predição **15**
1.4.3 Regressão (Estimação) **16**
1.5 TÉCNICAS PREDITIVAS 16
1.5.1 Árvore de Decisão **16**
1.5.2 Classificação Bayesiana **18**
1.5.3 Redes Neurais **19**
1.6 CRISP-DM 20
METODOLOGIA **22**
1.7 DESENHO DO ESTUDO 22
1.8 MATERIAIS 22
1.9 PROCEDIMENTOS 24
RESULTADOS E DISCUSSÃO **26**
1.10 ENTENDIMENTO DO NEGÓCIO, BUSINESS UNDERSTANDING 26
1.11 ENTENDIMENTO DOS DADOS, DATA UNDERSTANDING. 26
1.12 PREPARAÇÃO DOS DADOS, DATA PREPARATION 28
1.13 MODELAGEM, MODELING 29
1.14 AVALIAÇÃO, EVALUATION, E IMPLANTAÇÃO, DEPLOYMENT 41
CONSIDERAÇÕES FINAIS **42**
REFERÊNCIAS **44**

INTRODUÇÃO

O crescimento expressivo das operações e atividades computacionais, devido à evolução acelerada da tecnologia, contribuiu para o aumento significativo do volume dos dados gerados pelas mais diversas plataformas disponibilizadas ao acesso pessoal ou empresarial. Segundo a EMC (2014), o crescimento dos dados e das informações digitais poderá chegar a marca de 1.6 Zettabytes (1.600 Exabytes[1]), isto somente no mercado brasileiro.

Os dados gerados das mais variadas formas configuram um aglomerado de dados tão vasto e volumoso que torna sua análise inviável por métodos tradicionais. Neste contexto, surge a necessidade de técnicas e ferramentas que possam auxiliar na análise destes dados. O Processo de Descoberta de Conhecimento em Banco de Dados, traduzido do inglês *Knowledge Discovery in Databases* (KDD), muitas vezes é confundido com a sua principal etapa, denominada de *Data Mining*, que pode automatizar parte desse trabalho.

Uma das definições acerca do *Data Mining* foi dita por Usama Fayyad (FAYYAD *et al*, 1996): "*Data Mining* é um processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis". Outros autores possuem definições semelhantes, como a de Berry e Linoff, (1997): "Mineração de dados é a exploração e a análise, por meio automático ou semiautomático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativos". As definições apresentadas por autores diferentes são próximas e entregam um conceito semelhante sobre a mineração de dados, ressaltando que não é algo simples e que o seu objetivo é entregar conhecimento às organizações.

A mineração de dados é composta por várias tarefas, dentre elas tem-se a previsão/predição. De acordo com Carvalho (2005), a técnica de *Previsão* "resume-se na avaliação do valor de algum índice baseando-se em dados do comportamento passado deste índice". A tarefa de predição/previsão pode ser usada com diversas ferramentas. Dentre elas, pose-se citar a Árvore de Decisão, Análise de Regressão, Redes Neurais etc.

A tarefa de predição/previsão pode ser aplicada em diversos segmentos como agropecuária, saúde, *Business Intelligence* (BI), entre outros. Diante disto, este estudo buscou responder se é possível, por meio de utilização de *Data mining*, prever a taxa de germinação de sementes em situações adversas através de dados de estudos anteriores, utilizando para isto os dados coletados pelo Laboratório de Pós-colheita de Produtos Agrícolas, do CEULP/ULBRA.

Com base nesse questionamento foi inferida a seguinte hipótese: aplicando-se as tarefas de *Data Mining* voltadas para a previsão de dados, com o objetivo de analisar dados de

[1] A unidade Exabyte é equivalente, em números aproximados, a 1.000 Petabytes, ou a 1.000.000 de Terabytes, ou ainda a 1.000.000.000 de Gigabytes.

pesquisas relacionadas à germinação de sementes, é possível predizer as taxas de germinação para outras situações e experimentos. Para chegar a confirmação da hipótese inferida o objetivo geral foi utilizar ferramentas que possibilitem a execução de tarefas e técnicas de *Data Mining* focadas na previsão de dados que permitam predizer a taxa de germinação em sementes de milho, utilizando como base para análise os dados coletados no Laboratório de Pós-colheita de Produtos Agrícolas, no Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Para atingir o objetivo proposto, foram estabelecidos objetivos específicos, que são: apresentar uma revisão bibliográfica do processo de descoberta do conhecimento em base de dados e *Data mining*; apresentar de forma eficiente a aplicação das tarefas e técnica de previsão de dados; apresentar de forma gráfica e escrita os resultados da aplicação das tarefas e técnicas de predição/previsão.

A relevância deste estudo é justificada de várias formas onde é possível verificar que os dados são de fundamental importância para as organizações, sendo que com os dados certos e a análise adequada sobre eles é possível abrir um leque de grandes possibilidades para empresas, entidades governamentais, não governamentais, entre outros. Por isto, quem dispõe de dados e, consequentemente, tem ferramentas que lhes auxiliam nas análises, possui meios para tomar as melhores decisões, traçar metas e definir melhores estratégias, desde que estes dados sejam bem analisados.

Para extrair informações e gerar conhecimento que seja precioso aos gestores, se faz necessário que os dados sejam submetidos ao processo de Descoberta de Conhecimento em Banco de Dados, ou simplesmente Processo KDD. Esse processo é composto por algumas fases e entre elas está a Mineração de Dados (*Data Mining*), etapa essencial do processo no qual são aplicadas tarefas para se extrair padrões nos dados.

Dentre as tarefas de *Data Mining* as mais utilizadas são a Classificação, Estimação, Previsão e Predição, Análise de afinidade, Análise de agrupamento ou *cluster* e Descrição. Neste trabalho objetivou-se aplicar as técnicas que levem a predição e previsão de dados das pesquisas realizadas sobre germinação de sementes no Laboratório de Pós-colheita de Produtos Agrícolas, no Centro Universitário Luterano de Palmas (CEULP/ULBRA). A tarefa de previsão/predição consiste na determinação do futuro de uma grandeza, por meio de dados do comportamento passado dessa grandeza. Essa tarefa tem recebido atenção de grandes companhias, indústrias, governos e estudiosos que desejam prever tendências através dos dados que são coletados e armazenados diariamente.

Os dados que foram estudados são de grande importância para a agricultura e podem auxiliar os grandes e pequenos produtores, já que as sementes são de fundamental importância

para a continuação de muitas espécies e, geralmente, são armazenadas por produtores fora da época de plantio, ou comerciantes que atendem a esses produtores. As formas de armazenamento que as sementes podem ser submetidas podem variar para cada caso, o que pode refletir diretamente na qualidade da semente, já que podem influenciar na germinação de um futuro plantio.

REFERENCIAL TEÓRICO

Nesta seção são apresentados os conceitos que serão aplicados no decorrer deste trabalho, enunciando a importância e o conceito sobre os dados num contexto geral. É apresentado de forma sucinta nas subseções a seguir o Processo KDD, que faz ligação ao objetivo principal deste trabalho que é a análise dos dados por meio de técnicas de *Data Mining*.

1.1 DADOS

O conhecimento e as informações estão cada vez mais difundidos, e a internet tem participação fundamental neste novo cenário. Com o amadurecimento das tecnologias de comunicação e disseminação de informações, outros meios de comunicações foram sendo desenvolvidos e aprimorados. Os novos e mais interativos meios de comunicação e informação (tais como aplicativos mensageiros, redes sociais, e-mails etc.) vieram principalmente com a evolução da internet. Foram tomando um espaço e atenção do público, que passou a ser gerador de muitos dados que podem se transformar em informação. Segundo Isotani *et al*. (2008), na internet todos usuários são produtores de informação, além disso a informação é criada de forma coletiva e não individual.

Essa conduta recebeu a atenção de pesquisadores, governos e principalmente de empresas competitivas, que buscam difundir seus produtos e estender seus espaços e lucros em um mundo cada vez mais competitivo. Segundo Araújo (2007), "a informação tem se tornado aliada na redução e antecipação de riscos e crises e no aumento da vantagem competitiva. ", para se ter a informação que pode levar a um conhecimento sobre algo, é necessário ter uma base de dados o que se tornou viável no panorama informatizado atual.

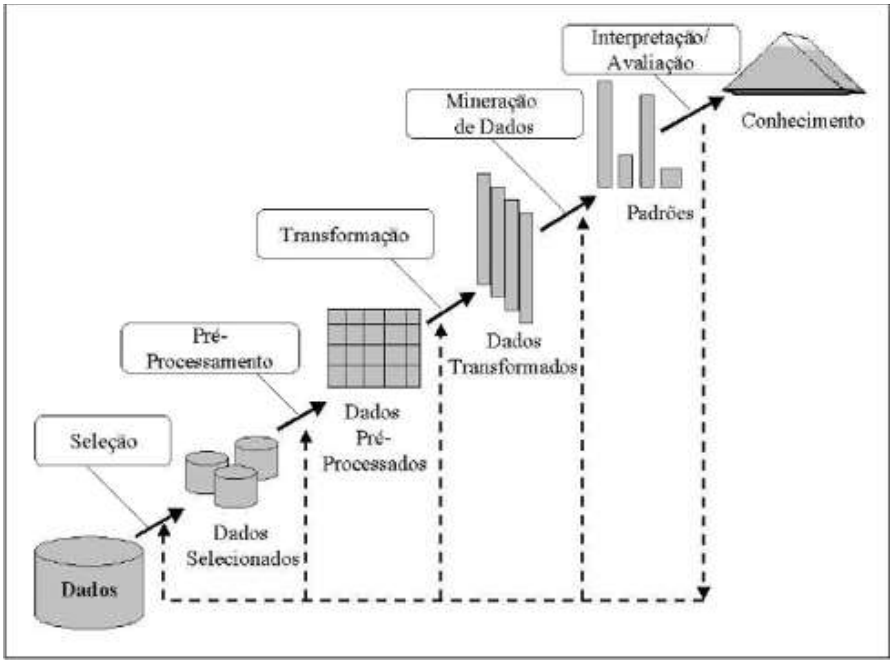
Dados, informação e conhecimento estão conceitualmente interligados de acordo com Resende (2005), "O dado é um elemento puro, quantificável sobre um determinado evento. Dados são fatos, números, texto ou qualquer mídia que possa ser processada pelo computador. [...] A informação é o dado analisado e contextualizado. Envolve a interpretação de um conjunto de dados". Resende (2005) afirma ainda que "a informação pode gerar conhecimento que ajude na análise de padrões históricos para conseguir uma previsão dos fatos futuros...".

Nesse contexto, para se entender os dados e gerar o conhecimento pode se fazer uso da mineração de dados, ou *Data Mining*, que por sua é parte do processo KDD, processo esse que será abordado mais detalhadamente na próxima seção.

1.2 PROCESSO KDD

Segundo Fayyad *et al*. (1996, p. 6), o "KDD é um processo, composto de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados". O processo KDD, cujas iniciais vêm do termo *Knowledge Discovery in Databases*, foi criado em 1989 no primeiro workshop sobre o tema, onde o objetivo era salientar que conhecimento é o produto final de uma descoberta baseada em dados (CAVALCANTE, 2014). O processo leva a extração de conhecimento útil aos seus detentores, por meio da extração de padrões de dados. O processo é interativo, pois há interferência humana na interpretação e na tomada de decisão, e é também iterativo, uma vez que pode haver repetições em todo processo ou em alguma das etapas que o compõe (FAYYAD *et al*, 1996). Quanto ao não trivial faz alusão a complexidade na execução do que o processo pode oferecer (BOENTE, 2008). Na Figura 1, a seguir, é apresentada a sequência das etapas do processo KDD (FAYYAD *et al*., 1996).

Figura 1 - Etapas do Processo KDD.



Fonte: FAYYAD et al. (1996, p. 10)

O processo KDD é composto por cinco etapas que levam o processo ao alcance do objetivo conforme apresentado na Figura 1. As etapas são: Seleção, Pré-Processamento, Transformação, Mineração de Dados e Interpretação/Avaliação.

A etapa inicial, **Seleção**, diz respeito a quais dados serão utilizados nas próximas etapas. Nessa etapa é necessário que o executor do processo já possua os objetivos definidos, e quais

tipos de conhecimento ele deseja extrair da base de dados. O resultado final da aplicação do processo KDD depende inicialmente desta etapa, pois a saída obtida vai depender de quais os dados foram selecionados, e nem todos os dados contidos nas bases de dados irão influenciar positivamente no objetivo do processo.

A etapa de **Pré-Processamento**, que pode ser nomeada como etapa de limpeza dos dados, tem por objetivo eliminar ruídos (dados fora dos padrões e erros), além de identificar e retirar valores inválidos, inconsistentes ou redundantes preparando os dados para a execução dos algoritmos da etapa de mineração de dados (THOMÉ, 2008). O pré-processamento é um processo semiautomático, sendo semiautomático ele depende do conhecimento do analista para identificar e eliminar os ruídos, valores inválidos e outras situações previamente citadas, que porventura podem estar presentes após a seleção dos dados. Segundo Batista (2003), o pré-processamento é tido como uma das tarefas mais trabalhosas e demoradas do KDD. Nesta fase despende-se cerca de 80% do tempo gasto em todo processo (PYLE, 1999). Sendo assim, requer uma atenção especial pois esta fase sendo bem aplicada culmina no sucesso ou fracasso de todo processo.

A terceira etapa, nomeada como **Transformação,** é a etapa onde os dados, pré-processados, passam por uma transformação que os agrupem em um formato adequado e os armazene de forma apropriada para a próxima etapa (CONTECSI, 2015). Semelhante a fase anterior, acontece também a limpeza e pré-processamento final dos dados. Por meio de critérios estabelecidos é feito o tratamento de dados faltantes e atributos ausentes para que os dados sejam levados a fase de mineração dos dados, *Data Mining*. Muitas organizações possuem dados de plataformas diferentes, em formatos diferentes. Por exemplo, dados sobre o sexo de uma pessoa que em uma base pode estar como “masculino”, pode estar apenas como “M”, ou como “Homem” em outra. Enfim, estes dados devem ser padronizados para um único formato, até para não haver confusão na aplicação da ferramenta de mineração.

Na sequência, ocorre a **Mineração dos Dados** ou *Data Mining.* É uma fase que tem grande importância para se obter um conhecimento útil dos dados, sendo considerada o núcleo do processo KDD (CONTECSI, 2015). Nesta etapa, os padrões serão encontrados utilizando as técnicas dos algoritmos envolvidos. Por ser uma fase de grande importância para o processo e para este trabalho esta etapa será melhor abordada na seção seguinte.

A quinta fase do processo KDD é a de **Interpretação e Avaliação**, que também é conhecida como pós-processamento, e ocorre no final da mineração dos dados. Esta fase consiste em visualizar os padrões e informações extraídos na fase anterior, estando incumbida também da organização e apresentação do conhecimento obtido.

Alguns autores acrescentam outras fases ao processo KDD, algumas dessas mudanças na organização das fases ou em sua nomenclatura devem-se ao fato da origem dos dados. Muitas organizações possuem bases de dados descentralizadas, tornando necessário a utilização de um *Data Warehouse* para unificar os dados (AMO, 2004). Não é o caso desse estudo por este motivo está etapa não será mencionada adequadamente. No entanto a fase do processo que terá maior ênfase devido a sua importância para o processo e sucesso deste trabalho é o *Data Mining*

e suas técnicas e algoritmos mais utilizados, que serão apresentados nas próximas seções e subseções.

1.3 DATA MINING

O *Data Mining* (DM) faz parte do processo KDD sendo a principal de suas etapas. Da primeira a terceira etapa do processo KDD, ou seja, da seleção até a transformação, os dados são preparados para o *Data Mining*. A mineração de dados pode ser muitas vezes confundida com processo KDD, isto acontece, pois ela pode tomar para si o protagonismo de todo processo do qual faz parte, já que é nesta etapa que os dados podem apresentar as novas relações. A mineração de dados é mais que parte do processo que converte dados brutos em informações úteis.

São muitas definições de *Data Mining* e algumas delas serão apresentadas, como a de Silva (2000), que afirma que "Mineração de Dados é técnica para determinar padrões de comportamento, em grandes bases de dados, auxiliando na tomada de decisão". Outra definição, sugerida por Carvalho (2005), é "definimos *datamining* como o uso de técnicas automáticas de exploração de grandes massas de dados de forma a descobrir novos padrões e relações que devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano. " Já Fayyad et al (1996), afirmar que *Data Mining* é a, "extração de conhecimento de base de dados (mineração de dados) é o processo de identificação de padrões validos, novos, potencialmente úteis e compreensíveis embutido nos dados. "

Há outras definições propostas por diversos autores, porém, estas são suficientes para ter um entendimento sobre a definição do DM. Embora diferentes as definições seguem um contexto semelhante sobre a MD que sinteticamente é, descobrir conhecimento de forma automatizada podendo ser supervisionada, em dados que podem estar relacionados ou não e entregar um novo conhecimento ou apresentar novas tendências.

Tem-se no processo de mineração conceitos importantes que são as *técnicas* e *tarefas* de mineração, que são escolhidas de acordo com o tipo de conhecimento almejado (CAVALCANATE, 2014). A diferenciação destes conceitos pode garantir um melhor entendimento das fases do DM, e assim prevenir que no projeto possa ter complicações na fase de planejamento da mineração.

É importante diferenciar **técnicas** de **tarefas** de mineração. Araújo (2009) afirma que "Tarefas de mineração de dados estão relacionadas às perguntas feitas na etapa de seleção dos dados [...]", as perguntas que o autor se refere são o que se deseja encontrar nos dados, são os

objetivos do processo. Relacionando as perguntas (objetivos) com os dados, tem-se a tarefa ou tarefas adequadas, sendo que, a partir da tarefa, é determinada a técnica ou técnicas a serrem aplicadas.

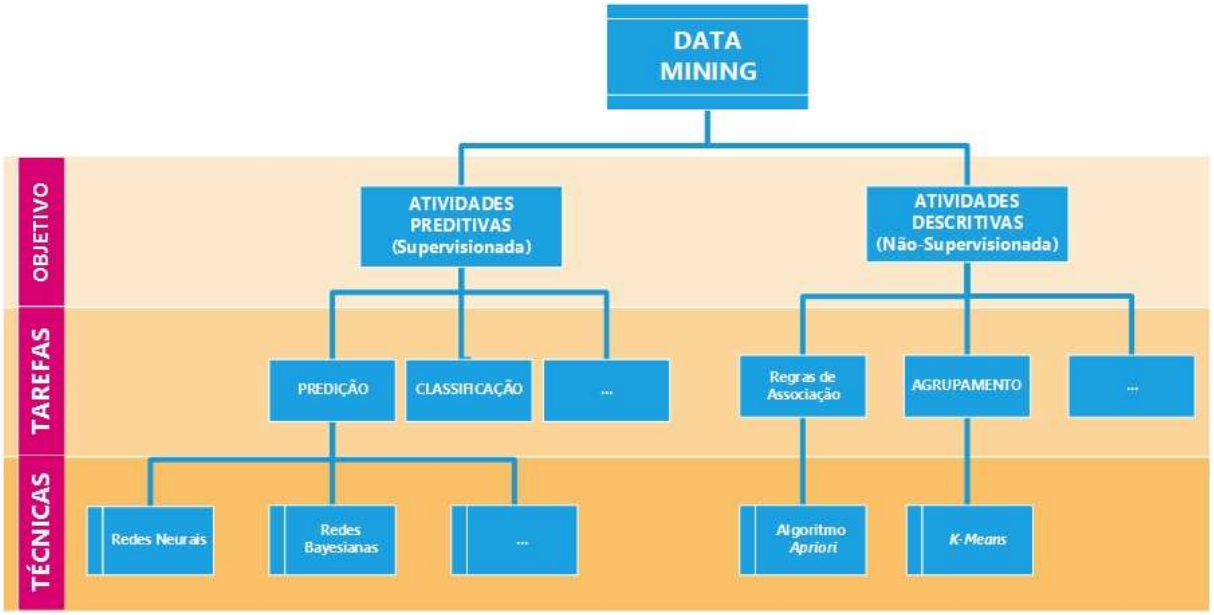
Segundo Amo (2004), técnica de mineração consiste na especificação dos métodos, que podem ser variados. Esses métodos são dispostos na forma de algoritmos computacionais que automatizam a técnica ou técnicas de DM. Na mineração não há uma técnica que resolva todos os problemas, e pode haver à aplicação de várias técnicas para um único problema de MD (DIAS, 2008). Sobre as técnicas de DM pode-se citar as mais comuns e abundantemente aplicadas para solução de vários problemas de DM são elas: Redes Neurais, Algoritmos Genéticos, Árvore de Decisão etc.

Em uma técnica podem ser usados vários métodos (ou algoritmos) que irão otimizar a descoberta de conhecimento. A exemplo da diversidade de métodos que podem ser aplicados em uma só técnica tem-se para a técnica de Redes Neurais os seguintes algoritmos: *Perceptron, Rede Counterpropagation, Rede* PNN*,* entre outros. Já para Algoritmos Genéticos é possível citar os seguintes; *CHC, Genitor, Algoritmo Genético* simples etc. E para Árvore de Decisão tem-se; *CART, SLIQ, Sprint* entre outros.

Tan et al (2009) afirma que as tarefas de DM geralmente são divididas em duas categorias principais que são: **tarefas de previsão**, que tem por objetivo prever os valores de um determinado atributo baseando-se em outros atributos, executada de forma supervisionada[2]; e **tarefas descritivas**, de natureza exploratória, tem por objetivo encontrar padrões, tendências, nos dados, não supervisionada[3] (CAVALCANTE, 2014). Na Figura 2 é possível verificar a divisão das tarefas e algumas técnicas que podem ser usadas de acordo com a tarefa.

[2] "Esta categoria de algoritmos possui esta denominação porque a aprendizagem do modelo é supervisionada, ou seja, é fornecida uma classe à qual cada amostra no treinamento pertence" (DOS SANTOS SILVA, 2004).
[3] "Nestes algoritmos o rótulo da classe de cada amostra do treinamento não é conhecido, e o número ou conjunto de classes a ser treinado pode não ser conhecido a priori, daí o fato de ser uma aprendizagem não-supervisionada" (DOS SANTOS SILVA, 2004).

Figura 2 - Categorias das tarefas de *Data Mining*



Fonte: Adaptado de BRANQUINHO (2015).

São apresentadas na Figura 2 as categorias do DM, sendo que as técnicas mudam conforme o objetivo a ser alcançado com o uso da mineração. Assim, se o objetivo for predizer algo, ou seja, o uso do DM para Atividades Preditivas usa-se as tarefas de Predição, Regressão ou a Classificação, sendo que as técnicas que visam este objetivo são variadas, e serão abordadas posteriormente. Para atividades Descritivas usa-se tarefas como Regra de Associação e Agrupamento entre outras e possuem técnicas diversas entre elas pode-se citar a Algoritmo *Apriori* e *K-Mens*.

Neste trabalho foram abordadas as tarefas e técnicas que tenham o objetivo de predizer através dos dados. E, por este motivo, as próximas subseções abordam esse tema.

1.4 TAREFAS PREDITIVAS

Segundo Rezende (2005), “as tarefas preditivas ou Mineração de Dados preditivo, consistem na generalização de exemplos ou experiências passadas com respostas conhecidas em uma linguagem capaz de reconhecer a classe de um novo exemplo. Os dois principais tipos de tarefas para predição são classificação e regressão”. As principais tarefas preditivas são a Classificação e a Regressão ou Estimação, porém alguns autores consideram a Predição como tarefa própria do DM. A Predição muito se assemelha à Classificação, diferenciando-se pelo fato de que na Predição os registros são classificados de acordo com alguma atitude a ser prevista (KREMER, 1999).

1.4.1 Classificação

Segundo Kremer (1999), “A Classificação é uma técnica que consiste na aplicação de um conjunto de exemplos pré-classificados para desenvolver um modelo capaz de classificar uma população maior de registros”. Tornou-se uma das tarefas mais usadas do *Data Mining* devido a sua aplicabilidade em diversos cenários, contribuiu também, a facilidade no entendimento desta tarefa, já que desde pequeno as pessoas aprendem a classificar objetos, pessoas, alimentos e etc.

Nesta tarefa os dados são inseridos e classificados previamente gerando regras, a partir dos atributos dos dados. Novos dados são analisados com base nas regras das classes, gerando como saída dados que podem ser da mesma classe e que possuem atributos semelhantes aos utilizados para definir as regras ou que são de classes diferentes. O objetivo da classificação é encontrar alguma relação entre os atributos de um conjunto de dados e uma classe, de forma que consiga predizer a classe de um novo objeto desconhecido.

A classificação consiste em obter um modelo baseado em um conjunto de exemplos que descrevem uma função não-conhecida. Esse modelo é então utilizado para predizer o valor do atributo-meta de novos exemplos (REZENDE, 2005). As técnicas mais utilizadas para a classificação são: árvores de decisão, classificação bayesiana e redes neurais. Entre outras aplicações, esta tarefa pode ser usada para detecção de fraudes e aplicações de risco e classificar pedidos de créditos como de baixo, médio e alto risco (KREMER, 1999).

1.4.2 Predição

A Predição, também chamada de Previsão. Kremer (1999), afirma que ela é uma variante do problema de agrupamento por afinidades, onde as regras encontradas entre as relações podem ser usadas para identificar sequências interessantes que serão utilizadas para predizer acontecimentos subsequentes. Conforme Carvalho (2005), “A previsão resume-se na avaliação do valor futuro de algum índice baseando-se em dados do comportamento passado deste índice”. Muito semelhante a classificação esta tarefa se diferencia, porque os registros são classificados de acordo com alguma atitude futura prevista (KREMER, 1999).

A predição não implica exclusivamente na previsão de um valor futuro, como afirma Souza (2003 *apud* PRADO, 1998), “a característica importante é que ela faz uma adivinhação educada sobre o valor de um ou mais atributos desconhecidos, dados os valores de outros atributos conhecidos”. A predição se baseia na média do número de acertos dos casos testado, daí é estabelecido se a predição é interessante ou não, para ser interessante os acertos devem estar acima da média. As técnicas mais utilizadas na predição são: árvores de decisão e redes

neurais. A predição pode ser aplicada para previsão da quantia de dinheiro que um cliente utilizará caso seja oferecido a ele um certo limite de cartão crédito entre outros cenários (KREMER, 1999).

1.4.3 Regressão (Estimação)

Conforme Fayyad *et al* (1996), " a regressão compreende a busca por uma função que mapeie um item de dado para uma variável de predição real". O objetivo dessa tarefa é encontrar uma função alvo que possa ajustar os dados com mínimo de erro possível (TAN *et al.* 2009). Possui grande similaridade com a classificação, mas se restringe apenas a valores numéricos outra característica que a diferencia é que o valor a ser predito é continuo em vez de discreto (CAVALCANTE, 2014).

Também conhecida como predição de valor real, predição funcional ou mesmo aprendizados de classes continuas, esta tarefa já é bastante estudada pela comunidade estatística, porém para *Data Mining* a maioria das pesquisas são voltadas para classificação (CASTANHEIRA, 2008). Na regressão os métodos geram modelos, que também podem ser conhecidos como repressores, esses modelos expressam o conhecimento obtido durante o processo de mineração (REZENDE, 2005).

A Regressão pode ser usada para estimar o número de filhos de uma família, a renda total de uma família, a demanda de um novo produto o tempo de vida de um cliente. Na regressão ou estimação a predição e feita baseando-se em dados semelhantes a situação a ser prevista (CASTANHEIRA, 2008). Para Amorim (2006), A arte de estimar é exatamente esta: determinar da melhor forma possível um valor, baseando-se em outros valores de situações semelhantes.

1.5 TÉCNICAS PREDITIVAS

A Mineração de Dados (MD) também envolve a utilização de diversas técnicas, materializada por algoritmos computacionais, necessárias para realizar as tarefas de mineração (CAVALCANTE, 2014). Entre as diversas técnicas utilizadas em DM destaca-se para atividades de predição as técnicas listadas nas seções seguintes.

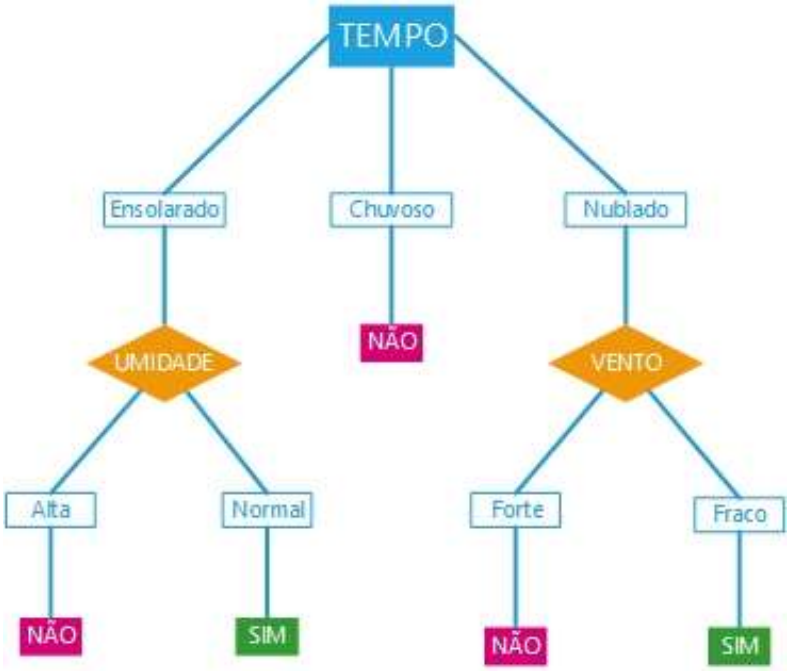
1.5.1 Árvore de Decisão

Segundo Rabelo (2007), árvore de decisão é uma técnica que utiliza a recursividade para o particionamento da base de dados. Cada nó não terminal desta árvore representa um teste ou

decisão sobre o item de dado. A abordagem dessa técnica é basicamente dividir para conquistar, levando para o tema proposto seria dividir para predizer. O objetivo da Árvore de Decisão é separar as classes, ela faz isso por meio de regras de classificação do tipo SE-ENTÃO (AMORIM, 2006).

Na árvore cada nó filho representa uma condição, e os nós folhas são as conclusões, as tomadas de decisão na árvore envolvem os atributos de maior relevância e isto é uma vantagem para a técnica. Outra vantagem é a facilidade no entendimento da mesma pois seus atributos são compreensíveis pela maioria das pessoas (CAVALCANTE, 2014). Um problema apresentado pela técnica da árvore de decisão é a necessidade de uma grande quantidade de dados para trabalhos complexos (AMORIM, 2006). Na Figura 3 é apresentada um exemplo de Árvore de Decisão.

Figura 3 - Árvore de Decisão Jogar Tênis



A árvore apresentada na Figura 3 apresenta o esquema básico de uma Árvore de Decisão. A árvore exemplifica o processo de decisão para saber se é possível jogar tênis em um dia qualquer. Na Árvore de decisão cada percurso é uma regra, se os valores dos atributos do dia em teste atender a algum valor da árvore, isto é, satisfaz á alguma regra da árvore, então ao final tem se a resposta se é possível ou não jogar tênis no dia.

Segundo Cavalcante (2014), “Para realizar uma classificação através de árvore de decisão é necessário dividir o conjunto de dados em dois conjuntos: conjunto de dados de teste e conjunto de dados de treinamento”, sendo que um é usado para construir o modelo a ser

utilizado. Este é o conjunto de treinamento e nele são inseridos os rótulos para os registros conhecidos. Já o outro é usado para testar os modelos gerados. No caso, neste não se insere rótulos, pois o objetivo é testar as árvores geradas.

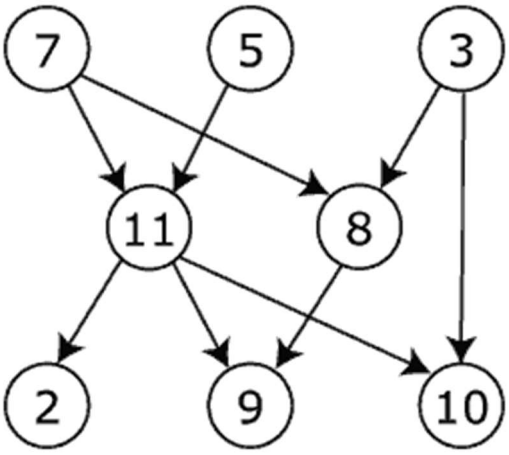
Os algoritmos que implementam esta técnica são: CART, CHAID, C5.0, Quest, ID-3, SLIQ, SPRINT (DIAS, 2001).

1.5.2 Classificação Bayesiana

A Classificação Bayesiana é uma técnica usada há bastante tempo e usa uma abordagem probabilística para associar uma classe a um objeto (MAIA, 2005). A técnica é baseada no teorema de Bayes. Camilo (2009), afirma que, com teorema de Bayes, é possível encontrar a probabilidade de um certo evento ocorrer, dada a probabilidade de um outro evento que já ocorreu: Probabilidade (B dado A) = Probabilidade (A e B) / Probabilidade (A).

Uma rede bayesiana é representada por um grafo, apresentado na Figura 4, mais especificamente o grafo acíclico direcionado.

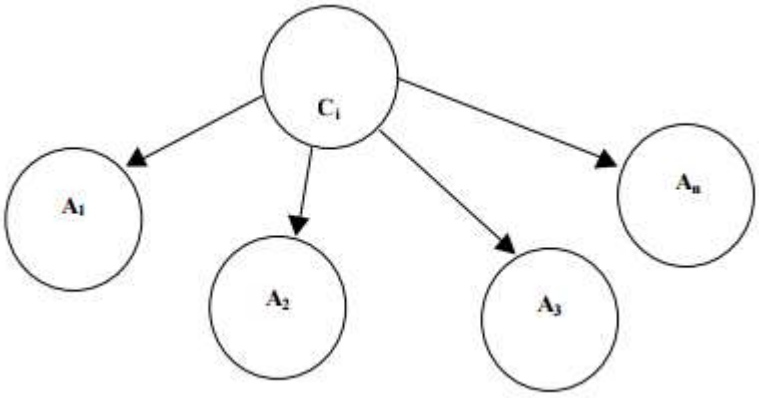
Figura 4 - DAG de uma Rede Bayesiana



Fonte: MAIA (2005)

Na Rede Bayesiana, como a representada na Figura 4, os nós representam as variáveis conectadas e os arcos representam uma dependência condicional entre os nós (MELLO, 2002). Os classificadores bayesianos assim como as redes bayesianas são representados por grafos, porém, a diferença é que os grafos que representam os classificadores são postos em forma de estrela, conforme representado na Figura 5.

Figura 5 - Estrutura de um Classificador Bayesiano em estrela



Fonte: Mello (2002)

O grafo em estrela apresentado na Figura 5 representa um classificador bayesiano, onde ao centro vai a classe a ser classificada, representada pela letra (C_1), e nas pontas são representados os atributos das classes (A_1 a An), que são as únicas conexões permitidas para a rede *Naïve Bayes*.

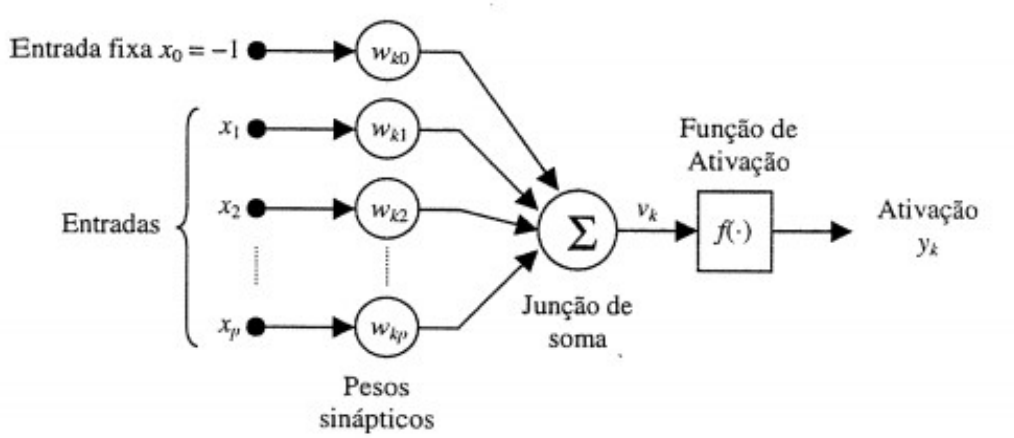
Comparando os algoritmos Bayesianos, também chamados de *Naïve Bayes*, com outras técnicas como de Árvore de Decisão, apresentada no item 1.5.1, e Redes Neurais, que é apresentada a seguir no item 1.5.3, detectou que o *Naïve Bayes* é compatível, em termos de precisão com os mesmos (CAMILO, 2009). Os Classificadores Bayesianos possuem alto poder preditivo e conta ainda com características que o fazem ser amplamente aplicado, que é a simplicidade, a rapidez e a fácil implementação (MAIA, 2005).

1.5.3 Redes Neurais

A rede neural é uma das técnicas mais utilizadas no *Data Mining*. Isso se deve a sua capacidade de se adequar as diferentes estratégias de DM. Segundo Cavalcante (2014), as redes neurais representam uma metáfora do cérebro humano para o processamento da informação. Estes modelos são biologicamente inspirados, uma vez que funcionam como réplicas do nosso cérebro.

Também conhecida como Rede Neural Artificial (RNA), a técnica mostra-se promissora para sistemas de previsão e classificação em suas diversas possibilidades de aplicações (CAVALCANTE 2014). A rede neural é composta por um número de elementos interconectados, denominados neurônios, apresentado graficamente na Figura 6.

Figura 6 - Modelo de neurônio



Fonte: Iyoda (2000)

No neurônio é possível identificar três elementos principais, as *sinapses* ou conexões de entrada, que são ponderadas pelos *pesos sinápticos*; a *Junção Soma* responsável pala junção das entradas acertadas; e a função de ativação, que processa a saída do neurônio a partir dos valores da *Junção da Soma*. A função de ativação é uma função de ordem interna e tem efeito sobre o próprio neurônio. A função de ativação mais utilizada é a função Linear, porém, há outras funções de ativação como a Esférica, Mahalanobis e Polinomial (THOMÉ, 2008). O processo de calibração dos pesos é conhecido como aprendizado ou treinamento, que é repetido até que a saída resulte na informação desejada (CAVALCANTE, 2014). Há uma série de modelos de RNA's como: Perceptrom, Rede MLP, Redes de Koholen etc.

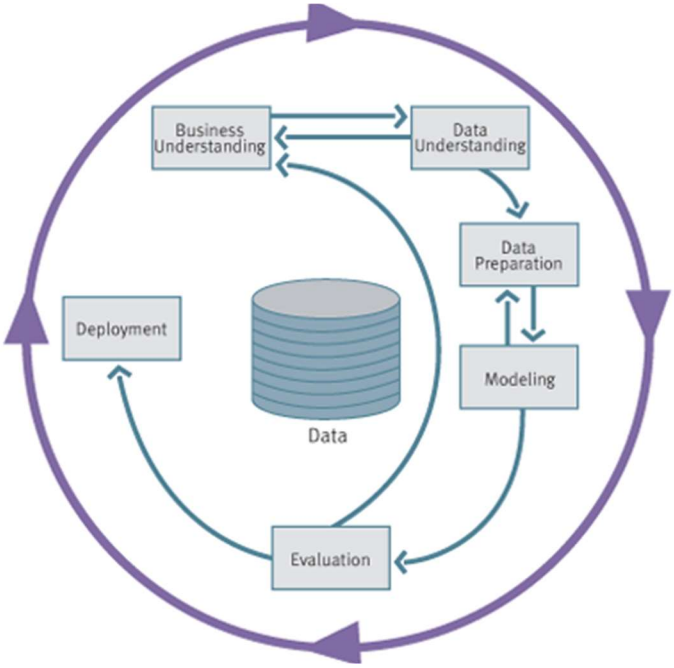
Na seção 1.6 é apresentado a metodologia para trabalhos envolvendo mineração de dados, trata-se de um conjunto de boas práticas para DM.

1.6 CRISP-DM

CRISP-DM é um acrônimo para *Cross - Industry Standard Process for Data Mining* que pode ser traduzido como Processo Padrão Inter-Indústrias para Mineração de Dados. O CRISP-DM é uma metodologia para implementação de projetos de DM e ajuda na resolução de problemas típicos de DM. Em 1996 quando a MD ainda era pouco conhecida, mas já chamava a atenção, um grupo encabeçado por Daimler-Benz, a Integral Solutions Ltd, a NCR e a OHRA que viram no cenário de DM a necessidade de criar um manual de boas práticas para a mineração de dados (GOMES, 2011). Uma das exigências do grupo é que esse manual fosse independente das industrias, gratuito e que pudesse ajudar as industrias no desenvolvimento dos seus projetos em DM.

Em 2000 foi lançada a versão 1.0 do CRISP-DM, que veio fiel a filosofia do processo KDD. O CRISP-DM padroniza os passos para a descoberta do conhecimento e pode ser usado em projetos independentemente de sua área de aplicação, seja na saúde, comércio, etc. A Figura 7 apresenta as fases do CRISP-DM.

Figura 7 - Fases do CRISP-DM



Fonte: Adaptado de (NOGUEIRA, 2014)

O CRISP-DM foi projetado para fornecer orientação para iniciantes em MD e fornece um modelo genérico para necessidades de quaisquer ramos dos utilizadores (Empresas) sendo uma das metodologias mais utilizadas para projetos de DM, do início até a conclusão de um projeto de *Data Mining*, que segue esta metodologia, se faz necessário a passagem por seis fases diferentes. As fases que compõem o CRISP-DM serão abordadas apropriadamente no item 1.9 que se encontra na próxima seção, e que trata dos procedimentos a serem executados no decorrer deste trabalho. Apresenta ainda, a finalidade e a abordagem deste estudo bem como os objetivos e procedimentos para realização do trabalho.

METODOLOGIA

1.7 DESENHO DO ESTUDO

Este estudo é um trabalho de pesquisa aplicada de natureza quantitativa, que visou aplicar tarefas de predição em uma base de dados sobre germinação de sementes. Possui uma característica exploratória pois busca através da análise do problema empregar a mineração de dados voltada para predição na germinação de sementes em novas situações. Com procedimentos metodológicos bibliográficos de estudo de caso direcionados na mineração de dados observando as tarefas e técnicas preditivas.

1.8 MATERIAIS

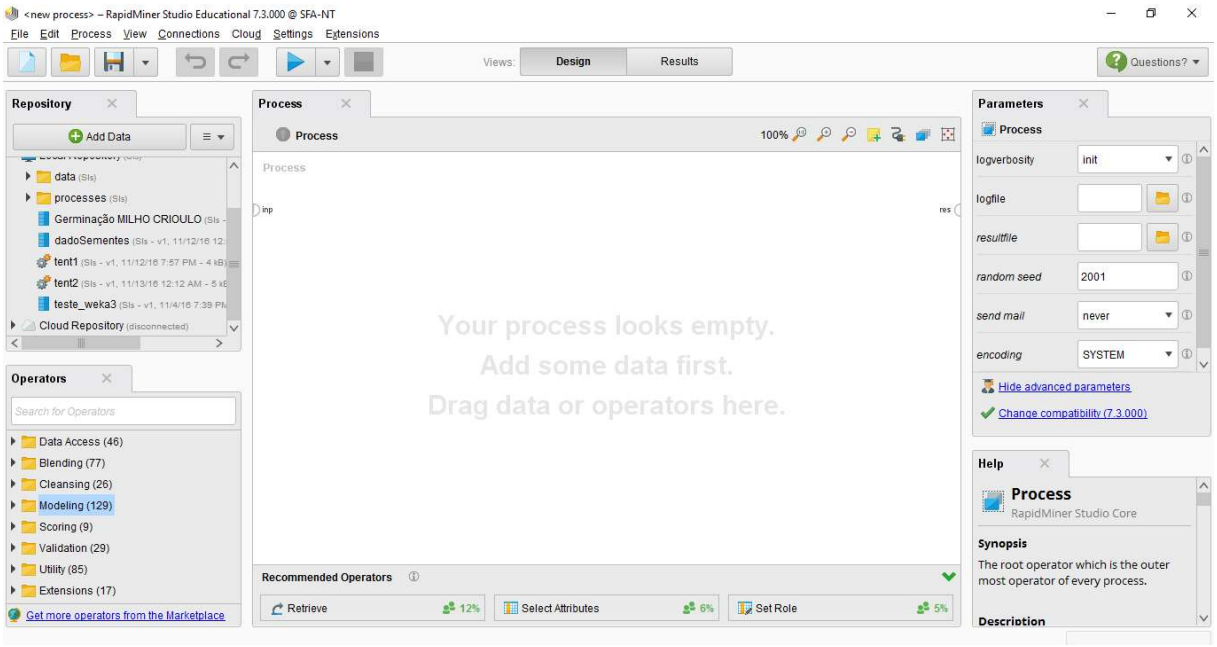
Para elaboração deste estudo foi efetuado o levantamento do referencial teórico, tendo como base diversas fontes bibliográficas oriundas de diversos artigos, teses, livros e dissertações. Esses materiais possibilitaram obter um conhecimento sobre o processo KDD e suas etapas, possibilitaram ainda um entendimento mais adequado sobre *Data Mining* e suas tarefas e técnicas que levam a predição, sendo essas descritas previamente no item 1.4 e 1.5 do referencial teórico.

Posteriormente, foram analisadas algumas ferramentas que trabalham com *Data Mining*. Existem uma variedade de ferramentas disponibilizadas no mercado, entre as principais e mais conhecidas principalmente no meio acadêmico destacam-se a **RapidMiner Studio** e **Weka**, **SAS Enterprise Miner Suíte** e **Clementine**, este último possui suporte ao processo CRISP-DM, que foi a metodologia a ser usada no desenvolvimento do trabalho, entre outras opções disponíveis no mercado de análise de dados.

Mediante as funcionalidades e o desempenho apresentado ante outras ferramentas, para execução do trabalho foi escolhida a **RapidMiner Studio**[4] na versão 7.3. Esta plataforma de software oferece um ambiente integrado para aprendizagem de máquina, mineração de dados, mineração de texto, análise preditiva e análise de negócios. A ferramenta foi selecionada por necessitar de um período de aprendizado menor em relação a outras ferramentas semelhantes disponíveis no mercado, além de possuir uma grande variedade de operações, poder ser usada para fins educacionais, comerciais e industriais e suporta todas as etapas de um projeto de DM, desde a fase inicial até a apresentação dos resultados. Na Figura 8, abaixo, é apresentada a interface padrão da ferramenta.

[4] https://rapidminer.com

Figura 8 - Interface RapidMiner Studio



O RapidMiner busca oferecer uma interface que facilite o entendimento e o uso de seus componentes apresentando graficamente todo processo da mineração, como pode ser visualizado na Figura 8, contendo funcionalidades como importação e exportação de dados e conta ainda com um conjunto de mais de 500 operadores que podem ser combinados com um sistema *Drag & Drop* (arrastar e soltar) sem a necessidade de programação. Em relação ao uso dos operadores a Figura 9, a seguir, apresenta um exemplo dessas funcionalidades em uso.

Figura 9 - Exemplo de operadores RapidMiner Studio

Os operadores podem ser utilizados sem a necessidade do conhecimento de seu processo interno, como um sistema caixa preta. No exemplo apresentado na Figura 9, é utilizado o operador ***Retrieve*** que faz o acesso a uma base de dados do repositório, neste caso a base de dados **Titanic**. O operador está ligando sua saída a entrada do operador de modelagem preditiva ***Decision Tree*** (Árvore de Decisão), que apresentará em sua saída uma árvore de decisão dos dados do Titanic.

O **RapidMiner Studio** oferece ainda a possibilidade de se relacionar o projeto em execução com algumas linguagens de programação como java, python e R, além de outras ferramentas de mineração como o WEKA. Outro ponto positivo a ser observado na ferramenta,

é que a mesma oferece ao usuário a possibilidade de criar seus próprios operadores de acordo com sua necessidade.

A próxima seção trata dos procedimentos que foram adotados para desenvolvimento e execução do projeto de *Data Mining* abordando suas principais tarefas, que forneceram durante a consumação do projeto a compreensibilidade e operacionalidade no cumprimento da descoberta de conhecimento em base de dados.

1.9 PROCEDIMENTOS

Para o desenvolvimento do projeto foram seguidas as etapas propostas pela metodologia CRISP-DM, apresentada brevemente na seção 1.6. Ela é composta por seis fases que foram graficamente apresentadas na Figura 7. A escolha desta metodologia foi baseada na apresentação de etapas com saídas concretas e voltadas especificamente para DM, além de ser uma metodologia madura e amplamente utilizada que serve tanto para iniciantes como experientes em DM.

Seguindo a metodologia proposta pelo CRISP-DM, o projeto será composto por 6 fases:

- A fase 1 do projeto refere-se ao Entendimento do Negócio, ***Business Understanding***, na qual acontecerá a definição dos objetivos, constituiu em efetuada uma avaliação preliminar nos dados, recursos disponíveis para projeto serão verificados e será determinada a tarefa ou tarefas de mineração que serão aplicadas no projeto. Esta fase visa "colocar as cartas na mesa" e avaliar quais os passos seguir para descoberta do conhecimento.
- A fase 2 consiste no Entendimento dos Dados, ***Data Understanding***, nela ocorreu a manipulação dos dados. Ainda nesta fase, foi realizada a descrição dos dados, verificando quanto a qualidade dos dados e foram efetuadas as limpezas iniciais sobre os dados. Neste trabalho não houve a necessidade, porém nesta fase é possível retornar à fase anterior para alterar quaisquer discordâncias do projeto com os dados, que podem ocorrer por haver dados que são irrelevantes para o projeto ou que não podem oferecer resultados compatíveis com o objetivo firmado na fase 1.
- Na fase 3, Preparação dos Dados, ***Data Preparation***, foi onde aconteceu uma maior manipulação dos dados. Esta etapa constituiu na execução da preparação dos dados para serem importados para a ferramenta de mineração, nela os dados foram efetivamente limpos de ruídos (dados fora dos padrões e erros) e *outliers*. Efetuou-se ainda na fase 3, a retirada de dados irrelevantes ao objetivo do projeto e, caso seja necessário, poderá haver a integração com outras bases de dados.

- Após a fase anterior, iniciou-se a Modelagem, ***Modeling***. Nesta fase, houve a seleção da técnica de DM que foi utilizada. Aqui na fase 4 foram escolhidos os dados que tiveram a incumbência de serem utilizados para o treinamento, validação e testes, caso seja necessário desta fase pode haver um retorno a fase anterior.
- Na fase 5, a Avaliação, ***Evaluation***, foi feita a conferência dos resultados obtidos, isto é, os dados minerados, após a aplicação da modelagem. Nesta etapa foi efetuada a avaliação da MD realizada analisando se a mesma foi satisfatória aos objetivos firmados na fase 1 e se os parâmetros utilizados na modelagem foram eficientes. Se na avaliação os resultados confrontados com os casos já relatados e com a base de dados de testes forem expressivamente negativos, é possibilitado o retorno a fase inicial, fase 1, para alterar todo projeto de DM.
- Baseando-se nos resultados obtidos na fase anterior, o projeto chegou a sua etapa final, a fase 6, ***Deployment***, ou Implantação, que tratou da entrega do produto, ou ainda dos resultados obtidos. No caso deste estudo foram entregues os resultados de forma gráfica e textual em forma de relatórios, além de ser efetuada a apresentação do projeto executado.

RESULTADOS E DISCUSSÃO

A presente seção tem por objetivo apresentar os resultados obtidos no trabalho. Seguindo a cronologia de fases do CRISP-DM, serão apresentados todos os desdobramentos resultantes da pesquisa onde serão demonstradas de forma concisa as fases seguidas e os procedimentos e ferramentas adotados no processo que levaram a predição dos dados sobre a germinação das sementes de milho.

1.10 ENTENDIMENTO DO NEGÓCIO, *BUSINESS UNDERSTANDING*

No CRISP-DM inicia-se um trabalho pelo entendimento do negócio. Nesta fase buscou-se estabelecer as metas e traçar os objetivos a serem alcançados. Neste trabalho, esta fase iniciou-se ainda na elaboração de proposta onde foram imputados os primeiros e principais objetivos. Foram também avaliadas as situações de seguimento do trabalho, bem como sua viabilidade. Diante da proposta aceita, iniciou-se o pré-projeto deste trabalho, denominado especificamente para este tipo de trabalho como Trabalho de Conclusão de Curso I (TCC I).

No projeto foram inseridas todas as informações necessárias para o andamento e embasamento do projeto, apresentado a hipótese de solução, os problemas gerais e específicos e a justificativa do trabalho proposto. O estudo sobre os dados e tarefas e técnicas de mineração, que constam no referencial teórico, acarretou em conhecimentos necessários para dar embasamento e determinar as ações a serem seguidas na busca pela predição dos dados.

Foram inseridas ainda no projeto as informações sobre as atividades a serem trabalhadas e o cronograma a ser seguido no decorrer do projeto, bem como as ferramentas a serem utilizadas e se as mesmas atendiam aos objetivos propostos. Diante do pré-projeto (TCC I) pronto, entregue e aprovado passou-se para o trabalho prático que foi a realização de todo projeto, seguindo o cronograma e a metodologia proposta, as próximas subseções abordaram a prática e consolidação de um projeto de *Data mining.*

1.11 ENTENDIMENTO DOS DADOS, *DATA UNDERSTANDING.*

Após a finalização do projeto (TCC I), foi iniciada a execução do trabalho na prática, sendo que nesta fase aconteceu o primeiro contato com os dados a serem analisados e minerados. Estes dados foram cedidos pelo Laboratório de Pós-colheita de Produtos Agrícolas, no Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Os dados são parte de pesquisas e experimentos realizados no referido laboratório e contêm informações sobre a germinação das sementes de milho em três (3) tipos de armazenamento em épocas diferentes, que aconteceram por 300 dias ou 10 meses. A Figura 10,

a seguir, representa graficamente a tabela de dados cedidos pelo Laboratório de Pós-Colheita do CEULP/ULBRA, recebidos em arquivo do Excel (*.exe).

Figura 10 - Imagem da tabela de dados

	A	B	C	D	E	F	G
8							
9	Amostra	Sementes	Sementes Anormais		Mortas	Percentual(%)	
10		Normais	Danificadas	Infeccionadas			Média
11	A1 -a	48	1		1		
12	A1 -b	48			2	96	
13	A2 -a	48			2		
14	A2 -b	48		2		96	
15	A3 -a	47		1	2		
16	A3 -b	47	1	1	1	94	
17	A4 -a	44	1	3	2		
18	A4 -b	47	2		1	91	94
19	B1 -a	44		1	5		
20	B1 -b	48		1	1	92	
21	B2 -a	47		1	2		
22	B2 -b	46	1		3	93	
23	B3 -a	48			2		
24	B3 -b	48			2	96	
25	B4 -a	49			1		
26	B4 -b	45		2	3	94	94
27	C1 -a	47		2	1		
28	C1 -b	48		2		95	
29	C2 -a	49		1			
30	C2 -b	47		3		96	
31	C3 -a	49			1		
32	C3 -b	45		4	1	94	

TEMPO 0 | 1ªEpoca | 2ªEpoca | 3ªEpoca | 4ªEpoca | 5ªEpoca | grafico

Os resultados dos testes realizados nas sementes, foram tabulados e colocados em planilha (Figura 10). As amostras são materiais genéticos de sete espécies de milho, Al Bandeirante, Al Avaré, Cati Verde 02, Saracura, Sol da Manhã e o Hibrido representados na tabela por letras do alfabeto que vão do A ao G e estão contidos na coluna A. Cada amostra foi submetida a tratamentos e a testes semelhantes realizados no início e a cada sessenta dias. Nos testes, as sementes poderiam ser classificadas em normais, danificadas, infeccionadas e mortas. Os testes foram executados por trezentos dias ou dez meses e iniciaram-se no tempo zero (0), percorrendo as épocas de 1 a 5, cada uma com um período de sessenta dias.

Munido dos dados e com o auxílio do especialista no domínio representado pela Bióloga Doutora em Pós-Colheita de Produtos Agrícolas, Coordenadora de pesquisa do

CEULP/ULBRA, Conceição A. Previero, foi possível ter um maior entendimento dos rótulos dos dados e seus significados para a pesquisa. Na seção a seguir é apresentada a preparação dos dados para sua utilização na ferramenta.

1.12 PREPARAÇÃO DOS DADOS, *DATA PREPARATION*

Após o entendimento dos dados, auxiliado por um especialista no domínio, os mesmos foram submetidos a preparação. Para execução da fase de modelagem (próxima etapa), esta preparação se faz necessária pois nas tabelas podem conter dados irrelevantes ao objetivo a ser alcançado. Outra tarefa que pode ser necessária é a inclusão de novos dados que enriquecem o conjunto de dados possibilitando uma maior amplitude de resultados e maior confiança na predição alcançada.

Os dados cedidos eram organizados originalmente conforme Figura 10, apresentada anteriormente, onde os mesmos também eram divididos em 5 planilhas diferentes que representavam as épocas de testes realizados nos estudos. No entanto, para melhor adequação dos dados para a ferramenta foi necessária a junção de todas as tabelas de dados em apenas uma tabela, apresentada em forma de imagem, exibida na Figura 11

Figura 11 - Imagem da nova tabela de dados

	A	B	C	D	E	F	G	H	I	J
1	Amostra	Sementes Normais	Sementes Danificadas	Sementes Infeccionadas	Sementes Mortas	Total de Sementes	TEMPO	EMBALAGEM	Temperatura	Umidade Relativa
2	A1 -a	48	1		1	50	0	NO	27,51	72,18
3	A1 -b	48			2	50	0	NO	27,51	72,18
4	A2 -a	48			2	50	0	NO	27,51	72,18
5	A2 -b	48		2		50	0	NO	27,51	72,18
6	A3 -a	47		1	2	50	0	NO	27,51	72,18
7	A3 -b	47	1	1	1	50	0	NO	27,51	72,18
8	A4 -a	44	1	3	2	50	0	NO	27,51	72,18
9	A4 -b	47	2		1	50	0	NO	27,51	72,18
10	B1 -a	44		1	5	50	0	NO	27,51	72,18
11	B1 -b	48		1	1	50	0	NO	27,51	72,18
12	B2 -a	47		1	2	50	0	NO	27,51	72,18
13	B2 -b	46	1		3	50	0	NO	27,51	72,18
14	B3 -a	48			2	50	0	NO	27,51	72,18
15	B3 -b	48			2	50	0	NO	27,51	72,18
16	B4 -a	49			1	50	0	NO	27,51	72,18
17	B4 -b	45		2	3	50	0	NO	27,51	72,18
18	C1 -a	47		2	1	50	0	NO	27,51	72,18
19	C1 -b	48		2		50	0	NO	27,51	72,18
20	C2 -a	49		1		50	0	NO	27,51	72,18
21	C2 -b	47		3		50	0	NO	27,51	72,18

Como é possível verificar, a estrutura da tabela é a mesma, porém, houve retirada dos dados com as informações de percentual e média para que realmente houvesse a predição sobre

esses dados. Também foi efetuada a adição de outros dados que não constavam nas tabelas originais, tais como:

- A coluna **Dias**, que retrata o tempo de teste em dias a iniciar no tempo zero com zero dias.
- A coluna **Armazenamento**, que contém o tipo de armazenamento de cada amostra. Neste caso, no tempo zero não houve armazenamento ficando assim representado na tabela de dados como NO (não). Já nos tempos subsequentes dos dados, as amostras foram armazenadas em PET, PAPEL e POLIETILENO.

Nesta fase de preparação também foram adicionadas as colunas **Temperatura** e **Umidade Relativa**, que foram preenchidas com os dados obtidos no resumo expandido sob título: **Influência da infestação de insetos na viabilidade de sementes de milho tipo variedade armazenadas em condição de ambiente natural** apresentado no XLII Congresso Brasileiro de Engenharia Agrícola - CONBEA 2013, que usou como base os mesmos dados cedidos para este trabalho pelo laboratório de Pós-Colheita do CEUP/ULBRA. Conforme o artigo escrito por Previero, Santos e Gonçalves (2013), "As médias de temperatura registradas durante o período de armazenamento foram de 27,51°C, com máxima de 28,30°C e mínima de 25,86°C e umidade relativa média de 72,18%, com máxima de 81,96% e mínima de 54,48%. ". Como os dados da Temperatura e Umidade não foram disponibilizados de forma individual para cada amostra, tendo em vista que estes valores poderiam apresentar uma melhora nos dados, foi necessária a inserção dos valores utilizando a média de Temperatura e Umidade Relativa, disponibilizada no artigo para todas as amostras, visando aproximar os dados da realidade da pesquisa.

Após a adequação dos valores e a inserção e exclusão dos dados mencionado acima, a nova tabela foi finalizada para ser inserida na ferramenta de mineração, ferramenta esta que será melhor detalhada na seção seguinte denominada Modelagem, que descreve todo processo de uso da ferramenta desde a inserção dos dados até a execução das tarefas e técnicas que proporcionaram alcançar os objetivos.

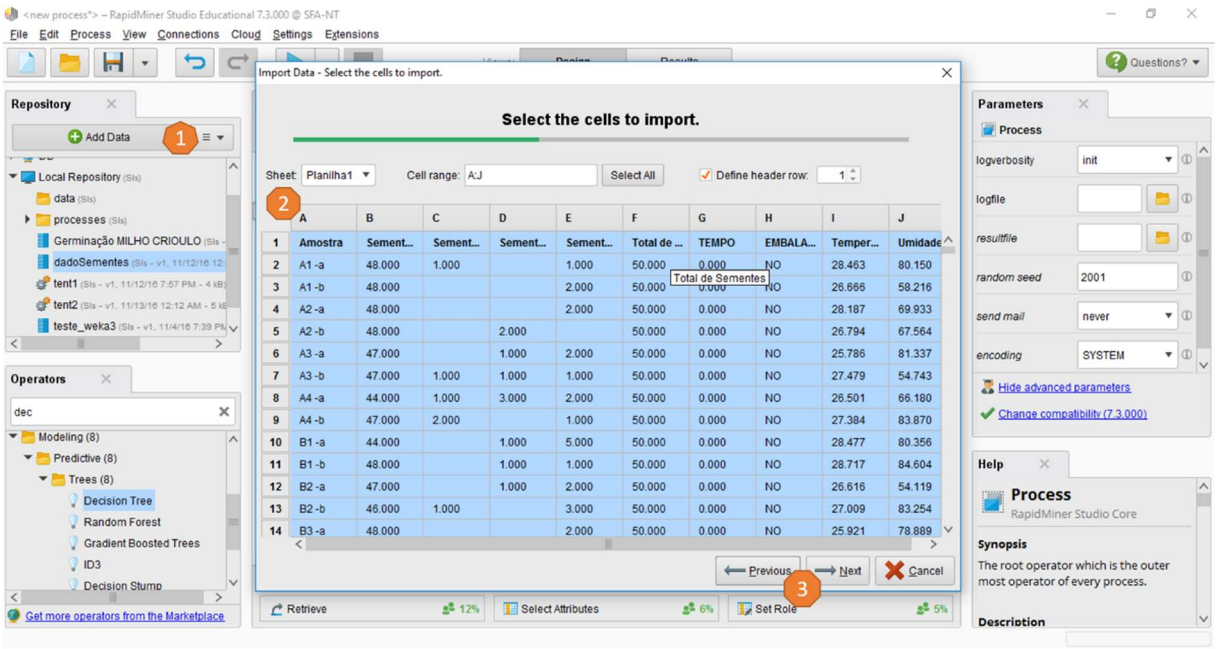
1.13 MODELAGEM, M*ODELING*

Esta fase, denominada pelo CRISP-DM de *Modeling* ou modelagem, preocupa-se com a construção e execução de "modelos", nomenclatura atribuída a esta etapa não por casualidade, mas pelo fato de que na mineração de dados é possível aplicar o mesmo conjunto de técnicas a vários conjuntos de dados, como um modelo, podendo ser aplicado para várias soluções. Usa-se nesta fase a ferramenta apresentada na seção 1.8 em conjunto com as técnicas mencionadas

no item 1.5. O processo de construção pode ser repetindo várias vezes até que se apresente o modelo ideal para mineração, nesta fase podem acontecer testes de eficiência do modelo.

Após o conhecimento básico da ferramenta foi iniciada a inserção dos dados no **RapidMiner Studio**. Na Figura 12, a seguir, são apresentados os passos (numerados) para o carregamento dos dados para a ferramenta.

Figura 12 - Carregando dados

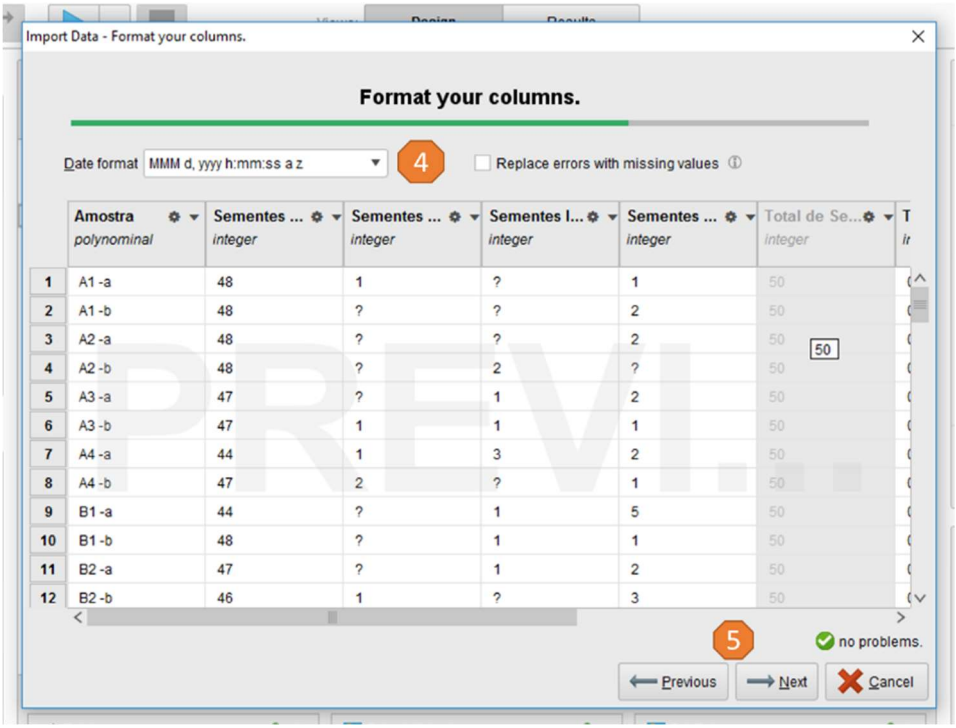


	A	B	C	D	E	F	G	H	I	J
1	Amostra	Sement...	Sement...	Sement...	Sement...	Total de ...	TEMPO	EMBALA...	Temper...	Umidade
2	A1 -a	48.000	1.000		1.000	50.000	0.000	NO	28.463	80.150
3	A1 -b	48.000			2.000	50.000	0.000	NO	26.666	58.216
4	A2 -a	48.000			2.000	50.000	0.000	NO	28.187	69.933
5	A2 -b	48.000		2.000		50.000	0.000	NO	26.794	67.564
6	A3 -a	47.000		1.000	2.000	50.000	0.000	NO	25.786	81.337
7	A3 -b	47.000	1.000	1.000	1.000	50.000	0.000	NO	27.479	54.743
8	A4 -a	44.000	1.000	3.000	2.000	50.000	0.000	NO	26.501	66.180
9	A4 -b	47.000	2.000		1.000	50.000	0.000	NO	27.384	83.870
10	B1 -a	44.000		1.000	5.000	50.000	0.000	NO	28.477	80.356
11	B1 -b	48.000		1.000	1.000	50.000	0.000	NO	28.717	84.604
12	B2 -a	47.000		1.000	2.000	50.000	0.000	NO	26.616	54.119
13	B2 -b	46.000	1.000		3.000	50.000	0.000	NO	27.009	83.254
14	B3 -a	48.000			2.000	50.000	0.000	NO	25.921	78.889

Para que os dados fossem carregados para o **RapidMiner Studio** foi necessário um clique em *Add Data* (Figura 12 - 1), onde então foi redirecionado para escolha de uma base de dados externa ou na própria máquina. Os dados selecionados são apresentados no formato de tabela (Figura 12 - 2), sendo possível verificar se todas as células poderem ser importadas. Esta escolha fica a cargo do usuário, que é o responsável por definir quais são os dados a serem utilizados. Na mesma tela o sistema ofereceu a opção de retroceder ou escolher novos dados, seguir adiante ou cancelar a operação (Figura 12 - 3).

Posteriormente, após ter sido efetuado o carregamento dos dados e a escolha das células a serem usadas na mineração, o **RapidMiner Studio** apresenta os dados pré-processados oferecendo várias opções de formatação das células selecionadas, conforme apresentado na Figura 13, adiante, sendo também possível efetuar a formatação de atributos com tipos de dados diferentes, conforme a necessidade do projeto.

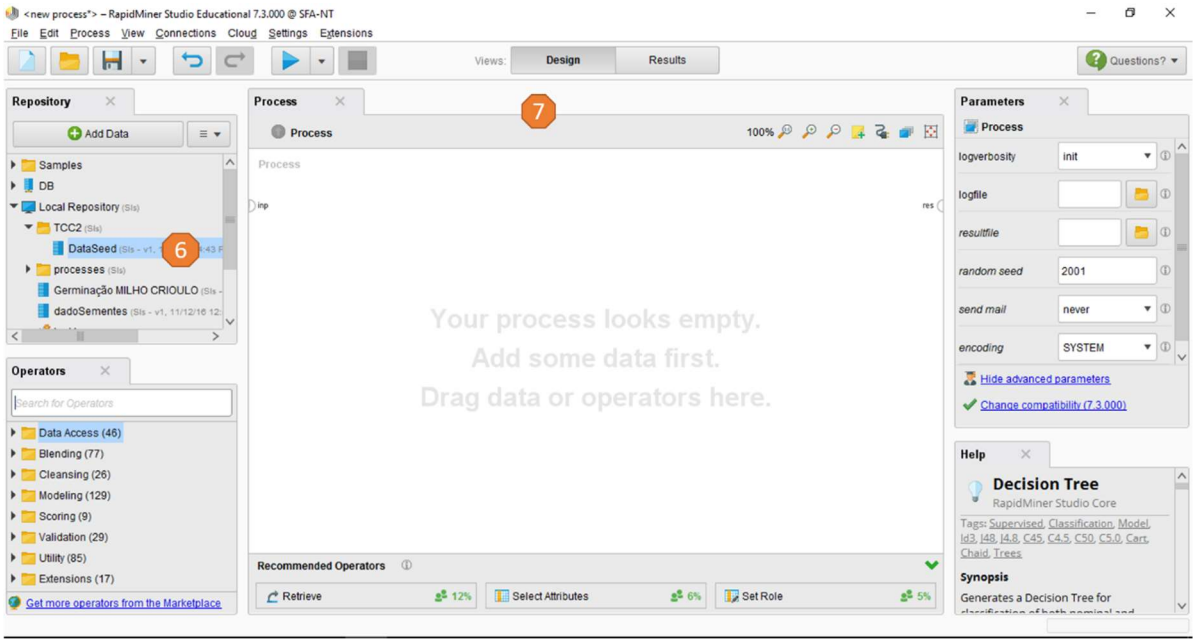
Figura 13 - Formato das colunas



Na Figura 13, a imagem exibida mostra que cada coluna pode ser renomeada e alterada individualmente, modificando os formatos dos dados originais para outros formatos, que podem ser inteiros, polinomiais, reais e outros (Figura 13 - 4). Ainda é ofertada a possibilidade de efetuar a exclusão de colunas que não contribuam para um bom resultado na mineração, como ocorreu com a coluna **Total de Sementes**. Após finalizadas as alterações, foi possível verificar que no canto inferior a direita há um indicador de estado dos dados para conversão, após a exibição da mensagem de confirmação foi permitido que os dados fossem salvos no formato suportado (Figura 13 - 5), o formato utilizado pela ferramenta **RapidMiner Studio** é o *.ioo.

As telas apresentadas anteriormente mostram os passos para importação dos dados para a execução no **RapidMiner Studio**. Porém, antes de prosseguir é necessário criar um local para o armazenamento dos dados importados e os processos executados na ferramenta. No caso, para este projeto foi criada uma pasta denominada TCC2, conforme apresentado na Figura 14.

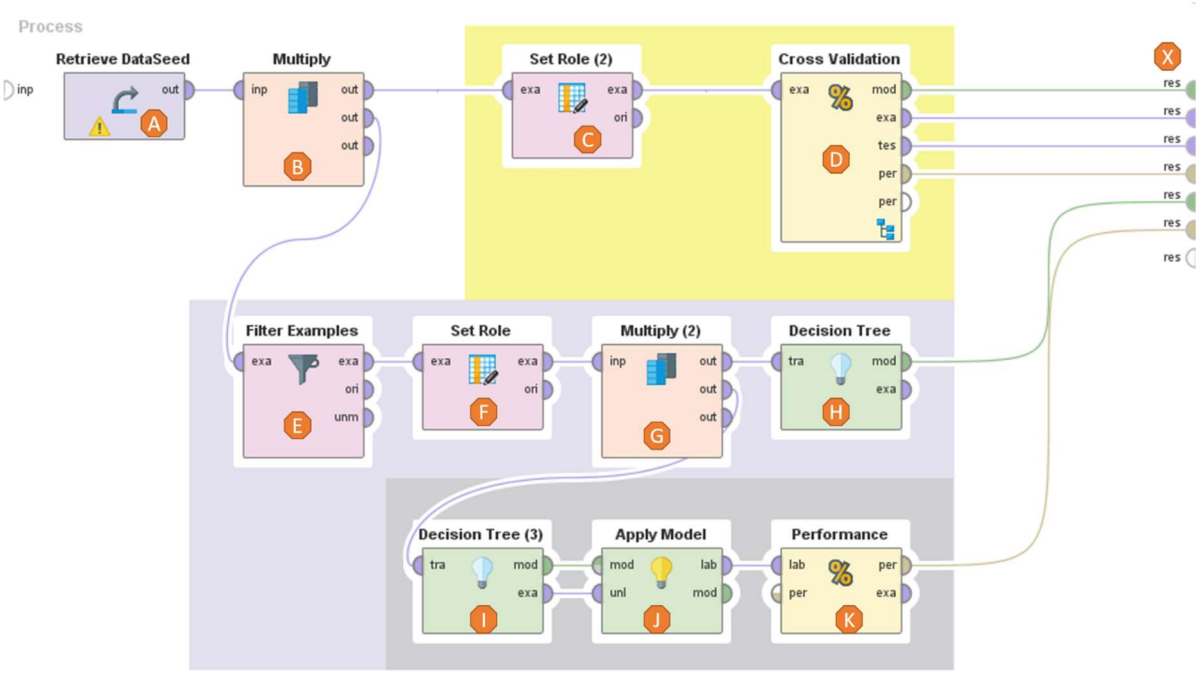
Figura 14 - Dados prontos para mineração



Observando a Figura 14 é possível verificar a esquerda, em *Local Repository*, que os dados estão prontos para mineração, os arquivos gerados foram dispostos na pasta criada, como mencionado anteriormente, com a finalidade de receber os documentos e processos gerados na mineração (Figura 14 - 6), os dados foram nomeados de **DataSeed**. Na tela exibida é possível observar o espaço para inserção dos operadores (Figura 14 - 7) que são as peças de montagem do modelo de mineração e são responsáveis pelo DM e o resultado final da descoberta do conhecimento.

O **RapidMiner Studio** possibilita, através da inserção dos operadores, executar várias técnicas de DM em um único processo, podendo assim ter resultados diferentes na mesma execução. Conforme apresentado nas seções 1.4 e 1.5, existem tarefas e técnicas preditivas que são utilizadas no DM. Dentre as apresentadas, para este trabalho foram utilizadas a classificação e a árvore de decisão como tarefa e técnica preditiva. Para construção do modelo forma utilizados alguns operadores para obter os resultados da classificação através da árvore de decisão. Assim, foi possível chegar ao modelo preditivo apresentado na Figura 15, a seguir. Os operadores exibidos na imagem serão apresentados conforme suas características e funções no modelo preditivo.

Figura 15 - Modelo Preditivo



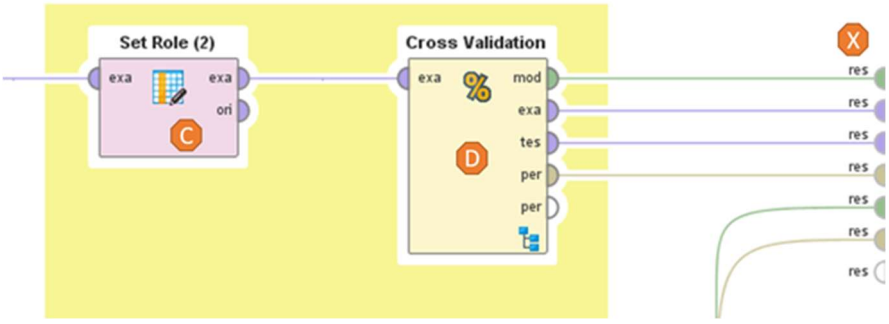
No RapidMiner um operador é um componente que pode ser visto como um sistema caixa-preta, onde o mesmo pode ser utilizado sem o conhecimento de como o processamento interno é realizado. Cada operador possui uma funcionalidade para o projeto, e esta funcionalidade vai de uma simples importação de dados a mineração de dados complexos. No modelo apresentado na Figura 15 os operadores são evidenciados com letras do alfabeto que vão do A ao K e a letra X para a saída dos resultados.

Iniciando o modelo, tem-se o operador ***Retrieve*** (Figura 15 - A) que é responsável pelo acesso aos dados que, neste caso, são os dados das sementes (denominado **DataSeed).** Este componete tem sua saída (*out*) conectada a entrada (*imp*) do operador ***Multiply*** (Figura 15 - B) a função do operador ***Multiply*** é criar cópias idênticas do objeto de entrada (**Retrieve DataSeed**) em suas saídas (*out*), que podem ser adicionadas de acordo com a necessidade. Com os dados copiados em várias saídas do operador foi possível construir modelos diferentes em um mesmo processo utilizando a mesma importação de dados, as abordagens variadas enriquecem a mineração e contribuem na tomada de decisão.

A primeira abordagem projetada refere-se a execução da mineração através da Validação Cruzada (*Cross Validation*), foi escolhida pois é um critério para avaliar modelos preditivos que priorizam sua capacidade de aproximar predições junto a dados não incluidos na análise, através da simulação e respostas futuras ainda não mensuradas, ajudando na verificação da probabilidade de ocorrerem falsos positivos. Esta abordagem se mostrou eficiente na

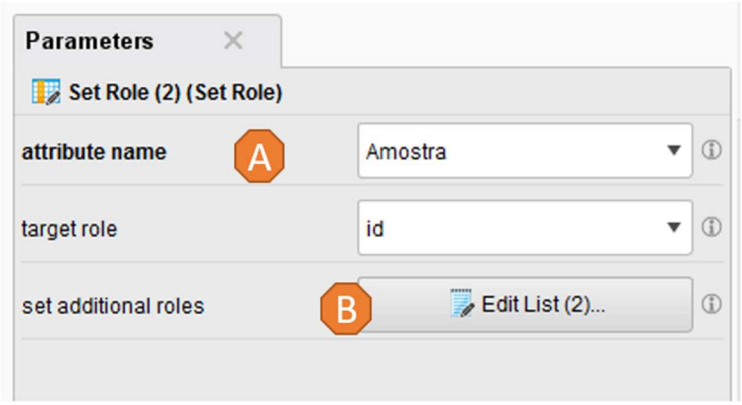
avaliação dos resultados e dos modelos utilizados na predição. Adiante, na Figura 16, recorte da Figura 15, são apresentados os operadores utilizados na Validação Cruzada.

Figura 16 - *Cross Validation*



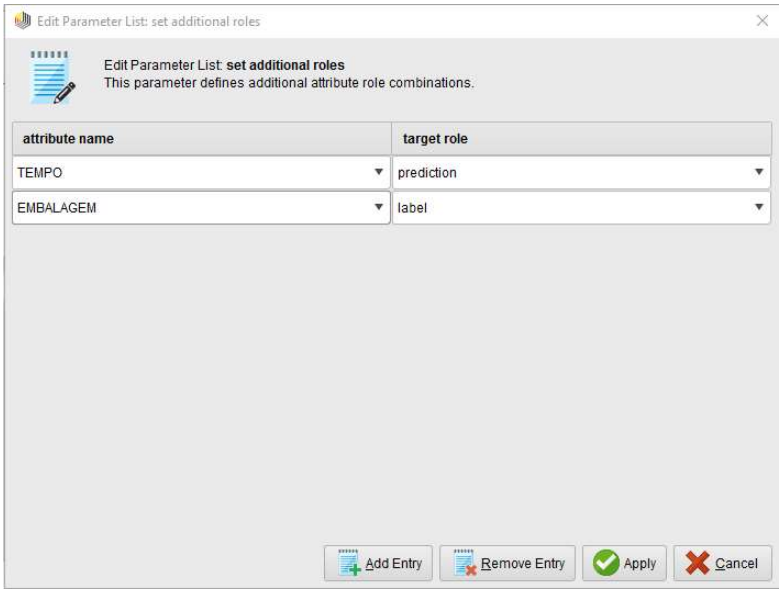
Para execução desse modelo, um dos operadores utilizados foi o ***Set Role*** (Figura 16Figura 15 - C), que tem a finalidade de alterar a função de um ou mais atributos. Quando o operador *Set Role* é selecionado, a ferramenta apresenta a direita da tela a aba de parâmetros, Figura 17, que permite modificar ou adicionar atributos conforme a necessidade do operador que é adicionado na saída do *Set Role*, neste caso as modificações ocorreram por exigênciada do operador deciosion tree, que necessita de um atributo do tipo label.

Figura 17 - Parâmetros *Set Role*



Inicialmente, o operador *Set Role* foi usado para alterar a função do atributo Amostra para o tipo Id (Figura 17 - A) , para que o atributo sirva apenas como identificador dos dados. Percebendo a necessidade de alteração de mais atributos foi necessario utilizar a opção *Edit List* (Figura 17 - B), onde é possível modificar todos os atributos de um conjunto de dados em apenas um operador *set role.* Em forma de uma lista foram adicionados espaços que carregam os atributos em *attribute name* e as possíveis novas funções a estes atributos em *target role*, como é apresentado na Figura 18.

Figura 18 - Edição de Parâmetros

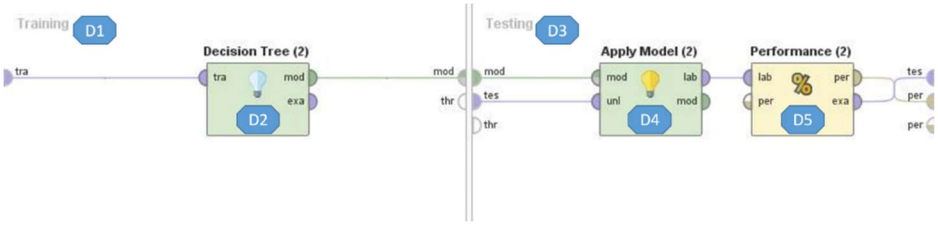


Através da edição de parâmetros, Figura 18, foi possível efetuar a modificação de dois atributos, sendo que:

- TEMPO recebeu a nova função de *prediction*, para agir como atributo previsto no modelo; e,
- EMBALAGEM passou a exercer a função de *label*, que age como atributo de destino para a aprendizagem, sendo um atributo exigido para árvore de decisão.

Os dados modificados em ***Set Role*** são enviados (por meio de sua saída (*exa*)) para o operador ***Cros Validation*** (Figura 16 - D), também denominado ***X-Validation***, o objeto (dados) é recebido pelo operador de validação cruzada por meio de sua entrada (*exa*), o ***Cross Validation*** executa sub-processos em segundo plano, que são acessados por meio de um duplo clique no canto inferior direito, esta ação leva a uma nova tela de processo do operador (Figura 19).

Figura 19 - Sub-Processo *Cross Validation*



Na Figura 19 é possível verificar que a tela do sub-processo é divida em duas partes, sendo uma denominada de *Training* (Figura 19 – D1), que foi utilizada para inserção do operador da Árvore de Decisão, ***Decision Tree*** (Figura 19 – D2). Na outra parte da tela, que é denominada de *Testing* (Figura 19 – D3), foram inseridos os operadores de avaliação do

modelo, os operadores utilizados para avaliar o modelo foram; o ***Apply model*** e ***Performance*** sendo que o primeiro, *Apply model* (Figura 19 – D4), foi utilizado para aplicar o modelo treinado em ***Training*** pela árvore de decisão, na base de dados. O segundo operador foi o ***Performance*** (Figura 19 – D5), utilizado para avaliar o desempenho do modelo com critérios determinados automaticamente que se ajustam ao tipo de tarefa executada, para os tipos de dados utilizados na classificação o operador utiliza critério de precisão, *accuracy* e Kappa, que mede a concordância entre os modelos testados.

Voltando à tela de processo principal, Figura 16, é notório que o operador ***Cross Validation*** (Figura 16 - D) possui várias saídas e cada uma delas com funções diferentes, tal como:

- *Mod,* que mostrou na saída o modelo treinado no sub-processo,
- *Exa,* que exibiu os dados, sem alteração, caso haja a necessidade de ser inserido em outro operador,
- *Tes,* que retornou os dados utilizados para testar o modelo,
- *Per,* que externou um vetor de desempenho com as estimativas obtidas nos testes,

  todas as saídas do operador são conectadas as portas, resultados (*res*) presentes no painel (Figura 16 - *Cross Validation*Figura 16 - X).

Após ser finalizada a construção do modelo, foi possível executar e verificar os resultados. No RapidMiner um processo pode ser executado ao clicar no botão no centro da interface ou teclar a tecla de função F11. Após a execução do processo, que é a mineração efetuada nos modelos inseridos no RapidMiner, a ferramenta apresenta os resultados em forma de tabela com a performance do modelo, árvore de decisão e tabela de dados usados no treinamento. Na Figura 20 é apresentada a tabela com os resultados de performance do modelo de validação cruzada.

Figura 20 - Tabela de performance

accuracy: 48.53% +/- 3.68% (mikro: 48.54%)

	true NO	true PET	true PAPEL	true POLIETILENO	class precision
pred. NO	25	1	0	5	80.65%
pred. PET	2	26	15	21	40.62%
pred. PAPEL	0	35	126	28	66.67%
pred. POLIETILENO	29	148	69	156	38.81%
class recall	44.64%	12.38%	60.00%	74.29%	

Na Figura 20, são apresentados os resultados da tabela de performance. Na tabela, é exibida a precisão global das previsões, ***accuracy***, no canto superior esquerdo, com um total de

48.53%. Este valor foi obtivo através da média de precisão calculada nos treinos definidos pelo operador ***Cross Validation*** (Figura 16 -D) no parâmetro *number of folds,* que foi fixado em 10 subdivisões. Esse parâmetro deve ser limitado e é utilizado, pois, o mesmo é essencial na validação cruzada, sua atribuição foi subdividir o conjunto de dados em 10 subconjuntos, cada subconjunto fruto da primeira divisão foi novamente subdividido em 10 subconjuntos. Após as subdivisões foi escolhida de forma aleatória uma amostra de cada subconjunto, que é avaliada quanto a sua *accuracy*, precisão. A média dos subconjuntos é a precisão global.

O resultado adquirido com a *accuracy*, precisão das previsões corretas, é um valor baixo para uma previsão o que gera a preocupação quanto a efetividade do modelo, pois o resultado apresentado indica que a probabilidade de ocorrer um falso positivo é alta. No entanto, é possível observar ao lado da *accuracy* o desvio padrão em +/- 3,68% que indica que este é um modelo estável.

Outro resultado que é possível ser verificado com a execução do modelo é a criação de uma nova tabela de dados usada para treinamento (Figura 21). Nesta foram adicionadas quatro novas colunas de forma automatizada pelo operador, sendo que as mesmas possuem a função de verificar a confiança nos rótulos utilizados no modelo criado para treinamento.

Figura 21 - Tabela usada para treinamento

Row No.	Amostra	EMBALAGEM	prediction(E...	confidence(NO)	confidence(PAPE...	confidence(...	confidence(...	Sementes N...	Sementes D...
1	C4 -b	NO	POLIETILENO	0.082	0.177	0.381	0.360	47	?
2	E2 -b	NO	POLIETILENO	0.082	0.177	0.381	0.360	49	?
3	E3 -b	NO	POLIETILENO	0.082	0.177	0.381	0.360	48	?
4	G1 -b	NO	NO	0.955	0	0.045	0	49	1
5	G3 -b	NO	NO	0.955	0	0.045	0	49	1
6	B21-3a	PAPEL	POLIETILENO	0.082	0.177	0.381	0.360	49	0
7	F21-1b	PAPEL	PET	0	0.095	0.238	0.667	27	8
8	D31-3b	POLIETILENO	POLIETILENO	0.082	0.177	0.381	0.360	47	1
9	E31-3b	POLIETILENO	POLIETILENO	0.082	0.177	0.381	0.360	44	2
10	G31-1a	POLIETILENO	POLIETILENO	0.082	0.177	0.381	0.360	48	0
11	G31-2a	POLIETILENO	POLIETILENO	0.082	0.177	0.381	0.360	49	1
12	B12-2B	PET	POLIETILENO	0.082	0.177	0.381	0.360	47	0
13	B12-3B	PET	POLIETILENO	0.082	0.177	0.381	0.360	50	0
14	E12-1A	PET	POLIETILENO	0.082	0.177	0.381	0.360	41	1
15	E12-1B	PET	POLIETILENO	0.082	0.177	0.381	0.360	44	1
16	G12-1B	PET	POLIETILENO	0.082	0.177	0.381	0.360	48	1
17	G12-2A	PET	POLIETILENO	0.082	0.177	0.381	0.360	48	2
18	A22-2A	PAPEL	POLIETILENO	0.082	0.177	0.381	0.360	37	4
19	E22-1A	PAPEL	PAPEL	0	0.677	0.133	0.190	32	4
20	G22-1A	PAPEL	POLIETILENO	0.082	0.177	0.381	0.360	46	2

É possível observar que uma das colunas criadas é a ***prediction*** (**EMBALAGEM**), que permite que seja efetuada a verificação manual da predição, comparando-a com a tabela **EMBALAGEM**. De forma que havendo a comparação, se os dados das duas colunas forem

iguais na mesma linha é uma predição correta, do contrário a predição para aquela linha é incorreta. Na Figura 21, é possível observar como isso ocorre. As linhas 4 e 5 apresentam as predições corretas, pois as colunas mencionadas possuem valores iguais, o que caracteriza como uma predição correta. Já os resultados conflitantes referem-se as predições incorretas, mas neste trabalho não foi necessário efetuar a verificação de forma manual pois a ferramenta efetuou isto de forma automatizada. As demais tabelas criadas são baseadas no rótulo EMBALAGEM, inserido no operador *Set Role* conforme a Figura 18, no qual apresenta a porcentagem de confiança para cada um dos tipos de embalagem.

Outro resultado da execução do processo com a validação cruzada é a árvore de decisão, gerada pelo operador *Decision Tree* (Figura 19 – D2), no RapidMiner é possível apresentar os resultados da classificação gerada pelo modelo, na forma descritiva, conforme exibida na Figura 22, e em forma de grafos, conforme Figura 23, possibilitando observar a classificação dos resultados de forma gráfica.

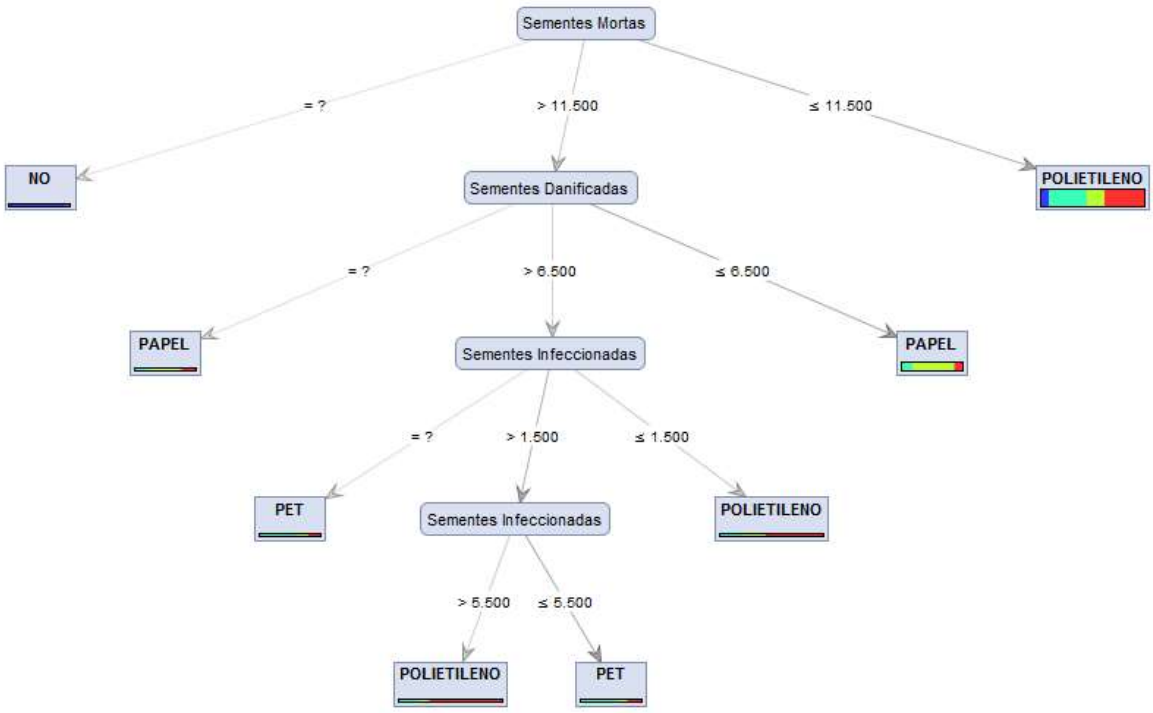
Figura 22 - Descrição da Árvore de Decisão

Tree

```
Sementes Mortas = ?: NO {NO=23, PET=0, PAPEL=0, POLIETILENO=1}
Sementes Mortas > 11.500
|   Sementes Danificadas = ?: PAPEL {NO=0, PET=5, PAPEL=11, POLIETILENO=5}
|   Sementes Danificadas > 6.500
|   |   Sementes Infeccionadas = ?: PET {NO=0, PET=3, PAPEL=1, POLIETILENO=1}
|   |   Sementes Infeccionadas > 1.500
|   |   |   Sementes Infeccionadas > 5.500: POLIETILENO {NO=0, PET=2, PAPEL=1, POLIETILENO=7}
|   |   |   Sementes Infeccionadas ≤ 5.500: PET {NO=0, PET=16, PAPEL=4, POLIETILENO=6}
|   |   Sementes Infeccionadas ≤ 1.500: POLIETILENO {NO=0, PET=4, PAPEL=4, POLIETILENO=10}
|   Sementes Danificadas ≤ 6.500: PAPEL {NO=0, PET=32, PAPEL=120, POLIETILENO=24}
Sementes Mortas ≤ 11.500: POLIETILENO {NO=33, PET=148, PAPEL=69, POLIETILENO=156}
```

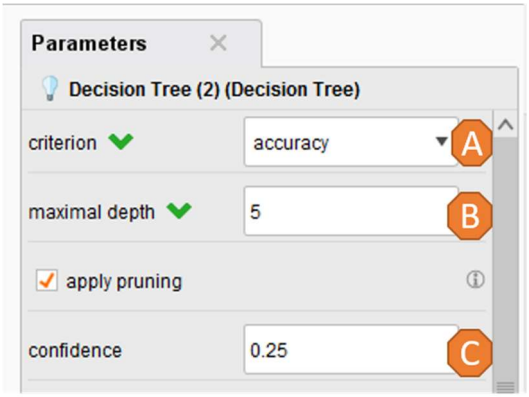
A árvore apresentada na Figura 22 apresenta de forma descritiva os resultados classificados, contudo os resultados são visualizados de forma mais clara se observados na árvore formada por grafos, que apresenta uma forma gráfica que se torna mais intuitiva, facilitando a compreensão e o entendimento dos resultados obtidos.

Figura 23 - Árvore de Decisão



A árvore de decisão apresentada na Figura 23 é o resultado da classificação do modelo por meio da validação cruzada. Na árvore de decisão apresentada, foi escolhido como critério a precisão, *accuracy*, o critério modifica o algoritimo a ser utilizado na classificação, ele é escolhido ou modificado ao ser selecionado o operador *Decision Tree*, assim é apresentada a tela de parâmetros do operador do lado direito da tela de processos, os parâmetros principais do operador *Decision Tree* podem ser observados na Figura 24.

Figura 24 - Parâmetros *Decision Tree*



Na Figura 24 é possível observar que há o parâmetro *criterion* (Figura 24 - A) onde foi informado o critério para divisão dos atributos. Afim de obter um resultado mais preciso e

confiável, neste trabalho foi escolhido como critério a *accuracy* e tal escolha possibilitou a maximização da precisão de toda árvore. É possivel observar na Figura 24 o parâmetro *maximal depth* (Figura 24 - B) que foi utilizado para restringir a profundidade da árvore, valor atribuído no parâmetro foi 5, fazendo com que a árvore tenha apenas 5 nós classificadores, esse número foi restringido por possuir uma quantidade limitada de dados e para facilitar o entendimento da árvore. Outro parâmetro do operador *Decision Tree* é o *confidence* (Figura 24 - C) esse parâmetro específica o nível de confiança, utilizado como base para o cálculo de erro, o valor de 0.25 ou 25%.

Retornando a árvore apresentada na Figura 23, é possível verificar que nos resultados apresentados mediante as configurações inseridas, a árvore gerada utilizou o atributo **Sementes mortas** como o preditor mais indicado para o modelo, sendo assim este atributo, **Sementes Mortas,** foi definido como raiz da árvore de decisão, os nós são os atributos classificadores e as folhas da árvore foram definidas pelo atributo *label* imputado na edição de paramêtros no operador *set role* (Figura 17).

Ao analizar a árvore é possível verificar que para o classificador raiz, **Sementes Mortas**, caso a amostra apresente até 11.500 sementes mortas, significa que o seu armazenamento principal era o POLIETILENO. Isto leva a inferir que este tipo de armazenamento apresenta a maior propensão das sementes morrerem. Há outros dois ramos que podem ser observados na primeira classificação, o ramo a esquerda não fornece informações sobre os valores das sementes, já o ramo central da árvore possui mais de 11.500 sementes que foram classificadas como danificadas, desse total até 6.500, a maior parte desse número está armazenda em PAPEL. As mais de 6.500 sementes que sobraram foram ainda classificadas como Infeccionadas sendo que grande parte dessas sementes eram armazenadas em POLIETILENO.

Um total maior que 6.500 foram classificadas como sementes infecionadas que, de acordo com a árvore, cerca de 1.500 estavam armazenadas em POLIETILENO e mais de 1.500 foram classificadas como infeccionadas. Conforme é observado na árvore, as sementes restantes classificadas foi é possivel contastar que um número maior que 5.500 de sementes infeccionadas estavam armazenadas em embalagens de POLIETILENO e um numero menor que 5.500 das sementes infeccionadas eram armazenadas em PET.

Através da classificação gerada pela árvore de decisão no modelo, utilizando a validação cruzada, ***Cross Validation***, tornou-se viável inferir que, conforme afirmado no parágrafo anterior, as sementes armazendas em POLIETILENO possuem uma probabilidade maior de morrerem ou apresentarem infecções que em outros armazenamentos. Foi possível constatar também que no armazenamento de PAPEL e POLIETILENO a chance das sementes serem

danificadas é maior que em outros armazenamentos. sendo assim o armazenamento de PET é a melhor opção entre os testados.

Os dados utilizados foram submetidos a mineração utilizando outros modelos de classificação, sendo executados em paralelo com a validação cruzada, como pode ser visto na Figura 15. Os modelos utilizaram os mesmos classificadores com abordagens diferentes, um dos modelos usou somente a árvore de decisão para classificação dos dados, como é apresentado na (Figura 15 – E-H), já o modelo subsequente utilizado emprega à árvore de decisão, mas com operadores que podem avaliar a classificação e performance do modelo, (-Modelo PreditivoFigura 15 - I-K) semelhante à validação cruzada. Após a execução, os modelos apresentaram resultados que diferem em alguns pontos da validação cruzada, como a *accuracy*, que obteve um valor mais baixo que na validação cruzada, no entanto após efetuar análises na árvore gerada pelo modelo, é possível perceber que o conhecimento obtido é semelhante.

Os modelos foram executados seguindo as diretrizes da metodologia de referência o CRISP-DM, que ainda conta com duas fases finais que são apresentados nas sessões seguintes.

1.14 AVALIAÇÃO, *EVALUATION,* E IMPLANTAÇÃO, *DEPLOYMENT*

Segundo o CRISP-DM na fase de avaliação, ***Evaluation***, os modelos utilizados na fase anterior são submetidos a uma avaliação técnica que visa verificar se os modelos cumpriram os objetivos do projeto. Da mesma forma, também são postos em avaliação os resultados da mineração e os processos a que foram submetidos. Já na Implantação, ***Deployment*****,** os modelos aprovados na avaliação são efetivamente utilizados para ajudar no processo de tomada de decisões. Nesta fase acontece a implantação do modelo, bem como a elaboração de um esquema para manutenção dos modelos.

Neste trabalho não foi possível executar as fases citadas, pois é necessário que os modelos sejam carregados com dados mais precisos, que tenham suas informações essenciais tabuladas para que haja uma previsão mais precisa, outro ponto é a necessidade de uma grande quantidade de dados, pois o data mining tem um resultado melhor sobre grandes volumes de dados.

CONSIDERAÇÕES FINAIS

Neste trabalho utilizou-se das tarefas e técnicas de Data Mining voltadas para predição de dados usando como estudo de caso a base de dados de germinação de sementes disponibilizada pelo Laboratório de Pós-Colheita do CEULP/ULBRA. Empregando para execução da mineração a ferramenta RapidMiner Studio, que é voltada para mineração e análise de dados, através da criação e execução de um modelo de mineração na ferramenta utilizada, foi possível obter conhecimentos que estavam implícitos nos dados.

Os modelos desenvolvidos para executar a mineração dos dados, ao serem executados, apresentaram saídas semelhantes. Por este motivo apenas o modelo de *Cross Validation* ou validação cruzada foi apresentado no item 1.13, que tratou da apresentação do modelo desenvolvido para efetuar a análise dos dados. O modelo de validação cruzada ou *Cross Validation*, ao ser executado foi efetivo para efetuar a descoberta de novos conhecimentos e predição de dados através da classificação utilizando a árvore de decisão nos dados analisados. Apesar da *accuracy*, que é a precisão da previsão dos dados ficarem abaixo de 50%, gerando desconfiança no modelo utilizado, o desvio padrão, que é utilizado para avaliar a condição de eficácia dos modelos, permaneceu abaixo de 5% o que torna o modelo eficiente na classificação voltada para previsão.

Observando os resultados obtidos com a participação da Co-orientadora, especialista no domínio, a Doutora em Pós-colheita de Produtos Agrícola, Conceição Aparecida Previero, foi possível averiguar que os mesmos estão dentro da realidade, possibilitando o levantamento de conhecimentos que vão de encontro aos identificados pela pesquisa efetuada pelo Laboratório de Pós-Colheita do CEULP/ULBRA. Nesta pesquisa, observou-se que para um armazenamento de sementes mais eficaz ou que assegure uma taxa de germinação aceitável pelo o maior tempo possível, as sementes devem ser armazenadas em embalagens de PET, os resultados das minerações possibilitaram inferir esta afirmação pois nas classificações efetuadas pela árvore de decisão os defeitos das sementes contidos nos dados e até a morte das sementes, ficaram evidentes em armazenamentos como PAPEL e POLIETILENO.

Com os modelos sendo avaliados positivamente fica em observação a qualidade dos dados utilizados, no desenrolar foi possível observar que houve dados faltantes que poderiam levar a uma predição com maior confiabilidade diminuindo a possibilidade de ocorrer um falso positivo na previsão. De acordo com as observações efetuadas, alguns dos dados que poderiam contribuir de forma edificante para um melhor resultado são os dados de temperatura e umidade. Se estes dados fossem apresentados na realidade para cada amostra, os resultados seriam bem

diferentes, podendo estes serem itens classificadores, levando a predições sob circunstâncias diferentes e não apenas sobre condições de material de armazenamento.

Levando-se em conta o problema de pesquisa apresentado e a hipótese inferida, é possível sim predizer a taxa de germinação das sementes através da aplicação de tarefas e técnicas de DM, mas para isto é necessário que os dados das sementes contenham mais atributos para serem utilizados nas ferramentas de mineração. Outro fator que pode contribuir para uma predição mais assertiva é a quantidade de dados, pois para mineração, quanto mais dados, melhor para a aplicação da mineração e criação de modelos mais eficientes.

REFERÊNCIAS

AMO, Sandra de. **Técnicas de Mineração de Dados**. Universidade Federal de Uberlândia. Faculdade de Computação.

AMORIM, Thiago. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados**. Monografia (Bacharel em Ciência da Computação). Universidade Federal de Pernambuco, 2006.

ARAÚJO, André Luiz Vale de. **Aplicação de Regras de Associação para Auxílio na Gestão de Vendas de uma Empresa Varejista Utilizando a ferramenta WEKA.** 2009. 63 f. TCC (Graduação) - Curso de Engenharia da Computação, Escola Politécnica de Pernambuco – Universidade de Pernambuco, Recife, 2009.

ARAUJO, José Marcelo Pereira de. **Processo de Descoberta de Conhecimento em Dados Não-Estruturados:** Estudo de Caso para a Inteligência Competitiva. 2007. 180 f. Dissertação (Mestrado) - Curso de Tecnologia da Informação, Universidade Católica de Brasília-ucb, Brasília, 2007.

BARREIRA, Rafael Gonçalves. **ANÁLISE DE SENTIMENTOS COM RAPIDMINER.** 2013. 71 f. TCC (Graduação) - Curso de Sistemas de Informação, Centro Universitário Luterano de Palmas, Palmas, 2013.

BATISTA, GEAPA. **Pré-processamento de dados em aprendizado de máquina supervisionado. 2003**. Tese de Doutorado. Tese (Doutorado) -ICMC–USP, São Paulo.

BERRY, M.J.A.; LINOFF, G. **Data mining techniques**. New York: John Wiley & Sons, Inc. 1997.

BOENTE, Alfredo Nazareno Pereira; GOLDSCHMIDT; ESTRELA, Vania Vieira. **Uma Metodologia de Suporte ao Processo de Descoberta de Conhecimento em Bases de Dados**. In: V Simpósio de Excelência em Gestão e Tecnologia, 2008, Resende - RJ. V SEGeT, 2008. v. 1. p. 4-5.

BRANQUINHO, Lucélia Pinto; BARACHO, Renata Maria Abrantes; ALMEIDA, Maurício Barcellos. **Descoberta de conhecimento com uso de ontologias na mineração de dados** DOI-10.5752/P. 2316-9451.2015 v4n1p20.Abakós, v. 4, n. 1, p. 20-33, 2015.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas**. Universidade Federal de Goiás (UFC), p. 1-29, 2009.

CARVALHO, Isamir Machado de et al. **Contribuições das Tecnologias KDD e DW como Ferramentas de Gestão do Conhecimento Aplicadas ao Processo de Compras do Governo Eletrônico.** Santa Catarina: Reseachgate, 2014. 14 p.

CARVALHO, Luíz Alfredo Vidal de**. Datamining**: **A Mineração de Dados, no Marketing, Medicina, Economia, Engenharia e Administração**. Rio de Janeiro: Editora Ciência Moderna Ltda., 2005.

CASTANHEIRA, Luciana Gomes. **Aplicação de técnicas de mineração de dados em problemas de classificação de padrões.** 2008. 95 f. Dissertação (Mestrado) - Curso de Engenharia Elétrica, UFMG, Belo Horizonte, 2008.

CAVALCANTE, Renata de Souza Alves Paula. **DESCOBERTA DE CONHECIMENTO NA PLATAFORMA LATTES**: **UM ESTUDO DE CASO NO INSTITUTO FEDERAL DE GOIÁS**. 2014. 219 f. Dissertação (Mestrado) - Curso de Engenharia de Produção e Sistemas, Pontifícia Universidade Católica de Goiás, Goiânia, 2014.

DIAS, Maria Madalena. **Parâmetros na escolha de técnicas e ferramentas de mineração de dados**. Acta Scientiarum. Technology, v. 24, p. 1715-1725, 2008.

DIAS, Maria. Madalena. Um modelo de formalização do processo de sistema de descoberta de Conhecimento em banco de dados. 2001. Tese (Doutorado) -Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina. Florianópolis, Santa Catarina, 2001.

DOS SANTOS SILVA, Marcelino Pereira. **Mineração de Dados-Conceitos, Aplicações e Experimentos com Weka**. 2004.

EMC. Brazil country brief. **The Digital Universe of opportunities. 2014**. Disponível em: Acesso em: 12 abril 2016.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R**. Advances in Knowledge Discovery and Data Mining**. California: AAAI Press/The MIT Press, 1996.

FAYYAD, Usama; Piatetski-Shapiro, Gregory; Smyth, Padhraic (1996). **The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: Communications of the ACM**, pp.27-34, nov.1996.

GOMES, Bruno Miguel Viana. **Previsão de Churn em Companhias de Seguros.** 2011. 151 f. Dissertação (Mestrado) - Curso de Escola de Engenharia, Departamento de Informática, Universidade do Minho, Braga, 2011.

INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS & TECHNOLOGY MANAGEMENT - CONTECSI, 12. 2015, São Paulo. **Identificação de Padrões em Registros de Doenças com Técnicas de Mineração de Dados.** São Paulo: Contecsi, 2015. 12 p.

ISOTANI, Seiji et al. **Web 3.0 - Os Rumos da Web Semântica e da Web 2.0 nos Ambientes Educacionais**. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 19, 2008, Fortaleza. **Anais.** Fortaleza-CE: Sbie, 2008. p. 1 - 11. Disponível em: <http://www.cs.cmu.edu/afs/cs/Web/People/sisotani/artigos/sbie2008_sw.pdf>. Acesso em: 22 abr. 2016.

IYODA, Eduardo Masato**. Inteligência Computacional no projeto automático de Redes Neurais Híbridas e Redes NeuroFuzzy heterogêneas**. Dissertação de M. Sc, Faculdade de Engenharia Elétrica e da Computação, UNICAMP, Campinas-SP, 2000.

KARCHER, Cristiane. **Redes Bayesianas aplicadas à análise do risco de crédito.** 2009. 103 f. Dissertação (Mestrado) - Curso de Engenharia Elétrica - Sistemas Eletrônicos, Escola Politécnica da Universidade de São Paulo, São Paulo, 2009.

KREMER, Ricardo. **Sistema de apoio à decisão para previsões genéricas utilizando técnicas de data mining**. Trabalho de Conclusão de Curso. Universidade Regional de Blumenau, 1999.

MAIA, Roberto Bomeny. **Detecção Da Intrusão Utilizando Classificação Bayesiana**. 2005. 158 f. Tese (Doutorado) - Curso de Engenharia Elétrica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2005.

MELLO, Luis Cesar de. **Um assistente de Feedback para o Serviço de Filtragem do Software Direto.** 2002. 115 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

NOGUEIRA, Diogo. **Agile Data Mining:** Uma metodologia ágil para o desenvolvimento de projetos de data mining. 2014. 109 f. Dissertação (Mestrado) - Curso de Mestrado Integrado em Engenharia Informática e Computação, Faculdade de Engenharia da Universidade do Porto, Porto - Portugal, 2014.

PIOVESAN, Pamela. **Validação cruzada com correção de autovalores e regressão isotônica nos modelos AMMI**. 2007. Tese de Doutorado. Escola Superior de Agricultura "Luiz de Queiroz.

PYLE, Dorian. **Data preparation for data mining**. San Francisco, CA: Morgan Kaufmann, 1999.

PREVIERO, Conceição Aparecida; SANTOS, Deise Laiz dos; GONÇALVES, Luanne Pereira. **Armazenabilidade de Sementes de Milho (Zea Mays L.) Tipo Variedade em Diferentes Embalagens, Em Palmas, Tocantins**. In: CONGRESSO BRASILEIRO DE ENGENHARIA AGRÍCOLA, 42., 2013, Fortaleza. **Resumo Expandido.** Fortaleza: Conbea, 2013. p. 1 - 4.

REZENDE, Solange Oliveira. **Mineração de Dados**. In: XXV CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 25. 2005, São Leopoldo-RS. **Anais....** São Leopoldo: CSBC, 2005. p. 397 - 433. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/enia/2005/0102.pdf>. Acesso em: 22 abr. 2016.

SILVA, E. M. **Avaliação do Estado da Arte e Produtos. Data Mining**. 2000. Tese (Dissertação de Mestrado). Universidade Católica de Brasília.

SILVA, Edilberto Magalhães. **Descoberta de Conhecimento com o uso de Text Mining:** Cruzando o Abismo de Moore. 2002. 175 f. Dissertação (Mestrado) - Curso de Gestão do Conhecimento e da Tecnologia da Informação, Universidade Católica de Brasília - Ucb, Brasília-DF, 2002.

SOUSA, Paulo de Tarso Costa de. **MINERAÇÃO DE DADOS PARA INDUÇÃO DE UM MODELO DE GESTÃO DO CONHECIMENTO.** 2003. 167 f. Dissertação (Mestrado) - Curso de Gestão do Conhecimento e da Tecnologia da Informação, Universidade Católica de Brasília - UCB, Brasília, 2003.

TAN, P. N. STEINBACH, M. KUMAR, V. **Introdução ao data mining: mineração de dados.** Rio de Janeiro: Editora Ciência Moderna Ltda., 2009. 900 p. Tradução de: Introduction to datamining

THOMÉ, Antônio Carlos Gay. **Redes neurais: uma ferramenta para KDD e data mining**. 2008.