UNIVERSIDADE FEDERAL DO MARANHÃO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

BERTHONE COLINS MARTINS

UMA DISCUSSÃO SOBRE DIFERENTES AMBIENTES DE SOFTWARE PARA MINERAÇÃO DE DADOS

SÃO LUÍS

2017

BERTHONE COLINS MARTINS

UMA DISCUSSÃO SOBRE DIFERENTES AMBIENTES DE SOFTWARE PARA MINERAÇÃO DE DADOS

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de bacharel em Ciência da Computação.

Orientador: Prof. Dr. Ivo José da Cunha Serra

SÃO LUÍS

2017

Martins, Berthone Colins.
Uma discussão sobre diferentes ambientes de software para mineração de dados / Berthone Colins Martins. - 2017.
47 f.

Orientador(a): Ivo José da Cunha Serra.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís, 2017.

1. Mineração de Dados. 2.RapidMiner. 3. Softwares para Mineração de Dados. 4. WEKA. I. Serra, Ivo José da Cunha. II. Título.

BERTHONE COLINS MARTINS

UMA DISCUSSÃO SOBRE DIFERENTES AMBIENTES DE SOFTWARE PARA MINERAÇÃO DE DADOS

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de bacharel em Ciência da Computação.

Aprovada em: 12/07/2017

BANCA EXAMINADORA

Prof. Dr. Ivo José da Cunha Serra (Orientador)
Centro de Ciências Exatas e Tecnologia - CCET
Universidade Federal do Maranhão – UFMA

Prof. MSc. Carlos Eduardo Portela Serra de Castro
Centro de Ciências Exatas e Tecnologia - CCET
Universidade Federal do Maranhão - UFMA

Prof. Dr. Samyr Béliche Vale
Centro de Ciências Exatas e Tecnologia - CCET
Universidade Federal do Maranhão - UFMA

À minha família que sempre acreditaram em meu potencial e determinação. Mãe, suas palavras de incentivo e apoio incondicional foram priomordiais. Pai, sua confiança em mim deram suporte a esta caminhada.

AGRADECIMENTOS

Em primeiro lugar, a Deus que me deu sabedoria e saúde ao longo de minha vida para que pudesse realizar todos os meus sonhos.

Aos meus pais, que independente de qualquer obstáculo sempre estiveram comigo, me dando toda a base necessária para minha formação acadêmica e pessoal.

Aos meus irmãos por estarem sempre comigo nas horas difíceis e por cuidarem de mim quando foi preciso.

Ao professor e orientador Ivo por seus esclarecimentos, conselhos e conhecimentos dados a mim ao longo desta jornada, cuja sua contribuição foi de suma importância para este trabalho.

Ao professor de física Barros, que de forma simples no ensino médio conseguiu despertar todo o meu interesse pela área de exatas, sua contribuição foi primordial para a minha vida pessoal e acadêmica.

À minha namorada pelos conselhos, apoio incondicional e pela paciência nas horas difíceis.

Aos meus amigos do curso de ciência da computação, que me acompanharam desde o início do curso, que me ajudaram academicamente quando precisei e que fizeram parte desta história.

Agradeço especialmente a todo o corpo docente que contribuíram com todo o conhecimento técnico e cientifíco que hoje possuo, no qual serão lembrados por toda a minha vida profissional.

"Seja você quem for, seja qual a posição social que você tenha na vida, a mais alta ou a mais baixa, tenha sempre como meta muita força, muita determinação e sempre faça tudo com muito amor e com muita fé em Deus, que um dia você chega lá. De alguma maneira você chega lá."

(*Ayrton Senna*)

RESUMO

Diante da evolução dos hardwares, os servidores de desevolveram se tornando capazes atualmente de suportarem um grande volume de dados. A mineração de dados surge com o propósito de identificar e extrair informações relevantes baseados nessa base de dados. A partir da evolução de fundamentos e conceitos de mineração de dados, ferramentas específicas para esta área foram desenvolvidas abordando tarefas e métodos de mineração de dados. O objetivo destas ferramentas é fornecer aos usuários um ambiente prático onde eles pudessem aplicar seus conhecimentos de forma prática, melhorando a precisão e visualização dos resultados. Este trabalho aborda dois softwares distintos WEKA e RapidMiner que são duas ferramentas de mineração de dados. Uma discussão foi realizada levando em consideração algumas especificidades de cada ferramenta como: tipos de distribuição, tarefas de mineração que os softwares abordam, arquivos de entrada que as ferramentas aceitam, entre outros critérios. O WEKA apresentou uma interface mais simples, tornando uma ferramenta mais prática para iniciantes, além disso aborda diversas tarefas de mineração de dados. O RapidMiner apresenta uma interface mais elaborada com layout mais elegante, apesar de possuir diversas tarefas de mineração de dados, possui menor praticidade e com quantidade de documentação menor em relaçao ao WEKA.

Palavras-chave: WEKA, RapidMiner, Mineração de Dados, Softwares para Mineração de Dados

ABSTRACT

Due to the hardware evolution, servers have become increasingly capable of supporting a large amount of data. In this context Data mining arises for the purpose of identifying and extracting relevant information based on these databases. From the evolution of fundamentals and concepts of Data mining, specific tools for this area have being developed. The purpose of these tools is to provide users with a practical environment where they can apply their knowledge in a practical way, improving the accuracy and visualization of the results. This work presents WEKA and RapidMiner which are two Data mining tools. A discussion was carried out taking into account some specificities of each tool such as: distribution types, mining tasks that the software supports, format of input files, among other criteria. WEKA presented a simpler interface, making it a more practical tool for beginners, and it addresses a variety of data mining tasks. RapidMiner presents a more elaborate interface with a more elegant layout, and although it has several tasks of data mining, it has less practicality and a smaller amount of documentation in relation to WEKA.

Keywords: WEKA, RapidMiner, Data mining, Softwares for Data Minning

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

KDD	Knowledge Discovery in Databases
WEKA	Waikato Environment for Knowledge Analysis
DM	Data Minning

SUMÁRIO

1 INTRODUÇÃO

Com o desenvolvimento da tecnologia da informação, servidores de banco de dados foram se desenvolvendo de tal forma que nos dias atuais conseguem suportar uma grande quantidade de dados. Atualmente diversos campos utilizam um banco de dados para guardar os mais variados tipos de dados, daí diante dessa enorme massa bruta de dados surgiu a necessidade de filtrar estes dados para partes menores, facilitando ações de quem utiliza estes dados, podendo com estes dados minerados elaborar estratégias viáveis ao seu ramo.

A mineração de dados surgiu da necessidade de processar grandes volumes de dados. As técnicas de mineração de dados são organizadas para agir sobre grandes bancos de dados com intuito de descobrir padrões úteis e recentes que poderiam, de alguma outra forma, permanecer ignorados (PANG-NING et al., 2009). Diante deste novo cenário, com uma maior quantidade de dados armazenados, mas que estavam sendo mal aproveitados, surgiram técnicas de mineração de dados com o objetivo de obter apenas dados relevantes.

Com o desenvolvimento destas tarefas de mineração de dados surgiu a necessidade de implementação de softwares que utilizem técnicas e fundamentos de mineração de dados para dar maior praticidade aos usuários que tivessem a necessidade de utilização destas técnicas. Entre os softwares desenvolvidos, está o WEKA (WAIKATO,) que foi iniciado em 1992, desenvolvido em java com código aberto, ele foi implementado na Universidade de Waikato (Nova Zelândia), por Ian Witten, Eibe Frank e colaboradores (WITTEN; FRANK, 2000).Temos também o RapidMiner que foi desenvolvido em Java, foi iniciado em 2001 por Ralf Klinkenberg, Ingo Mierswa e Simon Fischer na Unidade de Inteligência Artificial da Universidade de Dortmund (Alemanha) (MORAIS, 2012).

A partir do conhecimento em alguns fundamentos principais de técninas de mineração de dados, o objetivo é detalhar alguns conceitos e tarefas, para que estes fundamentos forneçam subsídio para o detalhamento das ferramentas. Será realizado um aprofundamento em características relevantes dos softwares, explicando alguns pontos como: endereço disponível para download; principais funcionalidades; tarefas de mineração de dados suportados; arquivos de entrada suportados além das especificidades de cada ferramentas.

Este trabalho propõe-se a fazer uma discussão focando em alguns critérios comum entre as ferramentas, para isso, um estudo de caso foi executado em ambas as ferramentas, observando seu comportamento e a facilidade de uso em aplicar um método de mineração de dados em um conjunto de regitros de entrada. Nesse contexto, será elaborada uma tabela

comparativa entre os softwares de mineração de dados abordados. O objetivo final é realizar uma discussão comparativa entre as ferramentas a partir de critérios definidos no contexto deste trabalho.

1.1 OBJETIVOS

O objetivo principal deste trabalho é realizar uma discussão sobre softwares para mineraçãos de dados, visando estabelecer a compatibilidade das técnicas e métodos de mineraçãos de dados. Partindo deste ponto, demonstrar os pontos positivos e limitações de cada uma.

1.1.1 Objetivos Específicos

Para que o objetivo geral seja atingido, alguns objetivos específicos devem ser alcançados:

- Apresentar alguns conceitos, técnicas e métodos de mineração de dados.
- Realizar um detalhamento das funcionalidades dos softwares WEKA e RapidMiner, destacando seus pontos positivos e limitações.
- Disponibilizar uma discussão atualizada sobre as ferramentas de mineração de dados Weka e RapidMiner.

1.2 ORGANIZAÇÃO DO TRABALHO

Esta monografia apresenta a seguinte organização:

No Capítulo 2, Fundamentação Teórica, seguem informações importantes para o contexto e entendimento do trabalho, tais como o entendimento sobre conceitos, técnicas e métodos de Mineração de Dados.

No Capítulo 3, realizar um detalhamento e conhecimento das funcionalidades dos softwares WEKA e RapidMiner com intuito de demonstrar sua usabilidade, interface gráfica e sua parte operacional em relação aos conceitos de mineração de dados.

No Capítulo 4, Resultados e Discussão, mostra-se algumas demonstrações das ferramentas, abordando pontos positivos e negativos.

No Capítulo 5, Conclusão, apresentam-se as conclusões obtidas, considerações finais e trabalhos futuros.

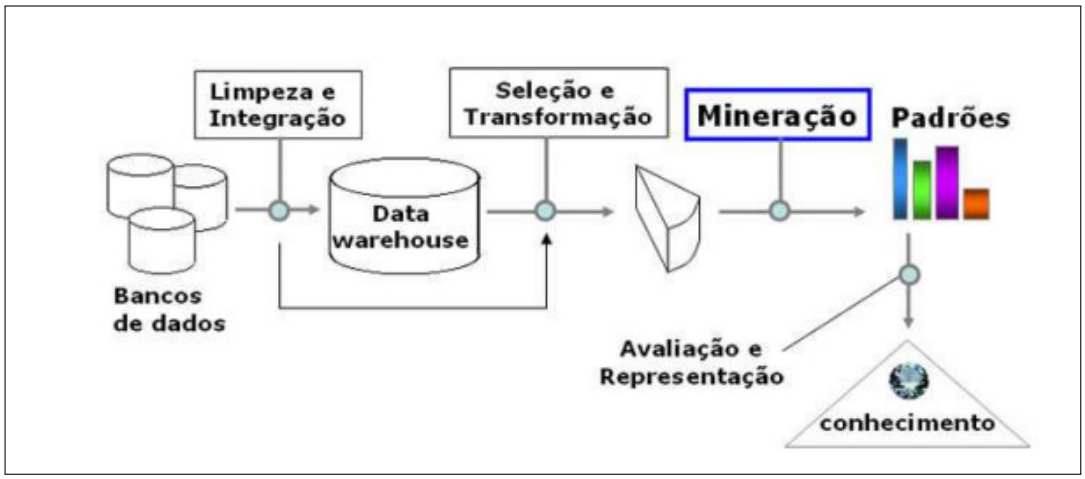
2 MINERAÇÃO DE DADOS

Mineração de Dados refere-se a disciplina que tem como objetivo descobrir “novas” informações através da análise de grandes quantidades de dados. A mineração de dados pode ser considerada como uma parte do processo de Descoberta de Conhecimento em Banco de Dados (KDD - Knowledge Discovery in Databases) (BAKER et al., 2010). Há vários conceitos para o termo "mineração de dados"que a grosso modo seria minerar ou filtrar os dados, ou seja, grandes volumes de dados armazenados seriam minerados para dados mais específicos.

A aplicação de métodos de mineração de dados se expandiram, tornando-se aplicável não somente em áreas científicas, progredindo para diversas áreas distintas. Devido a eficácia da mineração de dados, passou a ser utilizado na área financeira, marketing, comércio eletrônico, na medicina entre outras. Por exemplo, se pensarmos em um proprietário de um site eletrônico com diversos produtos em diversos seguimentos, é de suma importancia para este proprietário o conhecimento selecionado dos costumes de seus clientes, por exemplo, caso se queira saber para um produto X quais são os tipos de clientes que mais o compram, se os clientes para aquele produto são do sexo masculino ou feminino entre outras especificações que o proprietário julgar necessário. Assim como o exemplo citado há vários outros, informações selecionadas ou mineradas a partir de um grande volume de dados possui a capacidade como no caso acima de aumentar lucros.

Para que possamos entender as etapas do processo de mineração de dados é preciso conhecer sobre o KDD, que consiste nas seguintes etapas conforme a Figura 1:

Figura 1 – Etapas do processo KDD



Fonte: (AMO, 2004).

Limpeza e integração: Trata-se da retirada de dados inconsistentes para a melhora da aplicação dos algoritmos.

Seleção: etapa na qual é selecionado os parâmetros que o usuário julga necessário para o seu objetivo principal. Por exemplo pode não ser necessário o sexo do cliente para decidir se o mesmo paga em dias ou não.

Transformação: esta etapa ocorre após a seleção e limpeza dos dados, trata-se da transformação dos dados em dados compatíveis com os algoritmos de mineração.

Mineração: etapa na qual ocorre a utilização das técnicas de mineração para extrair padrões.

2.1 TAREFAS TRADICIONAIS DE MINERAÇÃO DE DADOS

A mineração de dados vem se expandindo no decorrer do tempo, se desenvolvendo nos mais diversos tipos de ambientes como nos negócios, bolsa de valores entre outros seguimentos. Segundo (HARRISON, 1998) as tarefas foram elaboradas no ínicio do processo de mineração de dados para nortear os algoritmos de acordo com o tipo de conhecimento extraído, ou seja caso se queira realizar uma predição de um fato futuro com base no passado, a tarefa a ser utilizada seria a classificação, porém não há uma técnica que resolva todos os problemas de mineração de dados . Diferentes tarefas servem para diferentes propósitos, a seguir veremos uma descrição das tarefas mais tradicionais, abordando em alguns casos suas aplicações.

2.1.1 Classificação

Uma das tarefas mais tradicionais, a classificação tem como objetivo identificar a qual classe um determinado registro pertence. Nesta tarefa, primeiramente há a indução do classificador a partir de um conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de "induzir"como classificar um novo registro (aprendizado supervisionado), tendo assim a classificação dos registros. Por exemplo, categorizamos cada registro de um conjunto de dados contendo as informações sobre os colaboradores de três classes: Perfil Técnico, Perfil Negocial e Perfil Gerencial. A partir daí, é realizado uma análise dos registros, posteriormente é possível estimar em qual categoria um novo colaborador ao ser cadastrado irá ser alocado, (CAMILO; SILVA, 2009). Na seção 2.2 veremos alguns métodos de classificação bastante utilizados atualmemte. A tarefa de classificação pode ser usada por

exemplo para:

• Determinar quando uma transação de cartão de crédito pode ser uma fraude;

• Detectar quando uma mensagem é spam em e-mails

• Diagnosticar quais doenças podem se manifestar neste paciente baseado em seus exames de rotina;

• Identificar ações de uma determinada pessoa como os sites que costuma visitar ou programas que costuma baixar, baseados nestas informções pode-se detectar se esta pessoa é ou não uma ameaça para a segurança.

2.1.2 Regressão

A tarefa de regressão é bem semelhante a de classificação, na regressão os valores das entradas são números e não em categorias, ou seja, um valor real diferentemente da classificação que utiliza valores discretos. O intuito é estimar um valor de uma determinada saída baseado em um conjunto de dados numéricos. Também conhecida como tarefa de estimativa, pois tem como função predizer um valor baseada em uma classe já conhecida. (CAMILO; SILVA, 2009) cita como exemplo um conjunto de registros contendo os valores mensais gastos por diversos tipos de consumidores e de acordo com os hábitos de cada um. Após ter analisado os dados, o modelo é capaz de dizer qual será o valor gasto por um novo consumidor. A tarefa de estimação pode ser usada por exemplo para: Estimar a quantia a ser gasta por uma família de quatro pessoas durante a volta às aulas Estimar a pressão ideal de um paciente baseando-se na idade, sexo e massa corporal.

2.1.3 Agrupamento ou *Clustering*

Diferentemente da classificação e regressão vistos anteriormente, a tarefa de agrupamento não se baseia em classes. O que acontece é a divisão de uma população geral em grupos e tendo como critério de formação de grupos a semelhança entre os elementos, o objetivo é dar subsídio para o início do processo de exploração, dividindo em padrões semelhantes.Para (CAMILO; SILVA, 2009) esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado). Por exemplo clientes de um site de comércio eletrônico, podemos dividi-los em grupos com hábitos de compras parecidos ou iguais e aplicar baseados nestes grupos, tarefas de mineração de dados afim de extrair um padrão.

2.1.4 Desvios

Visa analisar as exceções a regra, ou seja aqueles dados que seguem um comportamento padrão, ao se comportarem de forma diferente do padrão, é possível detectar este desvio e analisá-lo. Por exemplo, para detectar irregularidades em seguros, basta identificar aqueles que não estão de acordo com o padrão, para isso deve se adotar padrões de forma antecipada. Outro exemplo que podemos citar, é a análise comportamental de clientes de comércio eletrônico, verifica-se que estes apresentam um padrão, a partir de qualquer alteração neste comportamento, pode-se analisar e verificar estes clientes.

2.1.5 Associação

Tem como objetivo identificar as relações entre grande quantidade de registros de dados. Segundo (CAMILO; SILVA, 2009) estes registros se apresentam da seguinte forma: SE atributo X ENTÃO atributo Y, ou seja, se o atributo X for verdadeiro implica no acontecimeto do atributo Y. Por exemplo, clientes que na maioria das vezes, sempre que compram fraldas compram cerveja, há uma relação entre o ato da compra de fraldas com o ato da compra de cerveja. É uma das tarefas mais conhecidas devido aos bons resultados obtidos, principalmente nas análises da "Cestas de Compras"(*Market Basket*), onde identificamos quais produtos são levados juntos pelos consumidores. Por exemplo, identificar que determinada quantidade de clientes sempre que compram um par sapatos também compram um par de meias.

2.2 PRINCIPAIS MÉTODOS DE MINERAÇÃO DE DADOS

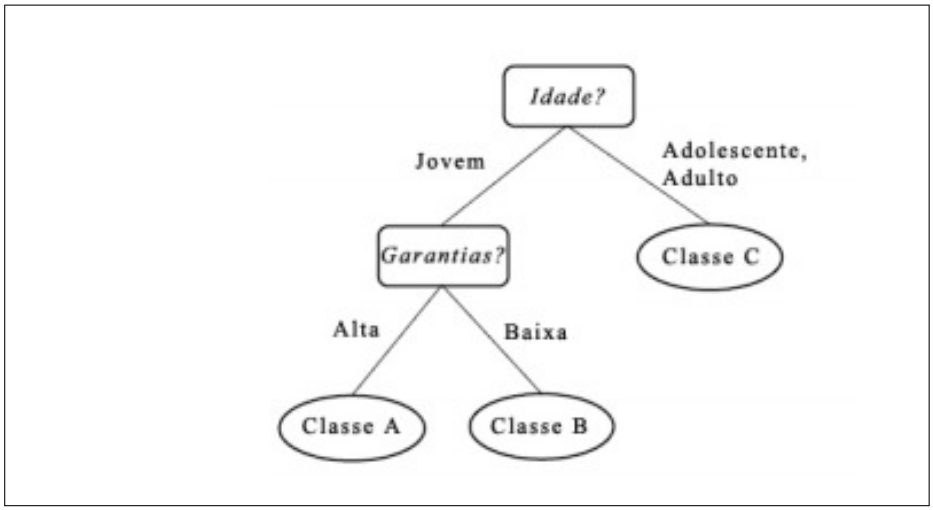
Em relação aos métodos ou técnicas para mineração de dados, como por exemplo: árvores de decisão, regras de associação, redes neurais, vetor máquina de suporte, algoritmos genéticos entre outros. Há diversos devido ao grande desenvolvimento deste campo, entretanto citaremos apenas alguns. Os métodos de Mineração de Dados que veremos neste trabalho estão ligados as tarefas vistas na seção 2.1, que podem ser supervisionadas ou não-supervisionadas, onde supervisionadas são aquelas conduzidas por um "supervisor"que ensina qual deve ser o comportamento de acordo com a situação, possuem o objetivo de predição e não-supervisionadas aprendem sozinhas as relações, padrões ou categorias, sem necessidade de um "supervisor". Por exemplo, ee encaixam no método supervisionado a classificação e regressão e no não-supervisionado, agrupamento e associação.

2.2.1 Classificação

2.2.1.1 Árvore de Decisão

A árvore de decisão funciona como um fluxograma em forma de árvore, onde cada nó (não folha) representa um teste feito sobre um valor (por exemplo, idade > 20), o resultado dos testes realizados indicam o fluxo a ser seguido. As ligações entre os nós representam os valores possíveis do teste do nó superior, e as folhas indicam a classe (categoria) a qual o registro pertence. Após a árvore de decisão montada, para classificarmos um novo registro, basta seguir o fluxo na árvore (mediante os testes nos nós não-folhas) começando no nó raiz até chegar a uma folha. E estrutura da árvore de decisão é baseada em condições que determinam o direcionamento do fluxo a ser tomado, desta forma através de decisões se chega ao objetivo final da árvore (CAMILO; SILVA, 2009). Para exemplificar temos a Figura 2 que trata da tomada de decisão nos nós não folhas, verificando primeiramente a idade, baseado na idade obtida o fluxo da árvore segue para o filho esquerdo ou direito até chegar a um nó folha que indica o fim da árvore de decisão.

Figura 2 – Árvore de Decisão



Fonte: (HAN; KAMBER, 2006)

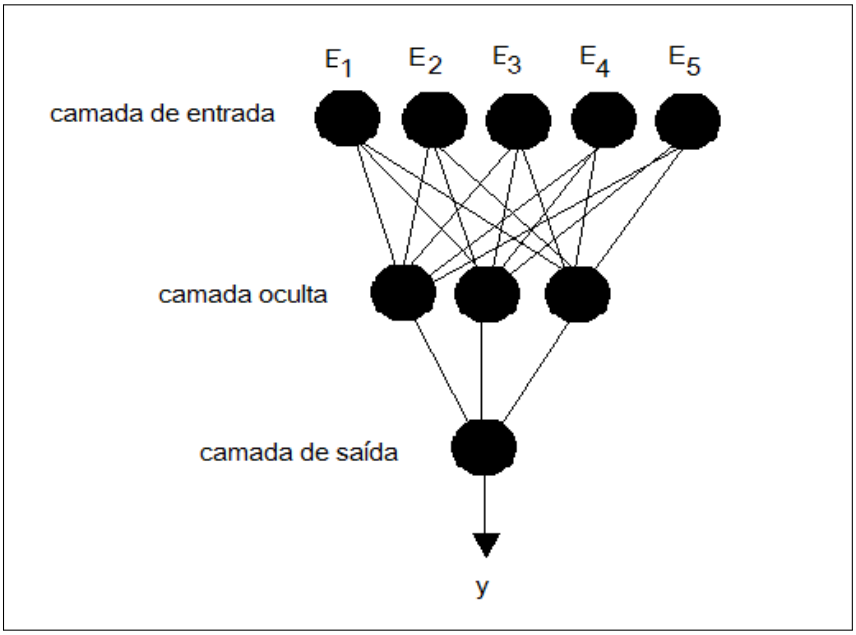
2.2.1.2 Redes Neurais

Redes neurais artificiais, como o próprio nome sugere, vem de uma comparação ao campo da biologia onde tenta-se uma semelhança com os sistemas neurais biológicos. De acordo

com (PANG-NING et al., 2009) é dividida em dois tipos de modelos: perceptron que se trata de um modelo mais simples, este modelo apresenta um conjunto de nodos de entrada que são usados para representar os atributos de entrada e um nodo se saída, que tem por finalidade, representar a saída do modelo, o outro modelo é mais complexo e é chamado de rede neural artificial multicamadas, esta rede pode conter diversas camadas intermediárias entre suas camadas de entrada e saída, além disso, apresentam um conjunto de unidades de entrada em forma de um vetor de entradas E1,E2,E3,E4,E5 com pesos relacionados a cada entrada, conforme a Figura 3, o sinal de entrada passa pela camada oculta (camadas intermediárias) até chegar ao nodo de saída.

As Redes Neurais foram originalmente projetadas por psicólogos e neurobiologistas que procuravam desenvolver um conceito de neurônio artificial análogo ao neurônio natural. Intuitivamente, uma rede neural é um conjunto de unidades do tipo entrada e saída, tais unidades são conectadas umas às outras e cada conexão tem um peso associado. Cada unidade representa um neurônio. Os pesos associados a cada conexão entre os diversos neurônios é um número entre -1 e 1 e mede de certa forma qual a intensidade da conexão entre os dois neurônios (AMO, 2004). O processo de aprendizado de um certo conceito pela rede neural corresponde à associação de pesos adequados às diferentes conexões entre os neurônios, desta forma, utilizando redes neurais é possível treinar e classificar a partir das entradas, que são os atributos do modelo que queremos obter a classificação.

Figura 3 – Exemplo de Rede Neural



Fonte: (PANG-NING et al., 2009) Adaptado.

2.2.1.3 Classificação Bayesiana

O classificador bayesiano está baseado no teorema de Bayes que se trata de um príncipio estatístico utilizando probabilidade. Este método aplica conceitos de probabilidade condicional onde se consegue prever a probabilidade de uma certa amostra ocorrer de acordo com hipóteses a priori. Para compreendermos com mais eficiência, podemos pensar em um exemplo bem simples: se comprarmos 100 peças em uma loja X, suponhamos que 80 destas peças estejam com defeito, então a probalidade para que as próximas peças compradas sejam defeituosas é de 0,8 enquanto que para peças perfeitas é de 0,2, ou seja consegue-se a partir do cálculo de probabilidades predizer fatos futuros.

A partir de um domínio desejado, identificando variáveis e as relações entre as varáveis do domínio podemos contruir redes bayesianas. De acordo com (PANG-NING et al., 2009) as redes bayesianas são grafos acíclicos direcionados, mostrando as relações de causalidade entre os atributos e possuem como finalidade representar graficamente os relacionamentos probabilísticos entre um conjunto aleatório de variáveis. Nestes grafos, as elipses representam os atributos e as ligações representam os relacionamentos de influência entre os atributos. Para (DIAS et al., 2008) a partir dos cálculos estatísticos, cada atributo terá uma tabela de valores de probabilidades para que suas possíveis ações sejam realizadas, dessa forma, na utilização de uma ferramenta de análise de Redes Bayesianas é possível definir hipóteses sobre um determinado atributo, tendo respostas sobre as influências dele de acordo com as ligações existentes.

2.2.1.4 Algoritmos Genéticos

Assim como o mundialmente conhecido Charles Darwin elaborou sua teoria da evolução das espécies, os algoritmos genéticos surgiram tendo como base estes conceitos, onde uma população se adapta a um ambiente através da seleção natural. Logicamente que este estudo criado por Darwin passaria por diversas mudanças científicas até chegar ao que conhecemos atualmente. O algoritmo realiza uma função similar ao da seleção natural, onde se tem um processo inicial e logo após uma série de processos como a seleção, cruzamento entre outros, por fim chegando a uma população final adaptada. Para exemplificar, podemos aplicar em gerenciamento de redes no qual realiza supervisão do tráfego nos links e das filas nos "buffers"de roteadores para descobrir rotas ótimas e para reconfigurar as rotas existentes no caso de falha de algum link, em computação evolutiva no qual gera programas que se adaptam a mudanças no

sistema ao longo do tempo (MIRANDA, 2007).

2.2.1.5 Análise Associativa

A análise de associação é o processo de interconexão de objetos na tentativa de expor caracteríticas e tendências, entende-se que a presença de um item implica necessariamente na presença de outro item na mesma transação (CÔRTES et al., 2002). Um exemplo muito conhecido de associação é o caso de compras de clientes em comércio varejista, conforme demonstra a Tabela 1:

Tabela 1 – Transação de cestas de compras

Itens	TID
Pão, Leite	1
Pão,Fraldas,Cerveja,Ovos	2
Leite,Fraldas,Cerveja,Cola	3
Pão,Leite,Fraldas,Cerveja	4
Pão,Leite,Fraldas,Cola	5

Fonte: (PANG-NING et al., 2009)

Como já dito anteriormente na seção 2.2, temos a associação, que relembrando é basicamente uma análise dos hábitos de compras de clientes quando aplicado ao comércio, verificando a associação entre um comportamento de compra implicar em outro comportamento. A Tabela 1 expressa a relação fraldas -> Cerveja, ou seja o comportamento do cliente em comprar fraldas implica em outro comportamento que é o de comprar cerveja. Com essa informação os varejistas podem usar para auxiliá-los a identificar novas oportunidades para vendas cruzadas dos seus produtos para os clientes. Ainda há um conjunto variado de representação da análise de associação, como por exemplo: representação binária onde os produtos compradossão sinalizados com o numeral 1 e caso não esteja é sinalizado com o numeral 0, fora esta representação, existem outras como: regra de associação, conjunto de itens e contador de suporte (PANG-NING et al., 2009).

2.2.2 Agrupamento

2.2.2.1 Algoritmos de Agrupamento

A necessidade da divisão de certos individuos ou elementos em grupos, teve por finalidade organizar populações e realizar um estudo destes grupos mais específicos. Visto

isso, a análise de grupos (*clusters*) ganhou força, pois passou a se tornar mais interessante analisar grupos ao invés de populações. A diferença entre classificação e agrupamento está na categorização dos registros e na predição, uma vez que no agrupamento não prediz valores e não possui registros em categorias, apenas realiza agrupamentos.

O algoritmo mais conhecido desta técnica se chama algoritmo K-means, esse algoritmo usa o conceito da centroide. Dado um conjunto de dados, o algoritmo seleciona de forma aleatória k registros, cada um representando um agrupamento. Para cada registro restante, é calculada a similaridade entre o registro analisado e o centro de cada agrupamento. O objeto é inserido no agrupamento com a menor distância, ou seja, maior similaridade. O centro do cluster é recalculado a cada novo elemento inserido (CAMILO; SILVA, 2009).

Entre alguns exemplos a serem destacados que utilizam agrupamento de organizações com características comuns estão a biologia com o agrupamento de seres parecidos, tais como reinos monera, fungi, animal e vegetal, temos também na sociedade atual as classes sociais que são agrupamentos de pessoas com as características monetárias em comum. Para que fique mais claro temos mais três exemplos citados por (PANG-NING et al., 2009):

• Recuperação de informações: Agrupar número de pesquisas na World Wide Web em um número pequeno de grupos.

• Clima: Agrupar padrões na pressão atmosférica de regiões polares e áreas do oceano que tenham impacto significativo sobre o clima da terra.

• Psicologia e Medicina: Agrupamento utilizado para identificar diferentes tipos de depressão.

3 FERRAMENTAS DE MINERAÇÃO DE DADOS

Com o desenvolvimento do DM *Data Minning*, surgiu um conjunto de ferramentas para auxiliar no processo de mineração de dados. Ferramentas como Oracle Data Mining (ODM)(ORACLE, 2017) , KNIME (KNIME, 2017), WEKA (WAIKATO, 2017), RapidMiner (RAPIDMINER, 2017) foram ganhando mercado e são utilizados por diversos profissionais. Neste capítulo dicorreremos sobre duas ferramentas que foram escolhidas por serem gratuitas e pela grande quantidade material disponível. Primeiramente, na seção 3.1 será apresentado o software WEKA e na seção 3.2 o software RapidMiner, a finalidade deste comparativo é realizar um detalhamento para que diante disto, possamos levantar uma discussão baseado em alguns critérios relevantes destas ferramentas, além disso utilizaremos um exemplo de uma partida de golf, observando duas classes: jogar ou não jogar e alguns atributos climáticos como: tipo de clima, temperatura, umidade e vento que será a nossa base de dados. Este exemplo será aplicado em ambas as ferramentas, analisando o comportamento dos softwares diante dos métodos de classsificação e agrupamento.

3.1 SOFTWARE WEKA

O software WEKA (Waikato Environment for Knowledge Analysis) (WITTEN; FRANK, 2000) é uma ferramenta de mineração de dados que contempla uma série de algoritmos de preparação de dados, de aprendizagem de máquina (mineração) e de validação de resultados. O WEKA foi desenvolvido na Universidade de Waikato na Nova Zelândia, sendo escrito em Java e possuindo código aberto disponível na Web (atualmente está na versão 3.8.0). Apresenta uma interface bem simples e auto explicativa, além disso conta com um conjunto de funcionalidades básicas apresentadas na Tabela 2:

Tabela 2 – Funcionalidades software WEKA

Funcionalidade	Descrição
Pré-processamento de dados	Realizar upload de arquivos de entrada e edição do arquivo.
Classificação	Métodos divididos em grupos (Bayesianos, vizinhança, regras etc.)
Agrupamento	Aprendizado não-supervisionado, incluindo EM, k-means
Visualização de dados	Gráficos de valores dos atributos por classe ou outros atributos.
Outras funções	Seleção de atributos, Regras de associação entre outros

Fonte: (WITTEN; FRANK, 2000)

Como visto na Tabela 2, o WEKA é uma ferramenta bem completa e que aborda

técnicas e termos já vistos neste trabalho, além disso conta com os métodos de classificação mais importantes que são árvore de decisão e Naive Bayes. A guia *preprocess* contempla um conjunto de funcionalidades, a partir dela é possível realizar a entrada de arquivos para exploracão, edição de arquivos, salvar arquivos editados, visualização gráfica entre outros.

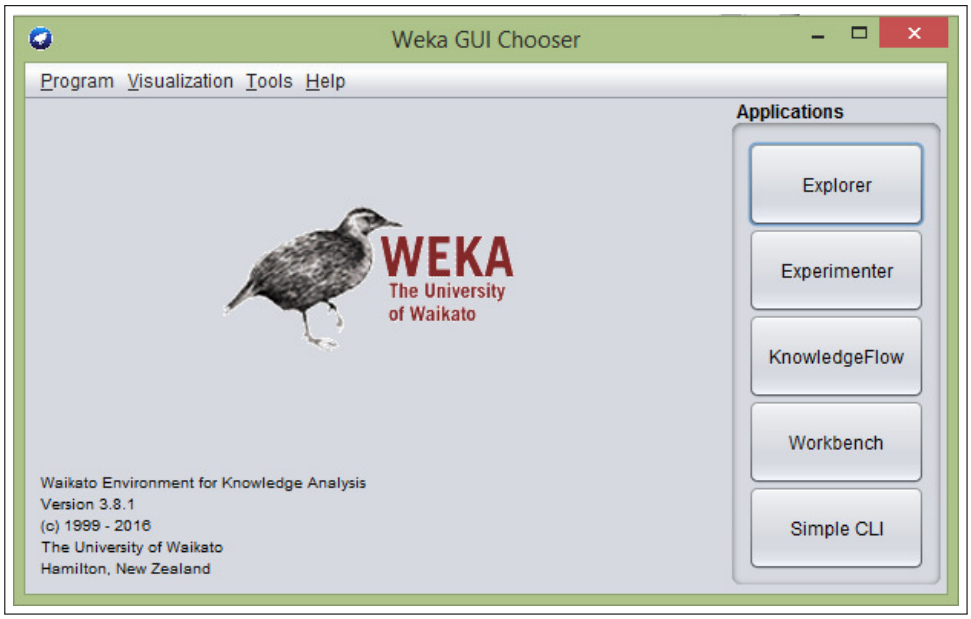
3.1.1 Instalação do software

A ferramenta está disponível para download no site: <http://www.cs.waikato.ac.nz/ml/weka/> juntamente com o código fonte. Possui diversas versões para diferentes sistemas operacionais como Windows, Linux e Mac OS X.

3.1.2 Tela inicial da ferramenta WEKA

Ao término do download e execução do software, a tela inicial possui esta interface demonstrada na Figura 4:

Figura 4 – Guia explorer WEKA



Fonte: (BOUCKAERT et al., 2010)

A inteface gráfica inicial consiste em uma tela bem amigável e com botões para iniciar as aplicações. Vejamos as definições para cada um destes botões segundo (BOUCKAERT et al., 2010):

• *Explorer*: Permite explorar dados para aprendizagem, suportando funcionalidades de upload de arquivos, visualização gráfica e escolha de algoritmos.

• *Experimenter*: Abre um ambiente para realizar experimentos e conduzir testes estatísticos entre esquemas de aprendizagem.

• *KnowledgeFlow*: Tem função similar ao do explorer entretanto possui uma interface de arrastar e soltar.

• *Workbench*: Similar ao explorer porém suportando menos funcionalidades

• *SimpleCLI*: Fornece uma interface simples de linha de comando que permite Execução de comandos WEKA para sistemas operacionais que não fornecem Sua própria interface de linha de comando.

Além dos botões e suas respectivas funcionalidades vistas acima, temos as abas *Program*, *Visualization*, *Tools* e *Help* no canto superior esquerdo da tela inicial. Vejamos algumas funcionalidades de cada aba de acordo com (BOUCKAERT et al., 2010):

Na aba *Program* temos a função Log Window, que possui a função de abrir uma janela de registro que captura tudo o que é impresso. Útil para ambientes como o MS Windows, onde WEKA normalmente não é iniciado a partir de um terminal.

Na aba *Visualization* temos as seguintes funções: *Plot* que possui a função de gráfico para plotar um gráfico 2D de um conjunto de dados; *TreeVisualizer* que exibe gráficos direcionados, por exemplo, uma árvore de decisão; *GraphVisualizer* que serve para visualizar gráficos de formato XML, BIF ou DOT, por exemplo, em redes bayesianas;*BoundaryVisualizer* que permite a visualização da decisão do classificador limites em duas dimensões;

Na aba *Tools* temos as seguintes funcionalidades: *PackageManager* que é uma interface gráfica para a gestão de pacotes do sistema WEKA; *ArffViewer* que tem a função de exibir o arquivo ARFF em formato de planilha, premite também a fucnionalidade de realizar alterações no arquivo; *SqlViewer* que representa uma planilha SQL, para consultar bancos de dados via JDBC e *ArffViewer* que é ma aplicação para edição, visualização e aprendizagem em redes Bayes.

3.1.3 Explorando Algoritmos de Mineração de Dados com WEKA

A partir da tela inicial, podemos ir para a guia explorer mencionada na subseção 3.1.2, nesta tela são realizados os testes em algoritmos de aprendizagem automática, permitindo sua exploração. Em relação as funcionalidades desta tela de acordo com (BOUCKAERT et al., 2010) podemos destacar as seguintes:

• *Preprocess*: Possibilita a escolha dos aquivos de entrada para exploração e modifi-

cação dos dados em execução.

• *Classify*: Permite treinar e testar os um conjunto de atributos e classes dos dados de entrada, utilizando algoritmos de aprendizagem baseados na técninca de classficação.

• *Cluster*: Aprender *clusters* para os dados, exemplo da seguimentação de clientes.

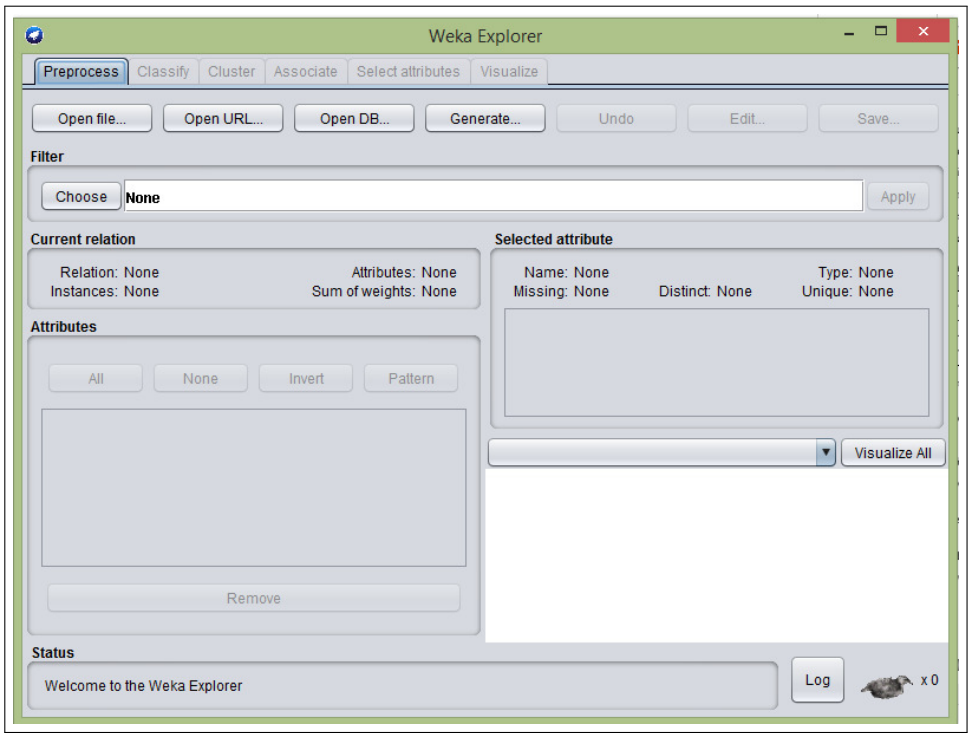
• *Associate*: Permite aprender regras de associação para os dados, como o exemplo da fralda e da cerveja visto na subseção 2.2.1.5.

• *Select attributes*: Selecionar os atributos mais relevantes nos dados.

• *Visualize*: Visualizar gráfico 2D interativo dos dados.

A seguir, a Figura 5 apresenta a tela principal da guia *explorer*, na parte superior de sua interface está o menu de navegação entre as abas, iniciando a exploração dos dados pela aba *preprocess* que é o início da exploração dos dados e nas abas seguintes os métodos de DM.

Figura 5 – Guia explorer WEKA



Fonte: (BOUCKAERT et al., 2010)

Para iniciar a exploração dos registros de entrada para exploração, devemos primeiramente abrir o arquivo com os registros de entrada, este arquivo pode ser nos formato JDBC, CSV e no formato nativo ARFF. A estrutura básica de um aquivo ARFF, constiste em uma seção de Cabeçalho (*Header*) e em uma seção de dados (*Data*), esta estrututra está detalhada a seguir de acordo com (WITTEN; FRANK, 2000):

• @*RELATION*: Define o nome do problema, que no caso acima o problema apresentado é o jogo.

• @*ATTRIBUTES*: define cada atributo e seu tipo de dado, que pode ser: numérico (real ou inteiro), string, nominal (lista entre) ou dado genérico (especifica formato).

• @*DATA*: define instâncias, sendo uma por linha, cada linha tem os valores de todos os atributos seguindo a mesma ordem.

A Figura 6, representa a estrutura básica do arquivo no formato ARFF, o arquivo apresenta um conjunto de atributos que implicam na tomada de decisão com obetivo de determinar se determinado dia está apropriado ou não para se jogar uma partida de golf. Por exemplo, caso esteja com clima de sol, temperatura quente, umidade alta e vento fraco, quer dizer que está favorável para o jogo.

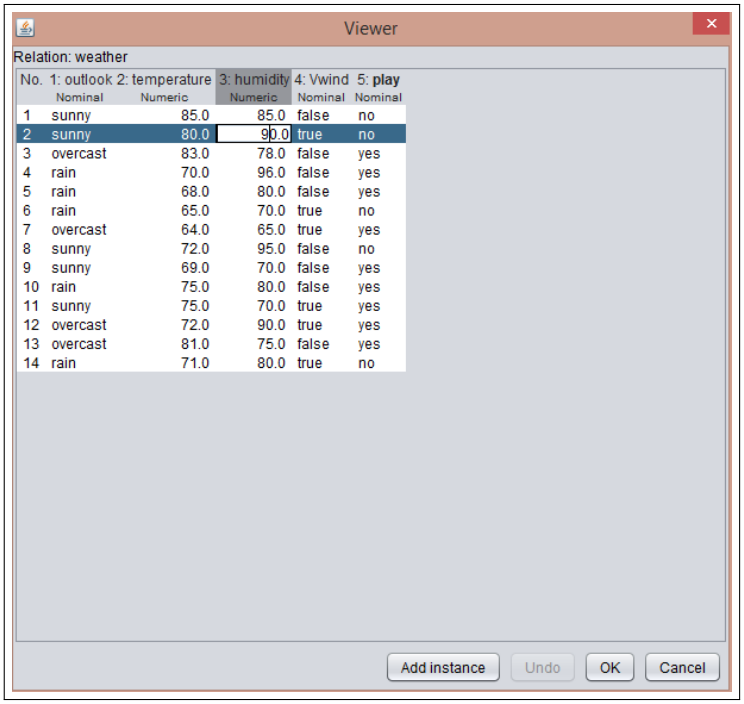
Figura 6 – Estrutura do Arquivo ARFF

```
@RELATION weather

@ATTRIBUTE outlook {sunny , overcast , rain}
@ATTRIBUTE temperature numeric
@ATTRIBUTE humidity numeric
@ATTRIBUTE wind {true, false}
@ATTRIBUTE play {yes , no}

@DATA

sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 78, false, yes
rain, 70, 96, false, yes
rain,68, 80, false, yes
rain, 65, 70, true, no
overcast, 64, 65, true, yes
sunny, 72, 95, false,no
sunny, 69, 70, false,yes
rain, 75, 80, false,yes
sunny, 75, 70, true,yes
overcast, 72, 90, true,yes
overcast, 81, 75, false,yes
rain, 71, 80, true,no
```

Fonte: (BOUCKAERT et al., 2010)

De acordo com Figura 7, a interface conta com a funcionalidade de edição do arquivo, abrindo uma nova guia de alteração, para isso basta clicar no botão *edit* no canto superior direito.

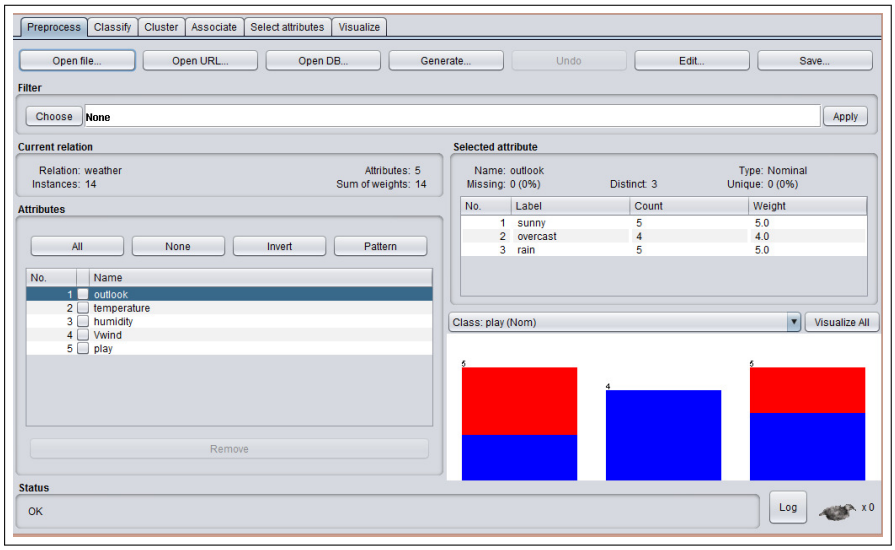
Figura 7 – Guia Explorer



Fonte: (BOUCKAERT et al., 2010)

Podemos acessar este arquivo diretamente na interface, permitindo algumas funcionalidades como: alteração dos atributos, remoção, número de instâncias, número de atributos, visualização de uma tabela de estatísticas para cada atributo selecionado e um gráfico baseado nos valores desta tabela. O gráfico pode ser visualizado separadamente ou em conjunto, a Figura 8 possibilita a visualização do aquivo na interface.

Figura 8 – Tela de edição

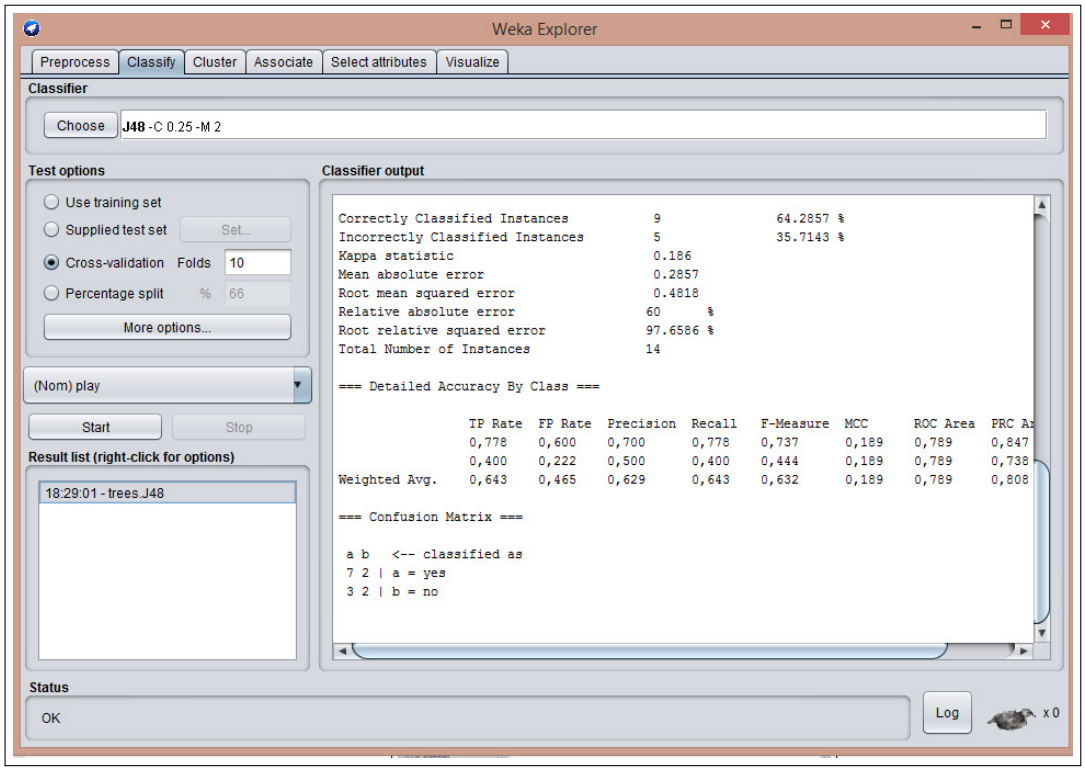


Fonte: (BOUCKAERT et al., 2010)

Como foi visto na subseção 3.1.3, temos as guias *classify*, *cluster*, *associate*, *select attributes* e *visualize* que permitem a exploração dos dados para aprendizagem automática e visualização de gráficos. A interface nos permite escolher diversos tipos de algoritmos nativos da ferramenta, a escolha do algoritmo varia de acordo com a necessidade a ser analisada. Todas as telas seguem o mesmo príncipio de usabilidade, para exemplificar apresentaremos os resultados obtidos utilizando algoritmos de classificação e algoritmos de agrupamento aplicados a um conjunto de registros de entrada, que no caso é a partida de golf. Após a execução dos algoritmos temos os resultados apresentados na Figura 9 e na Figura 10, ambas as telas permitem a escolha do algoritmo a ser utilizado.

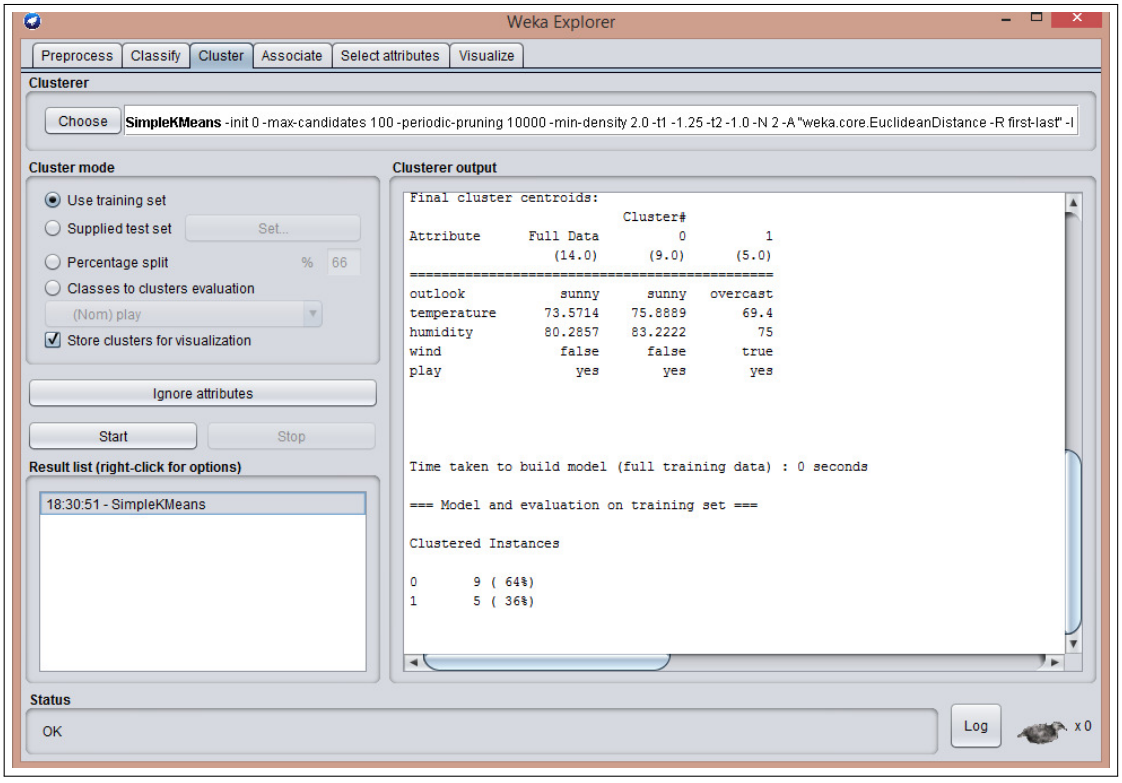
Para escolher o algoritmo basta clicar no botão *choose*, um conjunto de algoritmos serão listados, no caso da classificação entre um conjunto de algoritmos temos: na pasta *Bayes* o algoritmo NaiveBayes e na pasta *Tree* algoritmos de árvore de decisão, que tiveram seus fundamentos citados brevemente na seção 2.2. No caso do agrupamento o mecanismo de escolha dos algoritmos é análogo ao caso da classificação, variando os tipos de algoritmos a serem utilizados, uma vez que o objetivo final é agrupar os elementos com características semelhantes, como visto na subseção 2.1.2.

Figura 9 – Tela de Classifiação



Fonte: (BOUCKAERT et al., 2010)

Figura 10 – Tela de Agrupamento

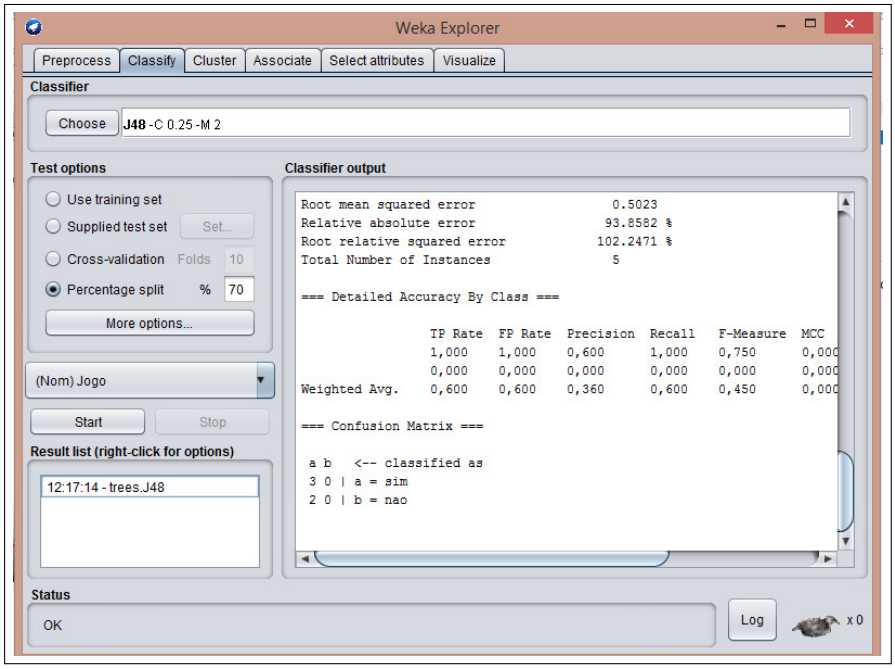


Fonte: (BOUCKAERT et al., 2010)

O princípio para exploração de algoritmos é bem intuitivo, apresenta na tela de pré-processamento opções de upload do arquivo de entrada e edição do arquivo após o upload, além disso possui um conjunto de tipos de algoritmos de classificação e agrupamento. Logo após a escolha do arquivo de entrada e do algoritmo, temos algumas opções de testes, no caso do *Percentage split*, serve para definir uma porcentagem de aprendizagem, e a porcentagem que falta será para testar o algoritmo, por exemplo se o valor da *Percentage split* for 70, esta porcentagem vai corresponder a quantidade a ser aprendida utilizando o algoritmo escolhido, o restante no caso 30 será para testar a aprendizagem.

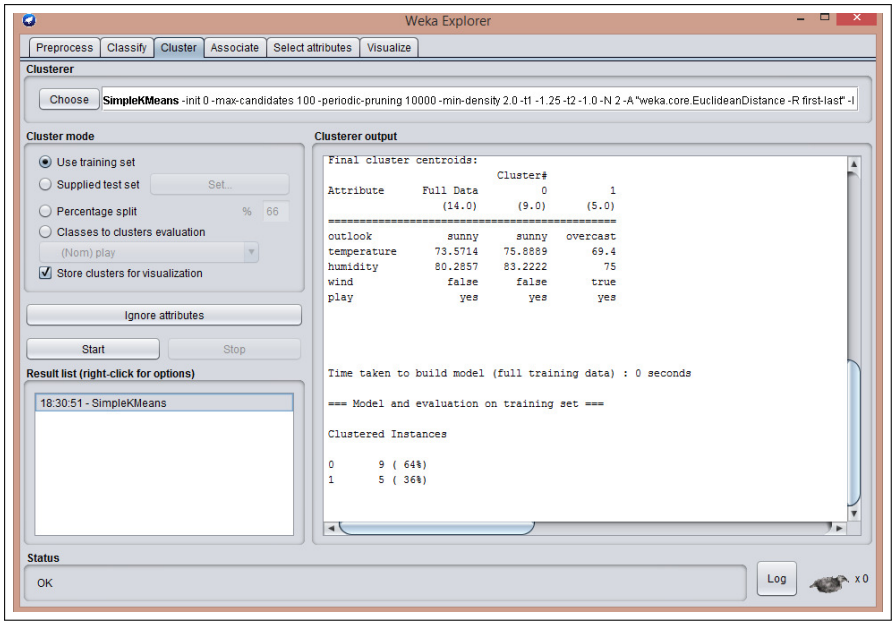
Para iníciar a exploração dos dados utilizando o algoritmo escolhido basta clicar em start, o algoritmo é executado sobre os registros de entrada fornecido no upload do aquivo, posteriormente a tela com as informações dos resultados aparecerá, conforme demonstra a Figura 11 para o exemplo da partida de golf utilizando o método de classificação e a Figura 12 demonstra o resultado do exemplo da partida de golf para o método de agrupamento.

Figura 11 – Tela de resultados *Classify*



Fonte: (BOUCKAERT et al., 2010)

Figura 12 – Tela de resultados *Cluster*



Fonte: (BOUCKAERT et al., 2010)

3.2 SOFTWARE RAPIDMINER

RapidMiner (RAPIDMINER, 2017) é uma ferramenta implementada em Java, open-source, que fornece a implementação de algoritmos utilizados em problemas de aprendizagem de máquina e uma interface gráfica para o desenvolvimento rápido de projetos para a criação de

modelos de predição (MIERSWA et al., 2006). Atualmente a ferrramenta se encontra na versão 7.4. Ao utilizar esta ferramenta é possível definir um processo de tratamento dos dados, inserindo operadores responsáveis por: I/O (entrada e saída); algoritmos de aprendizagem (supervisionados ou não-supervisionadas); funções de *on-line analytical processing*; pré-processamento; validação, e; visualização (BARTH, 2006).

O RapidMiner pode ser encontrado gratuitamente no seu próprio site<https://rapidminer.com/>, o software apresenta também versões profissionais, estas versões de acordo com o RapidMiner(RAPIDMINER, 2017) são destinadas a empresas por conter algumas melhorias em relação a versão gratuita, como por exemplo ganho de performance. A interface gráfica do RapidMiner apresenta um conjunto bem variado de funcionalidades de acordo com (BATISTA, 2010): apresenta uma interface gráfica muito intuitiva para o desenho de um processo de DM, além de mais de 500 operadores que implementam as mais diversas técnicas e algoritmos. Permite acesso a biblioteca de classes de aprendizagem automática Weka, uma das mais utilizadas na comunidade científica especializada; Acesso às mais diversas fontes de dados: Excel, Access, Oracle, IBM DB2,Microsoft SQL Server, Sybase, Postgress, SPSS, etc.

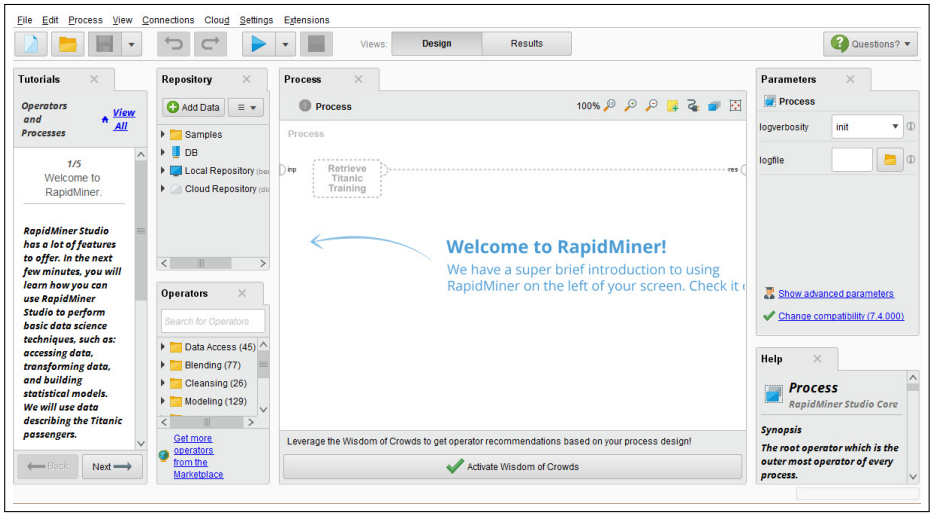
3.2.1 Instalação do software

A instalação do software RapidMiner é bem simples, o download está disponível no próprio site da RapidMiner <https://rapidminer.com/>, possui versões para Windows 32 bits, Windows 64bits, Mac 10.8+ e Linux.

3.2.2 Tela inicial da ferramenta RapidMiner

Ao término do download basta executar o software em uma sequência de telas até chegar ao botão *finish*, a Figura 13 demonstra a tela inicial da interface. A inteface gráfica inicial apresenta um layout bem elaborado, oferecendo um menu tradicional no canto superior esquerdo, série de botões e guias com diferentes funcionalidades, a aba *Design* apresenta todas as funcionalidades do software e a aba *Results* apresenta os resultados obtidos, ambas são acessadas separadamente sem necessidade de abrir uma nova janela.

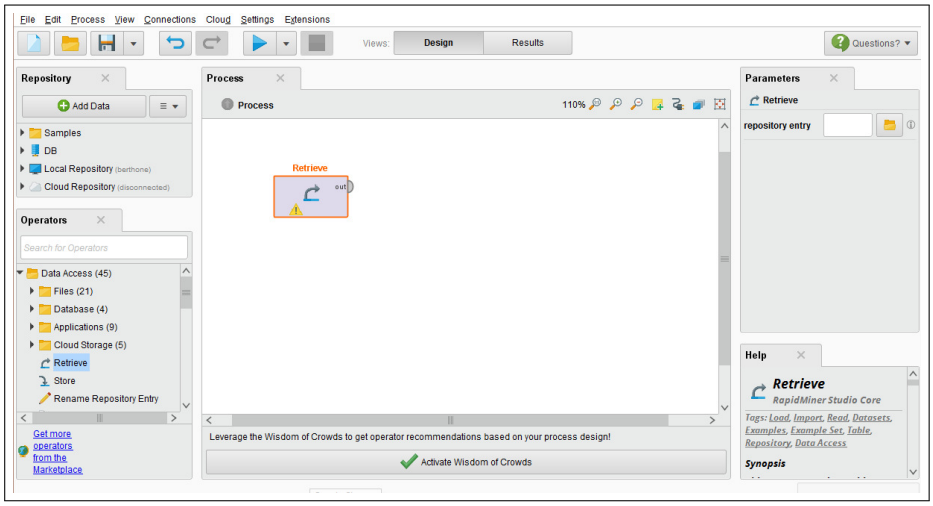
Figura 13 – Tela inicial RapidMiner



3.2.3 Explorando algoritmos de Mineração da Dados com RapidMiner

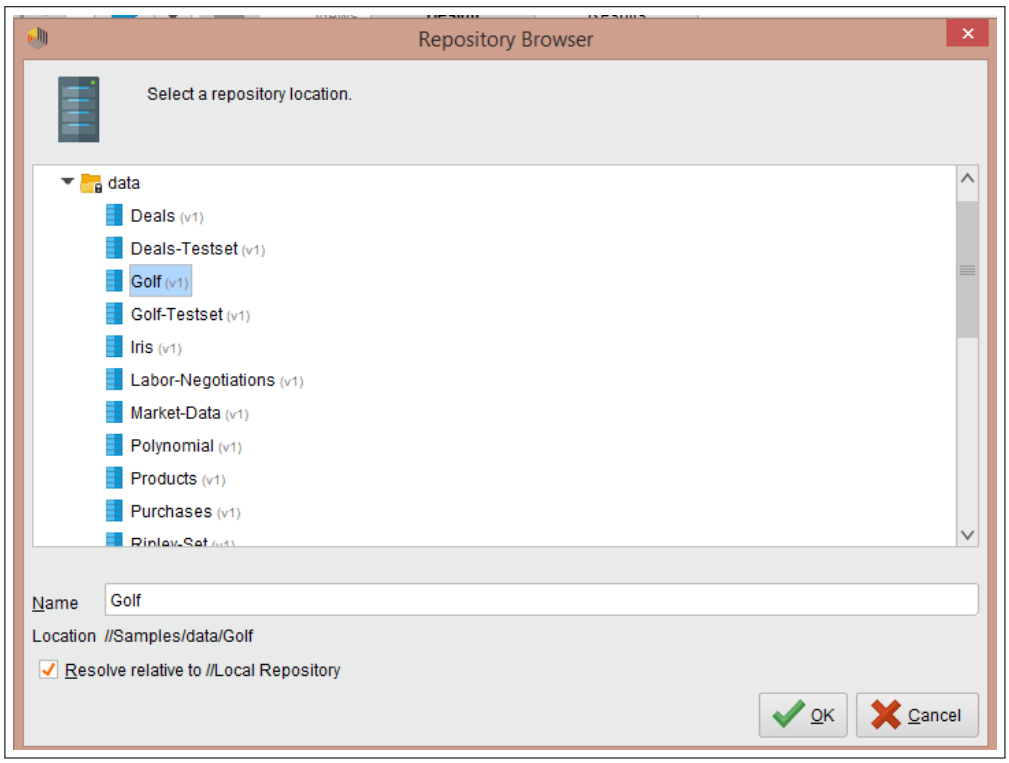
O primeiro passo para o uso da ferramenta é o upload do arquivo com extensão .CSV, para isso basta clicar no botão *Add Data*, para iniciar a análise primeiramente é preciso criar um novo processo no botão *New process*, caso já se tenha o processo criado, basta abri-lo através do botão *Open process*. A ferramenta conta com uma tecnologia de arrastar e soltar, no canto inferior esquerdo contém um conjunto de operadores, no qual o primeiro passo para se criar o modelo é criar um repositório de dados para aprendizagem, ou seja, no nosso caso é a partida de golf, os atributos e os dados serão a base de dados que permitirá o treinamento do modelo. A Figura 14 demonstra o repositório *retrieve*, este repositório tem a finalidade de receber os dados de entrada importados para exploração dos algoritmos de aprendizagem.

Figura 14 – Tela de processo RapidMiner



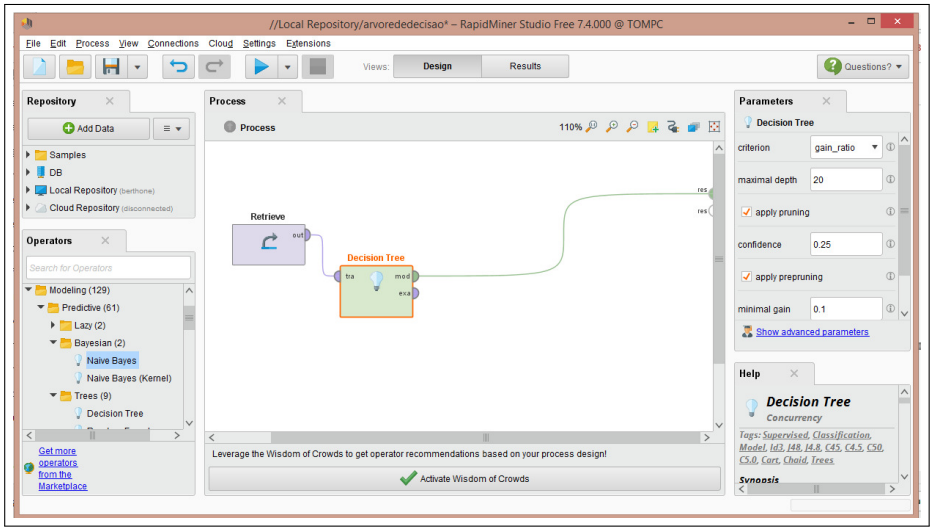
A próxima etapa é a escolha dos dados para o repositório, através dos *parameters* da interface é possível escolher a base de dados, a Figura 15 apresenta a seção de escolha deste repositório, podendo ser uma pasta local do usuário ou um repositório na nuvem.

Figura 15 – Escolha da base de dados



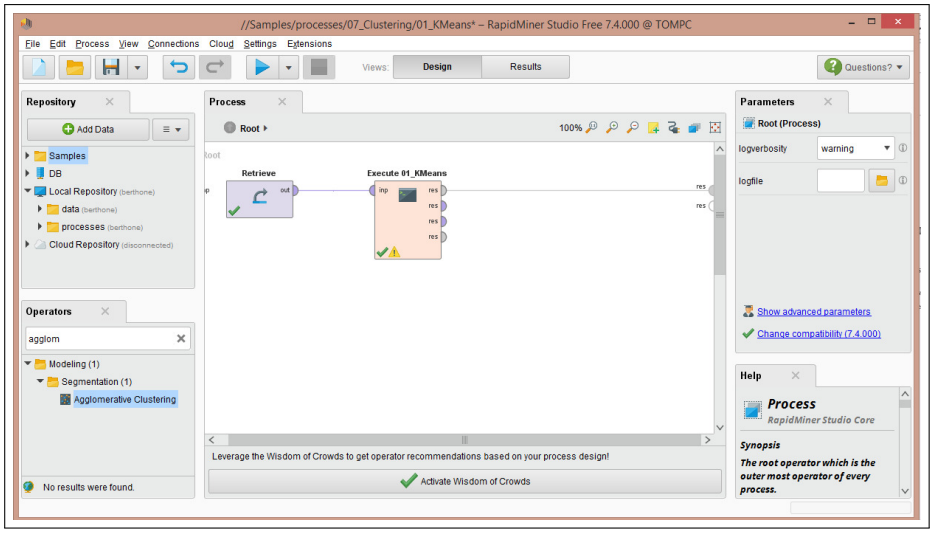
Realizada a escolha da base de dados (*retrieve*), analisaremos o comportamento da ferramenta para testar método de classificação e método de agrupamento utilizando os mesmos atributos e dados utilizados no software WEKA, para isto, devemos escolher primeiramente o método a ser aplicado para o nosso exemplo. Para o método de classificação, o exemplo utilizado foi uma partida de golf, observando duas classes: jogar ou não jogar. Diante dos registros de entrada fornecidos, a classificação tem a finalidade classificar a qual classe o novo registro pertence,ou seja, identificar baseado nas condições climáticas, a qual classe esse conjunto de condições pertence. Temos na janela de operadores uma lista de pastas que contém os métodos de mineração de dados, entre eles temos: *NavieBayes* e *DecisionTree*, ambos métodos de classificação. Os resultados obtidos são expressos em forma de árvore e em forma descritiva, apresentando de condições de jogo baseado nas condições do clima. Após a escolha do método, basta arrasta-lo e solta-lo na tela de processos, ligando-os a nossa base de dados utilizada, conforme a Figura 16:

Figura 16 – Escolha do método de *classify*



Em relação ao agrupamento, o processo é iniciado com a escolha da base de dados, utilizando os registros e classes do processo anterior de classificação. O método neste caso a ser aplicado ao exemplo é o agrupamento descrito na subseção 2.1.2 que tem o objetivo de agrupar registros com características similares, diferentemente da classificação, não posssui a função de predição, apenas identificação dos grupos. Para iniciar basta selecionar, arrastar e soltar na tela de processo, a partir disto, temos o algoritmo de agrupamento aplicado a nossa base de dados para análise, resultando na Figura 17.

Figura 17 – Escolha do método de *clustering*



Após a execução, temos a tela de resultados, neste caso da ferramenta RapidMiner assim como no caso do WEKA, foi o exemplo do jogo que se baseia em um conjunto de

registro de entrada. Os resultados do método de classificação utilizando árvore de decisão foram induzidos baseados nos dados e atributos do nosso caso de uso, visando estabelecer as condições para jogo de acordo com algumas condições climáticas. Os resultados obtidos se apresentam em forma de gráfico e em forma descritiva. A Figura 18 apresenta uma árvore de decisão no qual o fluxo vai para o filho esquerdo ou direito baseado nas condições climáticas, até chegar nas folhas, as folhas determinam duas opções para a partida, sim no caso de condições favoráveis para apartida ou não caso não haja condições para a partida. A Figura 19 demonstra o resultado obitido de forma descritiva, apresenta a quantidade de "sim"e "não"baseados nas condições climáticas necessárias para uma partida de golf.

Figura 18 – Gráfico do resultado de classificação

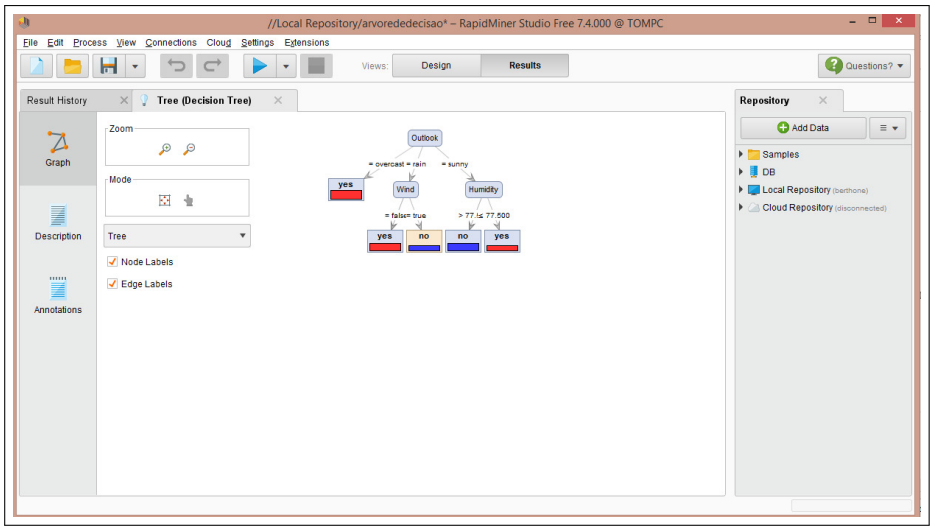
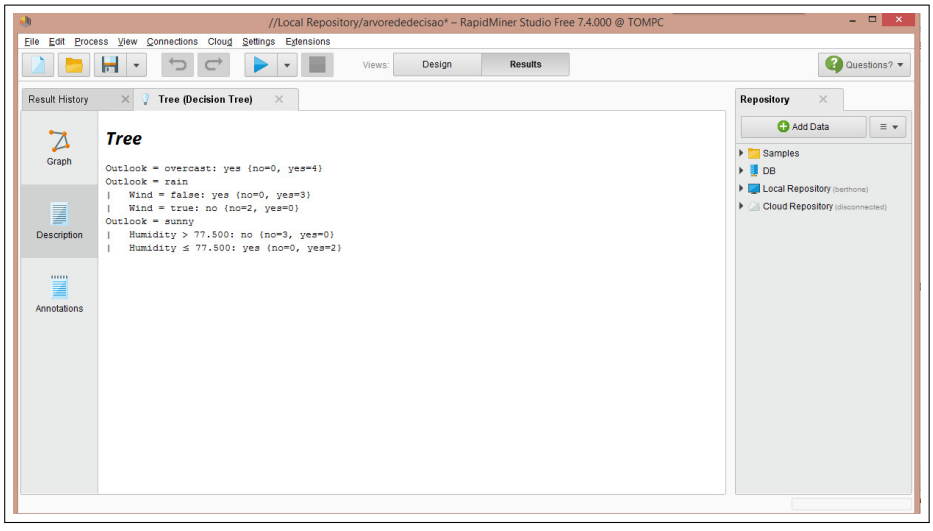


Figura 19 – Descrição do resultado de classificação



O método de *clustering* tem como objetivo agrupar elementos com características similares, para o nosso exemplo foram analisados um conjunto de registros de condições climáticas necessárias para uma partida de golf. A utilização do algoritmo pode ser ajustado, alterando os parâmetros do algoritmo de agrupamento, porém no nosso exemplo utilizamos os parâmetros padrão da ferramenta. O resultado foi a formação de 1 *cluster* baseados na similaridade dos registros. O agrupamento pode ser visualizado de duas maneiras, assim como na classificação, de forma gráfica Figura 20, no qual é possível analisarmos os elementos contidos em cada grupo ou em forma de descrição Figura 21, no qual é disponibilizado a quantidade de grupos, a quantidade de elementos em cada grupo e o total de elementos.

Figura 20 – Descrição do resultado utilizando método de agrupamento

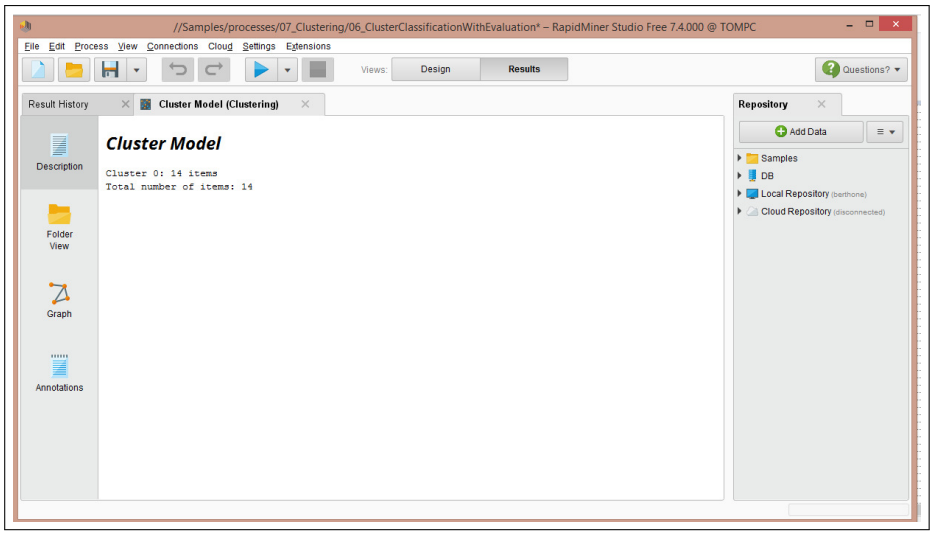
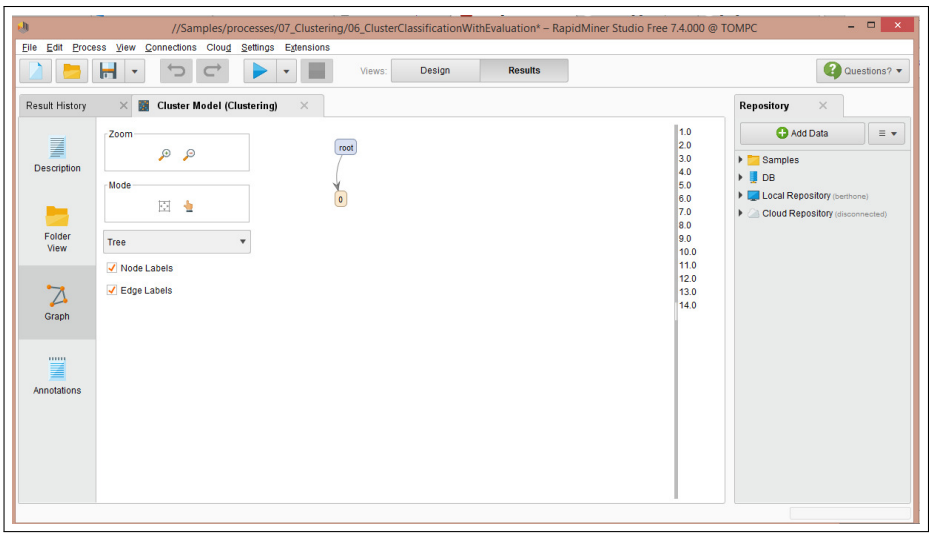


Figura 21 – Gráfico do resultado de agrupamento



4 DEMONSTRAÇÕES E DISCUSSÕES

No Capítulo 3 foi realizado um aprofundamento teórico e um exemplo prático acerca de duas ferramentas utilizadas para mineração de dados. A exploração das ferramentas teve o objetivo de detalhar alguns pontos importantes dos softwares como: o processo de instalação dos softwares, as funcionalidades mais relevantes dos softwares, a facilidade de uso das interfaces e o comportamento diante de um exemplo prático utilizando a mesma base de dados, esta base foi utilizada nas demonstrações das duas ferramentas. O software WEKA e o RapidMiner foram explorados com intuito de criar uma discussão sobre os seus pontos positivos e suas limitações. Em ambos, temos interfaces gráficas que facilitam a interação com o usuário, contribuindo de forma visual e intuitiva para uso de suas funcionalidades.

As informações obtidas no Capítulo 3 forneceram subsídio para a elaboração deste capítulo, ou seja, a discussão busca realizar um comparativo entre as ferramentas, este comparativo adota alguns critérios que serão detalhados com maior precisão no seção 4.1. O objetivo é apresentar um conjunto de critérios definidos no contexto deste trabalho de forma a permitir uma discussão comparativa entre as ferramentas.

4.1 CRITÉRIOS PARA DISCUSSAO DAS FERRAMENTAS

Nesta seção é demonstrado um conjunto de critérios relevantes para realizar uma discussão sobre as ferramentadas apresentadas no Capítulo 3. A base para o nosso objetivo é ter critérios específicos e que pudessem demonstrar a utilidade destas ferramentas. Entre os critérios a serem detalhados na seção 4.1, temos a forma de download dos softwares, verificando se seu download é disponibilizado gratuitamente ou se seu download é permitido apenas com a realização de alguma forma de pagamento. Analisaremos o material de apoio para o uso da ferramenta, com a finalidade de identificar a quantidade de documentação, exemplos disponíveis para aprendizado e tutoriais, além disso, será analisado os tipos de métodos de mineração de dados suportadas pelo WEKA e RapidMiner.

Além dos critérios citados acima, é verificado outros critérios como: a facilidade de uso das ferramentas, critério muito importante para os usuários que buscam por praticidade ou para fins didáticos por exemplo; os tipos de arquivos de entrada suportados, identificando qual ferramenta possui a maior ou menor quantidade de formatos de arquivo a serem explorados pelas ferramentas; verificar os dados gerados na saída do programa, se os dados são gerados apenas no

console ou se estes dados podem ser exportados em algum formato específico; verificar grupos que discutam sobre a utilização e funcionalidades dos softwares com o objetivo de fornecer a maior quantidade de informações possíveis com fóruns de discussão; apresentar desenvolvedores das ferramentas, buscando um histórico do seu surgimento; verificar o período de atualizações, a frequência com que estas atualizações são realizadas, pois ferramentas que não costumam se atualizar vão ficando com as funcionalidades ultrapassadas. A seção 4.1 guia essa discussão levando em consideração os critérios brevemente citados anteriormente, desta forma a discussão tem a finalidade de tornar relevante a comparativo entre as duas ferramentas de mineração de dados.

4.2 DISCUSSÃO SOBRE AS FERRAMENTAS DE MINERAÇÃO DE DADOS

Nesta seção é demonstrado como as ferramentas se apresentam diante de alguns critérios estabelecidos para a discussão entre elas. Nesse contexto, o objetivo é apresentar uma discussao de acordo com os critérios definidos neste trabalho, dando subsídio aos usuários que necessitam de uma ferramenta de mineração de dados.

O WEKA é um software livre (de código aberto) para mineração de dados, desenvolvido em Java, dentro das especificações da GNU (*General Public License*)(BOUCKAERT et al., 2010). O RapidMiner é um software que possui versão gratuita para download no seu próprio website, além disso possui versões pagas. A diferença entre a versão gratuita e paga são algumas melhorias como: suporte prestado e desempenho, entre outros. Os softwares apresentados possuem diversos materiais de apoio para os usuários que necessitam de suas utilizações, o RapidMiner apresenta um manual de usuário forncecido pelo próprio fabricante, neste manual consta a apresentação do software e funcionalidades, explicando de forma detalhada todas as suas funções, entretanto este manual possui apenas versão em inglês e pode ser acessado no seguinte endereço eletrônico <https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf>. Além disso, apresenta tutoriais de funcionamento elaborados pelos próprios usuários em forma de vídeos.

O WEKA possui uma extensa documentação, o fabricante fornece um manual de funcionamento assim como no RapidMiner, este manual explica todo o funcionamento do software, porém o manual possui versão em português, podendo ser fácilmente encontrado nos mecanismos de busca, o endereço para download do manual é: <https://users.info.uvt.ro/~dzaharie/dm2016/labEN/lab1/biblio/WekaManual-3-7-13.pdf>. O WEKA possui materiais

de apoio disponibilizados por universidades de todo o Brasil, tutoriais são encontrados com facilidade, abordando as suas funcionalidades e utilização de algoritmos de aprendizagem.

As tarefas e métodos de mineração de dados citados no Capítulo 3, são abordados pelas ferramentas, o RapidMiner traz um conjunto de algoritmos que auxiliam o usuário, entre os métodos suportados estão: algoritmos de classificação utilizando árvore de decisão, Naive Bayes, Redes Neurais, Associações, algoritmos de agupamento e um conjunto de outros métodos de mineração de dados. O WEKA em seu ambiente apresenta entre os métodos de classificação, associação e agrupamento que aborda, agoritmos de árvore de decisão, Naive Bayes, Kmeans e Apriori, possui uma relação bem extensa de algoritmos em seu console. Outro ponto importante seria a rapidez e usabilidade das ferramentas para criar modelos utilizando algoritmos de aprendizagem, o WEKA apresenta um ambiente intuitivo e simples, com interface gráfica auto-explicativa, com isso consegue ser menos pesado que o RapidMiner, por outro lado o RapidMiner possui uma interface gráfica mais elaborada e com riqueza de detalhes, layout com diversidade de cores e com recurso de arrastar e soltar.

A usabilidade de ambos os softwares é bem fácil, porém de acordo com (ROGERS et al., 2011), para uma análise mais aprofundada é necessário a aplicação de um questionário a um cojunto de usuários para verificar algumas metas de usabilidade como: eficácia no uso, fácilidade de lembrar seu uso, fácilidade de aprender. O WEKA apresenta mais simplicidade em sua interface gráfica, conta ainda com a edição do arquivo depois do upload na aba de pré-processamento. Nas abas que contém os métodos de mineração de dados possui muita praticidade, permitindo ao usuário a fácil navegação pelos algoritmos disponíveis na ferramenta, organizados em pastas. De modo geral o seu layout organizado em abas facilita a usabilidade e navegação do software.

O RapidMiner por sua vez, apresenta uma interface gráfica com maior quantidade de recursos em relação ao WEKA. Apesar do layout melhorar a visualização dos resultados e da ferramenta de forma geral, sua usabilidade se tornou um pouco mais complexa devido a sua riqueza de detalhes em seu layout. As ferramentas possuem a funcionalidade de realizar upload de arquivos de entrada que serão submetidos a mineração. O WEKA suporta uma série de formatos de arquivos de entrada aceitos, entre eles estão o ARFF que é um formato nativo da ferramenta, este formato foi apresentado na subseção 3.1.3, além do ARFF permite outros formatos como: CSV que é um arquivo de planilha, ASCII e JDBC. O RapidMiner por sua vez aceita arquivos de entrada no formato CSV, Excel e banco de dados SQL, diferentemente do

WEKA, não possui um formato nativo.

As ferramentas apresentam os resultados de suas operações de várias formas para os usuários, com o objetivo de facilitar a visualiazação e utilização dos dados. O RapidMiner apresenta a visualização de seus dados das seguintes formas: forma descritiva no próprio console, demonstrando os dados obtidos em forma de texto; visualização gráfica; dados em formato de tabela, o software não permite a exportação dos resultados. O WEKA apresenta seus resultados em forma descritiva no próprio console e não permite a exportação dos resultados.

Grupos de discussão que abordam a utilização das ferramentas são muito importantes para o desenvolvimento e aprendizado dos usuários. Nesse contexto, uma busca rápida sobre comunidades e fóruns que abordam dúvidas e discussão sobre os softwares foi realizado. O RapidMiner possui fóruns e comunidades facilmente encontradas, possui comunidade no seu próprio site, além disso possui fóruns criados por seus próprios usuários. O WEKA possui diversos fóruns espalhados pela internet, inclusive fóruns brasileiros, sua expansão e utilização é maior em relação ao RapidMiner, por essa razão apresenta maior número de comunidades e grupos de discussão.

Em relação ao surgimento dos softwares, o WEKA foi iniciado em 1992, software para mineração de dados desenvolvido em java com código aberto, desenvolvido na Universidade de Waikato (Nova Zelândia), por Ian Witten, Eibe Frank e colaboradores (WITTEN; FRANK, 2000). O RapidMiner foi desenvolvido em Java, permitindo a sua utilização versátil em qualquer sistema operativo e ambiente de trabalho. O projeto RapidMiner começou em 2001 por Ralf Klinkenberg, Ingo Mierswa e Simon Fischer na Unidade de Inteligência Artificial da Universidade de Dortmund (Alemanha). Em 2006, Ingo Mierswa e Ralf Klinkenberg fundaram a empresa Rapid-I1 (RapidMiner – brochura oficial) (MORAIS, 2012).

Ambos os softwares demonstram de acordo com suas funcionalidades boas ferramentas para auxiliio de usuários no processo de mineração de dados. As atualizações disponíveis para um software são importantes para a continuidade de uma ferramenta de mineração de dados, pois elas tem como objetivos principais corrigir falhas no sistema, melhorar o layout, acrescentar novas funcionalidades entre outros. O WEKA apresenta versões de seu software disponível para download, estas versões estão sendo lançadas devido ao crescimento das comunidades e grupos de discussão que cobram dos desenvolvedores atualizações periódicas, diante disto e de fatores internos o WEKA atualiza seu software com o objetivo de melhorar a sua interação com os usuários, melhorar as suas fucnionalidades e corrigir falhas.

O RapidMiner apesar de ser uma ferramenta mais recente que o WEKA se encontra na versão 7.5, o que demonstra que o software tem curtos períodos de atualização, por ter versões para empresas, a primordialidade de se atualizar é ainda mais necessária. Na opinião do autor a ferramenta WEKA devido a sua simplicidade e facilidade de uso se apresentou mais viável para usuários iniciantes, contudo o RapidMiner é um pouco mais complexo de forma geral abrangendo versões gratuitas e empresariais no qual a ferramenta é mais completa. A Tabela 3 apresenta um comparativo entre as duas ferramentas apresentadas, os critérios discutidos neste capítulo serve como base para realizar este comparativo, observando o comportamento dos softwares mediante os critérios escolhidos.

Tabela 3 – Comparativo entre WEKA e RapidMiner

Critérios	WEKA (WAIKATO, 2017)	Rapidminer (RAPIDMINER, 2017)
Gratuito/Pago	Gratuito.	Gratuito e Pago.
Documentação	Extenso Material de Apoio em relação ao RapidMiner.	Razoável quantidade de material de apoio em relação ao WEKA.
Tipos de Tarefas de Mineração	Classificação, Associação, Agrupamento entre outros.	Classificação, Associação, Agrupamento entre outros.
Usabilidade	Grande facilidade de uso em relação ao RapidMiner	Facilidade de uso razoável em relação ao WEKA
Arquivos de Entrada	ARFF, CSV, ASCII e JDBC.	CSV, Excel e banco de dados SQL .
Arquivos de Saída	Saída no prório console.	Saída no prório console.
Grupos de Discussão	Grande quantidade de grupos de discussão.	Menor quantidade de grupos de discussão.
Atualizações	Atualização periódica.	Atualização periódica.

5 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho introduziu uma discussão sobre duas ferramentas de mineração de dados muito utilizadas pelo campo científico. A discussão gira em torno de critérios relevantes para os softwares, baseado nos critérios estabelecidos no contexto deste trabalho, o WEKA teve como um ponto bastante positivo a sua distribuição gratuita. O WEKA é objeto de estudo em diversos artigos científicos que abordam exemplos práticos, o motivo de sua maior utilização está na seu custo-benefício, pois não há custo e o software entrega praticidade e simplicidade, além disso é bem leve, tornando-o uma ferramenta bem ágil. O WEKA apresenta uma interface gráfica mais simples e bem organizada, isso permite a exploração de tarefas de mineração de dados com maior rapidez e facilidade.

O RapidMiner apresenta como ponto bastante positivo um layout mais elaborado e com mais recursos que o WEKA, a apresentação dos resultados no RapidMiner é mais amigável pois pode ser visualizado de várias formas diferentes, inclusive em forma gráfica, enquanto no WEKA os resultados são visualizados com menos riqueza de detalhes. O WEKA apresenta com conjunto bem maior de material de apoio, tutoriais e exemplos práticos, podendo ser facilmente encontrado na internet, além disso, possui maior número de comunidades e grupos de discussão. Diante do conjunto de informações obtidas sobre os softwares, o WEKA é uma ferramenta com maior simplicidade e facilidade de uso sem perder funcionalidades relevantes para a mineração de dados, portanto se aplica perfeitamente a usuários iniciantes e para usuários que buscam por aprendizado. O RapidMiner devido a sua versão para empresas com, layout mais elaborado com menor facilidade de uso, é mais aplicável para usuários mais experientes e para fins comerciais em relação ao WEKA.

Este trabalho deixa sua contribuição, tendo em vista que promove uma discussão sobre o detalhamento de duas ferramentas de mineração de dados, este detalhamento teve a finalidade de apresentar para os usuários alguns pontos importantes, destacando os pontos positivos e negativos de cada software, desta forma os usuários podem analisar ambas as ferramentas com o objetivo de verificar a que melhor atende suas espectativas e a que melhor vai se comportar diante de suas problemáticas.

Propõe-se que em trabalhos futuros outras ferramentas de mineração de dados sejam utilizadas, visto que atualmente são diversos softwares gratuitos e profissionais disponíveis, desta forma seria válido o comparativo entre ferramentas que não foram objetos deste trabalho.

REFERÊNCIAS

AMO, S. de. Técnicas de mineração de dados. **Jornada de Atualizaçao em Informatica**, 2004.

BAKER, R. et al. Data mining for education. **International encyclopedia of education**, Oxford, UK: Elsevier, v. 7, n. 3, p. 112–118, 2010.

BARTH, F. J. Mineraçao de regras de associaçao em servidores web com rapidminer. **Disponível no site< http://fbarth. net. br/materiais/webMining/>. Brasil**, 2006.

BATISTA, P. R. L. **Data mining na identificação de atributos valorativos da habitação**. Dissertação (Mestrado) — Universidade de Aveiro, 2010.

BOUCKAERT, R. R.; FRANK, E.; HALL, M.; KIRKBY, R.; REUTEMANN, P.; SEEWALD, A.; SCUSE, D. Weka manual for version 3-7-3. **The university of WAIKATO**, 2010.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, p. 1–29, 2009.

CÔRTES, S. da C.; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de dados-funcionalidades, técnicas e abordagens**. [S.l.]: PUC, 2002.

DIAS, M. M.; FILHO, L. A. da S.; LINO, A. D. P.; FAVERO, E. L.; RAMOS, E. M. L. S. Aplicação de técnicas de mineração de dados no processo de aprendizagem na educação a distância. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2008. v. 1, n. 1, p. 105–114.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques , University of Illinois at Urbana-Champaign**. [S.l.]: Elsevier, 2006.

HARRISON, T. H. **Intranet data warehouse: ferramentas e técnicas para a utilização do data warehouse na intranet**. [S.l.]: Berkerley/ABDR, 1998.

KNIME. **KNIME**. 2017. <https://www.knime.org//>. [accessado em: 06 de junho de 2017].

MIERSWA, I.; WURST, M.; KLINKENBERG, R.; SCHOLZ, M.; EULER, T. Yale: Rapid prototyping for complex data mining tasks. In: ACM. **Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2006. p. 935–940.

MIRANDA, M. N. de. **Algoritmos Genéticos: Fundamentos e Aplicações**. 2007.

MORAIS, S. F. dos S. Sistemas de recomendação em rapid miner: um caso de estudo. 2012.

ORACLE. **Oracle**. 2017. <http://www.oracle.com/thecnology/products/bi/odm/index.html>. [accessado em: 06 de junho de 2017].

PANG-NING, T.; STEINBACH, M.; KUMAR, V. Introdução ao “data mining”. **Rio de Janeiro: Ciência Moderna**, 2009.

RAPIDMINER. **RapidMiner**. 2017. <http://https://rapidminer.com/>. [accessado em: 06 de junho de 2017].

ROGERS, Y.; SHARP, H.; PREECE, J. **Interaction design: beyond human-computer interaction**. [S.l.]: John Wiley & Sons, 2011.

WAIKATO. **WEKA**. <http://www.cs.waikato.ac.nz/ml/weka/>. Acessado em maio de 2017.

WAIKATO, U. **WEKA**. 2017. <http://http://www.cs.waikato.ac.nz/ml/weka/>. [accessado em: 06 de junho de 2017].

WITTEN, I. H.; FRANK, E. Weka. **Machine Learning Algorithms in Java**, p. 265–320, 2000.