USO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA INFERIR PADRÕES DE OCORRÊNCIA DA BRUSONE NO TRIGO[1]

Isabela dos Santos Corrêa[2]

Alexandre Tagliari Lazzaretti[3]

RESUMO

O aumento considerável nas bases de dados das mais diferentes áreas colaborou para a necessidade do uso de técnicas que auxiliassem a manipulação destes, uma vez que a busca tradicional de dados não era suficiente para o aproveitamento dos mesmos. O processo de mineração de dados procura extrair informações que estejam implícitas, e/ou previamente desconhecidas e as transformam em algo significativo para a tomada de decisões. Para auxiliar nesse gerenciamento, ferramentas computacionais descobridoras de novos conhecimentos são essenciais. A mineração de dados pode ser considerada como uma série de passos para obtenção de conhecimento dos padrões ocultos e úteis nesses dados através da aplicação de conceitos, métodos, ferramentas e técnicas. A Brusone do trigo é considerada uma doença de impacto e a sua ocorrência está associada à condições meteorológicas favoráveis. Neste sentido, este trabalho aplica algoritmos de mineração de dados em um conjunto de bases de dados com o objetivo de identificar padrões de ocorrência da doença.

Palavras-chave: Descoberta do conhecimento. Classificação. Pyricularia grisea.

1 INTRODUÇÃO

Das condições climáticas existentes no Brasil, as altas temperaturas e precipitações pluviais frequentes tendem a favorecer o desenvolvimento de inúmeras doenças, principalmente aquelas causadas por fungos. Segundo Goulart, Sousa e Urashima (2007),

> A severidade da Brusone do trigo varia grandemente em função da região, das condições climáticas e da cultivar em questão. A doença vem sendo considerada de importância econômica nos locais onde tem ocorrido, devido à intensidade dos sintomas que produz, principalmente nas espigas.

---

[1] Trabalho de Conclusão de Curso (TCC) apresentado ao Curso de Tecnologia em Sistemas para Internet do Instituto Federal Sul-rio-grandense, Câmpus Passo Fundo, como requisito parcial para a obtenção do título de Tecnólogo em Sistemas para Internet, na cidade de Passo Fundo, em 2017.

[2] Graduanda em Tecnologia em Sistemas para Internet do Instituto Federal de Educação, Ciência e Tecnologia Sul-Rio-Grandense. E-mail: izha_bela@hotmail.com.

[3] Orientador: professor do IFSul. E-mail: alexandre.lazzaretti@passofundo.ifsul.edu.br.

Tendo por objetivo inferir padrões que pré-dispõem o desenvolvimento da doença Brusone, e assim proporcionar um melhor manejo de defesa contra a mesma, pretende-se ao longo das fases que compõem a mineração de dados, poder juntamente com os algoritmos disponibilizados pelo framework WEKA[4], aplicar a mineração de dados em dados meteorológicos reais, a fim de se chegar ao mais preciso padrão para inferir a Brusone em cultivares de trigo.

2 REFERENCIAL TEÓRICO

Nesta seção encontram-se os conceitos que abrangem o tema abordado e os tópicos estudados para compreensão dos resultados obtidos.

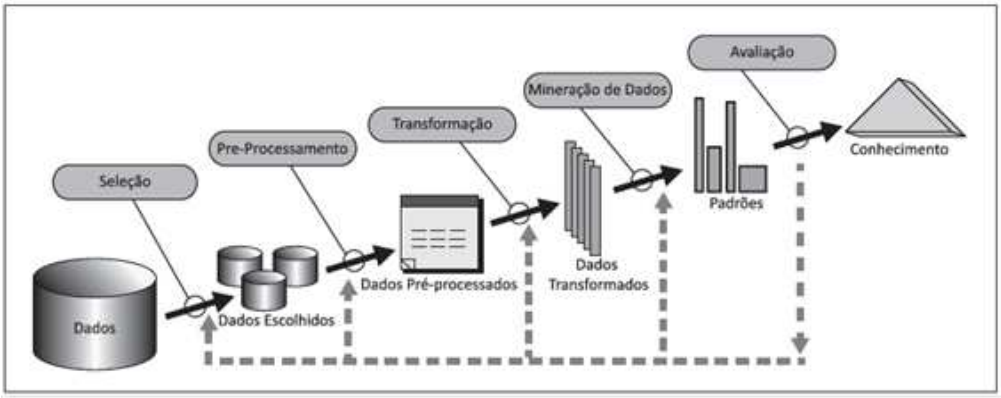
2.1 Processo de descoberta do conhecimento

A descoberta de conhecimento em bancos de dados (KDD) permite esclarecer a obtenção de relações entre dados de uma base. Silva (2004) aborda o KDD como:

> O processo não trivial de identificar em dados padrões que sejam válidos, novos (previamente desconhecidos),potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão."(SILVA apud FAYYAD et al. 1996).

De acordo com FAYYAD et al. (1996), o processo de KDD é constituído de diversas fases, observando-se a Figura 1 tem-se:

Figura 1. Fases do Processo de KDD.



Fonte: [Fayyad, Piatetsky-Shapiro e Smyth. 1996].

[4] Pacote de algoritmos de aprendizagem de máquina para resolver problemas reais de mineração de dados.

Após a fase inicial, o foco passa a ser a escolha ou seleção da massa de dados a ser minerada, podendo ser um conjunto de dados ou um subconjunto de variáveis onde a extração será realizada.

A fase de limpeza dos dados ou pré-processamento tem por objetivo assegurar a qualidade dos dados envolvidos no KDD realizando operações básicas como a remoção de ruídos, que podem ser, por exemplo, atributos nulos.

A fase seguinte consiste na seleção e transformação dos dados em que serão selecionados os atributos realmente interessantes ao usuário, além de transformados utilizando o padrão ideal para aplicar algoritmos de mineração.

Após a realização das fases anteriores, a mineração de dados é iniciada. Esta fase é a mais importante do KDD, sendo realizada através da escolha do método e do algoritmo mais compatível com o objetivo da extração, a fim de encontrar padrões nos dados que sirva de subsídios para descobrir conhecimentos ocultos.

A avaliação ou pós-processamento é a fase que identifica, entre os padrões extraídos na etapa de mineração de dados, os padrões interessantes ao critério estabelecido pelo usuário, podendo voltar à fase inicial para novas iterações.

Ao término da avaliação, o conhecimento descoberto deverá ser implantado e incorporado ao sistema, sempre documentando e publicando os métodos, a fim de apresentar o conhecimento descoberto ao usuário.

Tendo por base as fases relatadas por FAYYAD et al. (1996) pode-se perceber a importância da descoberta do conhecimento na escolha dos dados certos a serem minerados, uma vez que tornará a tarefa de mineração mais precisa e automatizada.

2.2 MINERAÇÃO DE DADOS

Segundo Camilo e Silva:

> "A manipulação dos dados e a análise das informações de maneira tradicional tornou-se inviável devido ao grande volume de dados (coletados diariamente e armazenados em bases históricas). Descobrir padrões implícitos e relacionamentos em repositórios que contém um grande volume de dados de forma manual, deixou de ser uma opção. As técnicas de mineração passaram a estar presentes no dia a dia." (CAMILO e SILVA. 2009).

Partindo-se desta afirmativa, entende-se mineração de dados como o conjunto de todas as técnicas que possibilitam extrair conhecimento de uma massa de dados.

2.2.1 Métodos, tarefas e técnicas de data mining

De maneira geral, as tarefas de mineração de dados podem ser separadas em duas categorias: descritiva e preditiva. As tarefas descritivas se caracterizam pelas propriedades gerais encontradas nos dados, enquanto as tarefas preditivas fazem uso das variáveis já conhecidas do banco de dados para predizer padrões, ainda desconhecidos (HAN; KAMBER, 2006).

Camilo e Silva (2009) destacam diversas tarefas de mineração de dados, a seguir é apresentada a tarefa mais relevantes a esta pesquisa:

- Classificação (Classification)- Uma das tarefas mais comuns, a Classificação, visa identificar a qual classe um determinado registro pertence. Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de 'aprender' como classificar um novo registro (aprendizado supervisionado);

2.3 WEKA

Silva (2004) define o framework WEKA como "um pacote de algoritmos de aprendizagem de máquina para resolver problemas reais de mineração de dados."

Contendo ambiente para experimentação, teste e comparação dos modelos de aprendizado, ainda permite a preparação dos dados a través de filtros, métodos de discretização, de maneira parametrizada.

Após a execução do algoritmo selecionado, através da área *Classifier output* é possível verificar os resultados obtidos. A Figura 2 demonstra a saída de execução de um algoritmo de árvore de decisão.

Figura 2: Resultado exibido na área Classifier output

```
Size of the tree : 1437

Time taken to build model: 0.39 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8775               66.9234 %
Incorrectly Classified Instances      4337               33.0766 %
Kappa statistic                          0.3384
Mean absolute error                      0.3694
Root mean squared error                  0.479
Relative absolute error                 73.8726 %
Root relative squared error             95.8071 %
Total Number of Instances           13112

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,693    0,355    0,663      0,693    0,678      0,339    0,729     0,715     SIM
               0,645    0,307    0,676      0,645    0,660      0,339    0,729     0,716     NAO
Weighted Avg.  0,669    0,331    0,670      0,669    0,669      0,339    0,729     0,715

=== Confusion Matrix ===

    a    b   <-- classified as
 4561 2018 |    a = SIM
 2319 4214 |    b = NAO
```

Fonte: Do autor

Segundo a documentação disponibilizada pela ferramenta, os valores obtidos em Correctly Classified Instances e Incorrectly Classified Instances são determinantes para acurácia prevista, uma vez que exibem os valores de classificação correta e classificação incorreta obtidos pelo algoritmo.

Em Kappa statistic encontra-se o valor referente a capacidade de aprendizagem do algoritmo.

Em Confusion Matrix é exibida a matriz de confusão de uma hipótese, a hipótese oferece uma medida efetiva do modelo de classificação, ao mostrar o número de classificações corretas versus as classificações preditas para cada classe, sobre um conjunto de exemplos T. O número de acertos, para cada classe, se localiza na diagonal principal M(Ci,Ci) da matriz, e os demais elementos M(Ci,Cj), para i ≠ j, representam erros na classificação. A matriz de confusão de um classificador ideal possui todos esses elementos iguais a zero uma vez que ele não comete erros.

2.3.1 Arquivo .ARFF

> *Attribute-Relation File Format* (ARFF) é um arquivo de texto ASCII que descreve uma lista de instâncias que compartilham um conjunto de atributos, desenvolvidos pelo Projeto de Aprendizado de Máquinas no Departamento de Ciência da Computação da Universidade de Waikato para uso com o software de aprendizado de máquinas Weka (WAIKATO, 2008).

Composto por um cabeçalho, o qual é formado pelo nome da relação dos dados, pelos atributos ou colunas e pelos seus tipos, conforme representado na Figura 3.

Figura 3 . Cabeçalho do arquivo ARFF

```
%temperatura,umidade,precipitacao,statusdoenca%
%Base1%
@relation BrusoneTrigo
@ATTRIBUTE temperatura REAL
@ATTRIBUTE umidade REAL
@ATTRIBUTE precipitacao REAL
@ATTRIBUTE statusdoenca{ BAIXO, MODERADO, SEVERO, MUITOSEVERO, NAODETECTADO }
```

Fonte: Do Autor

Após o cabeçalho constitui-se a segunda parte composta pelos dados, estes dispostos em linhas, separados por vírgula conforme a Figura 4.

Figura 4 . Disposição dos dados no arquivo ARFF

```
@data
%312855 registros%
18.600000000000001,81.5,0,NAODETECTADO
21.199999999999999,86.5,0,NAODETECTADO
23,77.900000000000006,0,NAODETECTADO
20.5,86,0,NAODETECTADO
28,65.799999999999997,0,NAODETECTADO
22.600000000000001,85.200000000000003,0,NAODETECTADO
15.4,90,0,NAODETECTADO
```

Fonte: Do Autor

Comentários podem ser adicionados entre os símbolos de "%", @RELATION representa o nome da relação de dado. Cada um dos atributos deve iniciar pelo marcador @ATTRIBUTE seguido pelo nome do atributo e o tipo, no caso de classes de classificação o tipo é substituído por possíveis classificadores, que devem ser colocados entre "{". A ordem em que os atributos são declarados deve ser respeitada quando são relacionados os dados, o WEKA sempre espera que todos os dados estejam presentes nas relações (WAIKATO, 2008).

2.3.2 Algoritmos utilizados

Levando-se em consideração o formato das bases para a mineração e pelo fato destas serem compostas de dados meteorológicos, optou-se por utilizar-se de quatro algoritmos disponibilizados pela ferramenta WEKA, sendo estes mostrados na Tabela 1.

Tabela 1 - Algoritmos utilizados.

Algoritmo	Descrição
JRip	Um algoritmo que implementa o principio de regras proposicionais, poda incremental repetida para produzir a redução de erros.
OneR	Um algoritmo de indução de conhecimento, onde o conhecimento inferido é representado na forma de uma árvore de decisão de um único nível, demonstrando através de um conjunto de regras para cada valor de um determinado atributo, usa o atributo de erro mínimo para predição, discretizando atributos numéricos.
J48	Um algoritmo que implementa a classe para gerar uma árvore de decisão podendo está ser formada com podas ou não podas, ou seja, classifica valores para atributos numéricos uma vez, os valores faltantes são tratados dividindo as instâncias correspondentes em peças.
REPTree	Um algoritmo aprendiz de árvore de decisão rápida, desenvolve uma árvore de decisão/regressão usando a variância da informação e usando as podas de erro reduzido. Classifica somente os valores para atributos numéricos uma vez, os valores faltantes são tratados dividindo as instâncias correspondentes em peças (semelhante ao J48).

Fonte: (http://weka.sourceforge.net/doc.stable/allclasses-noframe.html)

2.4 BRUSONE

Sendo o trigo considerado o segundo cereal mais produzido no mundo, este é cultivado principalmente nas regiões Sul, Sudeste e Centro-Oeste do Brasil, uma vez que as condições climáticas prevalecentes são pouco favoráveis à sua cultura.

Segundo Fernandes e Picinini, "as condições climáticas, onde predominam temperaturas altas e precipitações pluviais frequentes, favorecem o desenvolvimento de inúmeras doenças, principalmente aquelas causadas por fungos. Essas podem ser responsáveis por perdas elevadas no rendimento e na qualidade dos grãos de trigo" (2013). Essas características são comuns ao clima brasileiro, o que esclarece a vasta quantidade de doenças que vêm afetando os cultivares do Brasil.

O fungo *Magnaporthe grisea* (*Anamorfo Pyricularia grisea*) é o agente causal da Brusone do trigo (*Triticum aestivum*), uma doença limitante à cultura do trigo no Brasil em regiões produtoras localizadas acima do paralelo 24ºS. (ALVES & FERNANDES, 2006). Segundo TAKAMY (2011), os estados do Paraná, Mato Grosso do Sul, Rio Grande do Sul, Goiás, Minas Gerais e São Paulo registraram grandes perdas econômicas por conta da doença.

Na safra 2004 a brusone voltou a causar graves prejuízos, principalmente nas lavouras de trigo em áreas do Cerrado e no Mato Grosso do Sul, assim como nas regiões Norte e Oeste do estado do Paraná, provocando um

alerta geral aos produtores de trigo e cevada nessas regiões (TAKAMI apud CRUZ, 2008).

3 BASES DE DADOS UTILIZADAS

Partindo-se de dados meteorológicos reais fornecidos pelo IAPAR[5], obteve-se inicialmente dados relativos a três locais: Apucarana/PR, Londrina/PR e Maringá/PR, sendo estes compostos pelos valores de temperatura, umidade relativa e precipitação. Datados desde 1º de novembro de 1999 a 14 de junho de 2015, separados por hora, no total de 24hs/dia.

Realizou-se algumas modificações as bases de dados disponíveis, com o intuito de aumentar a precisão nos resultados a que se pretendia obter. Tais modificações se sucederam por meio das etapas propostas pelo KDD. Originou-se assim 3 bases de dados de características distintas, cada base possibilitou a geração de um arquivo de extensão .arff.

A Figura 5 apresenta o cabeçalho de especificações e as linhas iniciais dos dados, do arquivo Base1.arff.

Figura 5. Base1.arff

```
%temperatura,umidade,precipitacao,statusdoenca%
%Base1%
@relation BrusoneTrigo
@ATTRIBUTE temperatura REAL
@ATTRIBUTE umidade REAL
@ATTRIBUTE precipitacao REAL
@ATTRIBUTE statusdoenca{ BAIXO, MODERADO, SEVERO, MUITOSEVERO, NAODETECTADO }
@data
%312855 registros%
18.600000000000001,81.5,0,NAODETECTADO
21.199999999999999,86.5,0,NAODETECTADO
23,77.900000000000006,0,NAODETECTADO
20.5,86,0,NAODETECTADO
28,65.799999999999997,0,NAODETECTADO
22.600000000000001,85.200000000000003,0,NAODETECTADO
15.4,90,0,NAODETECTADO
```

Fonte:Do Autor

Base1: Composta pelos valores reais obtidos (temperatura, umidade relativa, precipitação) dos três locais, de todos os dias e horários aos quais se tinha acesso.

Contendo cinco possíveis níveis para a ocorrência da doença (baixo, moderado, severo, muito severo e não detectado), níveis estes inseridos conforme informações disponibilizadas pelas bibliografias que compõem este projeto.

[5] Instituto Agronômico do Paraná.

Porém optou-se pela remoção dos dados correspondentes aos anos em que não foi possível especificar a ocorrência ou falta de ocorrência da doença. Além disso, aos anos em que se soube não informarem a mesma optou-se por considerar como status de não detectado.

A Figura 6 apresenta as especificações do arquivo Base2.arff.

Figura 6. Arquivo Base2.arff

```
%tmax,tmin,urmax,urmin,precip,statusdoenca%
%Base2%
@relation BrusoneTrigo
@ATTRIBUTE tmax REAL
@ATTRIBUTE tmin REAL
@ATTRIBUTE urmax REAL
@ATTRIBUTE urmin REAL
@ATTRIBUTE precip REAL
@ATTRIBUTE statusdoenca{ SIM, NAO }
@data
%13112 registros %
24.5,17.7,99.9,65.1,0,SIM
26.6,15.2,93.8,57,0,SIM
25.9,14.2,91,49.1,0,SIM
27.9,15.1,83.4,55.8,0,SIM
29.3,16.9,90.9,56,29.6,SIM
```

Fonte: Do Autor

Base2: Composta por valores obtidos através da média/dia das variáveis: temperatura e umidade relativa, as quais transformaram-se em: *tmax* (temperatura máxima), *tmin* (temperatura mínima), *urmax* (umidade relativa máxima) e *urmin* (umidade relativa mínima) e ainda *precip* (precipitação do dia).

Optou-se pela remoção dos dados correspondentes aos anos em que não foi possível especificar a ocorrência ou falta de ocorrência da doença. Ainda, foi preferível classificar somente com valores de sim ou não o status em relação aos anos de ocorrência da doença.

A Figura 7 apresenta as especificações do arquivo Base3.arff.

Figura 7. Base3.arff

```
%tmax,tmin,urmax,urmin,precip,statusdoenca%
%Base3%
@relation BrusoneTrigo
@ATTRIBUTE tmax REAL
@ATTRIBUTE tmin REAL
@ATTRIBUTE urmax REAL
@ATTRIBUTE urmin REAL
@ATTRIBUTE precip REAL
@ATTRIBUTE statusdoenca{ SIM, NAO }
@data
%1821 registros%
25.8,18.2,68.7,45.7,0,SIM
23.6,17.7,76,49.1,0,SIM
25.3,17.2,87.7,47,0,SIM
26.4,17.9,68,42.7,0,SIM
```

Fonte: Do Autor

Base3: Igualmente composta por valores obtidos através da media/dia, sem a presença dos dados aos quais era desconhecido a existência ou inexistência da doença. Tendo esta também classificação da doença por somente sim ou não, em relação a sua ocorrência.

Porém optou-se pela utilização dos dados pertencentes a datas que referenciavam épocas de manejo do trigo, épocas está obtidas através das bibliografias usadas na composição deste projeto.

4 RESULTADOS

Utilizando-se das fases que compõem o processo de descoberta do conhecimento, após as modificações das bases de dados a serem utilizadas, foi necessário a escolha da tarefa de mineração de dados. Optou-se por utilizar-se da tarefa de classificação, pois está mostrou-se mais adequada aos tipos de dados utilizados. A sessões seguintes abordam os resultados obtidos.

4.1 RESULTADOS OBTIDOS

Os algoritmos JRip, OneR, J48 e REPTree foram igualmente executados no framework WEKA utilizando-se do método de teste cross-validation. Este método tem seu resultado baseado em um conjunto de dez execuções do algoritmo selecionado, onde cada uma destas execuções utiliza-se de 90% dos dados para o grupo de treinamento, dados estes escolhidos de forma aleatória e os outros 10% para a geração do modelo de teste, com a exclusão da class label, visa-se garantir que o algoritmo não venha a utilizar desta relação para obtenção de seus resultados.

A acurácia do modelo se dá através da porcentagem dos testes que foram corretamente classificados pelo algoritmo. A seguir, um parecer geral quanto ao resultado obtido na respectiva base, seguido dos resultados específicos de cada algoritmo.

4.1.1 Base 1

A Figura 8 exibe o resultado obtido com a execução do algoritmo JRip.

Figura 8. Resposta do algoritmo JRip

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      155418               49.6773 %
Incorrectly Classified Instances    157437               50.3227 %
Kappa statistic                          0.0041
Mean absolute error                      0.273
Root mean squared error                  0.3695
Relative absolute error                 99.8221 %
Root relative squared error             99.9187 %
Total Number of Instances           312855

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,000    0,000    0,000      0,000    0,000      0,000    0,501     0,169     BAIXO
               0,005    0,000    0,563      0,005    0,011      0,049    0,504     0,090     MODERADO
               0,000    0,000    0,000      0,000    0,000      0,000    0,501     0,084     SEVERO
               0,006    0,001    0,718      0,006    0,013      0,057    0,503     0,174     MUITOSEVERO
               0,999    0,996    0,496      0,999    0,663      0,036    0,502     0,496     NAODETECTADO
Weighted Avg.  0,497    0,494    0,414      0,497    0,332      0,032    0,502     0,318

=== Confusion Matrix ===

     a     b     c     d      e   <-- classified as
     0     0     0     0  52632 |      a = BAIXO
     0   142     0   103  26026 |      b = MODERADO
     0    12     0     0  26340 |      c = SEVERO
     0    39     0   341  52195 |      d = MUITOSEVERO
     0    59     0    31 154935 |      e = NAODETECTADO
```

Fonte: Do Autor

Obteve-se os valores de 49% para classificação correta e 50% para classificação incorreta dos dados. Partindo-se deste valores e também da matriz de confusão, consideram-se precários os valores obtidos para uma classificação acertada dos padrões possíveis ao tipo de base de dados utilizada, uma vez a margem para acerto e erro ficou em uma porcentagem bem próxima.

A Figura 9 exibe o resultado obtido com a execução do algoritmo OneR.

Figura 9. Resposta do algoritmo OneR

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      155237               49.6195 %
Incorrectly Classified Instances    157618               50.3805 %
Kappa statistic                          0.003
Mean absolute error                      0.2015
Root mean squared error                  0.4489
Relative absolute error                 73.6798 %
Root relative squared error            121.392  %
Total Number of Instances           312855

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,000    0,000    0,000      0,000    0,000      -0,002   0,500     0,168     BAIXO
               0,003    0,000    0,405      0,003    0,005      0,027    0,501     0,085     MODERADO
               0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,084     SEVERO
               0,006    0,001    0,542      0,006    0,012      0,042    0,502     0,170     MUITOSEVERO
               0,999    0,996    0,496      0,999    0,663      0,027    0,501     0,496     NAODETECTADO
Weighted Avg.  0,496    0,494    0,371      0,496    0,331      0,022    0,501     0,317

=== Confusion Matrix ===

     a     b     c     d      e   <-- classified as
     0     0     0     5  52627 |      a = BAIXO
     0    70     0   118  26083 |      b = MODERADO
     0     4     0    20  26328 |      c = SEVERO
     0    55     0   306  52214 |      d = MUITOSEVERO
     4    44     0   116 154861 |      e = NAODETECTADO
```

Fonte: Do Autor

Obteve-se os valores de 49% para classificação correta e 50% para classificação incorreta dos dados, classificação está bem próxima a do algoritmo

JRip. Valores estes também considerados precários para uma classificação acertada dos padrões possíveis ao tipo de dados utilizados.

A Figura 10 exibe o resultado obtido com a execução do algoritmo J48.

Figura 10. Resposta do algoritmo J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      155449               49.6872 %
Incorrectly Classified Instances    157406               50.3128 %
Kappa statistic                          0.0273
Mean absolute error                      0.2691
Root mean squared error                  0.369
Relative absolute error                 98.3695 %
Root relative squared error             99.785  %
Total Number of Instances           312855

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,016    0,007    0,313      0,016    0,031      0,036    0,577     0,200     BAIXO
               0,018    0,003    0,358      0,018    0,034      0,064    0,556     0,117     MODERADO
               0,007    0,002    0,232      0,007    0,014      0,028    0,553     0,107     SEVERO
               0,049    0,016    0,380      0,049    0,087      0,084    0,552     0,215     MUITOSEVERO
               0,976    0,949    0,503      0,976    0,664      0,071    0,533     0,524     NAODETECTADO
Weighted Avg.  0,497    0,475    0,415      0,497    0,353      0,063    0,547     0,348

=== Confusion Matrix ===

     a     b     c     d      e   <-- classified as
   846    77   131  1010  50568 |      a = BAIXO
   131   467    50   477  25146 |      b = MODERADO
   265    64   196   720  25107 |      c = SEVERO
   506   280   182  2573  49034 |      d = MUITOSEVERO
   957   417   285  1999 151367 |      e = NAODETECTADO
```

Fonte: Do Autor

Obteve-se os valores de 49% para classificação correta e 50% para classificação incorreta dos dados, classificação está bem próxima a dos algoritmos JRip e OneR. Valores estes também considerados precários para uma classificação acertada dos padrões possíveis ao tipo de dados utilizado.

A Figura 11 exibe o resultado obtido com a execução do algoritmo REPTree.

Figura 11. Resposta do algoritmo REPTree

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      150176               48.0018 %
Incorrectly Classified Instances    162679               51.9982 %
Kappa statistic                          0.0345
Mean absolute error                      0.2665
Root mean squared error                  0.3727
Relative absolute error                 97.4392 %
Root relative squared error            100.7726 %
Total Number of Instances           312855

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,059    0,039    0,234      0,059    0,095      0,037    0,603     0,214     BAIXO
               0,026    0,007    0,239      0,026    0,046      0,053    0,576     0,119     MODERADO
               0,015    0,006    0,177      0,015    0,027      0,028    0,571     0,110     SEVERO
               0,074    0,035    0,300      0,074    0,119      0,073    0,566     0,215     MUITOSEVERO
               0,917    0,884    0,505      0,917    0,651      0,055    0,554     0,543     NAODETECTADO
Weighted Avg.  0,480    0,451    0,375      0,480    0,365      0,053    0,568     0,360

=== Confusion Matrix ===

     a     b     c     d      e   <-- classified as
  3123   311   321  1937  46940 |      a = BAIXO
   992   671   153   951  23504 |      b = MODERADO
  1084   180   387  1367  23334 |      c = SEVERO
  2062   548   410  3885  45670 |      d = MUITOSEVERO
  6070  1101   917  4827 142110 |      e = NAODETECTADO
```

Fonte: Do Autor

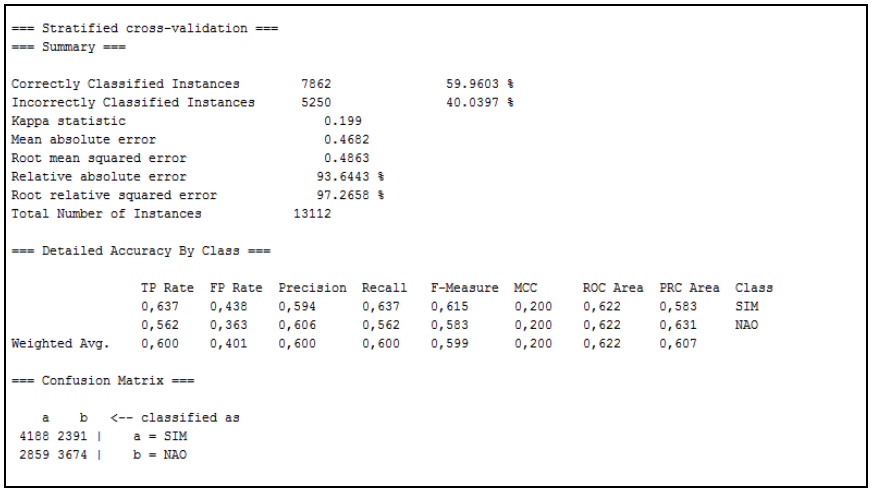
Obteve-se os valores de 48% para classificação correta e 51% para classificação incorreta dos dados, classificação está próxima a dos algoritmos

anteriores porém de acurácia ainda menor. Valores estes também precários para uma classificação acertada dos padrões possíveis ao tipo de dados utilizado.

4.1.2 Base 2

A Figura 12 exibe o resultado obtido com a execução do algoritmo JRip.

Figura 12. Resposta do algoritmo JRip



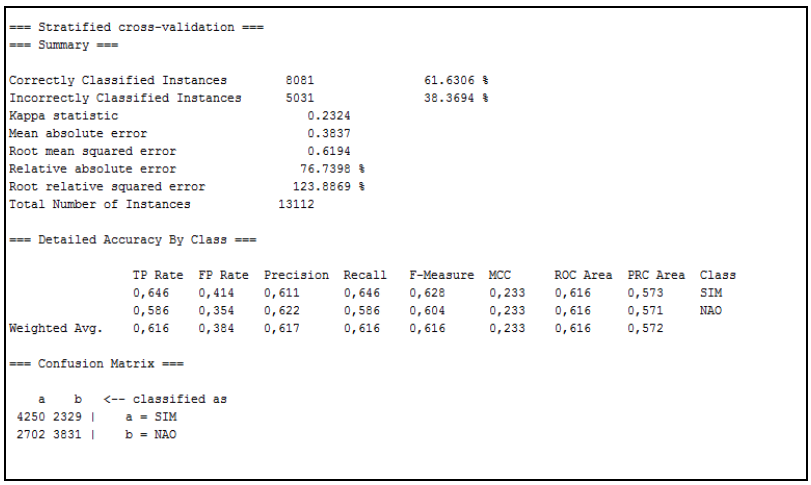
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7862               59.9603 %
Incorrectly Classified Instances      5250               40.0397 %
Kappa statistic                          0.199
Mean absolute error                      0.4682
Root mean squared error                  0.4863
Relative absolute error                 93.6443 %
Root relative squared error             97.2658 %
Total Number of Instances            13112

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,637    0,438    0,594      0,637    0,615      0,200    0,622     0,583     SIM
               0,562    0,363    0,606      0,562    0,583      0,200    0,622     0,631     NAO
Weighted Avg.  0,600    0,401    0,600      0,600    0,599      0,200    0,622     0,607

=== Confusion Matrix ===

    a    b   <-- classified as
 4188 2391 |    a = SIM
 2859 3674 |    b = NAO
```

Fonte: Do Autor

Obteve-se os valores de 59% para classificação correta e 40% para classificação incorreta dos dados, podendo-se assim considerar uma acurácia mediana, uma vez que seu valor de acertos foi maior que seu valor de erros.

A Figura 13 exibe o resultado obtido com a execução do algoritmo OneR.

Figura 13. resultado algoritmo OneR



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8081               61.6306 %
Incorrectly Classified Instances      5031               38.3694 %
Kappa statistic                          0.2324
Mean absolute error                      0.3837
Root mean squared error                  0.6194
Relative absolute error                 76.7398 %
Root relative squared error            123.8869 %
Total Number of Instances            13112

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,646    0,414    0,611      0,646    0,628      0,233    0,616     0,573     SIM
               0,586    0,354    0,622      0,586    0,604      0,233    0,616     0,571     NAO
Weighted Avg.  0,616    0,384    0,617      0,616    0,616      0,233    0,616     0,572

=== Confusion Matrix ===

    a    b   <-- classified as
 4250 2329 |    a = SIM
 2702 3831 |    b = NAO
```

Fonte: Do Autor

Obteve-se os valores de 61% para classificação correta e 38% para classificação incorreta dos dados, classificando-se também de acurácia mediana.

Valores estes ainda não tão bons para uma precisa classificação dos padrões possíveis ao tipo de base de dados utilizada, uma vez que, quanto maior os valores classificados corretamente menor o índice de classificações errôneas.

A Figura 14 exibe o resultado obtido com a execução do algoritmo J48.

Figura 14. Resultado do algoritmo J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8584               65.4667 %
Incorrectly Classified Instances      4528               34.5333 %
Kappa statistic                          0.3091
Mean absolute error                      0.3778
Root mean squared error                  0.4563
Relative absolute error                 75.5532 %
Root relative squared error             91.266  %
Total Number of Instances            13112

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,708    0,400    0,641      0,708    0,673      0,311    0,737     0,735     SIM
               0,600    0,292    0,672      0,600    0,634      0,311    0,737     0,741     NAO
Weighted Avg.  0,655    0,346    0,656      0,655    0,654      0,311    0,737     0,738

=== Confusion Matrix ===

    a    b   <-- classified as
 4661 1918 |    a = SIM
 2610 3923 |    b = NAO
```

Fonte: Do autor

Obteve-se os valores de 65% para classificação correta e 34% para classificação incorreta dos dados, classificando-se assim bem próximo de uma boa acurácia. Valores estes pouco mais precisos para obtenção de padrões possíveis ao tipo de dados utilizado.

A Figura 15 exibe o resultado obtido com a execução do algoritmo REPTree.

Figura 15. Resultado do algoritmo REPTree

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8775               66.9234 %
Incorrectly Classified Instances      4337               33.0766 %
Kappa statistic                          0.3384
Mean absolute error                      0.3694
Root mean squared error                  0.479
Relative absolute error                 73.8726 %
Root relative squared error             95.8071 %
Total Number of Instances            13112

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,693    0,355    0,663      0,693    0,678      0,339    0,729     0,715     SIM
               0,645    0,307    0,676      0,645    0,660      0,339    0,729     0,716     NAO
Weighted Avg.  0,669    0,331    0,670      0,669    0,669      0,339    0,729     0,715

=== Confusion Matrix ===

    a    b   <-- classified as
 4561 2018 |    a = SIM
 2319 4214 |    b = NAO
```

Fonte: Do autor

Obteve-se os valores de 66% para classificação correta e 33% para classificação incorreta dos dados, classificação está bem próximo de uma boa

acurácia. Valores ainda baixos para obtenção de precisos padrões possíveis ao tipo de dados utilizado.

4.1.3 Base 3

A Figura 16 exibe o resultado obtido com a execução do algoritmo JRip.

Figura 16. Resultado do algoritmo JRip

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1125               61.7792 %
Incorrectly Classified Instances       696               38.2208 %
Kappa statistic                          0.2355
Mean absolute error                      0.4624
Root mean squared error                  0.4862
Relative absolute error                 92.4889 %
Root relative squared error             97.2306 %
Total Number of Instances             1821

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,639    0,404    0,614      0,639    0,626      0,236    0,631     0,595     SIM
               0,596    0,361    0,622      0,596    0,609      0,236    0,631     0,616     NAO
Weighted Avg.  0,618    0,382    0,618      0,618    0,618      0,236    0,631     0,605

=== Confusion Matrix ===

   a   b   <-- classified as
 583 329 |   a = SIM
 367 542 |   b = NAO
```

Fonte: Do autor

Obteve-se os valores de 61% para classificação correta e 38% para classificação incorreta dos dados, valores estes considerados baixos para uma acurácia mediana, tendo em vista a quantidade bem menor em relação as outras bases de dados utilizadas. Valores estes não tão bons para uma precisa classificação dos padrões possíveis devido ao tipo de dados utilizado.

A Figura 17 exibe o resultado obtido com a execução do algoritmo OneR.

Figura 17. Resultado do algoritmo JRip

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1113               61.1203 %
Incorrectly Classified Instances       708               38.8797 %
Kappa statistic                          0.2224
Mean absolute error                      0.3888
Root mean squared error                  0.6235
Relative absolute error                 77.7596 %
Root relative squared error            124.7072 %
Total Number of Instances             1821

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,614    0,392    0,611      0,614    0,613      0,222    0,611     0,569     SIM
               0,608    0,386    0,611      0,608    0,610      0,222    0,611     0,567     NAO
Weighted Avg.  0,611    0,389    0,611      0,611    0,611      0,222    0,611     0,568

=== Confusion Matrix ===

   a   b   <-- classified as
 560 352 |   a = SIM
 356 553 |   b = NAO
```

Fonte: Do autor

Obteve-se os valores de 61% para classificação correta e 38% para classificação incorreta dos dados, semelhantes aos obtidos pelo algoritmo JRip, considerado baixo para uma acurácia de classificação mediana, uma vez que a quantidade de dados utilizados foi bem menor em relação as dados utilizados nas outras bases.

A Figura 18 exibe o resultado obtido com a execução do algoritmo J48.

Figura 18. Resultado do algoritmo J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1089               59.8023 %
Incorrectly Classified Instances       732               40.1977 %
Kappa statistic                          0.1961
Mean absolute error                      0.4587
Root mean squared error                  0.4902
Relative absolute error                 91.7349 %
Root relative squared error             98.035  %
Total Number of Instances             1821

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,590    0,394    0,600      0,590    0,595      0,196    0,632     0,620     SIM
                 0,606    0,410    0,596      0,606    0,601      0,196    0,632     0,611     NAO
Weighted Avg.    0,598    0,402    0,598      0,598    0,598      0,196    0,632     0,616

=== Confusion Matrix ===

   a   b   <-- classified as
 538 374 |   a = SIM
 358 551 |   b = NAO
```

Fonte: Do Autor

Obteve-se os valores de 59% para classificação correta e 40% para classificação incorreta dos dados, considerado baixo para uma boa acurácia. Valores estes também não tão bons para uma precisa classificação dos padrões possíveis devido ao tipo de dados utilizado.

A Figura 19 exibe o resultado obtido com a execução do algoritmo REPTree.

Figura 19. Resultado algoritmo REPTree

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1128               61.944  %
Incorrectly Classified Instances       693               38.056  %
Kappa statistic                          0.2389
Mean absolute error                      0.4355
Root mean squared error                  0.4992
Relative absolute error                 87.0979 %
Root relative squared error             99.8458 %
Total Number of Instances             1821

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,610    0,371    0,623      0,610    0,616      0,239    0,646     0,620     SIM
                 0,629    0,390    0,616      0,629    0,623      0,239    0,646     0,630     NAO
Weighted Avg.    0,619    0,381    0,620      0,619    0,619      0,239    0,646     0,625

=== Confusion Matrix ===

   a   b   <-- classified as
 556 356 |   a = SIM
 337 572 |   b = NAO
```

Fonte: Do Autor

Obteve-se os valores de 61% para classificação correta e 38% para classificação incorreta dos dados, semelhante ao obtido pelo algoritmo anterior OneR, considerado baixo para uma acurácia de classificação mediana. Valores estes também não tão bons para uma precisa classificação dos padrões possíveis devido ao tipo de base de dados utilizada.

Tendo em vista a necessidade de se comparar os algoritmos selecionados para melhor se chegar a um consenso quanto ao resultado final total, utilizou-se do método descrito por Han e Kamber (2006), que se utiliza de características bem definidas para a comparação entre algoritmos de classificação, destas serão utilizadas:

acurácia prevista: definida como a habilidade do algoritmo de classificar assertivamente novos dados quando aplicados ao modelo gerado;

robustez: a assertividade do modelo quanto aos dados incorretos, inexistentes ou com ruído;

escalabilidade: o desempenho do algoritmo quando aplicados a grandes conjuntos de dados.

4.1.4 Comparação de Algoritmos

Por meio dos critérios mencionados na sessão anterior, a Tabela 2 apresenta os principais valores utilizados para comparação dos resultados obtidos.

Tabela 2 - Resultado dos algoritmos utilizados.

Base	Algoritmo	Acurácia	Índice de aprendizagem	Matriz Confusão (Acerto/ Erro)
Base 1	JRip	49,6773 %	0,0041	DP=155.418 / DS=157.437
	OneR	49,6195 %	0,003	DP=155.237 / DS=157.618
	J48	49,6872%	0,0273	DP=155.449 / DS=157.406
	REPTree	48,0018%	0,0345	DP=150.176 / DS=162.679
Base 2	JRip	59,9603%	0,199	DP=7.862 / DS=5.250
	OneR	61,6306%	0,2324	DP=8.081 / DS=5.031
	J48	65,4667%	0,3091	DP=8.584 / DS=4.528
	REPTree	66,9234%	0,3384	DP=8.775 / DS=4.337
Base 3	JRip	61,7792%	0,2355	DP=1.125 / DS=696
	OneR	61,1203%	0,2224	DP=1.113 / DS=708
	J48	59,8023%	0,1961	DP=1.089 / DS=732
	REPTree	61,944%	0,2389	DP=1.128 / DS=693

Fonte: Do autor.

Utilizando-se da Tabela 2 é possível localizar o pior resultado em relação a Base 1 em: acurácia, robustez e escalabilidade, ao resultado fornecido pelo algoritmo REPTree tendo os menores índices de assertividade.

Em relação a Base 2 atribui-se as melhores acurácias e escalabilidade, sendo o algoritmo REPTree também responsável pelo melhor resultado quanto ao índice esperado .

A Base 3 no entanto mostra-se com padrões mais semelhantes entre os resultados obtidos com os algoritmos testados, tendo sua acurácia bem próxima em três algoritmos, porém o algoritmo J48 apresenta uma pequena diferença de assertividade quanto aos outros algoritmos.

Devido as comparações realizadas anteriormente, considera-se com isto os resultados obtidos por meio do algoritmo REPTree independente da base utilizada como os mais eficientes em relação aos outros algoritmos, pelo fato de que mesmo obtendo o pior resultado em relação a Base 1 seu índice de aprendizagem foi mais elevado se comparado aos outros algoritmos.

O anexo A deste projeto, traz o resultado retornado pela ferramenta WEKA após a execução do algoritmo REPTree na Base 3, optou-se pela inserção deste anexo a fim de se demonstrar o resultado completo gerado pelo algoritmo na ferramenta utilizada, optando-se pela Base 3 por ser a menor árvore gerada.

5 CONSIDERAÇÕES FINAIS

Acredita-se que uma das dificuldades em relação a previsão da doença pelos algoritmos caracterizou-se pela falta de dados concretos quanto a sua ocorrência ou não, considerando-se que os algoritmos submetidos a mineração utilizam-se de dados anteriores para criação do modelo de treinamento ao qual visa inferir a previsão. Uma vez que se soube apenas os anos em que ocorreu, e não precisamente sua duração, seja por épocas ou faixas, pode-se supor que estás relações vieram a interferir diretamente no resultado final.

Outro fator significativo aos resultados obtidos está diretamente ligado ao fato de se tratar de uma doença considerada de ocorrência esporádica. Tornando a precisão dos algoritmos um tanto irrelevante, uma vez que, a inferência de um

padrão baseado em dados anteriores deveria ser dividido na menor faixa possível, buscando-se o mais preciso padrão.

Tendo em vista uma das características da doença Brusone no trigo, como a sua ligação com fatores climáticos favoráveis em diferentes estágios de desenvolvimento da doença, pode-se pensar como trabalho futuro, a utilização da mineração de motifs que utiliza-se de conceitos de similaridade entre subseqüências de series temporais, dividindo-se entre a descoberta de padrões e pares de vizinhos mais próximos. Podendo assim proporcionar a localização de um padrão valido a previsão da doença.

ABSTRACT

The considerable increase in the databases of the most different areas collaborated to the need of the use of techniques that help the manipulation of these, since the traditional search of data was not enough for the use of them. The data mining process seeks to extract information that is implicit, and / or previously unknown and makes it meaningful for decision making. To assist in this management, computational tools discovering new knowledge are essential. Data Mining can be considered as a series of steps to obtain knowledge of the hidden and useful patterns in this data through the application of concepts, methods, tools and techniques. Wheat is considered an impact disease and its occurrence is associated with favorable weather conditions. In this sense, this work applies algorithms of data mining in a set of databases with the objective of identifying patterns of occurrence of the disease.

Keywords: Discovery of knowledge. Classification. Pyricularia grisea

REFERÊNCIAS

ALVES, Kalíbia Jane P.; FERNANDES, José Maurício C.; Influência da temperatura e da umidade relativa do ar na esporulação de Magnaporthe grisea em trigo. Passo Fundo, dez.2006. Disponível em: <http://www.scielo.br/scielo.php?pid=S0100-41582006000600007&script=sci_abstract&tlng=pt>. Acesso em: 30 mai. 2017.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. Goiás, ago. 2009. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 10 out. 2016.

FAYYAD, U.; PIATETSKY-SHAPIRO,G.; SMYTH, P. Data Mining to Knowledge Discovery: an overview. Em: FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Ed.). Advances in knowledge Discovery and data mining. Menlo Park: AAAI Press, 1996.

GOULART, SOUSA, URASHIMA. Danos em trigo causados pela infecção de Pyricularia grisea. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-54052007000400007> Acesso em: 10 mai. 2016.

HAN, Jiawei; KAMBER, Micheline. Data Mining: Concepts and Techniques. 2ª ed. San Francisco: Morgan Kaufmann, 2006.

NAVEGA, Sergio. Princípios Essenciais do Data Mining. São Paulo, ago. 2002. Disponível em: <http://www.intelliwise.com/reports/i2002.pdf>. Acesso em: 08 out. 2016.

SILVA, Marcelino Pereira dos Santos. Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka. São José dos Campos. 2004. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf> Acesso em: 08 out. 2016.

TAKAMI, LUCAS KENJI. Resistência de Genótipos de Trigo à Brusone (Pyricularia grisea). Disponível em: < http://locus.ufv.br/bitstream/handle/123456789/4546/texto%20completo.pdf?sequence=1&isAllowed=y> Acesso em 08 out. 2016.

WAIKATO. Attribute-Relation File Format. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/arff.html>. Waikato, 2008. Acesso em 08 out. 2016.

ANEXO

ANEXO A – Resultado exibido pela ferramenta WEKA na área *Classifier output*, na execução do algoritmo REPTree sobre a base de dados Base 3.

```
=== Run information ===
Scheme:       weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
Relation:     BrusoneTrigo
Instances:    1821
Attributes:   6
              tmax
              tmin
              urmax
              urmin
              precip
              statusdoenca
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
REPTree
============

urmin < 46.35
|   tmax < 28.25
|   |   tmin < 16.45
|   |   |   tmax < 18.7 : SIM (16/3) [8/1]
|   |   |   tmax >= 18.7
|   |   |   |   urmax < 87.4
|   |   |   |   |   tmax < 26.05
|   |   |   |   |   |   urmin < 39
|   |   |   |   |   |   |   tmin < 9.75 : SIM (4/1) [3/1]
|   |   |   |   |   |   |   tmin >= 9.75 : NAO (23/0) [12/2]
|   |   |   |   |   |   urmin >= 39 : SIM (68/32) [27/12]
|   |   |   |   |   tmax >= 26.05 : SIM (22/6) [15/7]
|   |   |   |   urmax >= 87.4
|   |   |   |   |   urmax < 99.85 : NAO (54/10) [15/4]
|   |   |   |   |   urmax >= 99.85 : SIM (5/1) [5/0]
|   |   tmin >= 16.45
|   |   |   urmin < 37.1 : NAO (25/0) [9/0]
```

```
|   |   |   urmin >= 37.1
|   |   |   |   precip < 1.8
|   |   |   |   |   tmax < 27.65
|   |   |   |   |   |   tmin < 17.35 : NAO (22/1) [6/2]
|   |   |   |   |   |   tmin >= 17.35
|   |   |   |   |   |   |   urmax < 60.65 : NAO (4/0) [1/0]
|   |   |   |   |   |   |   urmax >= 60.65
|   |   |   |   |   |   |   |   urmax < 65.6
|   |   |   |   |   |   |   |   |   urmax < 64.5 : NAO (14/5) [5/1]
|   |   |   |   |   |   |   |   |   urmax >= 64.5 : SIM (3/0) [1/0]
|   |   |   |   |   |   |   |   urmax >= 65.6
|   |   |   |   |   |   |   |   |   tmax < 25.65 : NAO (10/0) [1/0]
|   |   |   |   |   |   |   |   |   tmax >= 25.65
|   |   |   |   |   |   |   |   |   |   tmax < 25.9 : SIM (2/0) [1/0]
|   |   |   |   |   |   |   |   |   |   tmax >= 25.9
|   |   |   |   |   |   |   |   |   |   |   urmin < 42.2 : NAO (6/0) [3/1]
|   |   |   |   |   |   |   |   |   |   |   urmin >= 42.2
|   |   |   |   |   |   |   |   |   |   |   |   tmin < 19.45 : SIM (4/1) [3/1]
|   |   |   |   |   |   |   |   |   |   |   |   tmin >= 19.45 : NAO (4/0) [1/0]
|   |   |   |   |   tmax >= 27.65 : NAO (12/0) [5/1]
|   |   |   |   precip >= 1.8 : SIM (2/0) [0/0]
|   tmax >= 28.25
|   |   tmin < 17.05
|   |   |   urmin < 33.5 : SIM (7/0) [6/2]
|   |   |   urmin >= 33.5
|   |   |   |   tmin < 16.75 : SIM (42/19) [11/6]
|   |   |   |   tmin >= 16.75
|   |   |   |   |   tmax < 29.75
|   |   |   |   |   |   tmax < 29.1 : SIM (3/0) [1/0]
|   |   |   |   |   |   tmax >= 29.1 : NAO (2/0) [3/1]
|   |   |   |   |   tmax >= 29.75 : SIM (8/0) [5/2]
|   |   tmin >= 17.05 : NAO (140/60) [86/29]
urmin >= 46.35
|   urmax < 97.45
|   |   urmax < 75.55 : SIM (32/2) [20/5]
|   |   urmax >= 75.55
|   |   |   urmin < 54.05
|   |   |   |   urmin < 53.85
|   |   |   |   |   tmin < 8.8 : NAO (10/1) [5/2]
```

```
|  |  |  |  |  tmin >= 8.8
|  |  |  |  |  |  urmax < 89.85
|  |  |  |  |  |  |  tmin < 13.15 : SIM (16/1) [11/5]
|  |  |  |  |  |  |  tmin >= 13.15
|  |  |  |  |  |  |  |  tmax < 32.45
|  |  |  |  |  |  |  |  |  tmax < 29.75
|  |  |  |  |  |  |  |  |  |  tmax < 28
|  |  |  |  |  |  |  |  |  |  |  tmin < 13.35 : NAO (2/0) [1/1]
|  |  |  |  |  |  |  |  |  |  |  tmin >= 13.35
|  |  |  |  |  |  |  |  |  |  |  |  tmax < 22.3 : SIM (5/0) [3/1]
|  |  |  |  |  |  |  |  |  |  |  |  tmax >= 22.3
|  |  |  |  |  |  |  |  |  |  |  |  |  tmax < 26.05
|  |  |  |  |  |  |  |  |  |  |  |  |  |  tmax < 23.25 : NAO (2/0) [1/1]
|  |  |  |  |  |  |  |  |  |  |  |  |  |  tmax >= 23.25
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  tmax < 23.8 : SIM (3/0) [2/0]
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  tmax >= 23.8
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  urmin < 49.35 : NAO (12/4) [5/1]
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  urmin >= 49.35 : SIM (13/4) [8/2]
|  |  |  |  |  |  |  |  |  |  |  |  |  |  tmax >= 26.05 : SIM (14/2) [14/7]
|  |  |  |  |  |  |  |  |  |  tmax >= 28 : NAO (24/10) [9/4]
|  |  |  |  |  |  |  |  |  tmax >= 29.75 : SIM (25/4) [5/2]
|  |  |  |  |  |  |  |  tmax >= 32.45 : NAO (2/0) [1/1]
|  |  |  |  |  |  urmax >= 89.85
|  |  |  |  |  |  |  precip < 4 : NAO (44/17) [20/8]
|  |  |  |  |  |  |  precip >= 4 : SIM (6/0) [1/0]
|  |  |  |  urmin >= 53.85 : NAO (6/0) [0/0]
|  |  |  urmin >= 54.05 : SIM (142/35) [72/24]
|  urmax >= 97.45
|  |  urmax < 99.45
|  |  |  tmin < 9.8 : NAO (20/1) [3/0]
|  |  |  tmin >= 9.8
|  |  |  |  urmin < 70.9 : NAO (104/29) [51/12]
|  |  |  |  urmin >= 70.9
|  |  |  |  |  urmax < 98.95 : SIM (15/3) [9/4]
|  |  |  |  |  urmax >= 98.95
|  |  |  |  |  |  urmin < 74.55 : SIM (2/0) [7/1]
|  |  |  |  |  |  urmin >= 74.55 : NAO (14/3) [9/3]
|  |  urmax >= 99.45
|  |  |  tmin < 15.55 : SIM (85/16) [54/14]
```

```
|   |   |   tmin >= 15.55
|   |   |   |   urmax < 100.1 : SIM (102/50) [60/25]
|   |   |   |   urmax >= 100.1 : NAO (12/1) [3/0]

Size of the tree : 99
Time taken to build model: 0.01 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        1128               61.944  %
Incorrectly Classified Instances       693               38.056  %
Kappa statistic                          0.2389
Mean absolute error                      0.4355
Root mean squared error                  0.4992
Relative absolute error                 87.0979 %
Root relative squared error             99.8458 %
Total Number of Instances             1821
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area
Class
               0,610    0,371    0,623      0,610    0,616      0,239    0,646     0,620     SIM
               0,629    0,390    0,616      0,629    0,623      0,239    0,646     0,630     NAO
Weighted Avg.  0,619    0,381    0,620      0,619    0,619      0,239    0,646     0,625
=== Confusion Matrix ===

   a   b   <-- classified as
 556 356 |   a = SIM
 337 572 |   b = NAO
```