Descoberta de Conhecimento com Aprendizado de Máquina Supervisionado em Dados Abertos dos Censos da Educação Básica e Superior

Jonathan H. A. de Carvalho[1], Lisandra S. da Cruz[1], Roberta M. M. Gouveia[1]

[1]Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco (UFRPE) – Recife, PE – Brasil

{hcarvalho.jon,lisansouza,robertammg}@gmail.com

Abstract. *This article describes the activities performed and the results obtained in the process of Open Data Mining of the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira referring to the Census of Basic and Higher Education. The methodology of the work is based on the steps defined by Knowledge Discovery in Databases, with an emphasis on Data Mining, through the application of Supervised Machine Learning algorithms. The rules and standards found aim to analyze, within the state of Pernambuco, the infrastructure of the schools, the profiles of students and the profiles of Higher Education Institutions.*

Resumo. *Este artigo descreve as atividades desenvolvidas e os resultados obtidos no processo de mineração de dados abertos do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira referentes aos Censos da Educação Básica e Superior. A metodologia do trabalho está fundamentada nas etapas definidas pelo Knowledge Discovery in Databases, com ênfase na Mineração de Dados, por meio da aplicação de algoritmos do Aprendizado de Máquina Supervisionado. As regras e padrões encontrados objetivam analisar, no âmbito do estado de Pernambuco, a infraestrutura das escolas, os perfis de estudantes e os perfis das Instituições de Ensino Superior.*

1. Introdução

O processo de *Knowledge Discovery in Database* (KDD), segundo Frawley et al. (1992), consiste no desenvolvimento de métodos de aprendizagem de máquina para explorar conjuntos de dados coletados de diversos ambientes. Essa investigação objetiva extrair informações implícitas, previamente desconhecidas e potencialmente úteis dos dados armazenados [FAYYAD et al., 1996]. Sendo assim, este trabalho aplica o KDD nos dados abertos do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) nos Censos da Educação Básica e Superior dos anos de 2014 e 2015, com a finalidade de descobrir novas informações que podem ser úteis para a tomada de decisão.

Para isso, o foco foi dado na principal fase do processo, a Mineração de Dados (MD), uma vez que é caracterizada como uma área que possui diversas técnicas automáticas para classificação, *clustering* e associação de dados, que podem ser usadas para gerar conhecimento a partir de grandes volumes de dados [WITTEN et al., 2017].

O presente trabalho está fundamentado na técnica de classificação, cuja ênfase é a criação de um modelo para descrever e possibilitar um melhor entendimento do cenário educacional pernambucano. Logo, para a obtenção de resultados mais

DOI: 10.5753/cbie.wcbie.2017.674

específicos e fortemente direcionados, as análises foram divididas nos diferentes cenários: 1. Infraestrutura das escolas; 2. Perfil dos estudantes da educação básica; 3. Perfil dos universitários de cursos de TI; 4. Perfil das Instituições de Ensino Superior (IES).

A investigação desses dados permite que se observe tendências; que se detecte regiões cuja infraestrutura escolar possui melhores condições; que se constate perfis de alunos; entre outros diagnósticos. Após devidas análises, o conhecimento adquirido pode direcionar a aplicação de medidas corretivas e preventivas para minimizar, entre outros problemas, a evasão das instituições de educação básica e superior. Assunto esse que inquieta e incomoda os gestores das instituições, evidenciando a necessidade de sensibilização por parte da sociedade para prover ou auxiliar na concepção de soluções viáveis e eficazes, caracterizado como o propósito dos trabalhos de RIGO et al. (2012) e FERREIRA (2015), e um dos propósitos do presente trabalho.

O trabalho de SANTOS et al (2014) propôs, ao minerar dados educacionais para a previsão do desempenho acadêmico de alunos, auxiliar as instituições de ensino na tomada de decisões pedagógicas. Já a presente pesquisa, apesar de também possuir interesse em investigar e diagnosticar as deficiências e obstáculos existentes, ampliou o escopo da análise, por também abordar os ensinos fundamental e médio, e contemplar todo um estado, Pernambuco.

2. Metodologia

Esta seção descreve os procedimentos adotados para a realização do trabalho e o alcance dos objetivos definidos. A metodologia aplicada neste trabalho consiste em seguir as etapas definidas pelo KDD com o propósito de amenizar o problema definido por Han e Kamber (2006) de sermos ricos em dados, mas pobres em informações e conhecimentos. As etapas seguidas, desde a escolha dos dados até a interpretação dos resultados são: (I) Seleção; (II) Pré-processamento; (III) Transformação; (IV) Mineração de Dados; (V) Interpretação e validação dos resultados.

As três primeiras etapas do processo caracterizam a fase da Pré-Mineração, constituindo-se de tarefas responsáveis por definir os dados capazes de gerar conhecimentos e prepará-los para a aplicação dos algoritmos do Aprendizado de Máquina. Na fase da Mineração, são aplicados os algoritmos visando a construção de modelos computacionais para descobrir padrões ocultos nos dados. As tarefas na fase da Pós-Mineração são responsáveis por examinar os resultados obtidos, definir os conhecimentos encontrados e aplicá-los ou reportá-los para as partes competentes tomarem as medidas adequadas.

2.1. Seleção dos Atributos

Quanto maior for o número de características num conjunto de dados, mais difícil será o aprendizado por parte dos algoritmos de Mineração, e por isso, são obtidos modelos com acurácias muito baixas e assim a classe não pode ser prevista ou descrita, caracterizando o efeito conhecido como a Maldição da Dimensionalidade. Segundo Erbert (2001), este efeito deve-se ao aumento do número de parâmetros a serem estimados, especialmente na matriz de covariância. Então, para contornar essa situação, faz-se necessário selecionar os atributos mais relevantes e pertinentes para o descobrimento de padrões ocultos e potencialmente úteis nos dados.

Por isso, foram realizadas avaliações individuais nos atributos coletados a fim de identificar os principais contribuintes na determinação de perfis de alunos ou

instituições de ensino. Como exemplo, definiu-se que as características referentes ao local de nascimento do aluno não determinam, de maneira significativa, nos diferentes perfis existentes de alunos, e por isso, não gerariam conhecimentos tão proveitosos para o objetivo do trabalho.

Dessarte, as características dos alunos da educação básica escolhidas foram: idade; sexo; raça; utilização ou não de transporte público escolar, necessidade especial e a etapa de ensino. Já para as escolas, foram selecionadas as seguintes informações: região; dependência administrativa; presença de água, energia, esgoto, coleta de lixo, laboratório de ciências, laboratório de informática, quadra, biblioteca, dependências para deficientes, auditório e internet; e a quantidade de salas, *datashows*, computadores e funcionários na instituição.

Com relação aos alunos de educação superior, foram selecionadas as seguintes informações: turno, grau acadêmico do curso, modalidade de ensino, raça, sexo, idade, necessidade especial, situação do aluno no curso, participação em programa de reserva de vagas para ingresso no curso, recebimento de apoio social, participação em atividades extracurriculares, vínculo temporário em instituição internacional e o tempo de permanência na instituição. Já para as IES, foram escolhidas as seguintes informações: categoria administrativa, região, quantidade total de técnicos, quantidade de técnicos com pós-graduação e a nota no Índice Geral de Cursos (IGC).

2.2. Pré-processamento dos Dados

O pré-processamento determina a eficiência dos algoritmos de mineração, uma vez que eles dependem de uma base de dados íntegra e sem redundâncias. Nesta fase foram identificados os atributos que estavam codificados numericamente, os quais tinham os significados nominais definidos nos metadados e, visando uma interpretação mais eficiente do modelo gerado, os valores desses atributos foram substituídos pelos categóricos.

Além disso, o fato de que os dados coletados são, na verdade, microdados, possibilitou agrupar os atributos que tratam de uma mesma característica num único e novo atributo, a exemplo das necessidades especiais. Esse novo atributo possui como valores a denominação de cada um dos atributos que foram agrupados. E para finalizar as atividades desta fase, a base de dados foi enriquecida com o objetivo de abordar mais informações relevantes, como é o caso do atributo IGC, que não estava presente na coleta dos microdados e precisou ser coletado em um momento distinto.

2.3. Transformação dos Dados

A terceira etapa do processo de KDD é a última antes da aplicação dos algoritmos do aprendizado de máquina, e por isso finaliza as tarefas de ajustes e correções na base de dados, que têm como objetivo maximizar a qualidade dos dados a serem minerados. Neste estágio foram aplicados dois processos cruciais para o sucesso dos algoritmos de MD: a Estratificação dos atributos classe e a Discretização dos atributos numéricos.

A estratificação consiste na normalização dos atributos classe, isto é, na exclusão de registros até que as instâncias do treinamento estejam distribuídas de maneira uniforme entre os valores de classe. A realização dessa tarefa se faz necessária para que as regras geradas não sejam tendenciosas, ou seja, para que o algoritmo explore uma quantidade similar para cada valor do atributo classe e assim obtenha uma aprendizagem de forma igualitária.

O segundo processo aplicado foi a discretização dos atributos numéricos, que consistiu na distribuição desses dados em categorias de intervalos. Para o algoritmo não se tornar tendencioso em relação a um determinado valor do atributo, os intervalos foram definidos de maneira que cada um possuísse uma quantidade aproximada de ocorrências.

Por fim, devido ao amplo escopo dos dados do Censo da Educação Básica (CEB) e do Censo da Educação Superior (CES), fez-se necessário criar quatro cenários específicos, contendo subgrupos de atributos, com o objetivo de obter informações mais precisas.

2.4. Mineração de Dados

Para a Mineração de Dados foi utilizada a plataforma WEKA, software *open source* que contém uma coleção de algoritmos para a mineração de dados. Este trabalho contemplou o Aprendizado Supervisionado (AS) e os algoritmos: Árvore de Decisão (J48) e Classificação Bayesiana (Naive Bayes).

Segundo Patil e Sherekar (2013), o classificador J48 é baseado no algoritmo C4.5 e utiliza uma tecnologia *greedy* para induzir árvores de decisão, em que cada nó da árvore avalia a existência ou significância de cada atributo individual. As árvores de decisão são construídas do topo para a base (*top-down*), através da escolha do atributo mais apropriado para cada situação. Dentre os vários algoritmos do aprendizado de máquina supervisionado, a técnica de classificação por árvore de decisão gera regras explícitas, facilitando a interpretação humana dos resultados [WITTEN et al., 2017]. Dessa forma, visando facilitar a compreensão dos resultados gerados pelo modelo preditivo, o algoritmo J48 foi escolhido para os estudos de caso do presente trabalho. Outra razão para escolha da técnica de árvore de decisão diz respeito à obtenção de taxas de acurácia aceitáveis por meio do algoritmo J48.

Tanto Dimitoglou (2012) quanto Gonzalez (2012) descrevem o algoritmo *Naive Bayes* como um classificador probabilístico simples que calcula um conjunto de probabilidades contando a frequência e as combinações de valores em um certo conjunto de dados. O algoritmo usa o teorema de Bayes e assume que todos os atributos são independentes, dado o valor da variável de classe. Este teorema é um modelo estatístico que permite determinar a probabilidade de hipóteses ocorrerem em um determinado conjunto de registros. Seja P(H|X) a probabilidade da hipótese H estar correta dado X; P(X|H) a probabilidade de X ocorrer, dada a hipótese; P(H) a probabilidade da hipótese ocorrer; e P(X) a probabilidade de X ocorrer, o teorema de Bayes é definido por: $P(H|X) = (P(X|H)P(H))/P(X)$.

Partindo da teoria de que as decisões tomadas pelos gestores precisam ser embasadas em padrões constantes para terem impactos positivos e significantes, se faz necessário definir regras que estejam sendo válidas durante anos. Por isso, decidiu-se utilizar a opção *Supplied Test Set* para que fosse possível testar a aptidão do modelo com um conjunto de dados diferente do que foi utilizado no treinamento. Sendo assim, o modelo foi gerado a partir dos dados do ano de 2014 e testado com os do ano de 2015, possibilitando uma validação da continuidade das regras encontradas no trabalho, visto que uma taxa de acurácia elevada do algoritmo indicaria que os padrões encontrados nos dados de 2014 se mantiveram para o ano seguinte, e caso permaneçam para outros anos, podem ser utilizados para apoiar os gestores das instituições de ensino na tomada de decisão.

2.5. Interpretação dos Resultados

Esta é a última fase do processo KDD, também chamada de pós-processamento, a qual consiste em interpretar os padrões obtidos na MD e verificar a utilidade desses novos conhecimentos de acordo com os objetivos previamente traçados. Mesmo está sendo a última fase do processo KDD, vale destacar que caso não sejam obtidos resultados satisfatórios, há a possibilidade de se retornar a qualquer fase anterior do processo para o ajuste dos dados e/ou seleção de outros algoritmos de *Machine Learning*, na tentativa de obter resultados relevantes.

3. Resultados e Discussões

Nesta seção são expostos os resultados obtidos a partir da aplicação da metodologia apresentada na Seção 2, o qual seguiu o processo KDD com foco nos algoritmos J48 e *Naive Bayes*.

3.1. Cenário 1: Perfis das Escolas quanto à Infraestrutura

Este cenário foi definido na tentativa de captar e diferenciar as características referentes à infraestrutura de escolas privadas e públicas no estado de Pernambuco, e para as públicas, ainda existe a subdivisão indicando se a escola é administrada por órgãos de domínio municipal, estadual ou federal.

Para alcançar esse objetivo, aplicou-se o algoritmo de árvore de decisão J48, gerando um modelo com uma taxa de acurácia de 62%. Porém, as regras aprendidas pelo modelo para a classe Federal não se ajustaram ao conjunto de teste (ano 2015), fazendo com que a precisão do modelo para essa classe fosse de 48%, ou seja, as regras foram válidas em menos da metade das ocasiões em que o algoritmo classificou como sendo da classe Federal.

Num primeiro momento, essa situação pareceu indicar que as características comuns da infraestrutura das escolas federais no ano de 2014 sofreram alterações no tempo, resultando em escolas com configurações diferentes no ano posterior. Para validar essa teoria e anular a possibilidade do algoritmo não ter aprendido para as escolas federais, foi feito um experimento utilizando o conjunto de treino para também testar as regras encontradas, ou seja, o teste foi realizado com os dados de 2014, e não mais 2015.

Após a realização do experimento, foi constatado a ocorrência do fenômeno denominado de *overfitting*, uma vez que o modelo obteve uma taxa de acurácia de 97%. Segundo Monard e Baranauskas (2003), o *overfitting* é caracterizado pelo ajuste em excesso para o conjunto de treinamento das hipóteses induzidas pelo algoritmo. O resultado desse acontecimento é a impossibilidade de aplicar o algoritmo em novos dados, pois o seu aprendizado se tornou específico demais para os casos do treinamento. Entretanto, o objetivo do aprendizado de máquina é gerar modelos cada vez mais genéricos e capazes de obter bons desempenhos em novas situações.

Porém, o *overfitting* pode trazer benefícios em algumas aplicações, desde que elas tenham como objetivo descrever e entender os dados utilizados para o treino, já que esse fenômeno fornecerá detalhes intrínsecos aos dados analisados. Apesar dessa pesquisa ter como objetivo entender os dados usados no estudo, o foco foi dado em regras que tenham continuidade ao longo dos anos, e por isso o *overfitting* não foi benéfico.

3.2. Cenário 2: Perfis dos Alunos do Ensino Básico

Este cenário foi utilizado para traçar os perfis dos alunos da educação básica de acordo com a sua região escolar: RMR ou Interior, por meio da utilização do algoritmo J48. O modelo gerado obteve uma taxa de acurácia de 63%.

Dentre as regras obtidas, as mais relevantes foram: (I) 57% dos alunos analisados que não utilizam transporte público, que declararam ser da raça parda e tem idades entre 8 e 17, estudam na RMR. Correspondendo a um total de 24,4% de acertos do modelo; (II) 86% dos alunos analisados que utilizam transporte público escolar, estudam no interior. Correspondendo a um total de 16% de acertos do modelo; (III)57% dos alunos analisados do sexo feminino, que não utilizam transporte público escolar e não declararam raça, estudam na RMR. Correspondendo a um total de 16% de acertos do modelo; (IV) 66% dos alunos analisados do sexo masculino com idades maior do que 15, que não utilizam transporte público escolar, não declararam raça e que não possuem deficiência, estudam na RMR. Correspondendo a um total de 14% de acertos do modelo.

Ainda na análise dos resultados foi possível verificar e comprovar que, quanto maior for o número de atributos considerados em uma determinada regra, mais específica ela se torna, consequentemente a quantidade de instâncias que se adéquam a este padrão contribui infimamente no total de acertos do modelo. Essa situação pode ser exemplificada na seguinte regra, cuja relevância é inferior a 1%: 51% dos alunos analisados do sexo feminino, declarados como da raça branca, com idade maior do que 17, que não utilizam transporte público escolar, estão no ensino médio, não possui necessidade especial e estudam no interior.

3.3. Cenário 3: Situações dos Alunos dos Cursos de TIC

Este cenário foi utilizado para identificar as características em comum entre os estudantes dos cursos de Tecnologia da Informação e Comunicação (TIC) que estivessem numa mesma situação acadêmica. Porém, após a aplicação do algoritmo de árvore de decisão J48, foi visto que ocorreu o fenômeno chamado de *underfitting*, caracterizado por um aprendizado insatisfatório pelo modelo gerado, posto que suas classificações tiveram sucesso em, apenas, 27% dos dados utilizados para testá-lo. Segundo Monard e Baranauskas (2003), o *underfitting* acontece quando as hipóteses induzidas no aprendizado se ajustam muito pouco ao conjunto de dados apresentado.

Após a obtenção desse resultado, foi assumido que existiam inconsistências elevadas entre os conjuntos de dados. A primeira hipótese adotada para validar tal constatação foi de que o processo de exclusão de registros para estratificar o conjunto de treino adulterou os padrões existentes nos dados, e para confirmar isso foi necessário realizar o mesmo teste, mas com a base de treino antes do processo de estratificação. Com isso, foi obtido uma taxa de acurácia relativamente melhor, em torno de 60%.

Porém, como pode ser observado na Figura 1, ocorreu justamente o problema que justifica a realização do processo de estratificação, o modelo classificou a grande parcela das instâncias como sendo da classe que possuía a maior quantidade de registros, e por isso teve uma taxa de acerto melhor, mas sem aprendizado.

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,955    0,914    0,619      0,955    0,752      0,085    0,578     0,656     Cursando
               0,057    0,035    0,317      0,057    0,096      0,047    0,565     0,256     Desvinculado do curso
               0,049    0,018    0,222      0,049    0,080      0,063    0,532     0,113     Matrícula trancada
               0,000    0,000    0,000      0,000    0,000      -0,005   0,590     0,105     Formado
Weighted Avg.  0,599    0,566    0,468      0,599    0,486      0,068    0,572     0,475

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 6178  213   73    2 |    a = Cursando
 2126  134   95    1 |    b = Desvinculado do curso
  914   47   49    0 |    c = Matrícula trancada
  757   29    4    0 |    d = Formado
```

Figura 1. Hipótese do Problema na Estratificação

Analisando as taxas para cada classe individualmente, existe uma discrepância ampla entre a classe Cursando e as demais, e também entre taxas para a mesma classe Cursando. Como o modelo classificou praticamente todas as instâncias como sendo dessa classe, fez com que a taxa de verdadeiros positivos (*TP Rate*) para ela fosse de 95,5%, enquanto que para as outras não chegou a mero 1%. Além disso, a precisão do modelo para essa classe foi em torno de 62%, evidenciando 38% de instâncias classificadas como Cursando, mas que eram de outra classe. Todas essas discordâncias fizeram com que a hipótese adotada (isto é, alteração dos padrões em virtude da estratificação) fosse anulada.

Dando continuidade a investigação, foi feito mais um experimento com o conjunto de treino após a estratificação, mas utilizando ele mesmo como sendo o conjunto para testar as regras encontradas. Com essa configuração, foi obtido um modelo com acurácia em torno de 62%, caracterizando a existência de regras e um aprendizado por parte do modelo. Além disso, as taxas de desempenho para cada classe individualmente se mostraram bastante balanceadas, diferentemente do que foi encontrado no experimento anterior.

Devido a esses resultados, concluiu-se que existem padrões entre os alunos dos cursos de TIC no estado de Pernambuco no ano de 2014, mas essas regras não tiveram continuidade para o ano seguinte, e por isso o modelo não se ajustou bem aos dados de teste.

3.4. Cenário 4: Perfis dos Alunos do Ensino Superior

Este cenário foi utilizado para comparar os perfis entre os alunos das IES situadas na Região Metropolitana de Recife com as do Interior do estado. Dentre todos os cenários definidos e devidamente aplicados os dois algoritmos escolhidos para a fase da Mineração, este cenário com o algoritmo *Naive Bayes* foi o que gerou o modelo com a maior taxa de acurácia, em torno de 77%.

Como a saída deste algoritmo determina as contribuições dos valores de cada atributo em relação aos valores do atributo classe, são mais adequados para situações com o objetivo de prever a classe de novas instâncias, e não descrever as instâncias utilizadas no treinamento. Diante disso, foi feita uma análise do desempenho do modelo para ratificar ou anular o aprendizado obtido e a possibilidade de aplicação em situações futuras. Essa análise foi feita por meio de exames e interpretações das taxas calculadas pelo WEKA, conforme apresentado na Figura 2.

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 4.8 seconds

=== Summary ===

Correctly Classified Instances      212358               76.5147 %
Incorrectly Classified Instances     65181               23.4853 %
Kappa statistic                          0.47
Mean absolute error                      0.2652
Root mean squared error                  0.3825
Relative absolute error                 53.0524 %
Root relative squared error             76.5075 %
Total Number of Instances           277539

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,737    0,137    0,949      0,737    0,830      0,511    0,893     0,966     Metropolitana de Recife
                 0,863    0,263    0,486      0,863    0,622      0,511    0,893     0,751     Interior
Weighted Avg.    0,765    0,165    0,846      0,765    0,783      0,511    0,893     0,918

=== Confusion Matrix ===

      a      b   <-- classified as
 158825  56717 |      a = Metropolitana de Recife
   8464  53533 |      b = Interior
```

Figura 2. Desempenho do modelo do Cenário 4

Sob a perspectiva geral, aproximadamente 77% das instâncias estão localizadas na diagonal principal da matriz de confusão, caracterizando os acertos do modelo. Por essa acurácia, o modelo tem plenas condições para ser utilizado em pesquisas futuras, mas também se faz necessário uma análise quanto ao aprendizado para cada valor do atributo classe.

Iniciando a decomposição da taxa de acurácia final pela lembrança (*Recall*), o valor para a classe Interior foi superior ao valor obtido para a outra opção do atributo classe. Isso indica uma maior sensibilidade e completude por parte do modelo para as instituições do Interior, já que aproximadamente 86% dos registros do Interior foram classificados corretamente contra 74% por parte da RMR.

Porém, essa taxa elevada da lembrança para o Interior não é o resultado de um satisfatório aprendizado para essa classe, e sim devido ao baixo índice de erros nas vezes em que o algoritmo classificou as instâncias como sendo da RMR. Isso está representado na taxa de precisão do modelo, indicando que para a RMR sua exatidão foi de, aproximadamente, 95%. Essa taxa significa que apenas 5% das instâncias classificadas como RMR eram, na verdade, do Interior. De contrapartida, o algoritmo não obteve um aprendizado satisfatório para os registros do Interior, resultando numa precisão de 49%, isto é, das vezes que o modelo classificou as instâncias como sendo do Interior, menos da metade possuíam, de fato, este valor para o atributo classe.

Existe um cruzamento entre essas taxas para os valores da classe. A alta precisão para a RMR resulta num menor número de Falsos Positivos, que por sua vez, influencia positivamente na lembrança para o Interior. No caminho contrário, a baixa precisão para o Interior, contribuiu negativamente na lembrança para as instâncias da RMR.

A solução para uma análise de desempenho menos específica é avaliar a taxa *F-measure*. Por ela é feita uma combinação da lembrança com a precisão do modelo, através do cálculo da média harmônica ponderada a seguir: F = 2 * ((Precisão * Lembrança) / (Precisão + Lembrança)). A constante 2 é utilizada para fazer com que a medida F atinja seu valor máximo, 1 (um). Isso acontece quando ambas, precisão e lembrança, também são iguais a 1. Com o balanço feito por essa medida, a RMR obteve 83% contra 62% para o Interior, resultando numa média, ponderada pela quantidade de registros para cada uma das classes no conjunto de teste, de 78%.

4. Considerações Finais

As etapas do KDD que antecedem a fase da MD são responsáveis pela maior parte do

tempo gasto no processo e influenciam completamente na eficácia dos modelos gerados. Sendo assim, foi dado foco na obtenção de dados com níveis de qualidade elevados, ou seja, puros, íntegros e confiáveis. Então foram feitas repetidas iterações entre as etapas da Pré-Mineração a fim de garantir que os dados estavam realmente apropriados para serem minerados, revelando mais uma vez a importância e participação dos analistas do domínio do problema.

Portanto, as bases de dados contempladas têm totais condições para serem reutilizadas em trabalhos futuros, desde que também tenham como objetivo explorar os dados educacionais dos Censos da Educação Básica e Superior, coletados e disponibilizados pelo INEP. O grande benefício para esses outros trabalhos é a oportunidade de aplicar o processo de KDD passando rapidamente pelas etapas iniciais de limpeza dos dados, apenas precisando fazer ajustes específicos para o objetivo em questão, e já passar para as fases da obtenção, análises e interpretações dos resultados.

No que tange aos resultados encontrados, em duas situações fenômenos existentes na área afetaram o modelo gerado a ponto de impedir a sua utilização neste trabalho. Vale destacar que o nível da qualidade dos dados, de forma alguma, foi a causa dessas ocorrências, uma vez que se mostraram íntegros e consistentes. Entretanto, este fato serviu para elucidar, de maneira prática, sobre as situações inesperadas que podem ser encontradas durante o processo de mineração de dados.

Já no quarto cenário, apesar da considerável diferença entre o aprendizado para os dois valores do atributo classe, o modelo obteve taxas relativamente satisfatórias, assim como o modelo gerado pelo J48 no segundo cenário analisado. Então, outra contribuição científica da pesquisa foi conceder modelos, com comprovações de suas condições para serem aplicados em trabalhos futuros, que possibilitam a determinação da localização das instituições de alunos, tanto do ensino básico quanto do superior, do estado de Pernambuco.

Uma vez que as medidas tomadas pelos gestores para amenizar os problemas e melhorar a situação estudantil possuem abordagens e necessidades diferentes dependendo da localização geográfica da instituição de ensino, esses cenários devem ser utilizados como embasamento para as suas decisões. Pois, a utilização dos recursos tecnológicos desenvolvidos nesta pesquisa possibilita a investigação e detecção das principais carências com base nos diferentes perfis dos alunos e das instituições, em todos os níveis de ensino.

Cabe às pesquisas futuras, dar continuidade ao processo iniciado nesta, aplicando os modelos encontrados e realizando análises mais profundas das regras que forem geradas, no que se refere aos padrões que justifiquem as dificuldades presentes na educação, a fim de elevar a qualidade das escolas e cursos, contribuindo para o desenvolvimento tecnológico, econômico e social do estado de Pernambuco.

Referências

DIMITOGLOU, George; ADAMS, James A.; JIM, Carol M. Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability. arXiv preprint arXiv:1206.1121, 2012.

ERBERT, Mauro. Introdução ao sensoriamento Remoto. Master Tesis, Universidade Federal do Rio Grande do Sul, 2001.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37, 1996.

FERREIRA, Gisele. Investigação acerca dos fatores determinantes para a conclusão do Ensino Fundamental utilizando Mineração de Dados Educacionais no Censo Escolar da Educação Básica do INEP 2014. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2015. p. 1034.

FRAWLEY, William J.; PIATETSKY-SHAPIRO, Gregory; MATHEUS, Christopher J. Knowledge discovery in databases: An overview. AI magazine, v. 13, n. 3, p. 57, 1992.

GONZALEZ, L. F. P. Uma abordagem para mineração de dados e visualização de resultados em imagens batimétricas. Pontifícia Universidade Católica do Rio Grande do Sul, 2012.

HAN, J. and KAMBER, M. (2006), Data Mining Concepts and Techniques, Morgan Kauffmann Publishers, Second Edition.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. Sistemas Inteligentes-Fundamentos e Aplicações, v. 1, n. 1, 2003.

PATIL, Tina R.; SHEREKAR, S. S. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications, v. 6, n. 2, p. 256-261, 2013.

RIGO, Sandro J.; CAZELLA, Silvio C.; CAMBRUZZI, Wagner. Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. In: Anais do Workshop de Desafios da Computação Aplicada à Educação. 2012. p. 168-177.

SANTOS, Rodrigo et al. Uso de Séries Temporais e Seleção de Atributos em Mineração de Dados Educacionais para Previsão de Desempenho Acadêmico. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2016. p. 1146.

WITTEN, Ian. H; FRANK, Eibe; HALL, Mark A. Data mining: practical machine learning tools and techniques. São Francisco: Morgan Kaufmann. 4rd edição, 2017.