XVII COLÓQUIO INTERNACIONAL DE GESTÃO UNIVERSITÁRIA
Universidade, desenvolvimento e futuro na Sociedade do Conhecimento
Mar del Plata – Argentina
22, 23 e 24 de novembro de 2017
ISBN: 978-85-68618-03-5

SOFTWARE SDBAYES: UM AUXÍLIO PARA A PREDIÇÃO DE EVASÃO DISCENTE

JACQUES NELSON CORLETA SCHREIBER
Universidade de Santa Cruz do Sul
jacques@unisc.br

ALVIN LAURO BESKOW
Universidade de Santa Cruz do Sul
alvinbeskow@hotmail.com

ELPIDIO OSCAR BENITEZ NARA
Universidade de Santa Cruz do Sul
elpidio@unisc.br

JULIANA IPÊ DA SILVA
Universidade de Santa Cruz do Sul
Juliana_ipe@hotmail.com

JAQUELINE DE MORAES
Universidade de Santa Cruz do Sul
jaque.moraes.93@hotmail.com

VERÔNICA MEINHARDT NAJDZION
Universidade de Santa Cruz do Sul
vnajdzion@mx2.unisc.br

RESUMO

Esse artigo apresenta um software que é resultado de um projeto de pesquisa que visa a predição de evasão discente no ensino superior. O software foi criado com o intuito de prever os discentes mais propensos a evadir de uma instituição de ensino superior, apresentando a probabilidade individual e o motivo mais forte que está conduzindo o discente a evadir, auxiliando os gestores acadêmicos na tomada de decisão proativa. O software utiliza redes bayesianas que é um dos métodos de classificação descritos na literatura de *Data Mining*, muito usado para geração de estimativas discretas sobre o problema abordado. O software possui dois modos de operação: o primeiro, calcula a probabilidade de um único discente evadir, além de ser um modo mais completo e intuitivo de analisar os riscos de evasão do discente, já o segundo modo, analisa um conjunto de discentes, atribuindo a cada um deles um valor de probabilidade de evasão, além do que é possível exibir as variáveis que mais influenciam para tal probabilidade.

Palavras chave: Redes Bayesianas, Predição, Evasão estudantil, Software.

1. INTRODUÇÃO

A gestão em uma organização é um dos principais fatores que permitem a longevidade e estabilidade da mesma, para que tais feitos sejam alcançados, informações processadas são necessárias, não somente informações discretas, mas sim estatísticas e dados complementares que auxiliem o gestor na tomada de decisão. As organizações costumam armazenar dados de suas transações, produtos, ocorrências, clientes, fornecedores, etc. e com isso pensam estarem informadas, mas não estão; pois sem um método de descoberta de conhecimento, a gestão não irá conseguir processar esses dados de forma/tempo hábil. Algumas técnicas de mineração de dados (*Data Mining*) são exatamente o que a organização precisa para processar esses dados Cheng (2009) e conseguir com isso, um ganho de informação e ter uma gestão proativa, evitando gastos desnecessários, aplicando onde tem o melhor retorno financeiro, evitando a perda de clientes, e afins.

O objetivo desse trabalho foi desenvolver uma ferramenta que analisa uma base de dados de uma Instituição de Ensino Superior (IES), para determinar quais as probabilidades que o discente tem de permanência e de evasão do curso em que está matriculado. Essa ferramenta usa Redes Bayesianas (RB) para fazer a previsão de suas probabilidades, que é uma das inúmeras técnicas de mineração de dados usadas para classificação Friedman (1997), e atualmente é denominada como *SDBayes: a System to Predict Student Drop-out* (*SDBayes*).

O trabalho está dividido da seguinte maneira: na próxima seção uma abordagem sobre os trabalhos relacionados, posteriormente, a proposta de solução desenvolvida pela equipe, depois os resultados obtidos, e as considerações finais.

2. TRABALHOS RELACIONADOS

Rovira, *et al.,* (2017) propôs um sistema baseado em aprendizagem de máquina para auxiliar o tutor quando deve-se intensificar o auxílio ao discente. Para isso o sistema aborda duas principais tarefas: 1) a previsão precoce de abandono do aluno e, 2) a previsão de notas de cursos subsequentes para cada aluno, bem como recomendações de cursos personalizados. Os cursos analisados foram Ciência da Computação, Direito e Matemática, onde o índice de evasão beira os 30%. Foram testados diversos métodos, no entanto, o método, de geração de redes, com melhor índice de acerto foi o *Naive Bayes*, juntamente com a Regressão Logística, atingido acerto de até 82%.

No trabalho de Marbouti (2016), foram testados sete diferentes métodos de predição, com o principal objetivo, identificar o melhor método de predição capaz de identificar o risco de evasão estudantil, levando em consideração o motivo de um método ser melhor que outro. As técnicas usadas foram, Regressão logística, *Support Vector Machine*, Árvore de Decisão, *Multi-Layer Perceptron*, *Naive Bayes*, *K-Nearest Neighbor*, e o Modelo *Robustness*. Os dados de aprendizagem foram os mesmos para todos os testes. O problema local, é que a quantidade de casos testados é de 780 e apenas 39 evadiram, o que torna a possibilidade de previsão muito difícil. Então foram criadas duas categorias; a dos métodos que predisseram melhor a assertividade de evasão, e os que previram melhor a assertividade de permanência. O modelo que apresentou a melhor precisão de assertividade de permanência foi o *K-Nearest Neighbor*, que identificou 94,9% dos estudantes. O segundo melhor foi o *Multi-Layer Perceptron*, que identificou 93,1% dos estudantes, e a Árvore de Decisão foi a terceira colocada. O modelo que melhor previu o abandono escolar, foi o *Naive Bayes*, com 86,2%; considerando que houve apenas 39 casos aprendidos de evasão dentre os 780 casos, foi um índice muito bom, já

o segundo melhor foi o *Support Vector Machine*, com probabilidades semelhantes ao *Naive Bayes*.

Tekin (2014) trabalhou em comparar três técnicas de *Data Mining* para prever as médias de notas dos alunos, se os estudantes fossem preditos com baixas médias, medidas deveriam ser tomadas para implementar uma nota maior. Os dados usados foram adquiridos de um curso de computação, com 127 alunos que abandonaram a graduação e outros 49 que concluíram, o que torna um grande agravante para fazermos previsão, pois a cada estudante que concluiu o curso, outros 2,6 que evadiram, são aprendidos pelo algoritmo. As técnicas foram validadas com o método *k-fold* onde o k=5, e o melhor método de predição nesse problema foi o *Support Vector Machine* com 97,98% de assertividade, o segundo foi o *Extreme Learning Machine* com 94,92%, e por fim, uma rede neural com 93,76%.

O trabalho de Maria (2016), traz uma proposta semelhante com a ferramenta desenvolvida nesse projeto. Eles desenvolveram um software para a predição de evasão, baseado em Redes Bayesianas, que resultou em ótimas médias de acerto com os dados coletados, no entanto, a técnica usada se assemelha com o *Naive Bayes*, exceto por 3 nodos dentre os 18 usados, com isso, avaliamos que poderiam ter sido usadas outras técnicas de montagem da rede bayesiana, como por exemplo o algoritmo *tree augmented naive bayes*, que gera uma árvore aumentada a partir do *naive bayes*, obtendo probabilidades de acerto ainda maiores.

3. PROPOSTA DE SOLUÇÃO

A abordagem desse trabalho tem o intuito de auxiliar a gestão universitária especialmente com o problema de evasão discente, visto que os investimentos feitos para o longo prazo, podem ser superiores à demanda prevista, podendo acarretar em dificuldades na sustentabilidade econômica do curso. Com isso, foram pesquisados métodos de análise de grandes volumes de dados, visando prover poder de decisão ao gestor acadêmico. A técnica escolhida foi as Redes Bayesianas (RB) que conforme Kelly (2013) é dividida em duas partes principais, a parte Qualitativa e a parte Quantitativa;

A parte Qualitativa são as ligações entre as variáveis e, cada estado que a variável pode assumir. Já a parte Quantitativa, é composta pelas probabilidades à priori e posteriori, geradas a partir das relações determinadas na parte Qualitativa. O método de aprendizagem de máquina, é o responsável pela geração dessas tabelas, dentre os métodos de aprendizagem de máquina, utilizou-se o algoritmo *Expectation Maximization* (EM), embora métodos como *Learn and Count* e *Gradient* tenham sido considerados. O EM foi utilizado para gerar as tabelas de probabilidades da RB, até porque a RB é uma evolução do método *Naive Bayes* (NB), visto que o NB é uma das etapas da criação da parte Qualitativa da RB pelo método do *Three Augmented Naive Bayes* (TAN) Friedman (1997).

O método de criação das RB, descrito acima foi adotado como o núcleo de previsão do projeto, no entanto, para um gestor que precisa se preocupar com outras tarefas, desfrutar unicamente de uma RB não ajudará muito, pois ele terá grande dificuldade em extrair qualquer informação. Com isso, criou-se um software que em seu núcleo existem três RB de diferentes cursos, além de possuir uma interface amigável e de fácil utilização. Na seção 3.1 será abordada a funcionalidade do software e na seção 3.2, a criação das tabelas de dados, que podem ser usadas como parâmetro para as redes bayesianas.

3.1. FUNCIONALIDADES DO SOFTWARE

Durante o desenvolvimento das redes de previsão, após entrevistas com os coordenadores dos respectivos cursos, concluiu-se que cada curso tem seu próprio perfil de aluno. Com isso, um aluno de um curso X, que cursa 5 cadeiras/semestre em média, tem uma probabilidade diferente de evadir, que outro aluno de outro curso com a mesma quantidade de cadeiras/semestre. Com essa notação, foram criadas 3 redes bayesianas, para três cursos distintos, onde cada rede usa o histórico discente do respectivo curso. A primeira rede desenvolvida, foi para o curso de Ciência da Computação, onde haviam cerca de 1800 casos, e somente 266 casos eram de alunos que concluíram o curso, o que foi um complicador para criar a tabela de treino da rede, pois apenas, cerca de 15%, dos alunos efetivamente terminaram o curso, entretanto, no final, conseguiu-se reduzir para 905 casos, onde 266 desses casos eram de alunos que realmente concluíram o curso. A segunda rede desenvolvida, foi para o curso de Administração, onde a rede foi treinada com 799 casos, dentre os quais 383 foram de alunos que efetivamente concluíram o curso. Já a terceira rede desenvolvida, foi para o curso de Engenharia de Produção, onde as redes foram treinadas com 659 casos, dentre os quais, 314 eram sobre informações de discentes que concluíram o curso, abaixo uma demonstração da tela inicial do software (Figura 1).

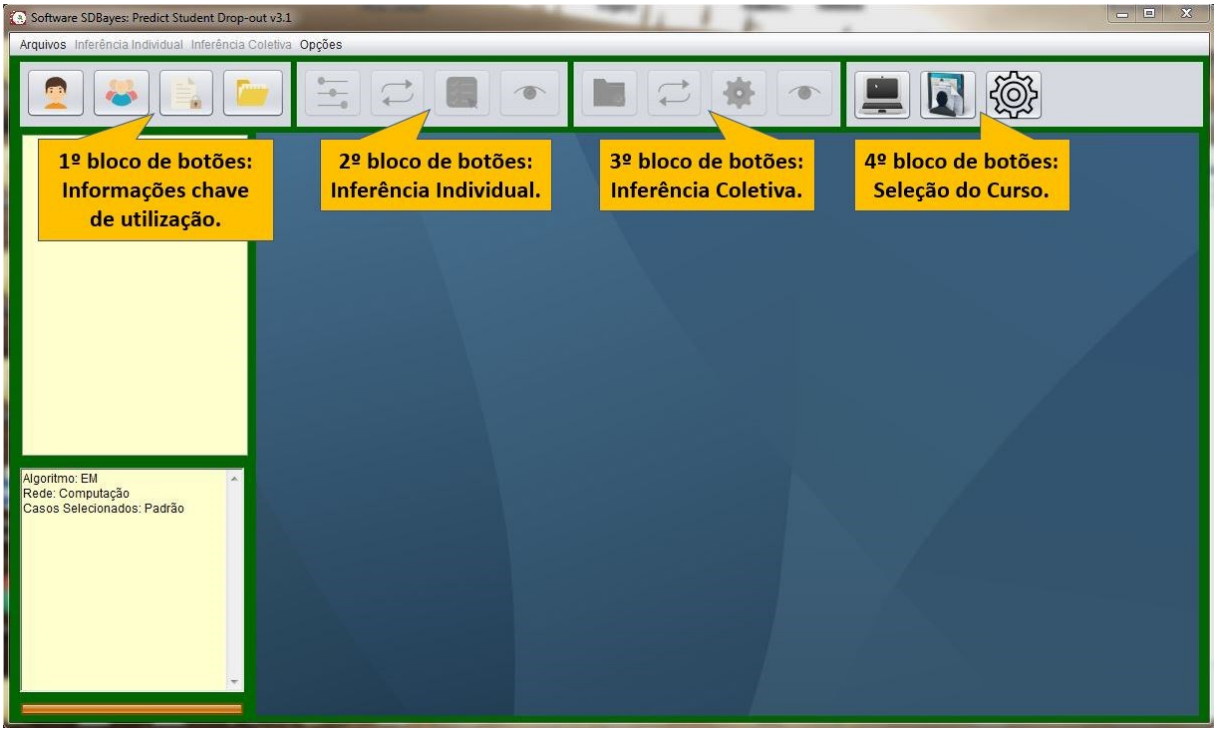

Figura 1: Software *SDBayes: a System to Predict student drop-out*
Fonte: Autores, 2017.

O software *SDBayes* ilustrado na Figura 1, conta com quatro blocos de botões, onde o primeiro bloco diz respeito às informações chave de utilização, no segundo e terceiro bloco, tem-se o modo de uso do sistema, que são: "Inferência Individual", e "Inferência Coletiva", e por fim, o quarto grupo é referente ao curso que será testado (Computação, Administração, ou engenharia de Produção). Abaixo, subtópicos 3.1.1 e 3.1.2, uma explicação sobre os modos de operação do sistema.

3.1.1. Interferência Individual

O método de inferência Individual, consiste em ter controle de apenas um discente, porém, um controle preciso e facilitado. Como pode ser notado na Figura 2, quando clicado no botão de Inferência Individual, do primeiro bloco de botões, o método de inferência Individual é ativado, com isso, quatro botões de controle são disponibilizados, referentes ao segundo bloco de botões, além de aparecer uma árvore à esquerda contendo informações sobre os comandos de operação.

As funções de controle disponibilizadas são abstrações da usabilidade da rede bayesiana, clicando no primeiro botão do segundo bloco de botões, será feita a abertura do ajustador das variáveis do discente, na arvore, tem o nome do "Alterar *LikeliHood*". Essa opção é a mais importante para esse modo de operação, como pode ser notado na Figura 2, são disponibilizadas nove variáveis, para que o gestor entre com os dados do aluno, e caso não haja certeza absoluta, pode-se controlar a confiança dessas informações ajustando a barra contida em cada variável.

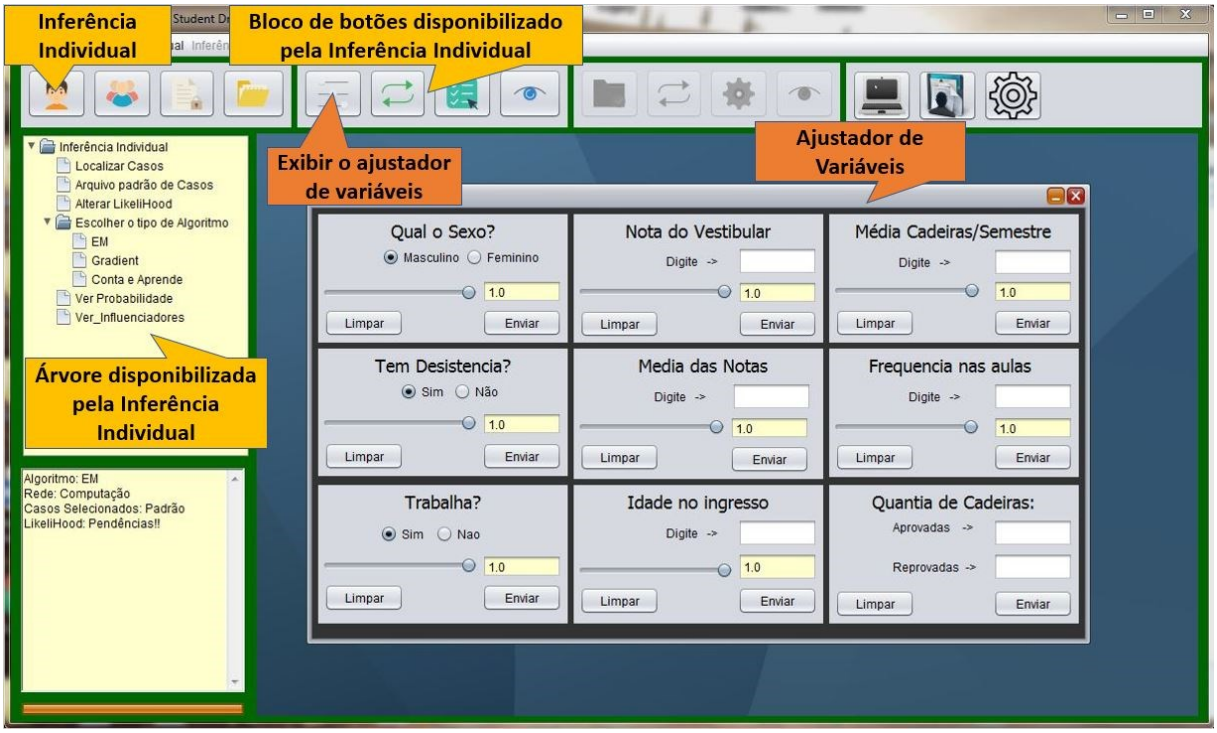

Figura 2: Software *SDBayes: a System to Predict student drop-out*, Inferência Individual, preenchimento das variáveis
Fonte: Autores, 2017.

Além da entrada de dados, esse modo de inferência permite que façamos a escolha de dois modos de visualização de resultados, um modo que exibe somente a probabilidade de evasão, denominado "Teste Comum", e outro modo, no qual tem-se a probabilidade de evasão, mas além disso, exibe as variáveis que mais influenciam para que tal probabilidade chegasse a esse valor, que é denominado "Teste Avançado". O motivo de haverem dois modos, é que o modo que mostra as variáveis mais influenciadoras, precisa de um tempo computacional maior para ser executado, cerca de 5x. Abaixo, na figura três o modo de operação Inferência Individual, exibindo o resultado avançado de um teste.

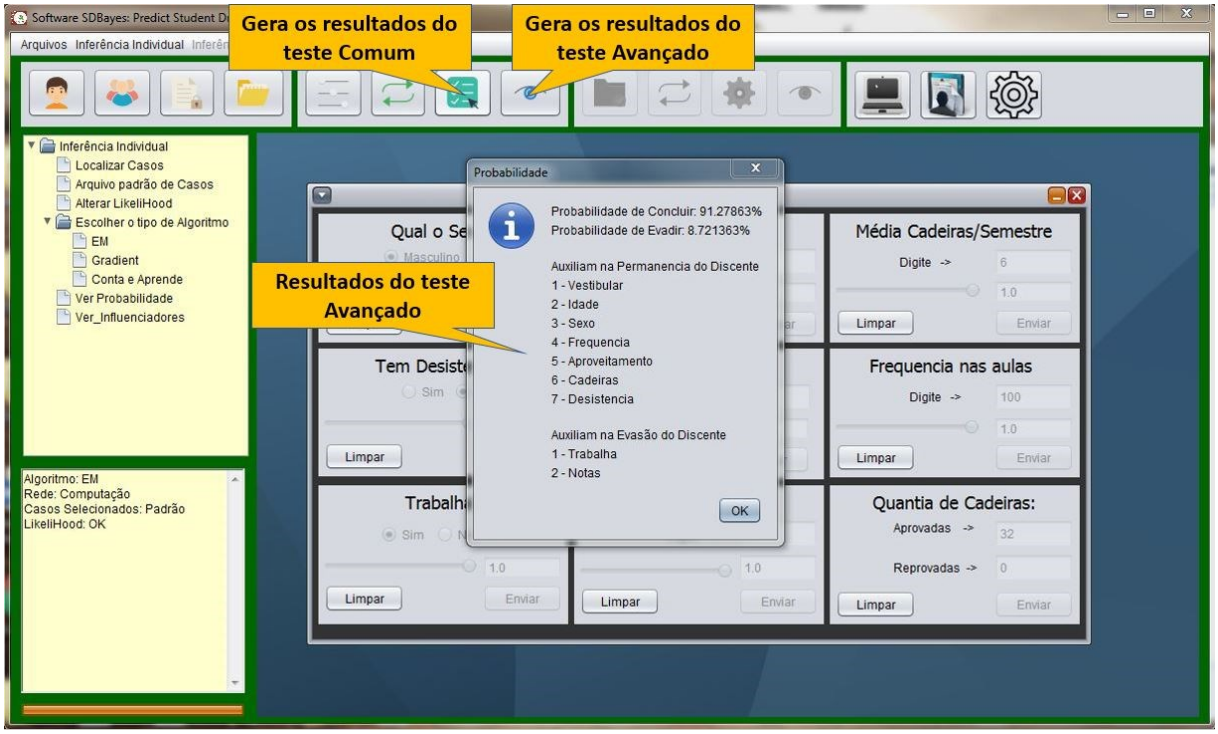

Figura 3: Software *SDBayes: a System to Predict Student Drop-Out*, Resultados da inferência individual, com o teste avançado
Fonte: Autores, 2017.

3.1.2. Interferência Coletiva

O modo de operação denominado inferência coletiva, é o modo de operação que fornece a maior comodidade para o gestor, levando em conta que ele pode analisar todos os alunos de um curso e determinar com isso, quais são os alunos mais propensos a abandonar a instituição. Abaixo uma imagem ilustrando o software no modo de inferência coletiva.

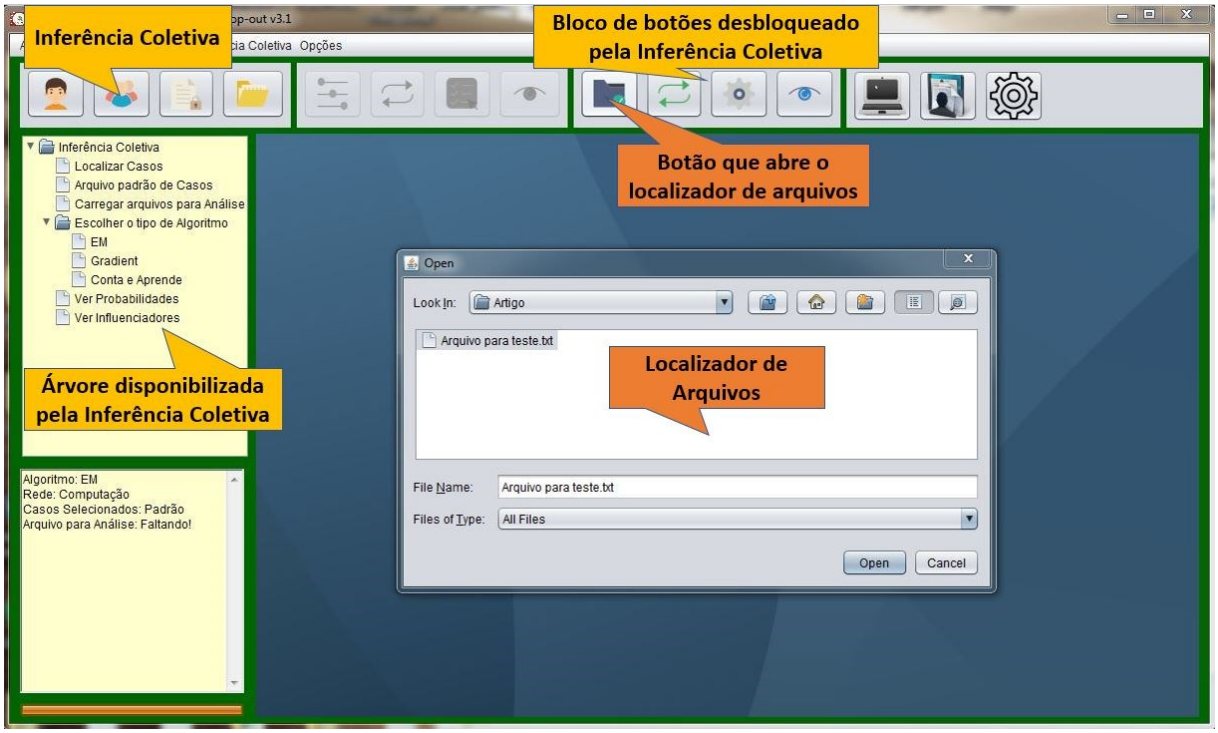

Figura 4: Software *SDBayes: a System to Predict Student Drop-Out*, Inferência coletiva
Fonte: Autores, 2017.

Como pode ser observado na Figura 4, o terceiro bloco de botões destacado na figura está habilitado, isso foi ocasionado pelo clique no botão de Inferência Coletiva. Após a liberação do terceiro bloco, também foi disponibilizada uma árvore, à esquerda, contendo as informações de controle. Nesse modo, podemos selecionar uma tabela contendo os dados dos alunos, clicando no primeiro botão do terceiro bloco de botões, além disso, temos também dois modos de exibição das informações. Um no qual mostra apenas a probabilidade de evasão/permanência podendo ser exibido pelo terceiro botão do terceiro bloco, e outro que mostra as duas variáveis que mais influenciam em ter essa probabilidade, quarto botão do quarto bloco, os modos muito semelhantes aos desenvolvidos para a Inferência Individual, ambos podem ser vistos na Figura 5.

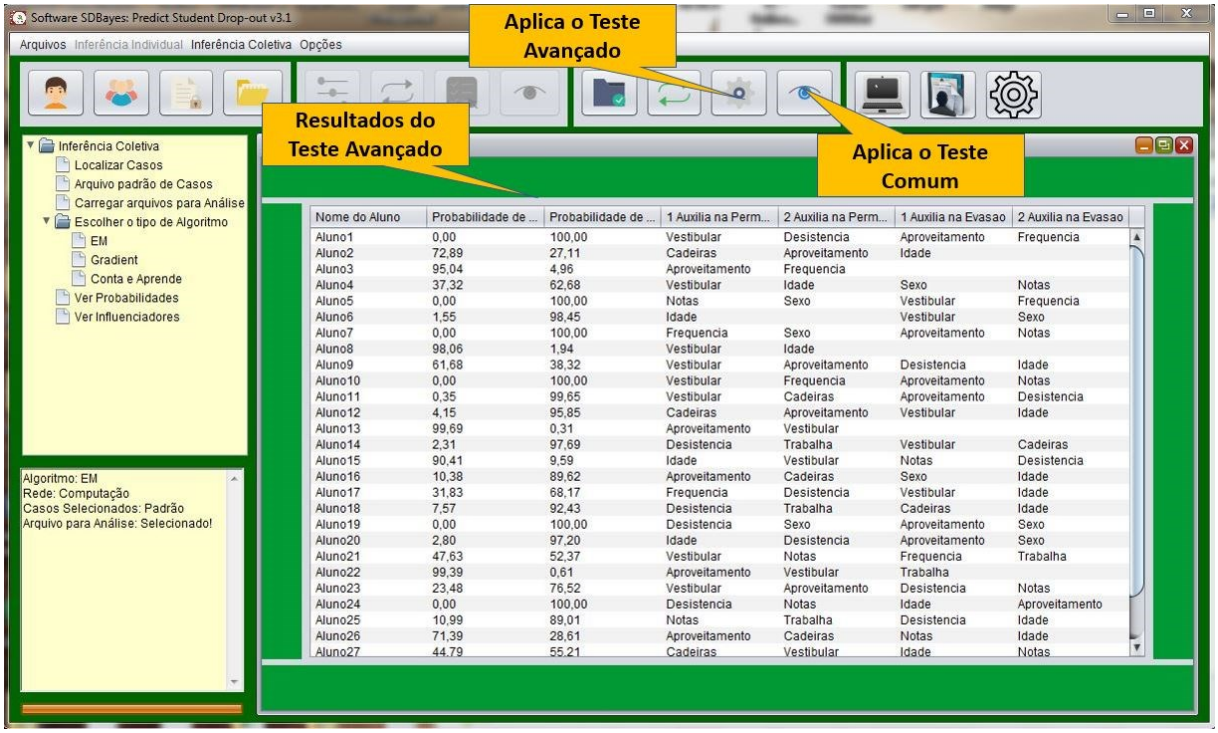

Figura 5: Software *SDBayes: a System to Predict Student Drop-Out*, inferência coletiva, com teste Avançado
Fonte: Autores, 2017.

3.2. PREPARAÇÃO DE DADOS

Antes da utilização da ferramenta é preciso criar o arquivo de casos, no qual as redes irão se basear. Esse arquivo de dados faz parte do processo de KDD, que é o processo de busca de conhecimento que contém uma série de passos: seleção, pré-processamento e limpeza, transformação, mineração de dados (data mining) e interpretação/avaliação. Simplificando, pode-se dizer que o processo de KDD compreende, na verdade, todo o ciclo que o dado percorre até virar conhecimento, conforme pode ser visto na figura 6.

Figura 6: Processo de KDD
Fonte: adaptado de GONÇALVES, 2001.

Esse arquivo é chamado arquivo de aprendizagem (ou treino), e é gerado a partir do banco de dados da Instituição. Somente contendo algum arquivo desse gênero, essa ferramenta é capaz de operar, pois são com esses dados que permitimos a rede ter um parâmetro para calcular as probabilidades dos discentes. A ferramenta já possui um arquivo de aprendizagem para cada curso, mas como cada instituição tem sua realidade, o arquivo gerado em uma instituição pode ser totalmente inválido em outra. A formatação do arquivo tem o seguinte layout de colunas.

- A primeira coluna é referente a situação real do discente, no entanto é um dado booleano, atribuindo S para quem concluiu e N para quem não concluiu o curso.
- A segunda coluna novamente possui dados booleanos, referentes ao sexo do aluno (M, F).
- Terceira coluna contém dados booleanos, para informar se já houve ao menos uma desistência de disciplina por parte do estudante (S, N).
- Quarta coluna novamente um dado booleano indicando se o discente está ou não trabalhando (S, N).
- Quinta coluna, é referente a nota do vestibular, onde tem-se dados nominais, os quais são limites de valores que determinam uma nova palavra.
- Na sexta coluna, usa-se uma estrutura semelhante a quinta, no entanto essa é para a média aritmética de todas as notas das disciplinas.
- O mesmo se aplica para a sétima coluna onde tem-se a idade do discente.
- Na oitava coluna foi usada a média de cadeiras por semestre que o discente cursou.
- Na nona, a frequência em aula do discente.
- E por fim, a décima coluna é um cálculo matemático, criado por nós para determinar o aproveitamento das cadeiras cursadas, onde tem-se uma relação entre as cadeiras concluídas, com as cadeiras onde houveram reprovações, e é expressa pela fórmula:

  Nomenclaturas:

  AP = Aproveitamento

  DA = Soma das disciplinas aprovadas pelo discente

  DR = Soma das disciplinas reprovadas pelo discente

$$AP = \frac{DA - DR}{DA}$$ **Fórmula 1:** Equação de Aproveitamento

Esse arquivo gerado pode substituir o arquivo original, enviado com o software, selecionando o quarto botão do primeiro bloco de botões, figura 1, então será exibido um Localizador de Arquivos. Caso seja escolhido usar os dados originais, pode-se ignorar essa opção, pois já está habilitado como padrão, ou clicando no terceiro botão do primeiro bloco de botões, também figura 1.

Outra vantagem da geração desse arquivo de treino, é a possibilidade de testar todos esses dados usando a Inferência Coletiva figura 4, no entanto, a primeira coluna de dados desse arquivo, será o nome do discente e não a situação atual dele.

4. RESULTADOS

A validação efetiva é obtida pelo gestor ao comparar os prognósticos emitidos pelo software e comparado aos fatos. No entanto, essa prática não é viável momentaneamente, então usou-se como parâmetro de validação, o método de validação cruzada *k-fold*, que de acordo com Hadihardaja (2014) é um método muito bom, com o k=10, com isso simulamos o desempenho do software no aspecto preditivo. Conforme Helma (2004), método de validação cruzada 10-fold divide a base de dados em dez partes, sendo que 9 dessas partes são usadas para aprendizagem da rede bayesiana e a parte restante é usada para teste. Isso é feito 10 vezes de modo que cada parte tenha servido como teste, gerando 10 resultados.

Testando o *k-fold* nas três redes inerentes ao software, a Redes de Computação apresentou probabilidades de 60,98%, 90,04% e 82,01% para permanência, evasão e no geral respectivamente. Com isso temos que 60,98% das vezes que o software analisou um estudante do curso de computação, que realmente permaneceu no curso, ele acertou. Também, tem-se que 90,04% das vezes que o software analisou um estudante do curso de computação, que realmente evadiu do curso, ele acertou. E 82,01% das vezes que o software analisou um estudante do curso de Computação, independentemente de ter sido aprovado ou reprovado, ele acertou.

Além disso o desvio padrão foi de 11,53%, 1,91% e 3,47%, respectivamente para permanência, evasão e no geral. Abaixo temos o gráfico 1, que mostra o resultado das investidas:

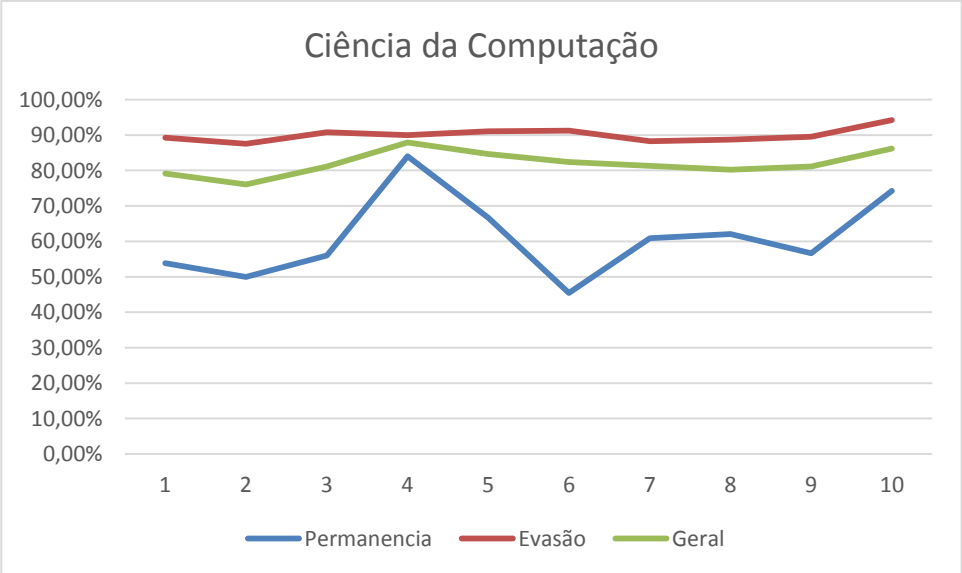

Gráfico 1: Acerto das Investidas na rede de Computação
Fonte: Autores, 2017.

A rede de Administração comportou-se de forma mais estável nos testes em relação a Ciência da Computação. O *k-fold* na rede de Administração, apresentou probabilidades de 79,14%, 73,02% e 75,79% para permanência, evasão e no geral respectivamente. Com isso temos que 79,14% das vezes que o software analisou um estudante do curso de Administração, que realmente permaneceu no curso, ele acertou. Também, tem-se que 73,02% das vezes que o software analisou um estudante do curso de Administração, que realmente evadiu do curso, ele acertou. E 75,79% das vezes que o software analisou um

estudante do curso de Administração, independentemente de ter sido aprovado ou reprovado, ele acertou.

Além disso o desvio padrão foi de 7,15%, 7,60% e 6,10% para permanência, evasão e no geral respectivamente. Abaixo temos o gráfico 2, que mostra o resultado das investidas:

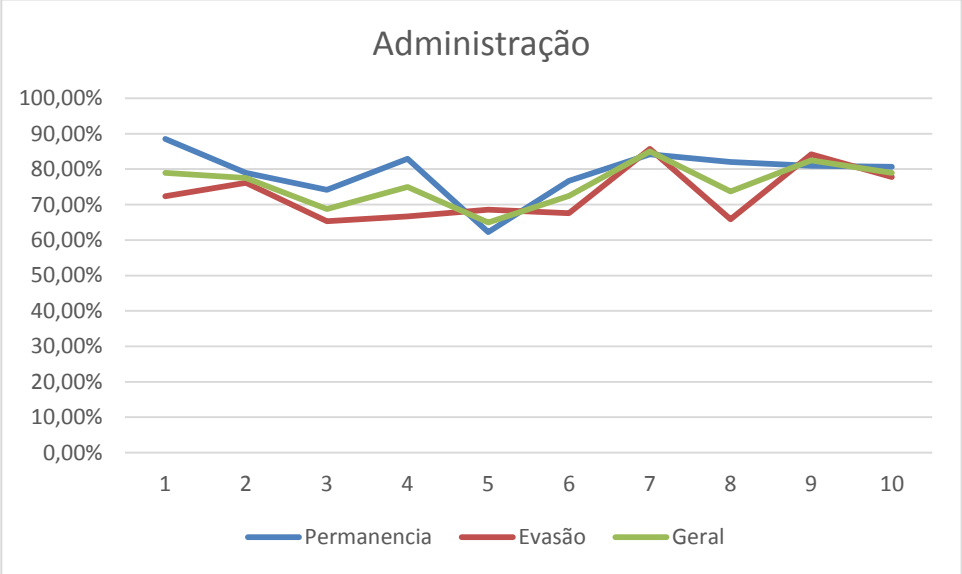

Gráfico 2: Acerto das Investidas na rede da Administração
Fonte: Autores, 2017.

Por fim a Rede de Engenharia de Produção que apresentou probabilidades de 57,39%, 69,29% e 63,94% para permanência, evasão e no geral respectivamente. Com isso temos que 57,39% das vezes que o software analisou um estudante do curso de Engenharia de Produção, que realmente permaneceu no curso, ele acertou. Também, tem-se que 69,29% das vezes que o software analisou um estudante do curso de Engenharia de Produção, que realmente evadiu do curso, ele acertou. E 63,94% das vezes que o software analisou um estudante do curso de Engenharia de Produção, independentemente de ter sido aprovado ou reprovado, ele acertou.

Além disso o desvio padrão foi de 21,90%, 16,27% e 17,48% para permanência, evasão e no geral respectivamente. Abaixo temos o gráfico 3, que mostra o resultado das investidas:

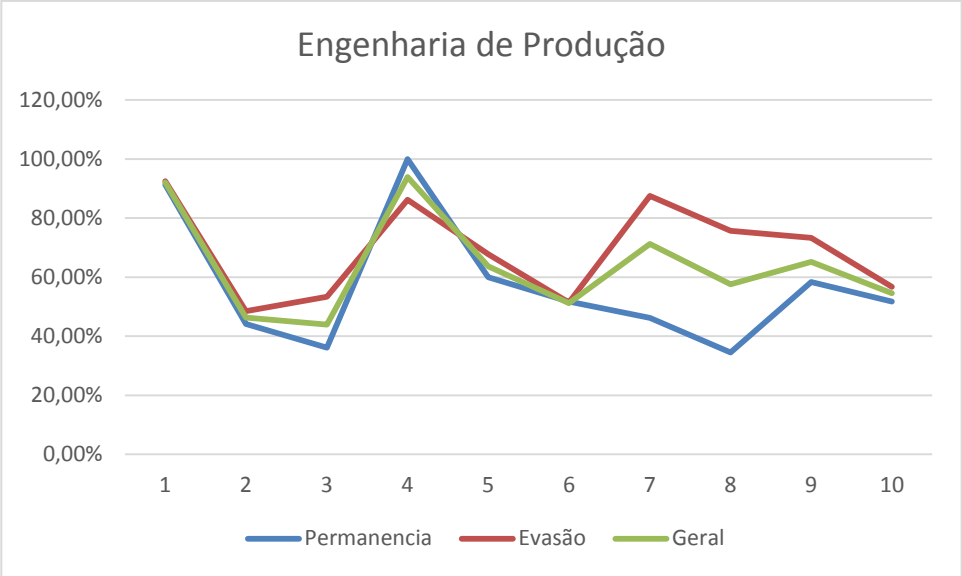

Gráfico 3: Acerto das Investidas na rede de Produção
Fonte: Autores, 2017.

Com isso, podemos ver que a rede de administração tem o percentual de acerto mais estável de todos os testes, seguido por Computação e Engenharia de Produção. No entanto o curso que mais acerta as investidas feitas sobre os discentes, é o curso de computação que chega a ter acerto geral de até 82%. Já analisando as investidas feitas sobre os discentes que permaneceram na instituição, o curso que apresenta a melhor estimativa é o Curso de Administração que acertou 79,14% das investidas. E por fim, o curso que melhor apresentou resultados sobre os alunos que evadiram do curso, foi o curso de computação que acertou 90,04% das investidas.

5. CONCLUSÃO

A evasão discente no ensino superior é um problema que afeta o resultado dos sistemas educacionais. As perdas de estudantes que iniciam, mas não terminam seus cursos são desperdícios sociais, acadêmicos e econômicos. Nas universidades públicas são recursos investidos sem o devido retorno, nas universidades privadas é uma importante perda de receita. Em ambos casos, a evasão é fonte de ociosidade de professores, funcionários, equipamentos e espaço físico.

Neste artigo um dos resultados do nosso projeto de pesquisa, o software *SDBayes*, foi apresentado em detalhes. A metodologia, a descrição detalhada do processo de construção e inferência da rede bayesiana subjacente estavam fora do escopo deste artigo, outro artigo em processo de redação no momento da escrita desse artigo elucidará dúvidas sob estes aspectos. Nesse artigo o foco foi apresentar aspectos da funcionalidade do aplicativo e principalmente demonstrar os resultados promissores apresentados pelo método de validação *K-fold*, um dos mais utilizados na área de mineração de dados.

As previsões de quais alunos estão propensos a evadir e por qual motivo permitirão uma gestão acadêmica proativa. Acreditamos que essa informação contribua para uma diminuição dos índices de evasão. A escolha dos três cursos deveu-se ao fato de cursos de computação possuírem os índices mais altos de evasão no Brasil. Os dois cursos adicionais, Engenharia de Produção e Administração de Empresas foram utilizados como *benchmark* do software. Explicitando: o curso de ciência da computação na IES avaliada possui 76% de

média de evasão nos últimos 24 anos enquanto que o curso de Administração e Engenharia de Produção possuem índices bem menores. No primeiro caso, a tarefa difícil do software foi predizer quem completaria o curso, visto que a maioria evade; em contraposição os dois cursos adicionais possuem índices de evasão baixos, Administração com um índice de 52% de evasão e Engenharia de Produção com um índice de 53% de evasão, nesse caso a tarefa difícil foi predizer quem teria propensão à evasão.

O projeto de pesquisa continua investigando novos métodos de predição e ao mesmo tempo utilizando outras técnicas de validação, como métodos como o *F-Measure, Leave-one-out, Monte-Carlo cross-validation*, e o *Receiver Operating Characteristics*.

REFERÊNCIAS

CHENG, Hilary; LU, Yi-Chuan; SHEU, Calvin. An ontology-based business intelligence application in a financial knowledge management system. Expert Systems with Applications, v. 36, n. 2, p. 3614-3622, 2009.

FRIEDMAN, Nir; GEIGER, Dan; GOLDSZMIDT, Moises. Bayesian network classifiers. Machine learning, v. 29, n. 2, p. 131-163, 1997.

GONÇALVES, Lóren Pinto Ferreira. Avaliação de ferramentas de mineração de dados como fonte de dados relevantes para a tomada de decisão: aplicação na rede Unidão de supermercados, São Leopoldo-RS. 2001.

HADIHARDAJA, Iwan K. et al. A study of hold-out and k-fold cross validation for accuracy of groundwater modeling in tidal lowland reclamation using extreme learning machine. In: Technology, Informatics, Management, Engineering, and Environment (TIME-E), 2014 2nd International Conference on. IEEE, 2014. p. 228-233.

HELMA, Christoph et al. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. Journal of chemical information and computer sciences, v. 44, n. 4, p. 1402-1411, 2004.

KELLY, Rebecca A. et al. Selecting among five common modelling approaches for integrated environmental assessment and management. Environmental modelling & software, v. 47, p. 159-181, 2013.

MARBOUTI, Farshid; DIEFES-DUX, Heidi A.; MADHAVAN, Krishna. Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education, v. 103, p. 1-15, 2016.

MARIA, Willian; DAMIANI, João Luccas; PEREIRA, Max. Rede Bayesiana para previsão de Evasão Escolar. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2016. p. 920.

ROVIRA, Sergi; PUERTAS, Eloi; IGUAL, Laura. Data-driven system to predict academic grades and dropout. PloS one, v. 12, n. 2, p. e0171207, 2017.

TEKIN, Ahmet. Early Prediction of Students' Grade Point Averages at Graduation: A Data Mining Approach. Eurasian Journal of Educational Research, v. 54, p. 207-226, 2014.