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

TÉCNICAS DE VALIDAÇÃO DE DADOS PARA SISTEMAS INTELIGENTES: UMA ABORDAGEM DO SOFTWARE SDBAYES

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

JEAN CARLOS TORRES MÜLLER
Universidade de Santa Cruz do Sul
jeanctm00@gmail.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

JÚLIA WEBER REUTER
Universidade de Santa Cruz do Sul
juwreuter@hotmail.com

RESUMO
Nesse artigo é abordado a validação de métricas de Mineração de dados, referentes a um software, denominado SDBayes, que foi desenvolvido em um projeto de pesquisa. O software faz a predição dos discente mais propensos a evadir ou permanecer em uma Instituição de Ensino Superior apresentando probabilidades de permanência e probabilidades de evasão, também utiliza Redes Bayesianas, que são métricas de classificação muito usadas para a área médica, pois simula muito bem o raciocínio humano. No entanto, as classificações feitas pelas Redes bayesianas nem sempre correspondem com a realidade do problema, com isso, foram abordadas, cinco técnicas de validação de dados, para estimar a real capacidade de predição do sistema desenvolvido. Os métodos usados foram: *F-measure, K-fold, Hold-out, Leave-one-out* e o *Receiver Operating Characteristics (ROC)*.

Palavras chave: Métodos de validação, Rede Bayesiana, Predição de Evasão, Discente.

1. INTRODUÇÃO

O poder de tomada decisão de gestores Universitários sofre com a falta de recursos, geralmente um coordenador de curso baseia sua tomada de decisão em seus antigos feitos, o que pode ser arriscado e muitas vezes deixa-lo em situações arriscadas.

Em uma versão anterior desse mesmo projeto de pesquisa, foi desenvolvido uma ferramenta capaz de auxiliar o gestor na tomada de decisão, possibilitando-o tomar decisões com base em fatos e não em hipóteses, denominado SDBayes. Essa ferramenta conta com métricas de mineração de dados, mais especificamente as redes bayesianas, então a ferramenta carrega os dados de histórico discente dos anos anteriores e aplica esses dados nessas métricas. O software gerado como resultado desse projeto de pesquisa tem seu objetivo principal informar a probabilidade de evasão de cada discente, juntamente com as variáveis que mais influenciam nessa para tal probabilidade. Porém, apesar dos bons resultados de validação obtidos, em média 75% de acerto, as métricas de validação dos resultados adotadas possuem falhas, muitas vezes errando a probabilidade de evasão de um discente, mostrando informações não condizentes com a realidade. Com isso, esse trabalho visa validar as redes Bayesianas desenvolvidas com cinco métodos de validação; *F-measure, K-fold, Hold-out, Leave-one-out* e o *Receiver Operating Characteristics (ROC).*

O artigo está organizado da seguinte maneira: na próxima seção, tem-se as referências bibliográficas, seguindo com a metodologia abordada, posteriormente os resultados, e por fim as conclusões.

2. FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta as principais métricas de validação da eficácia de um procedimento de mineração de dados, nessa seção são encontrados os seguintes métodos*: F-measure, K-fold, Hold-out, Leave-one-out* e o *Receiver Operating Characteristics*.

2.1 F-MEASURE

O *F-measure* é utilizado para situações em que se deseja ter apenas um resultado ao invés de dois para medir a performance. Por exemplo, ao invés de termos um resultado para precisão e outro para revocação, e interpretar cada um separadamente, junta-se estes dois resultados pela média ponderada da precisão e revocação, sendo possível interpretar apenas um resultado. A pontuação do *F-measure* chega a 1 como um bom resultado e 0 como um resultado ruim. Um valor alto de *F-measure* significa resultados de precisão e revocação balanceados.

F-measure é a média ponderada dos resultados de precisão e revocação. A fórmula de precisão é uma medida que mede a relevância dos resultados retornados, também pode ser chamada de Predição de valor Positivo.

O *F-Measure* utiliza de métricas derivadas da Matriz de confusão binária, onde tem-se os seguintes componentes: TP = Verdadeiro Positivo, TN = Verdadeiro Negativo, FP = Falso Positivo e FN = Falso Negativo Stehman et al. (1997), Pontius Jr et al. (2006). Com essas métricas, derivam-se as seguintes fórmulas para o desenvolvimento do *F-measure*:

$$TPV = \frac{TP}{TP+FP}$$ **Fórmula 1: Predição de valor Positivo**

$$TNV = \frac{TN}{TN+FN}$$ **Fórmula 2: Predição de valor Negativo**

$$TPR = \frac{TP}{TP+FN}$$ **Fórmula 3: Taxa de Verdadeiro Positivo**

$$TNR = \frac{TN}{TN+FP}$$ **Fórmula 4: Taxa de Verdadeiro Negativo**

Abaixo, a tabela 1 apresenta os dados que serão analisados para exemplificação do *F-measure*, onde a primeira coluna refere-se à situação real do discente, a segunda coluna refere-se ao resultado de métricas referentes ao software, e por fim a terceira coluna apresenta a classificação dos valores das colunas anteriores.

Valor real	Resultado do software	Classificação
Sim	Sim	Verdadeiro positivo
Sim	Sim	Verdadeiro positivo
Sim	Não	Falso negativo
Não	Não	Verdadeiro negativo
Não	Não	Verdadeiro negativo

Tabela 1: Classificação dos dados
Fonte: Autores, 2017.

Levando em consideração os cinco registros da tabela acima, o resultado final totalizado ficou como verdadeiros positivos = 2, falsos positivos = 0, verdadeiros negativos = 2 e falso negativo = 1. Considerando a fórmula 5 dos valores verdadeiros positivos: O resultado da fórmula da precisão levando em consideração os totais da tabela acima da coluna resultado final é:

$$TPV = \frac{2}{2+0} = 1$$ **Fórmula 5: Valores verdadeiros positivos**

Uma precisão alta (como o 1 da fórmula acima) representa um baixo número de falsos positivos, e uma revocação alta representa um baixo número de falsos negativos. As pontuações elevadas para ambos mostram que o classificador está retornando resultados precisos (alta precisão), além de retornar à maioria de todos os resultados positivos (revocação elevada).

A fórmula da revocação levando em consideração os totais da coluna resultado final:

$$TPR = \frac{2}{2+1} = 0{,}666$$ **Fórmula 6: Revocação sendo aplicada**

Um sistema com revocação elevado, mas baixa precisão retorna muitos resultados, mas a maioria dos resultados são incorretos quando comparados aos resultados de treinamento. Um sistema com alta precisão, mas baixa revocação é exatamente o oposto, retornando poucos resultados, mas a maioria dos resultados previstos são corretos quando comparados aos resultados de treinamento.

A medida que combina precisão e revocação é a média harmônica de precisão e revocação, a tradicional *F-measure* ou *F-score* balanceada:

$$F1 = 2 * \left(\frac{TPV*TPR}{TPV+TPR}\right)$$ **Fórmula 7: *F-Measure* para valores Verdadeiros**

$$F2 = 2 * \left(\frac{TNV*TNR}{TNV+TNR}\right)$$ **Fórmula 8: *F-Measure* para valores Falsos**

O motivo de ser utilizado a média harmônica ao invés de média aritmética para cálculo da F-measure é porque tende para um menor número de resultado da média. Com isso, minimiza o impacto de grandes *outliers* e maximiza o impacto de pequenos *outliers*, na *F-*

measure, portanto tende a privilegiar sistemas equilibrados Nadeau et al. (2007). Com os totais apresentados da tabela acima utilizando a fórmula da *F-measure:*

$$F1 = 2 * \left(\frac{1*0,666}{1+0,666}\right) = \frac{1,33}{1,66} = 0,8$$ **Fórmula 9: *F-Measure* sendo aplicada.**

Os resultados da precisão e revocação no contexto acima são mais intuitivos de interpretar do que *F-measure*, isso porque o mesmo é uma mistura desses dois resultados. O valor de resultado do *F-measure* é utilizado quando é necessário medir a performance a partir de um resultado apenas. Por exemplo, com um resultado alto da *F-measure* conclui-se que precisão e revocação estão igualmente balanceados, porém poderia ser feito ao invés de interpretar a *F-measure*, interpretar os resultados de precisão e revocação

2.2 K-FOLD

A validação cruzada *K-fold* é uma técnica computacional intensiva, que usa todas as amostras disponíveis como amostras de treinamento e teste Duchesne et al. (2005). Com isso, em relação a outros métodos de validação cruzada como *Hold-out* e *Leave-One-Out* consegue-se chegar a resultados mais precisos, muitas vezes superior ao *Leave-One-Out* que em muitos casos não é utilizado por exigir um desempenho maior de processamento de recursos computacionais.

Dado uma base de dados hipotética em que conste 100 registros, e definindo o k=10 a base de dados será dividido em 10 subconjuntos onde cada subconjunto terá 10 registros cada. Após a divisão em subconjuntos, será utilizado um subconjunto, para ser utilizado na validação do modelo e os conjuntos restantes são utilizados como treinamento. O processo de validação cruzada é então repetido K (10) vezes, de modo que cada um dos K subconjuntos sejam utilizados exatamente uma vez como teste para validação do modelo.

Por exemplo, dados 10 subconjuntos B1, B2... B10 o primeiro passo do *K-Fold* é utilizar B1 para teste e de B2 a B10 para treino. No segundo passo, B2 é utilizado para teste e todo o restante para treino, incluindo B1 que foi usado para teste no primeiro passo, no terceiro passo até o décimo será aplicada a mesma lógica sucessivamente. O resultado final da validação *K-Fold* é o desempenho médio do classificador nos K testes. O objetivo de repetir os testes diversas vezes é com o intuito de aumentar a confiabilidade da estimativa da precisão do classificador.

2.3 LEAVE-ONE-OUT

A validação *Leave-One-Out* ocorre da mesma maneira que o método *K-Fold* com a principal diferença, é de que o treinamento é realizado com n-1 dados e o teste com 1 dos registros somente. O método *Leave-One-Out* define o número de subconjuntos igual ao número de registros da base de dados. Então, se a base de dados tiver 100 registros dentro dela, serão definidos 100 subconjuntos cada um com 1 registro. Após a divisão dos subconjuntos o mesmo processo do *K-Fold* é realizado, utiliza-se o subconjunto B1 para teste e o restante para treinamento, no caso do exemplo seriam 99 subconjuntos para treinamento, assim sucessivamente.

2.4 HOLD-OUT

Na validação *Hold-out* o método assemelha-se com o *K-Fold* onde o k=2, porém com uma particularidade, a base de dados é dividida em duas partes, com isso uma das partes é usada para treino e a outra parte para teste, sem a alternação que ocorre com o *k-fold*. Este processo é realizado uma vez apenas, diferente do processo de *K-Fold* em que os dados são

divididos em K partes, e cada parte é usada tanto para treino como para teste, de tal forma que todas as partes passem por ambos os lados. Uma vantagem do modelo *hold-out* é que o tempo necessário para aprender o modelo é relativamente menor do que o tempo necessário para a aprendizagem do modelo usando a validação cruzada *k-fold* Yadav et al. (2016).

2.5 RECEIVER OPERATING CHARACTERISTICS

As curvas ROC (*Receiver Operating Characteristics*) têm sido usados na teoria da detecção de sinal para descrever o *tradeoff* entre taxas de sucesso e taxas de falsos alarmes de classificadores Fawcett (2006), na sequência, a área da saúde começou a usufruir dessas métricas, para estimar o acerto de patógenos, doenças e a fins Zweig et al. (1993), Metz et al. (1978), porém, o modelo de validação trabalha com diversos parâmetros de entrada, como dados discretos, e sua exibição gráfica é uma curva sobre um plano cartesiano, então, começou-se a usar também com o intuito de validar o acerto de predições feitas em técnicas de aprendizagem de máquina e Data Mining, que é o foco desse trabalho.

O ROC, usa como parâmetro de validação, dados referentes a tabela de confusão, que pode ser lido em: Stehman et al. (1997), Pontius Jr et al. (2006), onde os principais dados são referentes à resultados considerados Verdadeiros Positivos (TP), Falsos Positivos (FP), Verdadeiros Negativos (TN) e Falsos Negativos (FN). O valor de N é referente à soma dos falsos positivos com falsos negativos, e o valor de P é a soma dos TP com TN, com isso, derivam-se as fórmulas:

$$fp\ rate = \frac{FP}{N}$$ **Fórmula 10: Percentual de Falsos Positivos**

$$Sensitivity = \frac{TP}{P}$$ **Fórmula 11: Percentual de Verdadeiros Positivos**

$$Specificity = \frac{TN}{FP+TN} = 1 - fp\ rate$$ **Fórmula 12: referente ao restante dos Falsos Positivos para completar 1.**

$$Precision = \frac{TP}{TP+FP}$$ **Fórmula 13: Precisão**

$$Accuracy = \frac{TP+TN}{P+N}$$ **Fórmula 14: Acerto**

Com essas fórmulas, já pode-se fazer várias estimativas para determinar a precisão de acerto de um teste, no entanto, o ROC tem seu resultado de forma gráfica, onde tem-se um plano cartesiano, como a figura 1, onde o eixo Y é referente aos Verdadeiros Positivos, e o eixo X, aos Falsos Positivos, com isso, o ideal, é buscar-se o ponto (0,1), que se encontra na parte superior à esquerda, o que significa que não haveriam falsos positivos. No entanto, quando um ponto vai para a parte inferior a Diagonal secundária, temos que o resultado do teste é pior que um teste aleatório, porém, simplesmente alterando o sinal da saída, caso booleano, tem-se o acerto da investida.

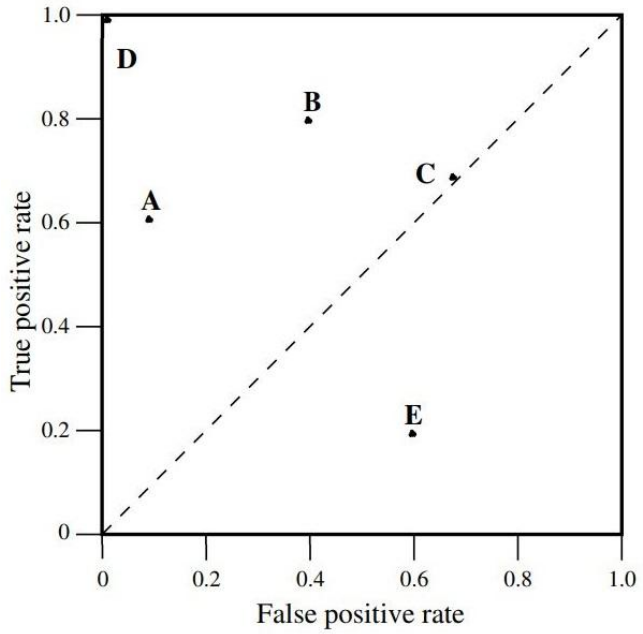

Figura 1: Gráfico ROC com 5 classificações discretas
Fonte: Fawcett (2006)

GERAÇÃO DE PONTOS

Para a geração dos pontos, e consequentemente da curva, é necessário estipular o percentual de escala no qual deseja marcar o ponto. Por exemplo, caso deseja-se identificar o ponto que se encontra em 90% da curva, usa-se a variável *Prevalence* com o valor de "0,9". Essa variável é o percentual de dados estimados, e usa valores entre 0 e 1, para calcular os pontos intermediários. Para isso, podem ser usadas as seguintes fórmulas derivadas do teorema Bayesiano:

$$PPV = \frac{Sensitivity * Prevalence}{(Sensitivity * Prevalence) + \big((1 - specificity) * (1 - prevalence)\big)}$$

Fórmula 15: Identificação do Y

$$NPV = \frac{specificity * (1 - prevalence)}{\big((1 - sensitivity) * prevalence\big) + \big(specificity * (1 - prevalence)\big)}$$

Fórmula 16: Identificação do X

Ambas as fórmulas 15 e 16 são usadas para determinar se os pontos (X, Y), mas no caso PPV para o Y e NPV para o X.

3. METODOLOGIA

Para chegar aos resultados dos métodos e posteriormente avaliá-los, foram feitas de duas formas, manualmente e via algoritmos. Os algoritmos se fizeram necessário em alguns dos métodos, pois a realização manualmente se tornaria inviável, tanto pelo número de registros, como pela complexidade do método. O projeto desenvolvido, conforme descrito brevemente na introdução, resultou no desenvolvimento de uma ferramenta capaz de prever a probabilidade de evasão dos discentes, considerando para isso, um conjunto de variáveis previamente definidos. Além disso o software *SDBayes*, usava em seu núcleo de previsão 3

redes bayesianas, cada uma com cerca de 700 casos, com isso, alguns métodos como o *Leave-one-out*, não foram viáveis testá-los manualmente, com isso, desenvolveu-se alguns algoritmos para auxiliar no processo de validação das redes bayesianas do *SDBayes*.

Os métodos que foram feitos manualmente são: *F-measure*, *K-Fold* e *Hold-out*. O ROC, foi utilizado uma ferramenta de terceiros, o MedCalc Schoonjans et al. (1995).

Abaixo, tem-se a metodologia de execução dos métodos de validação.

3.1 F-MEASURE

O F-measure possui dois resultados sobre uma mesma amostragem, uma em relação aos Verdadeiros Positivos, e outra em relação aos verdadeiros negativos. No caso atual, associam-se aos Alunos que efetivamente evadiram, e aos alunos que efetivamente não evadiram em comparação com a realidade e correlacionando esses resultados aos resultados do SDBayes. Com isso foram geradas três tabelas de Confusão, onde estão contidos os dados que serão analisados pelo f-measure.

Valor Testado	Valor Original	
	Positivos	Negativos
Positivos	208	58
Negativos	39	600

Figura 2: Tabela de Confusão referente ao Curso de Ciência da Computação
Fonte: Autores, 2017.

Como pode ser notado na Figura 2, que é referente ao curso de Ciência da Computação, tem-se os valores da diagonal principal com a maior parte dos dados, isso é interessante, pois significa que a quantidade de outliers é baixa. Com esses dados, são feitas as validações do f-measure, pois aqui estão contidos os Verdadeiros Positivos, Verdadeiros Negativos, Falsos Positivos e Falsos Negativos.

Valor Testado	Valor Original	
	Positivos	Negativos
Positivos	339	44
Negativos	82	333

Figura 3: Tabela de Confusão referente ao Curso de Administração
Fonte: Autores, 2017.

Na Figura 3, tem-se os dados referentes ao curso de Administração, onde a maior parte dos dados se encontra na diagonal principal, isso é interessante, pois significa que a quantidade de outliers é baixa. Com esses dados, são feitas as validações do f-measure, pois aqui estão contidos os Verdadeiros Positivos, Verdadeiros Negativos, Falsos Positivos e Falsos Negativos.

Valor Testado	Valor Original	
	Positivos	Negativos
Positivos	293	21
Negativos	47	297

Figura 4: Tabela de Confusão referente ao Curso de Engenharia de Produção
Fonte: Autores, 2017.

Como pode ser notado na Figura 4, que é referente ao curso de Engenharia de Produção, tem-se os valores da diagonal principal com a maior parte dos dados, isso é interessante, pois significa que a quantidade de outliers é baixa. Com esses dados, são feitas as validações do f-measure, pois aqui estão contidos os Verdadeiros Positivos, Verdadeiros Negativos, Falsos Positivos e Falsos Negativos.

3.2 K-FOLD

Para o método *K-Fold* com o k=10, inicialmente foi feita uma randomização nas linhas, para tornar a validação o mais precisa possível. Sequencialmente, dividiu-se em dez partes iguais, onde havia cerca de 91 casos para cada bloco, para então dar início as entradas no programa SDBayes: *a system to Predict Student Drop-Out* que é responsável por realizar a predição. O parâmetro k=10 foi escolhido porque segundo Witten et al. (2005), testes extensivos em vários conjuntos de dados, com diferentes técnicas de aprendizagem, mostraram que 10 é o número certo para obter a melhor estimativa de erro, e também há algumas evidências teóricas que apoiam isso. Todavia ainda há bastante discussão quanto ao melhor parâmetro k, porém na prática o k=10 se tornou o método padrão em termos práticos Witten et al. (2005). Os 905 registros da base de dados foram divididos em 10 partes e o processo foi passar para o software os dados de treinamento de teste. Após executar no programa são retornadas três colunas, uma indicando a situação real do discente, outra mostrando se o aluno tem probabilidade de permanecer no curso, e por fim a última coluna mostrando a probabilidade de evasão. Esse processo foi realizado 10 vezes. No final, foi obtido a média dos resultados de cada iteração do método.

3.3 LEAVE-ONE-OUT

Para conseguir chegar ao resultado do método *Leave-One-Out* foi necessário desenvolver um software, o qual usa alguns trechos de código do sistema SDBayes. O treinamento realizou-se com 904 casos enquanto o teste foi feito com 1 dos registros, dessa maneira o número das divisões realizadas foi o total de registros da base de dados, ou seja, 905 divisões. Após a divisão dos dados em subconjuntos o registro que ficou sozinho serviu de teste e os outros 904 registros serviram de teste. Assim foi feito sucessivamente até as 905 linhas, cada linha isoladamente, ter servido para teste.

3.4 HOLD-OUT

Para chegar no resultado do método *Hold-out*, a base foi dividida em duas partes, sendo uma parte dos dados referente a teste e a outra parte dos dados para treino. A diferença do método *Hold-out* para o *K-Fold* é que foi necessário realizar apenas uma vez todo o processo. O processo se resumiu em dividir a base 50% dos dados para cada lado, sendo que em uma

das partes ficaram os dados de teste e a outra parte os dados de treino. Após, utilizou-se uma parte desses dados para treinar a rede bayesiana, e outra parte para testar a precisão de acerto do conjunto de dados usado para treino

3.5 RECEIVER OPERATING CHARACTERISTICS

A metodologia adotada para a geração dos resultados do ROC, foi baseada em aplicar os resultados do software, que por sinal retorna valores discretos, ao lado o estado de predição acertada ou com erro. Para isso foram usados alguns algoritmos da ferramenta MedCalc. Primeiramente, testou-se os dados diretamente com a rede bayesiana, na sequência, tinha-se a descrição do estado real, e ao lado o valor que a rede acusa de evasão, por exemplo: "aluno realmente evadido" e "probabilidade de evadir 80%", o que é um Verdadeiro positivo, com isso foi criado uma nomenclatura, para todos os dados e adaptados para o padrão da tabela de Confusão, para então introduzir essa entrada, nos algoritmos do MedCalc.

4. RESULTADOS

Nesta seção do artigo serão apresentados os resultados obtidos com os métodos de validações descritos no referencial e na metodologia. É apresentado em cada gráfico o percentual de precisão de acerto referente a predição de cada método. A predição é em relação a permanência e evasão cada curso, ou seja, é o percentual que cada método acertou da predição com base nos dados reais para o curso em foco.

4.1 F-MEASURE

Abaixo, no gráfico 1, tem-se os resultados obtidos a partir da aplicação das métricas referentes ao f-measure em relação ao curso de Ciência da Computação. Pode-se notar que os resultados foram bem interessantes em relação ao acerto de predição em relação aos alunos que evadem do curso, segundo os testes, 92,53% dos alunos efetivamente foram previstos com assertividade, já para os casos onde os alunos permanecem no curso, 82,32% foram previstos com acerto.

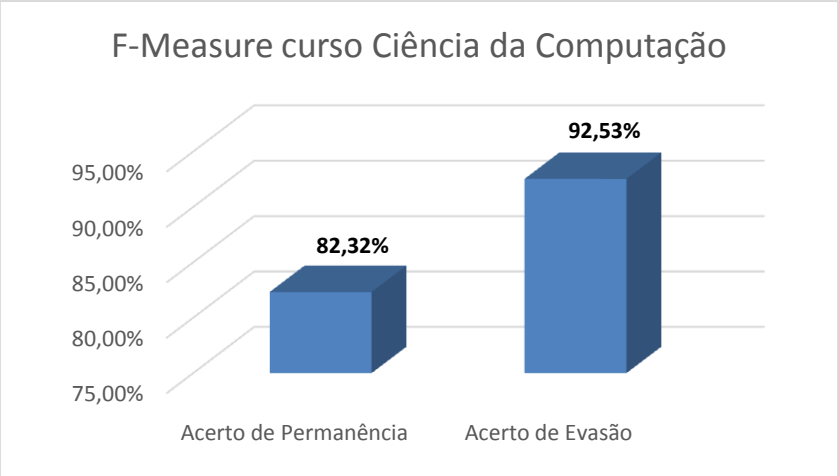

Gráfico 1: Resultado do F-Measure para o curso de Ciência da Computação
Fonte: Autores, 2017.

Pode-se notar no gráfico 2, que os resultados do curso de Administração não foram tão marcantes como os do curso de Ciência da Computação, no entanto, um acerto mais equilibrado, onde o acerto de predição em relação aos alunos que evadem do curso é de 84,1%, e para os casos de acerto por predição de conclusão do curso é de 84,34%.

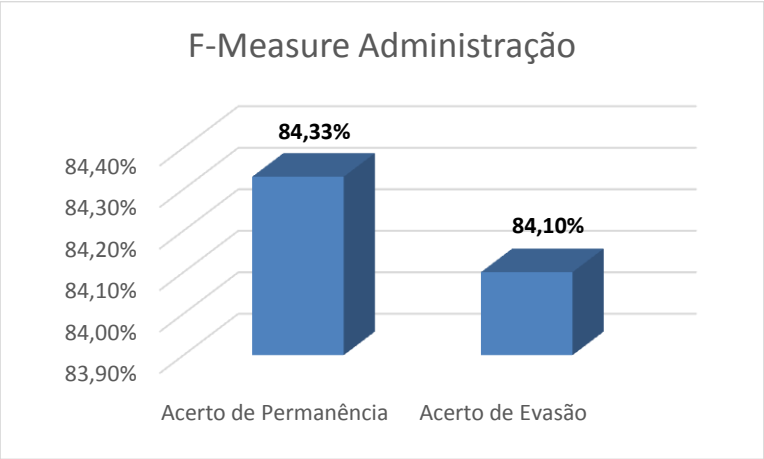

Gráfico 2: Resultado do F-Measure para o curso de Administração
Fonte: Autores, 2017.

No gráfico 3, os resultados do curso de Engenharia de Produção seguiram a tendência do curso de Administração, com um acerto mais equilibrado, no entanto, mais alto, onde o acerto de predição em relação aos alunos que evadem do curso é de 89,73%, e para os casos de acerto por predição de conclusão do curso é de 89,6%.

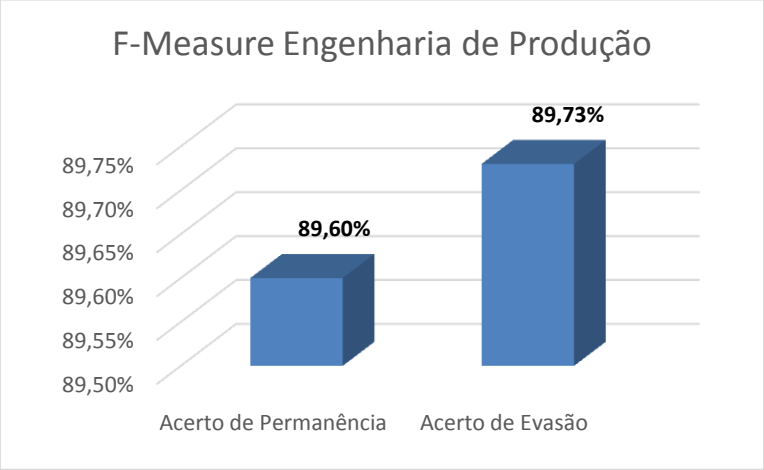

Gráfico 3: Resultado do F-Measure para o curso de Engenharia de Produção
Fonte: Autores, 2017.

4.2 K-FOLD

No gráfico 4 demonstra-se o resultado referente ao método *K-Fold* aplicado nos dados do curso de Ciência da Computação. Pode-se notar que nessa metodologia, os resultados referentes ao acerto de permanência, foram um tanto quanto duvidosos, pois em alguns testes, os pontos chegaram a baixar de 50%, no entanto o acerto médio de permanência foi de 60,98%, com desvio padrão de 11,53%. Já a média de acerto para a predição de evasão é de 90,34%, com um desvio padrão de 2,31%. Além disso, foi feita a média aritmética dos acertos de predição, e conseguiu-se uma média de 82,01%, com um desvio padrão de 3,47%.

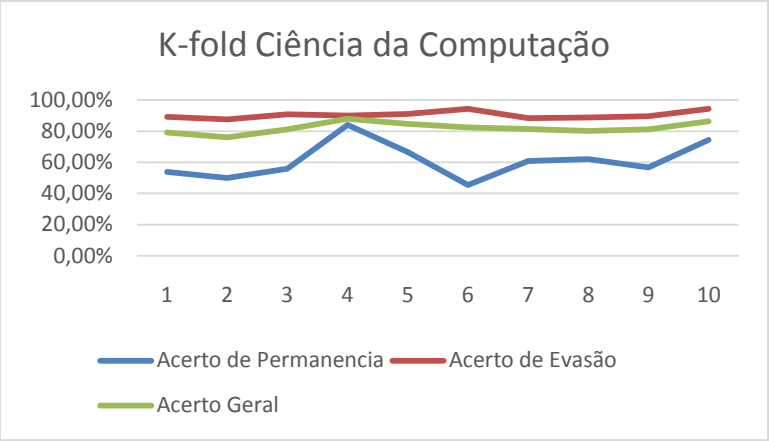

Gráfico 4: Resultados K-Fold do curso de Ciência da Computação
Fonte: Autores, 2017.

No gráfico 5 demonstra-se o resultado referente ao método *K-Fold* aplicado nos dados do curso de Administração. Pode-se notar que as certificações em relação ao acerto em relação a esse curso foram mais estáveis que em comparação com o curso de Ciência da Computação. O acerto médio de permanência foi de 79,14%, com desvio padrão de 7,15%. Já a média de acerto para a predição de evasão é de 73,02%, com um desvio padrão de 7,6%. Além disso, foi feita a média aritmética dos acertos de predição, e conseguiu-se uma média de 75,79%, com um desvio padrão de 6,1%.

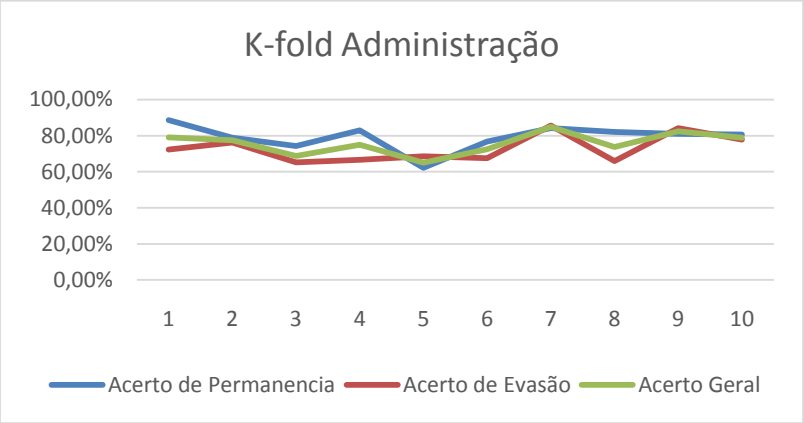

Gráfico 5: Resultados K-Fold do curso de Administração
Fonte: Autores, 2017.

No gráfico 6 demonstra-se o resultado referente ao método *K-Fold* aplicado nos dados do curso de Engenharia de Produção. Pode-se notar que as oscilações ocorridas no curso de Ciência da Computação, voltaram a ocorrer no Curso de Engenharia de Produção. O acerto médio de permanência foi de 56,89%, com desvio padrão de 22,49%. Já a média de acerto para a predição de evasão é de 69,39%%, com um desvio padrão de 16,26%. Além disso, foi feita a média aritmética dos acertos de predição, e conseguiu-se uma média de 63,94%, com um desvio padrão de 17,48%.

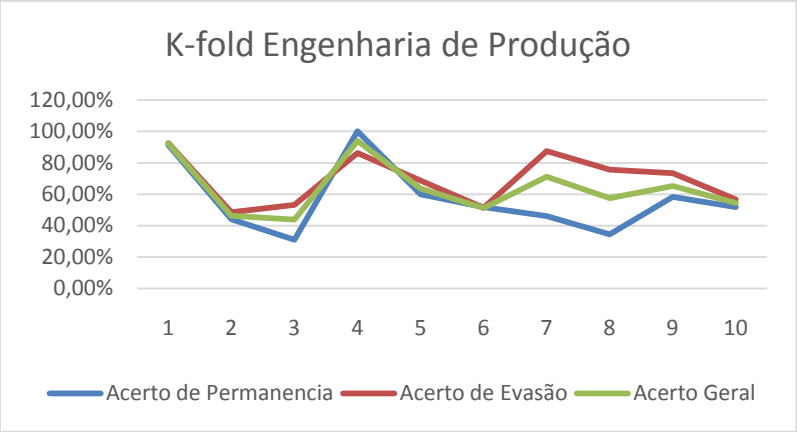

Gráfico 6: Resultados K-Fold do curso de Engenharia de Produção
Fonte: Autores, 2017.

4.3 LEAVE-ONE-OUT

No gráfico 7 demonstra-se o resultado referente ao método *Leave-one-out* aplicado nos dados do curso de Ciência da Computação. Pode-se notar que os resultados de acerto em relação à permanencia ficaram bem abaixo dos acertos em relação à evasão dos discentes. O acerto médio de permanência foi de 60%, com desvio padrão de 33,84%. Já a média de acerto para a predição de evasão é de 91,55%, com um desvio padrão de 21,63%. Além disso, foi feita a média aritmética dos acertos de predição, e conseguiu-se uma média de 82,3%, com um desvio padrão de 28,81%.

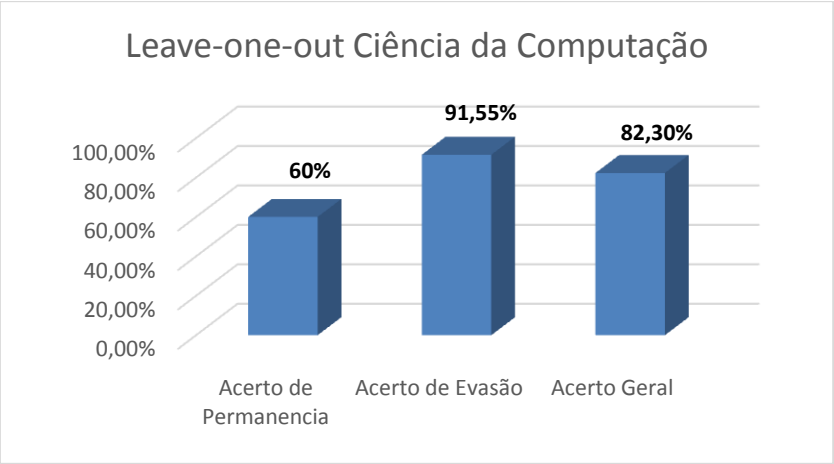

Gráfico 7: Resultados Leave-One-Out do curso de Ciência da Computação
Fonte: Autores, 2017.

No gráfico 8 demonstra-se o resultado referente ao método *Leave-one-out* aplicado nos dados do curso de Administração. Pode-se notar que diferente do curso de Computação, os resultados de acerto em relação à permanencia ficaram acima dos acertos em relação à evasão dos discentes. O acerto médio de permanência foi de 80,42%, com desvio padrão de 24,07%. Já a média de acerto para a predição de evasão é de 73,98%, com um desvio padrão de 32,44%. Além disso, foi feita a média aritmética dos acertos de predição, e conseguiu-se uma média de 77,07%, com um desvio padrão de 28,72%.

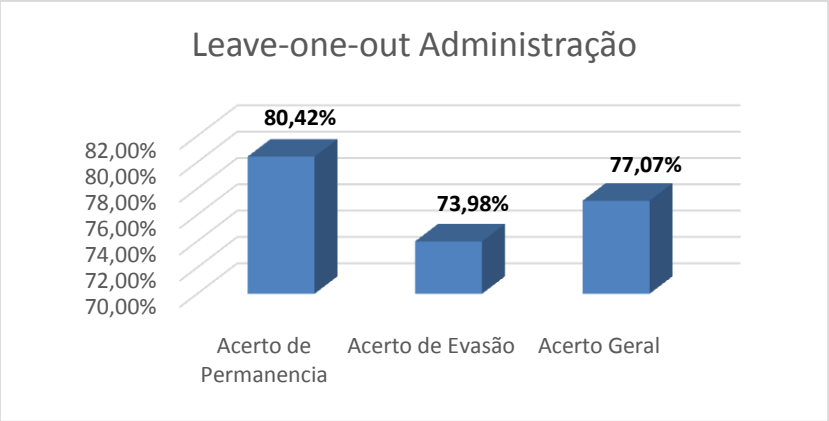

Gráfico 8: Resultados Leave-One-Out do curso de Administração
Fonte: Autores, 2017.

No gráfico 9 demonstra-se o resultado referente ao método *Leave-one-out* aplicado nos dados do curso de Engenharia de Produção. Pode-se notar que esse curso apresentou resultados mais próximos, no entanto o acerto de permanência, foi superior ao acerto de Evasão. O acerto médio de permanência foi de 64,33%, com desvio padrão de 17,80%. Já a média de acerto para a predição de evasão é de 62,21%, com um desvio padrão de 18,62%. Além disso, foi feita a média aritmética dos acertos de predição, e conseguiu-se uma média de 63,22%, com um desvio padrão de 18,28%.

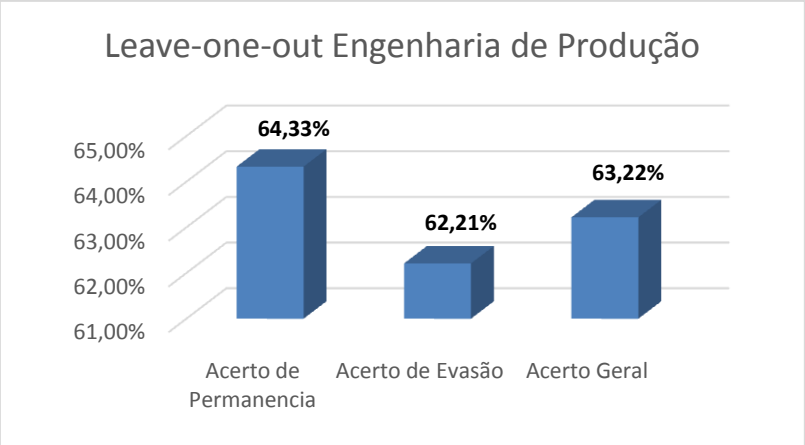

Gráfico 9: Resultados Leave-One-Out curso de Engenharia de Produção
Fonte: Autores, 2017.

4.4 HOLD-OUT

No gráfico 10, é possível notar os resultados em relação à métrica de teste do *Hold-out*, em relação ao Curso de Ciência da Computação. Observa-se que o Acerto de permanência ficou bem abaixo de acerto de evasão, com os seguintes resultados: Acerto de Permanência 61,06%, Acerto de Evasão 90,68% e além disso foi feito o acerto geral, dos alunos que evadem e permanecem na instituição, que foi de 78,32%.

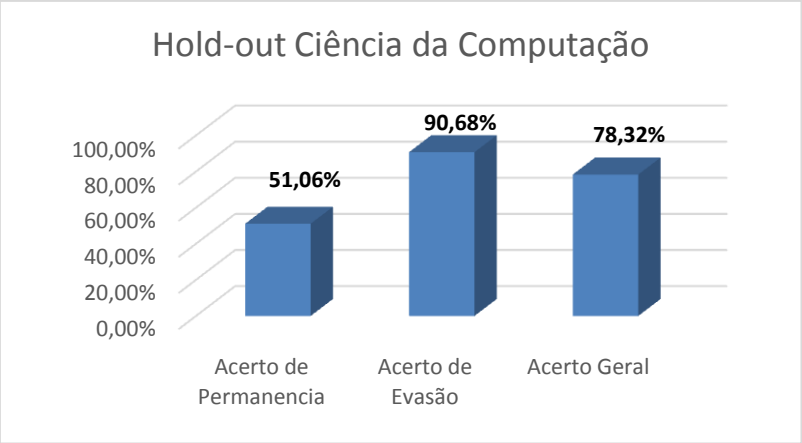

Gráfico 10: Resultados Hold-Out do curso de Ciência da Computação
Fonte: Autores, 2017.

No gráfico 11, é possível notar os resultados em relação à métrica de teste do *Hold-out*, em relação ao Curso de Administração. Observa-se que diferente do curso de Ciência da Computação, o Acerto de permanência ficou acima do acerto de evasão, com os seguintes resultados: Acerto de Permanência 80,1%, Acerto de Evasão 70,67% e além disso foi feito o acerto geral, dos alunos que evadem e permanecem na instituição, que foi de 75,19%.

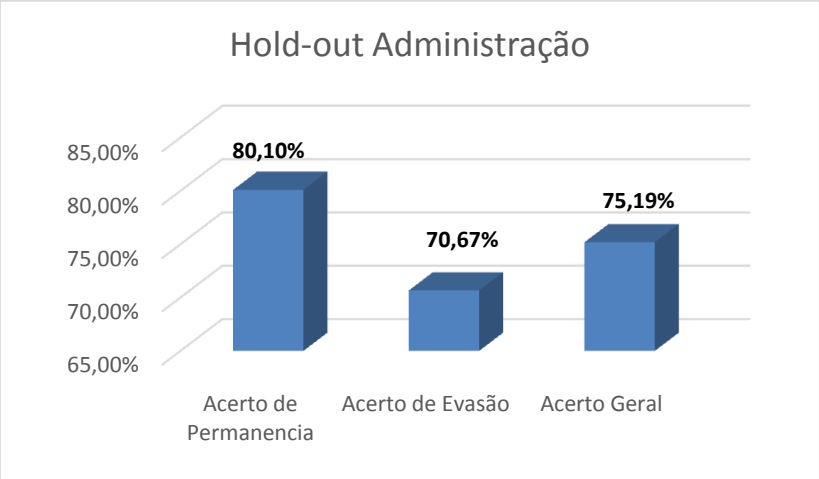

Gráfico 11: Resultados Hold-Out do curso de Administração
Fonte: Autores, 2017.

No gráfico 12, é possível notar os resultados em relação à métrica de teste do *Hold-out*, em relação ao Curso de Engenharia de Produção. Observa-se que o Acerto de permanência ficou acima de acerto de evasão, lembrando os resultados em relação ao curso de Administração, com os seguintes resultados: Acerto de Permanência 88,54%, Acerto de Evasão 76,16% e além disso foi feito o acerto geral, dos alunos que evadem e permanecem na instituição, que foi de 82,07%.

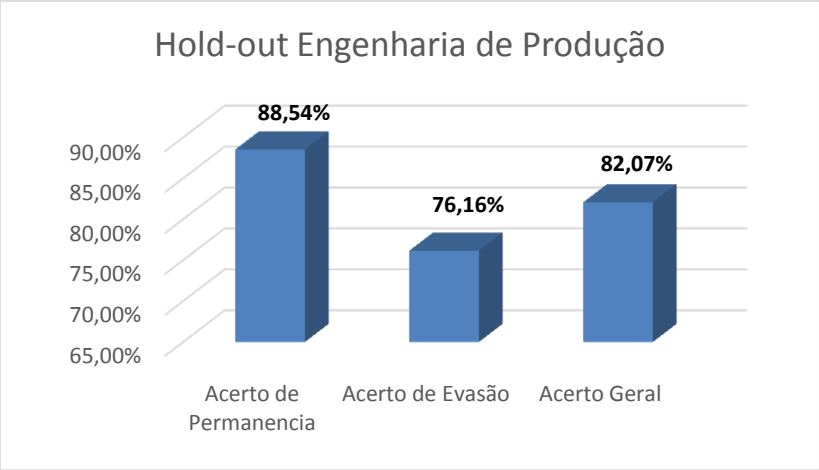

Gráfico 12: Resultados Hold-out do curso de Engenharia de Produção
Fonte: Autores, 2017.

4.5 RECEIVER OPERATING CHARACTERISTICS

Abaixo, tem-se os resultados dos testes com as três redes bayesianas:

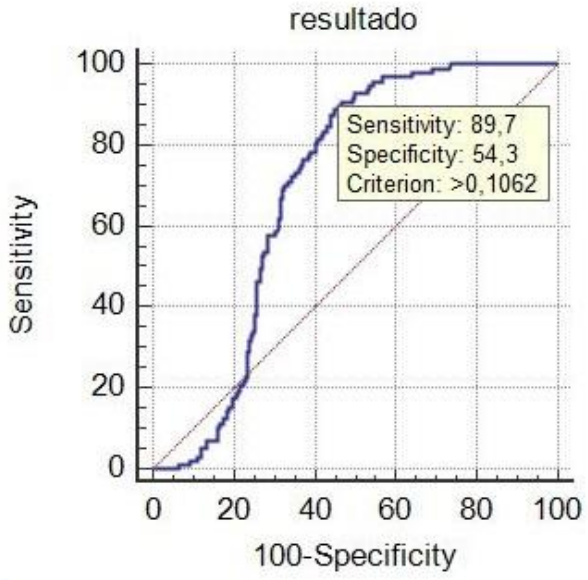

Figura 5: Resultados de acerto da rede do curso de Ciência de Computação
Fonte: Autores, 2017.

O curso de Ciência da Computação, apresentou valores duvidosos, em relação a testes como o *K-fold*, no entanto no ROC, os resultados foram muito bons, visto que quase toda a curva se encontra sobre a diagonal secundária, no entanto alguns os pontos iniciais ficaram abaixo da diagonal, o que é considerado ruim.

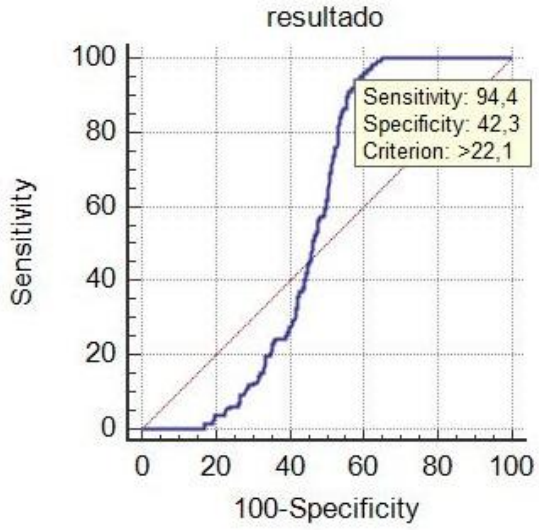

Figura 6: Resultados de acerto da rede do curso de Administração
Fonte: Autores, 2017.

O curso de Administração, apresentou valores estáveis, em relação a testes como o *K-fold*, no entanto no ROC, os resultados foram aceitáveis, visto que em média metade da cursa está situada acima da diagonal secundária, e metade abaixo.

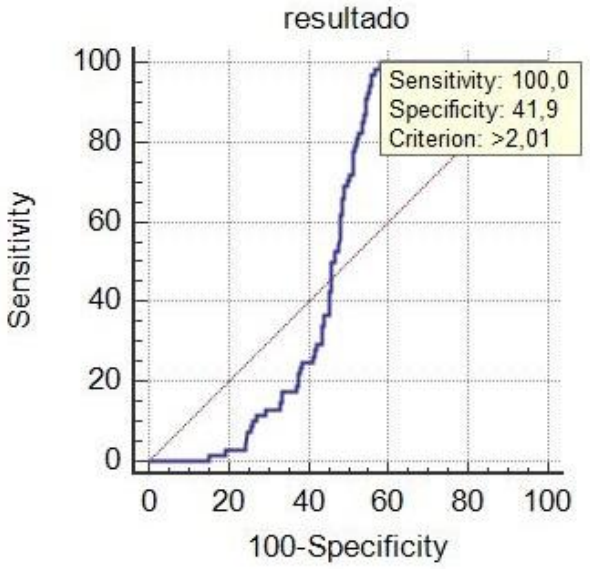

Figura 7: Resultados de acerto da rede de Engenharia de Produção
Fonte: Autores, 2017.

Por fim, o curso de Engenharia de Produção, apresentou valores duvidosos, em relação a testes como o *K-fold*, no entanto no ROC, os resultados foram assemelhados com o teste no curso de Administração, onde praticamente metade da curva situa-se sobre a diagonal secundária, metade abaixo.

5. CONCLUSÃO

A necessidade de se ter resultados condizentes com a realidade do problema, é decorrente na área de mineração de dados, por isso, é fundamental validar as técnicas de Inteligência Artificial, Data mining e a fins. Resultados que não correspondem com o problema, podem

acarretar em sérias dificuldades, como informações decisivas erradas para a tomada de decisão por exemplo.

Nesse artigo foi avaliado o acerto dos resultados gerados a partir do Software SDBayes, que possui três redes bayesianas, de três cursos de graduação, Ciência da Computação, Administração e Engenharia de Produção, onde cada rede foi testada em cinco métodos de validação de dados, *F-measure, K-fold, Hold-out, Leave-one-out* e o *Receiver Operating Characteristics (ROC)*.

Os acertos de previsão demostraram que os métodos *K-fold, Hold-out* e *Leave-one-out*, são muito semelhantes, tendo resultados bem próximos, onde o acerto médio dos resultados em relação ao acerto de previsão em relação a permanência de alunos no curso de Computação foi aproximadamente 57%, para Administração foi de aproximadamente 80% e para o curso de Engenharia de produção foi de aproximadamente 70%. Já o acerto de previsão em relação a evasão de alunos no curso de Computação foi aproximadamente 90%, para Administração foi de aproximadamente 73% e para o curso de Engenharia de produção foi de aproximadamente 70%. E o acerto geral de previsão em relação ao curso de Computação foi aproximadamente 81%, para Administração foi de aproximadamente 76% e para o curso de Engenharia de produção foi de aproximadamente 70%.

Além disso o método F-Measure apontou que as redes possuem uma ótima taxa de acerto, onde todos os resultados superaram 80% e o método ROC, tem-se os resultados graficamente, no entanto esse método é sem dúvida um dos mais importantes, visto que, diferente dos outros métodos, ele analisa discretamente os resultados, por isso, toda a informação que está sobre a diagonal secundária é considerada correta, com isso, o curso de Ciência da Computação tem um acerto médio de 80%, já os cursos de Administração e de Engenharia de Produção, um acerto médio de 70%.

Por fim, o software SDBayes possui margem para melhorar seu desempenho, então para futuros projetos e/ou linhas de pesquisa, pode-se apontar como um norte, a busca de taxas de acerto superiores à 90%.

REFERÊNCIAS

DUCHESNE, Pierre; RÉMILLARD, Bruno (Ed.). Statistical modeling and analysis for complex data problems. **Springer Science & Business Media**, 2005.

FAWCETT, Tom. An introduction to ROC analysis. **Pattern recognition letters**, v. 27, n. 8, p. 861-874, 2006.

METZ, Charles E. Basic principles of ROC analysis. In: **Seminars in nuclear medicine**. WB Saunders, 1978. p. 283-298.

NADEAU, David; SEKINE, Satoshi. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, v. 30, n. 1, p. 3-26, 2007.

PONTIUS JR, Robert Gilmore; CHEUK, Mang Lung. A generalized cross- tabulation matrix to compare soft- classified maps at multiple resolutions. **International Journal of Geographical Information Science**, v. 20, n. 1, p. 1-30, 2006.

SCHOONJANS, FRANK et al. MedCalc: a new computer program for medical statistics. **Computer methods and programs in biomedicine**, v. 48, n. 3, p. 257-262, 1995.

STEHMAN, Stephen V. Selecting and interpreting measures of thematic classification accuracy. **Remote sensing of Environment**, v. 62, n. 1, p. 77-89, 1997.

WITTEN, Ian H. et al. **Data Mining: Practical machine learning tools and techniques**. Morgan Kaufmann, 2016.

YADAV, Sanjay; SHUKLA, Sanyam. Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. In: **Advanced Computing (IACC), 2016 IEEE 6th International Conference on**. IEEE, 2016. p. 78-83.

ZWEIG, Mark H.; CAMPBELL, Gregory. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. **Clinical chemistry**, v. 39, n. 4, p. 561-577, 1993.