Revista dos Mestrados Profissionais

MODELO PREDITIVO PARA CLASSIFICAÇÃO DE ESTUDANTES COM RETENÇÃO UNIVERSITÁRIA

PREDICTIVE MODEL TO STUDENTS CLASSIFICATION WITH UNIVERSITY RETENTION

Othon Luiz Teixeira de Oliveira[1], Marcus Souza[2], Orlando Bandrao[3], Eric Sales[4]

[1] Universidade de Pernambuco - otluiz@gmail.com

[2] Universidade de Pernambuco - marcosd3souza@gmail.com

[3] Universidade de Pernambuco - orlabrandao@gmail.com

[4] Universidade de Pernambuco - ericsales0@gmail.com

INFORMAÇÕES

Palavras-chave:

Sistema de suporte a decisão
Retenção e evasão de estudantes
Mineração de dados
Regressão logística.

RESUMO

A realidade encontrada nos processos acadêmicos nas universidades passa pela busca por alternativas à problemática da evasão e da retenção de alunos. Para dar suporte às decisões dos gestores dessas universidades, técnicas de mineração de dados vêm sendo utilizadas. O presente trabalho propõe um modelo de predição de desligamento ao final do segundo semestre de uma Universidade Pública Federal em Pernambuco. Para o experimento foi utilizada a metodologia CRISP-DM. Para extração de conhecimento, foram utilizados os algoritmos de regressão logística e árvore de decisão. Os dados referem-se ao período entre 2002 a 2008, possui 180.357 registros e 108 variáveis a partir dos dados originais. O algoritmo com melhor comportamento para árvore de decisão foi o ―C4.5‖, e para regressão logística foi o ―Forward Stepwise‖, que obtiveram uma taxa de acerto de 55% e 58%, respectivamente. O desempenho do modelo alcançou MAX_KS=0,66 e AUC_ROC=0,80.

INFORMATION

Keywords:

Decision support system
Student retention and evasion
Data mining
Logistic regression.

ABSTRACT

A reality found among the academic processes in universities is to search for alternatives to solve the problems of dropout and student retention. To support these managers higher education decision making, knowledge discovery technique has been gaining momentum in several areas has been helping. This paper proposes a exclusion forecasting model at the end of the second half. For the experiment was used KNIME tool, based on CRISP-DM methodology. For extracting knowledge logistic regression and decision tree algorithms were used. The data refer to the period 2002-2008, having 180.357 records and 108 variables from the original data. The algorithm with the best performance for decision tree was the "C4.5", and logistic regression was the "Forward Stepwise" who obtained a hit rate of 55% and 58%, respectively. The estimated risk model reached MAX_KS = 0.47 and AUC_ ROC=0.75.

1 INTRODUÇÃO

A retenção e a evasão de estudantes é um sério problema que afeta tanto as universidades públicas quanto as universidades e faculdades privadas. A retenção é caracterizada como a permanência do estudante além do tempo máximo de integralização do curso, enquanto que a evasão é definida como o abandono do aluno do curso ou da universidade (1). Sendo considerado como um marco teórico, o modelo desenvolvido por Tinto (3) é um dos estudos mais antigos que aborda a evasão em universidades americanas, no qual ele sugere que a evasão ocorre pela falta de integração do estudante com o ambiente acadêmico e/ou social da universidade.

No Brasil, as autoridades do sistema educacional começaram a tratar do assunto retenção/evasão universitária em 1989, quando o Ministério da Educação e Cultura (MEC) criou um grupo de estudos formado pelos reitores das universidades federais para debater o assunto. Como resultado do trabalho deste grupo de estudo, foi lançado em 1996 um relatório que reunia um conjunto de dados sobre o desempenho das universidades públicas descrevendo o cenário da realidade brasileira sobre os índices de diplomação, retenção e evasão dos estudantes de graduação, naquela época (1).

Com o objetivo de ampliar o acesso e permanência na educação superior, para os cursos de graduação, o MEC criou o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais – REUNI (6), em 2007. O REUNI estabeleceu como meta global alcançar, ao final de cinco anos, a taxa de conclusão média dos cursos de graduação presenciais em 90%, bem como, a relação de 18 alunos por professor.

A UFPE aderiu ao REUNI criando o Projeto REUNI/UFPE 2007 (7) de reestruturação e expansão universitária. Naquele ano a UFPE possuía 70 cursos de graduação distribuídos por 12 centros acadêmicos em três campi, totalizando 25.425 alunos na graduação. Dentre as duas metas globais estabelecidas pelo MEC, a UFPE já havia alcançado uma até aquele ano, a relação de 18,2 alunos por professor. As diretrizes do REUNI foram estruturadas em seis dimensões, cada uma com seus próprios conjuntos específicos de aspectos, que devem ser abordados nos projetos das universidades aderentes ao programa.

Para o contexto deste trabalho interessou-nos a dimensão que trata da redução das taxas de evasão e retenção. Após um diagnóstico inicial verificou-se que os cursos com as maiores taxas de desligamento naquela época, foram das áreas de Engenharia, Ciências Sociais Aplicadas e Ciências Humanas, que apresentaram em média 7,21% de taxa de evasão no período de 2006 a 2007. Quanto à retenção, foi identificado que 58,4% dos alunos estavam atrasados com relação à conclusão de seus cursos. Dessa forma, a UFPE estipulou como meta para essa dimensão a redução gradativa da taxa de evasão para 2% em 2012, bem como, a redução da taxa de retenção para 20% em 2012.

Esse artigo está organizado em mais seis seções, além da Introdução. A seção 2 descreve a modelagem e as técnicas utilizadas. A seção 3 descreve o entendimento do negócio, dos dados objetos desta pesquisa e a análise descritiva dos dados. A Seção 4 descreve o pré-processamento e a transformação desses dados. A Seção 5 aborda os cenários de simulação e a interpretação do grupo sobre os resultados obtidos. Finalizando, a Seção 6 apresenta uma conclusão para este trabalho e indica futuras pesquisas que possam ser conduzidas a partir dele.

2. PROCESSOS E TÉCNICAS DE MINERAÇÃO DE DADOS

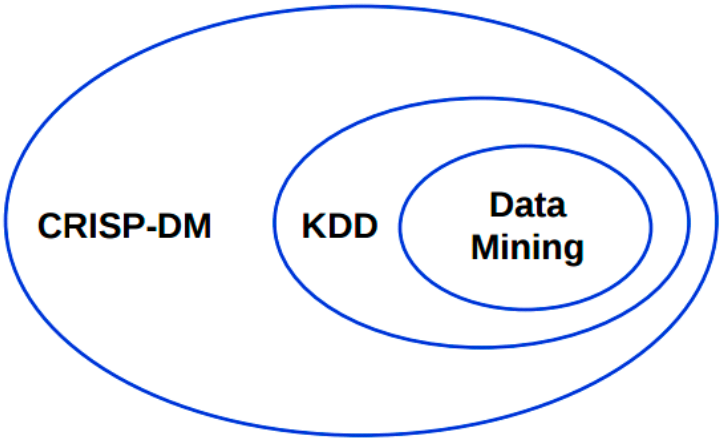
A inteligência artificial é uma área que pode, através de algoritmos eficientes, propor soluções adaptativas para dar conta das mais diversas necessidades humanas, sobretudo aquelas relacionadas ao contexto social, como gestão de tempo, de logística relacionada ao acesso dos estudantes ao conhecimento disponível na universidade, dentre outros. A Mineração de Dados (MD) vai buscar na Inteligência Artificial algoritmos para a descoberta de padrões e automatizar tarefas de investigação dos dados. Essa automatização também conhecida como "Machine Learning", aplica-se a quase todos os caminhos na descoberta do conhecimento oferecida pela MD, porém para aplicar as técnicas de MD há um modelo, com etapas bem definidas, o CRISP – DM.

O CRISP – DM (Cross Industry Standard Process for Data Mining) é um modelo que é descrito em forma de processos para mineração de dados que descreve como especialistas neste campo aplicam as técnicas de mineração de dados para obter os melhores resultados. A aplicação do CRISP – DM (9) é guiada desda o nível mais genérico até o nível mais especializado, sendo explicado normalmente em quatro dimensões:

- Domínio da aplicação – a área específica onde os projetos de mineração de dados acontecem;
- Tipo de problema – descreve as classes específicas do objetivo do projeto de mineração de dados;
- Aspecto técnico – cobrem as questões específicas acerca dos desafios usualmente encontrados durante o processo de mineração de dados;
- Ferramentas e técnicas – dimensão específica que cada ferramenta/técnica de mineração de dados é aplicada durante o projeto.

Um projeto de Mineração de Dados a ser desenvolvido deverá estar contextualizado nestas quatro dimensões. Na aplicação dos processos o modelo CRISP – DM (Figura 2.3) descreve as fases onde estas dimensões se encaixam. A aplicação das técnicas de mineração de dados identificam padrões ocultos nos dados, inacessíveis pelas técnicas tradicionais, como por exemplo, consultas a banco de dados, técnicas estatísticas, dentre outras. Além disto, possibilita analisar um grande número de variáveis simultaneamente, o que não acontece com o cérebro humano. A análise deste processo permite extrair novos conhecimentos a partir dos dados, que é tratado na literatura como KDD – Knowledge Discovery Database (Figura 2.2). Fayyad, P., Piatetsky-Shapiro, U., & Smith,G., (22) destacam a natureza interdisciplinar da KDD que contempla a intersecção de campos de pesquisa tais como a Aprendizagem de Máquina (Machine Learning), Reconhecimento de Padrões, I. A., Estatística, Computação de Alto Desempenho e outros, propõe que o objetivo principal é extrair um conhecimento de alto nível a partir de dados de baixo nível, num contexto de grandes bases de dados. O CRISP – DM, por sua vez, engloba todos estes elementos como pode ser visto na figura 2.2.

Figura 2.2. Domínio das técnicas aplicadas a mineração de dados



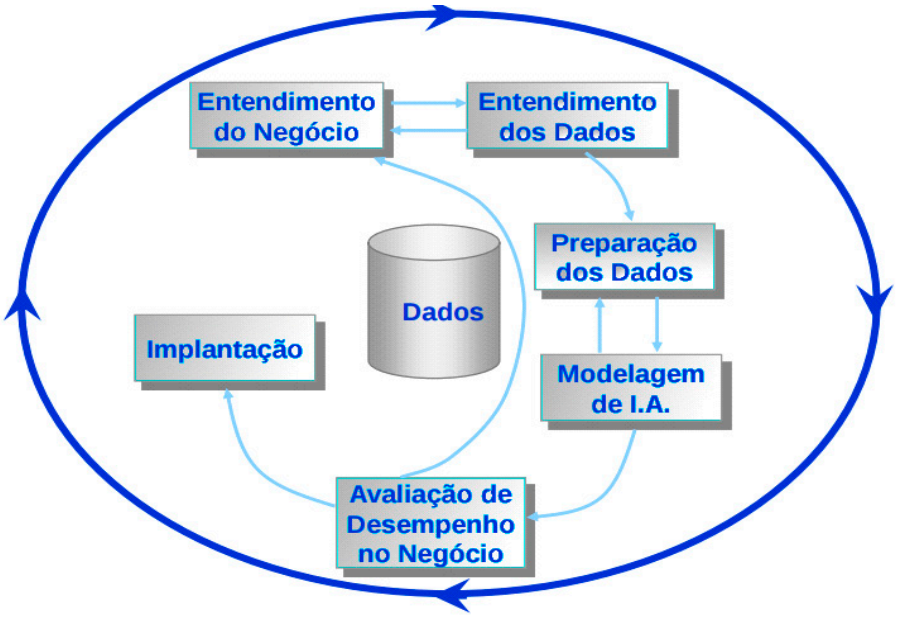
Fonte: Neurotech – 2012

Neste contexto a aplicação da das técnicas de Algoritmos de aprendizagem de máquina (Machine Learning) se dá de dentro para fora do modelo descrito acima, (Figura 2.2).

O Cross Industry Standard for Data Mining (CRISP – DM) (9) portanto é um modelo que é implementado em forma de um processo recursivo, em que cada etapa deve ser revista até que o modelo apresente resultados satisfatórios, senão, deve-se, fazer ajustes e "rodar" o ciclo novamente. O analista de dados é o profissional que acompanha e executa o modelo de MD aplicando o CRISP – DM. A figura 2.3, abaixo, descreve o Ciclo de Vida do CRISP – DM:

O modelo CRISP-DM contempla seis fases para um projeto de mineração de dados, sendo assim determina-se um ciclo de vida compreendido para cada uma dessas fases:

Figura 2.3. Modelo CRISP – DM



Fonte: CRISP – DM 1.0

A primeira fase, conhecida como Entendimento do negócio, ou "fase de entendimento dos objetivos é dos requerimentos sob a perspectiva do negócio" segundo CHAMPAN (9) é uma fase crucial da mineração. A fase Entendimento dos dados caracteriza-se pelo exame acurado dos dados, procurando identificar a sua qualidade. Dados ausentes "missing data" – são comuns em base de dados não estruturadas, configurando-se como um problema a ser considerado. A terceira fase, Preparação dos dados diz respeito sobre a construção final do conjunto dos dados, cada conjunto de dados é "explicado" por um atributo, para selecionar quais dados serão mais relevantes, para variáveis numéricas calcula-se o coeficiente de correlação entre os atributos (variáveis). Outra forma de qualificar os dados é calculando a quantidade de informação que cada atributo possui. A máxima entropia de cada atributo pode fornecer informações sobre a qualidade da variável quando esta estabelece ganho de informação (24), vide a equação da Entropia (1).

$$H_x = \sum_{\forall x \in X}^{\square} P(x)\log_2 P(x) \qquad = \qquad (1)$$

Onde H_x é a medida de entropia, x um atributo do conjunto X de variáveis. H_x pode variar entre 0 e 1, quando o valor da entropia tende para 1 significa que maior ganho de informação. A entropia condicional, formalizada na equação (2), é a entropia restante dos atributos de Y no valor y quando o atributo X é dado como x (25):

$$H_{Y \vee X} = -\sum_{\forall x \in X}^{\square} P(x) \sum_{\forall x \in Y}^{\square} P(y \vee x)\log_2 P(y \vee x) \qquad (2)$$

Na quarta fase do CRISP-DM, Modelagem de I.A. a tecnologia deve ser escolhida de forma criteriosa, baseada na experiência do analista de dados. Na fase cinco, Avaliação de desempenho um ou mais modelos devem ter sido construídos e testados de forma que o modelo esteja adequado aos objetivos do negócio. A sexta e última fase, Implementação caracteriza-se pela conclusão do processo, pode ser necessário retomar todo o ciclo até que o modelo esteja adequado as necessidades especificas determinadas previamente.

No domínio da UFPE, estudos realizados por especialistas na área de I.A. corroboram para com o que está descrito na 4ª fase do CRISP-DM, dos critérios de escolha da tecnologia.

Os mais recentes trabalhos realizados na área sobre retenção universitária destacamos o trabalho de análise e tratamento da evasão/retenção universitária, Campelo e Lins (4) estabelecem como domínio o curso de graduação em engenharia de produção da Universidade Federal de Pernambuco (UFPE). Utilizaram os dados dos alunos que ingressaram no curso no período de 2000 a 2006 através do sistema de gestão acadêmica da UFPE e do COVEST, antigo vestibular realizado pela UFPE. Para construção do modelo e análise do problema de evasão e retenção, utilizaram técnicas de mineração de dados com o uso de clusters, rotinas de processamento analítico OLAP e o software WEKA, identificando seis tipos de clusters diferentes de alunos com perfis de evasão/retenção. Auxiliados por professores e alunos da UFPE, identificaram as principais causas do problema e propuseram várias estratégias de ação visando minimizá-las. Chegaram à conclusão de que cada curso de graduação possui características e perfis próprios, "não sendo correto generalizar o estudo da evasão/retenção, seja para a mesma instituição de ensino ou até para o mesmo departamento" (4).

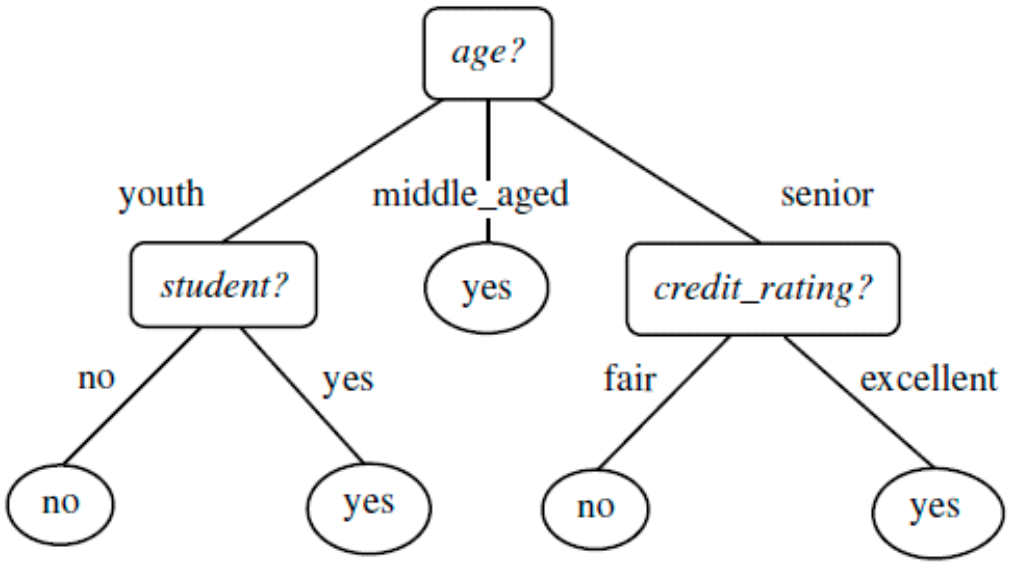
Outro trabalho que abordou o tema evasão/retenção na UFPE, foi o de Hadautho e Adeodato (5). O estudo foi realizado a partir da base de dados de alunos da UFPE, compreendendo o período entre 1998 a 2008, dos cursos de Direito, Ciências Econômicas, Engenharia Civil,

Pedagogia, Medicina e Línguas. Os dados originais da base continham 415.327 registros de 11.036 estudantes dos diversos cursos de graduação. Foi utilizada regressão logística e regras de indução, deixando os dados após o pré-processamento na granularidade estudante com 5.793 registros. Posteriormente, houve transformações com a inclusão de mais 21 novas variáveis somadas às originais. Após a aplicação dos algoritmos, observou-se que a base de dados em estudo possuía 47% de estudantes retidos. Para avaliar o risco da solução proposta, Hadautho e Adeodato (5) utilizaram Kolmogorov-Smirnov (KS-2) e a área sob a curva ROC (AUC_ROC), tendo o desempenho de Max_KS = 0,51 e AUC_ROC = 0,84.

As **árvores de decisão** aglutinam essas características anteriormente descritas, por isso são empregadas no universo das Engenharias, Ciência da Computação, Medicina, Economia e em sistemas de suporte à decisão, como diagnósticos de doenças, investigação de fraudes, dentre outros. De maneira sintética, uma árvore de decisão tem como entrada um conjunto de atributos e como saída uma dimensão. Os atributos podem, ainda, ser classificados de duas maneiras: discretos e contínuos. Em virtude dos atributos de entrada, tem-se, como resultado um função de valores discretos – aprendizagem de classificação – ou de valores contínuos – aprendizagem de regressão (15). A decisão gerada aparece em função de uma sequência de testes executados, estando cada um deles relacionados a um nó da árvore. As ramificações que ocorrem dos testes são resultados encontrados a partir da realização do experimento.

O exemplo a seguir (figura 1.3) ilustra uma árvore de decisão simples, onde se vê seu nó-raiz e suas ramificações.

Figura 1.3 – Árvore de decisão



Fonte: Han, J. and Kamber, M.

Uma árvore de decisão obedece à regra "se-então", de maneira que, parte-se do nó (se) até as folhas (então). O conhecimento é representado em cada nó, que apresenta pelo menos duas saídas ou ramificações possíveis, que pode, ou não, coverter-se em um novo nó, relacionado a um novo nível. Embora seja um algoritmo simples e de fácil interpretação, uma das mais importantes questões a ser considerada é como propor as regras de forma adequada e relevante para a geração da árvore. É necessário identificar o melhor atributo, que será responsável por criar o nó de decisão. As ligações entre os nós representam os valores possíveis do teste do nó

superior e as folhas indicam a classe (categoria) a qual o registro pertence, segundo Camilo, C. O.; Silva, J. C. (23)

A origem das árvores de decisão, data da segunda metade do século XX. A literatura aponta que as árvores de decisão foram propostas por Ross Quinlan, pesquisador australiano, no final da década de 70 e início dos anos 80, sendo o ano de 1983 aquele em que foi apresentado o primeiro algoritmo para geração de árvores de decisão: O ID3 (Iterative Dichotomiser), utilizado até hoje e considerado um dos mais importantes.

Todavia, Hssina; Merbouha & Mohamed, sugerem que autores no campo da estatística na década de 60, com Sonquist e Morgan, utilizaram árvores de decisão para predição e explicação, com o algoritmo AID (Automatic Interation Detection), sem saber das pesquisas de Quinlan. A partir desse modelo houve uma expansão para problemas de classificação e discriminação, cuja abordagem teria culminado no CART(Classification and Regression Tree).

Arvores de decisão (e outras técnicas) são construídas a partir de **variáveis** encontradas nos dados a serem pesquisados. Uma variável dependente pode ser explicada pelo conjunto de variáveis independentes, elas podem ser resultados de medições em um experimento, ou de dados coletados pelo pesquisador. Na equação (3) "x" é a variável independente e "Y" a variável dependente. Podemos entender que a variável "Y" pode ser explicada pelos valores de "x".

$$Y_x = 2 + 5x - x^2 \quad (3)$$

Se extrapolarmos a equação (3) para um modelo de predição, em que dados provém de uma base de dados de desempenho acadêmico de estudantes universitários, os dados podem ser explicados conforme um modelo de distribuição normal. Porém, para nossa pesquisa os dados foram transformados em variáveis binárias (assumem somente valores "0" ou "1", ou verdadeiro ou falso).

Modelos de regressão linear são para problemas de natureza contínua. A regressão logística é semelhante à regressão linear, contudo a variável dependente não é contínua, é discreta ou categórica (10). A aplicação da regressão logística foi utilizada com sucesso na decisão de oferta de crédito nos anos seguintes ao fim da 2a guerra mundial para tomar decisão de oferecer crédito a terceiros (11).

A regressão logística analisa dados binários de uma distribuição binária que têm a fórmula segundo a equação (4).

$$Y_i = B(p_i, n_i), i = 1,2,3,\dots m, \quad (4)$$

Regressão Logística é definida como um logaritmo. Após algumas transformações à equação (4), aplicando-se logaritmos, podemos chegar a fórmula da regressão logística definida na equação (5)

$$log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$\text{onde } \pi(x) \text{ é: } \pi(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}}$$

e $x_1, x_2, \ldots, x_n$ são as variáveis a serem exploradas. (5)

A regressão logística é uma técnica estatística algoritmo de fácil implementação e relativa rapidez, contudo seu resultado carece de boa interpretação por humanos, necessitando de outra ferramenta para tal. Para auxiliar na interpretação do resultado desta pesquisa utilizamos as árvores de decisão e para explicar o modelo a regressão logística.

Para validar o modelos adotado utilizou-se métricas amplamente difundidas para classificação binária. Para esta construção, a matriz de contingência (ou matriz de confusão) classifica as probabilidades como:

Verdadeiro positivo, Falso positivo, Falso negativo e Verdadeiro negativo, respectivamente (True Positive – TP, False Positive – FP, False Negative – FN e True Negative – TN), também conhecido como matriz de confusão descrita na tabela 1 e 2:

Tabela 1. Matriz de confusão

	Predito		
Real	TP	FN	Positive: POS
	FP	TN	Negative: NEG
	PP	PN	

Frequência absoluta

Tabela 2. Matriz modelo de confusão

	Y	$\bar{Y}$	
X	P(X,Y)	P(X, $\bar{Y}$)	P(X)
$\bar{X}$	P($\bar{X}$,Y)	P($\bar{X}$, $\bar{Y}$)	P($\bar{X}$)
	P(Y)	P($\bar{Y}$)	

Probabilidade condicional

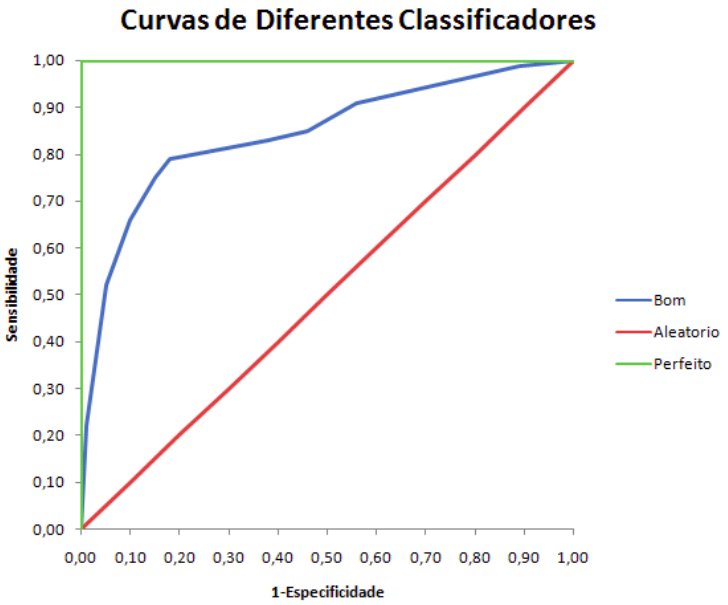
Fonte: os autores

A matriz da tabela 1 sintetiza a matriz da tabela 2, portanto, são equivalentes. De acordo

com as probabilidades condicionais temos: P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X). Então a taxa de verdadeiros positivos é P(Y|X) e a probabilidade de falsos alarmes ou taxa de falsos positivos será P(Y|), a barra sobrescrita em “”‘(ou “Y”) é a negação

Para analisar a qualidade dos resultados encontrados o método gráfico é eficiente e de rápida interpretação. Para detecção e avaliação da qualidade de sinais conhecido como curva ROC (Receiver Operating Characteristic) (20), foi um método criado e desenvolvido na década de 50 do século XX para detecção da qualidade de transmissão de sinal em um canal com ruído (20). A curva ROC é construída cruzando-se a taxa dos verdadeiros positivos (tpr = P(Y|X)) com a taxa dos falsos positivos (fpr = P(Y|X)). Esta curva é capaz de medir a qualidade dos classificadores utilizados nos modelos preditivos.

Fig. 1.4 – Curva ROC



Fonte:http://crsouza.com/2009/07/13/analise-de-poder-discriminativo-atraves-de-curvas-roc/

3. ENTENDIMENTO DO NEGÓCIO E DOS DADOS

A UFPE possui três campi nas cidades; Recife, Caruaru e Vitória de Santo Antão. São 12 centros acadêmicos totalizando 100 cursos de graduação regular presencial, dentre outros como na modalidade à distância e pós-graduação, com 29.502 estudantes na graduação no segundo semestre de 2014. (Fonte: http://www.ufpe.br). Os dados utilizados neste estudo foram extraídos do Sistema de Informação e Gerenciamento Acadêmico da UFPE (SIG@). A base de dados disponibilizada para este estudo possuía originalmente 24 planilhas totalizando 3.388.309 registros de dados históricos e 230 variáveis. Para esta pesquisa foi considerado o período entre 2002 e 2015, de todos os cursos de graduação da UFPE, exceto medicina, por ser o curso que apresenta índices irrisórios de retenção/evasão universitária. A tabela a seguir é uma amostragem da estrutura de dados original.

Tabela 3. Estrutura de dados original

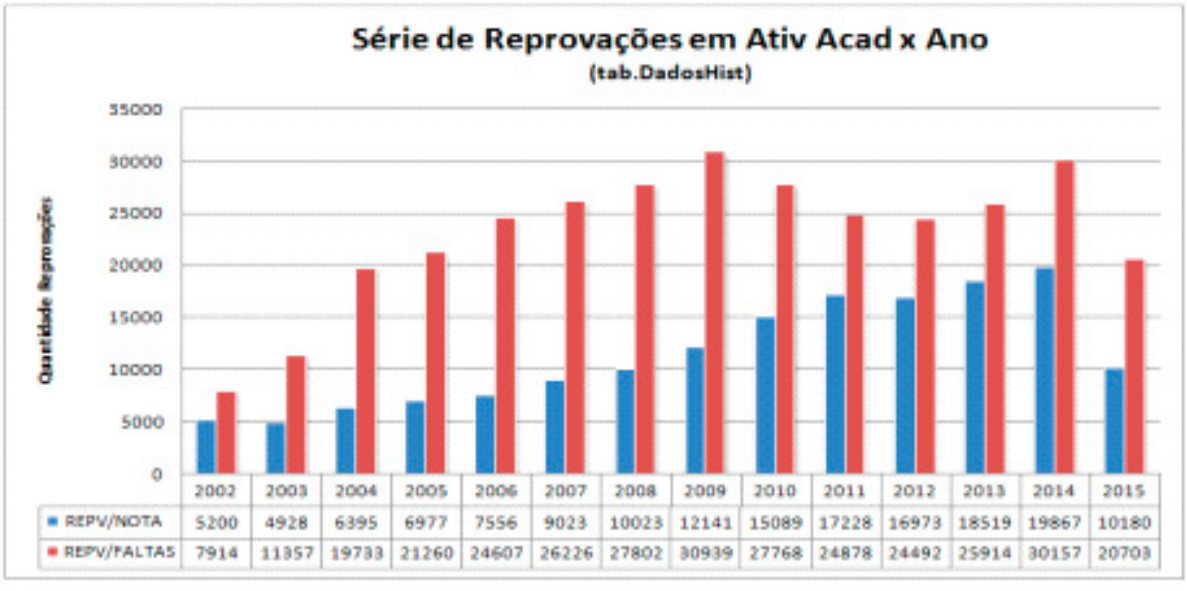
Atributos	Descrição
Matrícula	Identifica o aluno
Sexo	Gênero do aluno
Estado Civil	Estado Civil do aluno
Data de Nascimento	Data de Nascimento do aluno
Curso	Nome do curso do aluno
Carga horária	Carga horária total do curso
Entrada	Ano e semestre que o aluno iniciou o curso
Nota ENEM	Nota obtida no Exame Nacional do Ensino Médio
Cota	Tipo de cota adquirida pelo aluno
Ano de Conc. EM	Ano de conclusão do ensino médio
Código da Disciplina	Código identificador da disciplina
Disciplina	Nome da disciplina
Status da Disciplina	Disciplina sendo cursada pelo aluno no semestre recomendado
Semestre	Semestre em que o aluno se encontra matriculado
Situação	Situação do aluno na disciplina (aprovado, reprovado, evadido,..)
Nota	Nota registrada para o aluno na disciplina

Fonte: os autores

3.1 Análise Descritiva dos Dados

A análise descritiva foi realizada a partir dos valores apresentados nos dados originais. O gráfico abaixo faz uma comparação da quantidade de reprovações por notas e por falta em atividades acadêmicas nas quais o aluno é matriculado a cada ano. Um aluno pode estar vinculado a uma ou a várias atividades acadêmicas no ano. Portanto, o gráfico não se refere à quantidade de reprovações por alunos e, sim, por atividade acadêmicas.

Fig. 1. Série de reprovações em atividades acadêmicas

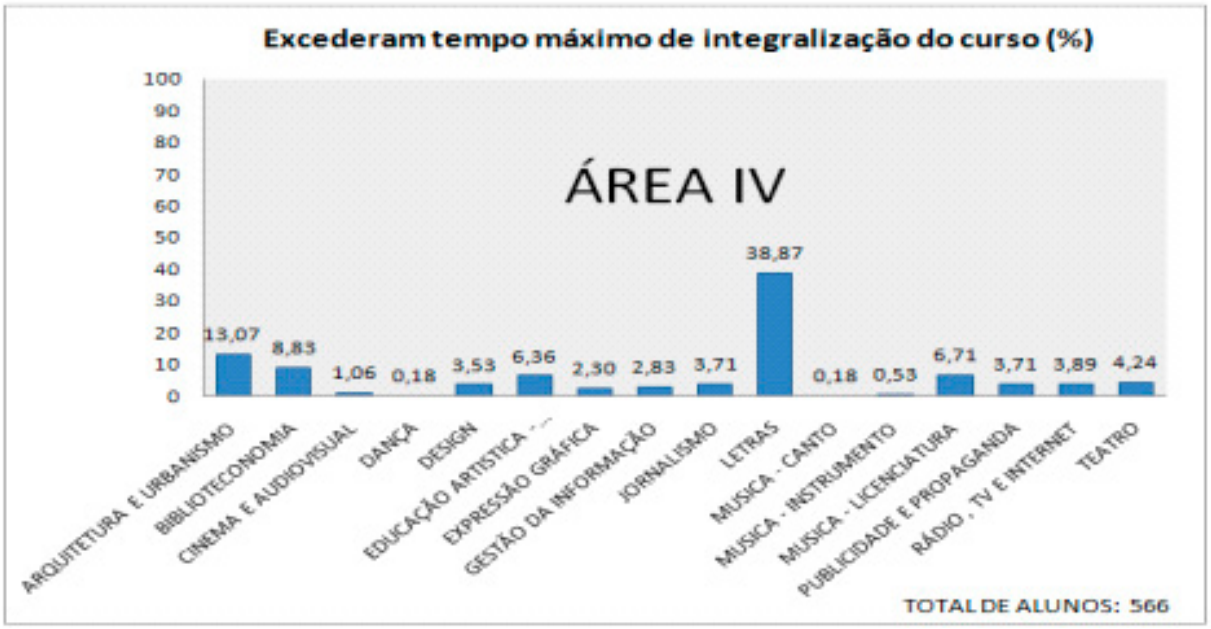


Fonte: os autores

O próximo gráfico apresenta o percentual de alunos que atingiram o critério de recusa

de matrícula por terem excedido o tempo máximo de integralização do curso ao qual estão matriculados, na área acadêmica IV da UFPE.

Fig. 2. Percentual dos alunos que excederam o tempo máximo de integralização do curso na área acadêmica IV



Fonte: os autores

4. PREPROCESSAMENTO E TRANSFORMAÇÃO DOS DADOS

A etapa de preprocessamento é fundamental para o processo de descoberta de conhecimento em mineração de dados. No mundo real os dados chegam para o especialista em mineração com algum tipo de ―sujeira‖ que são; dados incompletos, errados, ausentes, desatualizados e inconsistentes, precisam ser tratados ou ―limpos‖ como é conhecido este processo e em seguida reorganizados.

Foi realizada a junção das 24 planilhas numa grande tabela contendo aproximadamente 3,5 milhões de registros de dados históricos com 230 atributos, seguindo os seguintes critérios:

- Seleção dos atributos (subconjuntos das 230 variáveis de entrada):
- Limpeza nos dados (completude, veracidade, integridade, outliers;)
- Discretização (variáveis transformadas em atributos categóricos não contínuos);
- Binarização (variável alvo do tipo 0 ou 1, sim ou não, verdadeiro ou falso);
- Construção de novas variáveis (totais, médias, taxas);
- Transformação de variáveis (normalização dos dados, numa mesma escala).

O custo computacional tornou-se muito elevado para o processamento dessa quantidade de informações pelos algoritmos de mineração de dados, nos computadores comuns, disponíveis para esta pesquisa. Dessa forma, considerou-se uma seleção/redução na dimensionalidade dos dados para que, de alguma maneira, sob alguns critérios, não prejudique a eficiência dos métodos de mineração de dados.

4.1 Critérios de Exclusão

Alguns critérios de exclusão foram utilizados na amostra original. Efetuou-se remoções com base nos seguintes critérios:

- Discentes de curso EAD: Por ser de natureza diferente dos cursos presenciais, os cursos EAD foram excluídos da amostra;
- Ingressantes extra vestibular: Mantiveram-se apenas os alunos que ingressaram pelo vestibular, os demais foram removidos da base;
- Curso de Medicina: Este curso possui um índice de evasão muito reduzido. A taxa no período entre 2000 e 2014 foi de 1,6%. Por ser muito distinto dos demais, os seus discentes não fizeram parte do escopo; (13)
- Ingresso antes de 2002: Apesar de ter obtido dados cadastrais dos alunos a partir do ano 2000, o registro de dados comportamentais só existe a partir de 2002. Então, os anos anteriores foram excluídos da amostra;
- Vínculos diferenciados: Alunos estavam registrados como em cooperação internacional, mobilidade e disciplinas isoladas foram retirados do escopo;
- Situações acadêmicas diferenciadas: Discentes que não cursaram regularmente os dois primeiros períodos foram removidos. Por exemplo, os que realizaram trancamento ou matrícula vínculo nos semestres em questão;
- Recusa de matrícula: Aqueles que atingiram os critérios de recusa de matrícula antes do momento da decisão (final do período 2) também foram excluídos da amostra;
- Tempo para conclusão do curso: Para garantir que todos os discentes da amostra tiveram tempo de concluir o curso, foi estipulado um período letivo limite de ingresso para cada curso. Para isso, tomando como base 2015.1 foram diminuídos o número de semestres de cada curso mais quatro períodos.

4.2. Crítérios de Seleção

A seleção aplicada aos dados originais foi a redução de dimensionalidade, que consiste na eliminação de atributos, deixando apenas os atributos que são estatisticamente relevantes para o problema em questão. Cabe ressaltar que a realização desta atividade deve, também, ser feita com o acompanhamento do especialista no negócio.

Algumas variáveis foram retiradas do modelo por apresentarem altas taxa de “missings values”. Outras foram removidas por se tratarem de variáveis a posteriori, como data de expedição de diploma, ano e semestre de conclusão do curso e data de expedição de diploma. Ao final da atividade de seleção/redução dos dados originais o data set ficou com 27504 registros no grão aluno. Destes 7.857 atingiram os critérios de jubilamento. A tabela a seguir exprime esse resultado também em percentuais, para cada uma das áreas acadêmicas da UFPE:

Tabela 4: Percentual de jubilados por área acadêmica

Área Acadêmica	Alunos	% de Jubilados
Área I	11244	28%
Área II	5233	43%
Área III	5242	15%
Área IV	3738	29%
UFPE – Vitória	479	9%
UFPE – Agreste	1568	31%

Fonte: os autores

4.3. TRANSFORMAÇÃO DOS DADOS

A atividade de transformação dos dados consiste na criação, consolidação, transformação, normalização de atributos com a finalidade de adequar o data set para execução dos métodos de mineração de dados escolhidos, nesta etapa foram geradas novas variáveis para representarem conceitos relacionados com o problema. Como os dados comportamentais se encontravam na granularidade "período letivo / componente curricular / discente" foi necessário realizar a transformação para a granularidade aluno, gerando a variável alvo Jubilado e demais variáveis.

A variável alvo visa identificar os discentes que atingiram os critérios de recusa de matrícula estabelecidos pela Universidade. Essa regra também é conhecida como jubilamento. Todos os discentes que atingirem algum dos seguintes critérios terá sua matrícula recusada definitivamente (14):

I. Prazo máximo: esgotar o prazo máximo do curso;

II. Reprovações em componente: obter 4 reprovações, independente de tipo e p e - ríodo, em uma mesma componente curricular;

III. Reprovação total: obter reprovações em todos as componentes curriculares do período letivo por 2 vezes;

IV. Coeficiente: obter 2 vezes o Coeficiente de Rendimento Escolar inferior a 3

Tabela 3. Percentual de alunos com critério de jubilamento

Ano admissão	Deficit
2002	50%
2003	47,82%
2004	46%
2005	40%
2006	40,74%
2007	31,7%
2008	33,07%
2009	33,33%

Fonte: os autores

5. CENÁRIOS DE SIMULAÇÃO, RESULTADOS E INTERPRETAÇÃO

Os dados desta pesquisa, bem como a aplicação das técnicas descritas anteriormente foram separados em amostras de treinamento e de testes, para validação dos modelos preditivos, na proporção de 75% – 25%. Contudo devido ao espaço reduzido esta informação foi omitida.

Também foi utilizada na "framework" Weka, a validação cruzada, conhecida como "cross-validation" com 10 pastas, mas foram semelhantes aos encontrados acima e pelos mesmos motivos referidos acima os resultados omitidos.

A Regressão Logística aplicada nesta pesquisa foi para um problema de classificação binário, onde a granularidade está em alunos e o resultado do modelo de regressão proposto interpretado numa coluna denominada Jubilado com os valores "1" ou "0" (um ou zero) como sim ou não. Após definição da variável alvo e construção das variáveis transformadas descritas no item III, foi aplicada Regressão Logística. Manualmente foram retiradas as variáveis que possuíam significância (p>|Z|) inferior a 5% (0,05) sendo aplicada a avaliação através do processo de cross-validation em 10 "folds". O resultado final da regressão logística é ilustrado na figura a seguir:

Tabela 4. Resultado da regressão logística.

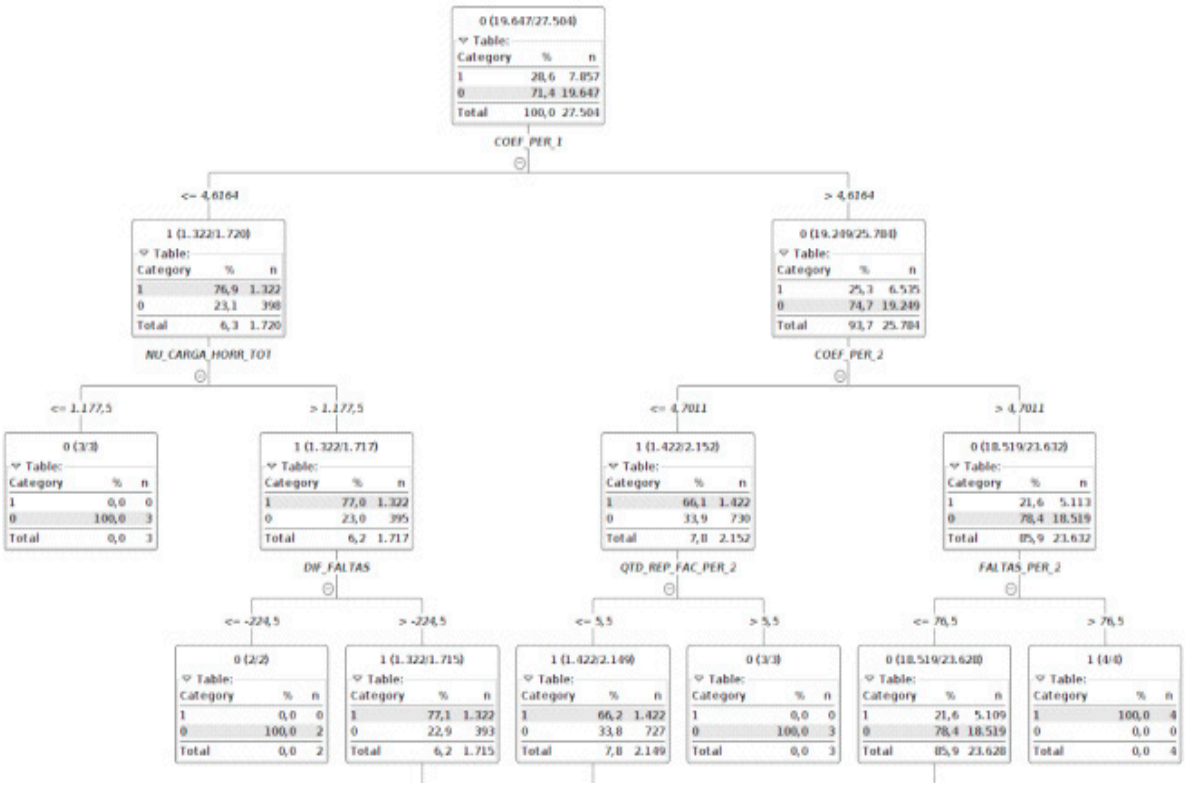
Variável	Beta	Z-escore	p>|z|
Total de aprovações no 2 período	0,439	11,938	0
Fez o ENEM	0,353	10,947	0
Coeficiente de rendimento do 2 período	0,334	18,510	0
Coeficiente de rendimento do 1 período	0,289	14,449	0
Quantidade de reprovação em disciplinas fáceis no 1 período	0,240	4,426	0
Quantidade de reprovação em disciplinas fáceis no 2 período	0,129	2,564	0,01
Total de aprovações no 1 período	0,075	4,484	0
Total de Faltas no 2 periodo	0,006	4,005	0
Carga horária do curso	0,000	11,473	0
Diferença entre ano de conclusão do ensino médio e entrada na UFPE	-0,016	-3,925	0
Total de reprovações no 1 periodo	-0,130	-3,903	0
Total de cancelamentos de disciplinas com ônus no 2 periodo	-0,202	-3,131	0,002
Quantidade de disciplinas do 2 periodo	-0,333	-10,531	0
Sexo masculino	-0,412	-12,799	0
UFPE - AREA III	-1,044	-5,510	0
UFPE - AREA II	-1,404	-7,457	0
UFPE - AGRESTE	-1,442	-7,412	0
UFPE - AREA I	-1,788	-9,509	0
Taxa de aprovações no 2 periodo	-2,092	-10,215	0
UFPE - AREA IV	-2,146	-11,298	0

Fonte: os autores

Foram criadas também, para cada um dos dois primeiros períodos, as variáveis: coeficiente de rendimento, aprovações, reprovações, quantidade de disciplinas, faltas, cancelamentos com ônus, aprovações em componentes com alta taxa de reprovação e reprovações naqueles com alta taxa de aprovação. Os dois últimos foram gerados a partir da identificação do 4-quartil de maiores taxas de aprovação e reprovação dos períodos letivos. Também foram criadas variáveis para representar a diferença entre faltas, coeficiente, taxa de aprovação, taxa de reprovação e quantidades de disciplinas de cada período. Ao final restaram 37 variáveis. A tabela a seguir descreve o percentual de alunos que apresentam critérios de jubilamento por ano de admissão:

A Árvore de decisão gerada demonstra com clareza a identificação dos resultados. Foi utilizado o algoritmo C4.5, considerado um exemplo clássico de método de indução de árvores de decisão. O C4.5 (QUILLAN, 1993) foi inspirado no algoritmo ID3 (QUILLAN, 1979), que produz árvores de decisão a partir de uma abordagem recursiva de particionamento de um conjunto de dados, utilizando conceitos e medidas da Teoria da Informação (19).

Fig. 5. Árvore de decisão gerada.



Fonte: os autores

Analisando os 3 primeiros níveis da árvore de decisão entende-se que o coeficiente acadêmico do 1º período exerce maior influência com relação ao jubilamento do discente.

O principal resultado pode ser identificado quando este coeficiente é maior do que 4,61 e o do 2o período é maior do que 4,7. Caso o número de faltas do 2o período seja menor que 77 representada uma confiança de que 78,4% (18519 de 23628 discentes) de não haver jubilamento.

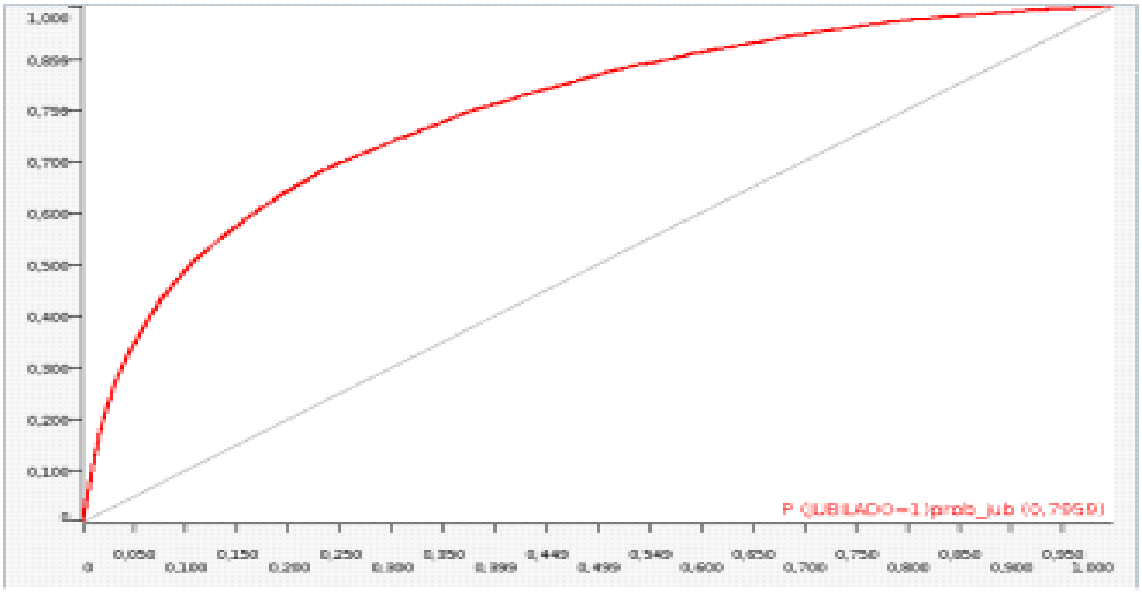
Com relação à tendência ao jubilamento são identificados dois pontos críticos. Um deles é o coeficiente do 1o período ser maior do que 4,61 e o do segundo ser menor ou igual a 4,7, não sendo a quantidade de reprovações em disciplinas consideradas fáceis para o 2 o período maior do que 5, a identificada confiança de 66,2% (1422 de 2149 discentes) do aluno atingir algum dos critérios de jubilamento.

Outro ponto crítico verifica-se quando o coeficiente do 1o período não é maior que 4,61. Quando a carga horária total do curso é maior do que 1177,5 e a diferença de faltas entre períodos é maior ou igual a 225 (esse número é negativo quando o aluno teve mais faltas no 1o período) encontramos que 1322 de 1715 discentes (77,1%) são enquadrados em algum critério de recusa de matrícula.

Um método gráfico eficiente para detecção e avaliação da qualidade de sinais conhecido como ROC (Receiver Operating Characteristic) ou curva ROC (20) foi criado e desenvolvido na década de 50 do século passado para detecção da qualidade de transmissão de sinal em um canal com ruído (20).

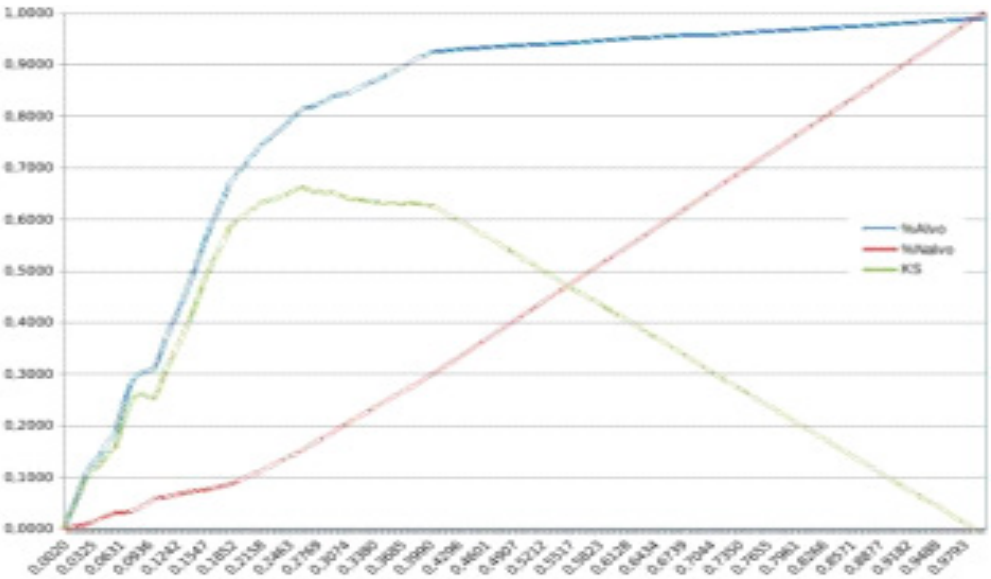
O gráfico abaixo, figura 6 é a curva ROC do modelo construído, na qual se obteve uma área de 0,80. Para verificar a máxima diferença absoluta entre as distribuições acumuladas foi utilizado o teste de Kolmogorov-Smirnov. É realizada a medição das distâncias verticais entre as duas distribuições (verdadeiros positivos e falsos positivos). O valor varia entre 0 a 1, quanto maior esse valor maior a distância entre as distribuições e, consequentemente, maior poder discriminante. O max_ks para as distribuições em questão foi identificado como 0,44837, como pode ser verificado na figura 7.

Fig. 6 – Curva ROC do modelo construído



Fonte: os autores

Fig. 7 – Curva KS do teste Kolmogorov – Smirnov



Fonte: os autores

6. CONCLUSÃO

Após a validação do modelo proposto nesta pesquisa foi verificado que o total de reprovações no segundo período é o principal fator que explica o jubilamento dos alunos pesquisados, sugere-se assim programas de aconselhamento para os alunos no término do segundo período.

Vale ressaltar que o resultado apresentado na validação do modelo na curva ROC foi satisfatório, tendo em vista que o trabalho realizado por Hadautho e Adeodato (5) fez uso do curso de medicina em seu modelo, na qual o mesmo possui baixos índices de retenção/evasão.

O trabalho realizado poderá servir como base para trabalhos futuros que analisem outras épocas, mas que tomem como fundamento as mesmas variáveis utilizadas neste modelo, pois novos fatores possam ser avaliados, tais como a junção com outras bases de dados que contenham informações socioeconômicas sobre os alunos, ora não utilizadas neste trabalho.

Outras questões foram levantadas pela pesquisa, como a possibilidade da tomada de decisão poderia ser no primeiro semestre. Os dados apontaram que isto também é possível, contudo este trabalho de pesquisa foi divido com outra equipe que fez o fluxo temporal baseado no primeiro semestre. Assim este trabalho reflete somente a posição referente ao segundo semestre.

Não foi realizado uma pesquisa sugerindo o terceiro semestre como fluxo temporal à tomada de decisão, haja vista que ainda seria possível "salvar" o aluno de perder o vínculo com esta universidade, já que o jubilamento (ou o início processo deste) ocorre somente após os resultados auferidos pelo aluno a partir do quarto semestre, pelo qual não haveria o início do processo de jubilamento impossibilitando a reversão da situação, segundo as regras atuais.

As árvores de decisão mostram-se algoritmos robustos e tolerante a falhas, de fácil interpretação por um ser humano, mesmo os problemas com "missing data", que tomaram mais de 80% do tempo, na fase de preparação dos dados.

A regressão logística por ser um algoritmo de fácil implementação e relativa rapidez, tornou o trabalho de pesquisa mais confiante, sobretudo quando se traçou as curvas ROC do modelo proposto, por estar próximo dos 80%, contudo seu resultado carece de boa interpretação por humanos, necessitando de outra ferramenta para tal.

Por fim, pretende-se estender tal estudo para averiguar que após as decisões tomadas sobre a avaliação de alunos no término do segundo semestre, se o efeito causado foi atendido com os índices de desempenho acadêmicos almejados pelo REUNI.

REFERÊNCIAS

Ministério da Educação. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. 1996. Disponível em: http://www.andifes.org.br/diplomacao-retencao-e-evasao-nos-cursos-de-graduacao-em-instituicoes-de-ensino-superior-publicas. Acessado em dezembro de 2015.

DEKKER G., PECHENIZKIY M.; VLEESHOUWERS J. Predicting Students Drop Out: A Case Study: Presented at International Conference on Educational Data Mining, Cordoba, Spain, 2009. BARNES, T.; DESMARAIS, M.; ROMERO, C.; VENTURA, S. Ed. p 41-50.

TINTO, V. Limits of Theory and Practice in Student Attrition: The Journal of Higher Education.Cap.53(6).p. 687–700, 1982. Disponível em http://doi.org/10.2307/1981525. Acessado em dezembro de 2015.

CAMPELLO, A. V. C.; LINS, L. N. ―Metodologia de análise e tratamento da evasão e retenção em cursos de graduação de instituições federais de ensino superior‖, 2008.

HADAUTHO, R. B. S.; ADEODATO, P. J. L. A Data Mining Approach for Preventing Undergraduate Students Retention, 2012.

REUNI. Reestruturação e Expansão das Universidades Federais: Diretivas Gerais, 2007.

UFPE. Programa de reestruturação e expansão das universidades federais: Projeto REUNI/ UFPE, 2007.

UFPE. Resolução no 11/2015: Disciplina a recusa definitiva de matrícula nos cursos de graduação oferecidos pela UFPE, modalidade presencial, 2015. Disponível em: https:// www.ufpe.br/proacad/images/DGA_PROACAD/RESOLUCOES/res_11_2015_recusa%20de%20matrcula_bo67_28_07_2015.pdf . Acessado em dezembro de 2015.

CHAPMAN, P.; CLINTON, J. et. Al. CRISP-DM 1.0: Step-by-step data mining guide, 2000.

UFPE. Evasão. Disponível em: http://www.proplan.ufpe.br/images/indicadores/evasao_coorte.pdf1. Acessado em dezembro de 2015.

UFPE.Resolução-11/2015/CCEPE, 2015.Disponível em: https://www.ufpe.br/proacad/images/DGA_PROACAD/RESOLUCOES/res_11_2015_recusa%20de%20matrcula_bo67_28_07_2015.pdf. Acessado em: dezembro de 2015.

RUD, O. P. Data Mining CookBook. Modeling: Data for Marketing, Risk, and Customer Relationship Management, 15p.

WEST, D. Neural network credit scoring models: Computers and Operations Research, cap 27, p. 1131 –

1152, 2000. In: HILBE, J. M. Logistic Regression Models: Chapman & Hall/CRC Press , 2009.

HAN, J.; KAMBER, M. Data Mining: Concepts and Techniques. Elsevier, San Francisco, 2 ed. 2006.

IAN, H.; Witten and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, San Francisco, 2 ed. 2005.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações. Ed. Elsevier, São Paulo, 2 ed. 2015.

EGAN. J. P. Signal detection theory and ROC analysis. New York, USA. Academic Press, 1975.

Disponível em: http://www.anaesthetist.com/mnm/stats/roc/Findex.htm. Acessado em: dezembro de 2015.

Disponível em: http://crsouza.com/2009/07/analise-de-poder-discriminativo-atraves-de-curvas-roc. Acessado em: dezembro de 2015.

GREEN D.; SWETS J. A. Signal Detection Theory and Psychophysics. Los Altos, USA: Peninsula Publishing, 1989.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Curva ROC para avaliação de classificadores, 2014.

FAYYAD, P.; PIATETSKY-SHAPIRO, U. & SMYTH, G. From data mining to knowledge discovery in databases. Advances in Knowledge Discovery and Data Mining, v. 17, n. 3, p. 1–36, 1996.

CAMILO, C. O.; SILVA, J. C. da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. [S.l.], 2009. 1–29 p.

RUSSEL, S.; NORVIG, P., “Inteligência Artificial.” .Elsevier, Rio de Janeiro, 3rd ed, pp. 716 – 721, 2004.

SRIVASTAVA,A.; KATIVAR, V; SINGH, N. “Review of Decision Tree Algorithm: Big Data Analytics”, v. 2, 2015.