Análise dos problemas da Evasão e Retenção: Uma abordagem através de Mineração de Dados Educacionais

Diego da Costa do Couto, Ádamo Lima de Santana

[1]Laboratório de Inteligência Computacional e Pesquisa Operacional (LINC)
Universidade Federal do Pará (UFPA)
Caixa Postal 479 – 66.075-110 – Belém – PA – Brasil

{diegocouto,adamo}@ufpa.br

Abstract. *This paper applies classification algorithms in a large database with the purpose of diagnosing the causes of two problems faced in Brazilian universities, college dropout and retention. The accuracies of many algorithms were measured with a focus on verifying the ability to correctly classify available instances. Results showed that the Bayesian Network method reached an overall precision approximately 86 % and it is considered a very satisfactory solution for the discovery and representation of knowledge about academic performance of undergraduate students, especially those who are willing to give up or extrapolate the deadline for completing to the course.*

Resumo. *Este artigo aplica algoritmos de classificação em uma grande base de dados com finalidade de diagnosticar as causas de dois problemas enfrentados em universidades brasileiras, a evasão e a retenção. Foram mensuradas acurácias de diversos algoritmos com foco em verificar a capacidade de classificar corretamente as instâncias disponíveis. Os resultados apontaram que o método Rede Bayesiana atingiu precisão geral de aproximadamente 86% sendo considerada uma solução bastante satisfatória para descoberta e representação do conhecimento acerca do desempenho acadêmico dos alunos da graduação, especialmente aqueles propensos a desistir ou extrapolar o prazo para conclusão do curso.*

1. Introdução

O Censo da Educação Superior revelou que entre os anos de 2013 e 2014 houve um acréscimo de, aproximadamente, 2,5% no número dos cursos ofertados em 2.368 Instituições de Ensino Superior (IES). Vale destacar que entre 2003 e 2014, a matrícula na educação superior registrou aumento de 96,5% [INEP 2014]. Estas constatações corroboram os avanços em termos quantitativos da educação superior no país nas iniciativas privada e estatal. Sobretudo, ressalta-se que gestores devem continuamente avaliar se este aumento em quantidade se converteu igualmente em qualidade, ao estudante, à instituição de ensino superior e à sociedade.

Os levantamentos realizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), também apontam descompasso entre os números de matrícula, ingressantes, cursos e concluintes [INEP 2014]. Essas informações denotam um importante diagnóstico: o aumento na quantidade de vagas não está impactando diretamente na permanência do aluno até a sua formatura. Esta problemática, conhecida

como evasão resulta em vagas ociosas ou remanescentes, as quais se destinam a outros processos de seleção. A retenção, por sua vez, é outro entrave verificado em IES brasileiras, que caracteriza-se quando o aluno excede o tempo máximo permitido para conclusão da graduação. Ambas as problemáticas, em algumas instituições de ensino, estão vinculadas, visto que alguns regimentos preveem a prescrição da vaga após certo período de retenção, consequentemente o estudante é, obrigatoriamente, desvinculado do curso.

Considerando que os problemas da evasão e retenção possuem inúmeras causas e consequências negativas para estudantes, instituições de ensino e comunidades nas quais esses indivíduos estão inseridos, este trabalho tem como objetivo a criação de subsídios que auxiliem gestores da instituição de ensino superior a identificar alunos, dos cursos de graduação, em situação de vulnerabilidade à evasão ou à retenção dentro dos seus ambientes de aprendizagem. Dentre os subsídios importantes à gestão, destacam-se: previsão de quais alunos são propensos a desistir ou permanecer além do tempo estipulado pelo currículo; representação desta informação; e identificar quais atributos, dentre os disponíveis, são mais relevantes durante a classificação desse aluno.

Pretende-se alcançar estes objetivos pela utilização da Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Database* – KDD) que representa um "processo não-trivial de extração de informações implícitas, previamente desconhecidas e potencialmente úteis a partir de dados" [Frawley et al. 1992]. Uma das suas etapas, denominada de Mineração de Dados (*Data Mining*) [Fayyad et al. 1996], onde ocorre a extração de padrões dos dados através do uso de algoritmos específicos, foi empregada para verificar a relação entre as variáveis e a problemática explicitada. A etapa de *Data Mining* pode ser aplicada em diversas áreas [Han et al. 2012] [Fayyad et al. 1996] desde que estas possuam razoáveis volumes de dados históricos.

Foram testados algoritmos, durante a etapa de *Data Mining*, a partir da tarefa de classificação, que define-se como "o processo de atribuir, a uma determinada informação recebida, o nome de uma classe à qual ela pertence" [Rich e Knight 1993] ou ainda constrói um modelo ou classificador [Han et al. 2012]. Dentro do contexto aplicado, a classificação permite presumir a situação (classe) do estudante na universidade, dado um conjunto de atributos a respeito desse aluno. Avaliaram-se métricas relativas ao desempenho dos classificadores, cujas características possam atender aos requisitos associados ao objetivo deste trabalho, com o intuito de testá-los e, posteriormente, selecioná-los à resolução do problema pesquisado.

Este trabalho está organizado da seguinte forma: A Seção 2 apresenta os trabalhos correlatos. Por sua vez, na Seção 3 são apresentados os materiais e métodos utilizados nesta pesquisa. Em seguida, Seção 4 serão discutidos os resultados. Na Seção 5 apresentação das considerações finais.

2. Trabalhos Correlatos

A Mineração de Dados Educacionais (*Educational Data Mining* – EDM) é uma disciplina emergente cujo objetivo está no desenvolvimento de métodos para explorar os dados provenientes de cenários educacionais e como essas metodologias são empregadas para compreender os alunos nos seus ambientes de aprendizagem [JEDM 2016]. Argumenta-se a existência de um aumento considerável no interesse por pesquisas valendo-se de EDM [Sachin e Vijay 2012]. Nesta perspectiva, [Romero e Ventura 2010] elaboraram um

trabalho relativo ao estado da arte da Mineração de Dados Educacionais, no qual são discutidos 235 publicações mais relevantes até o ano de 2009.

[Cortez e Silva 2008] obtiveram dados no período letivo de 2005 e 2006 de escolas públicas de Portugal. Os atributos constituíram-se de registros coletados de relatórios emitidos pelo sistema escolar e questionários com perguntas sobre aspectos sociais, demográficos e emocionais dos estudantes. A finalidade dos autores era predizer o desempenho escolar nas disciplinas básicas de Matemática e Português. Os resultados atingidos foram satisfatórios em testes com árvores de decisão, onde conseguiu-se a melhor taxa de acerto $(93,0\%)$. [Cortez e Silva 2008] priorizaram a geração de conhecimento especialista, os autores descobriram importantes regras das árvores de decisão.

No Brasil, as investigações em Mineração de Dados Educacionais se consolidaram em 2012, na ocasião, [Manhães et al. 2012, Manhães et al. 2014] elaboraram um estudo de caso para avaliar a evasão em 155 cursos de graduação ofertados por 28 unidades da UFRJ. Para a pesquisa em discussão, foram selecionados dados acadêmicos dos discentes que ingressaram nos dois semestres letivos dos anos de 2003 e 2004. O classificador *Naive Bayes* foi elegido pela interpretabilidade dos resultados aliada a precisão global superior a 80%.

O nosso trabalho, proposto neste artigo, possui similaridades com aqueles discutidos anteriormente, visto que, por exemplo, vale-se de algoritmos classificadores para detecção de variáveis associadas à evasão e retenção em âmbito acadêmico. Contudo diferencia-se dos demais nos seguintes aspectos: i) aplicação em uma grande base de dados, composta por quase 100 mil amostras, pois a maioria dos trabalhos usam *data sets* com algumas centenas de registros; ii) análise sobre todos os cursos de graduação, enquanto muitos trabalhos avaliam cursos ou disciplinas de maneira isolada. Além disso, este trabalho visa fortalecer o campo de EDM, uma vez que esta área é nova, há poucos estudos nacionais e exerce grande influência na resolução de problemas atrelados ao desempenho escolar.

3. Materiais e Métodos

3.1. Base de Dados

Os dados selecionados à pesquisa são registros acadêmicos, oriundos do sistema informatizado da Universidade Federal do Pará (UFPA), referentes aos discentes de graduação ingressantes até 2016. Inicialmente foram consideradas 175.779 amostras, porém removeram-se da base os alunos com situação indefinida quanto às classes ou com tuplas inconsistentes, permanecendo 98.698 linhas. A Tabela 1 apresenta os 31 atributos selecionados e os seus respectivos significados, alguns deles serão discutidos na Seção 3.2.

3.2. Pré-Processamento e Transformação de dados

Os significados dos atributos de 1 a 12, considerados intuitivos, podem ser consultados nas descrições dispostas na Tabela 1. Os atributos de 20 a 22 denotam a probabilidade de um discente formado nos últimos cinco anos possuir um dos índices acadêmicos igual ou superior aos demais alunos pertencentes ao mesmo curso e matriz curricular. Foram usados os indicadores MC, IRA e IEA, uma vez que estes, em suas definições matemáticas e conceituais, aferem a eficiência do aluno durante o seu percurso acadêmico.

Tabela 1. Atributos selecionados à pesquisa

Número	Variável	Descrição
1	sexo	sexo que o discente pertence
2	idade	idade que o aluno ingressou no curso
3	interior	informa se o discente estuda no *campus* capital ou em um dos *campi* do interior do estado
4	turno	turno no qual o discente estuda
5	forma_ingresso	forma de seleção pela qual o discente ingressou na universidade
6	numero_trancamento	Número de vezes que o discente trancou a matrícula
7	numero_vinculos	Número de vezes que o discente fez outras graduações (vínculos) até o ingresso no curso atual
8-10	perc_ch_{tipo}	Percentual das cargas horárias prática, teórica e de estágio
11	sem_ordem	O percentual das disciplinas cursadas fora da ordem proposta pelo currículo do discente
12	primeiro_semestre_ocorr	Informa qual o semestre que o discente cursou pela primeira vez uma disciplina fora de ordem
13-19	indices_academicos	representam os indicadores de rendimento acadêmico acumulado, a saber: Média de Conclusão (MC), Média de Conclusão Normalizada (MCN), Índice de Rendimento Acadêmico (IRA), Índice de Eficiência em Carga Horária (IECH), Índice de Eficiência em Períodos Letivos (IEPL), Índice de Eficiência Acadêmica (IEA) e Índice de Eficiência Acadêmica Normalizado (IEAN). Essas métricas quantificam o desempenho dos alunos da graduação e nos cálculos consideram-se dados do histórico acadêmico, tais como: quantidades de reprovações, aprovações, trancamentos, cargas horárias acumuladas e esperadas para integralização do curso, entre outros.
20-22	prob_índice	Refere-se a probabilidade de um discente formado nos últimos 5 anos possuir o índice acadêmico maior ou igual ao aluno avaliado (Teste z)
23-30	perc_{conceito}_{avaliação}	Refere-se ao percentual de um conceito conseguido pelo discente dentro do período avaliado
31	*status*	Denota a situação (classe) a qual o estudante pertence

A média das notas obtidas pelo estudante em cada disciplina, em um período letivo, é convertida em conceito, definido segundo a escala apresentada na Tabela 2. As variáveis indexadas de 23 a 30 referem-se ao percentual de um determinado conceito de acordo com o período de avaliação, seja este geral (acumulado por todo o curso) ou para o primeiro ano cursado. Por exemplo, a variável $perc\_ins\_primeiro\_ano$ denota o percentual de conceitos do tipo INS referentes ao primeiro ano de graduação.

Tabela 2. Correspondência ente a média das notas e o conceito

Conceito	Intervalo da média
Insuficiente (INS)	[0-4,99]
Regular (REG)	[5-6,99]
Bom (BOM)	[7-8,99]
Excelente (EXC)	[9-10]

Finalmente, o atributo 31 representa a classe a qual o discente pertence, cujos possíveis valores são: “Formado”, “Evadido” e “Retido”. Os alunos considerados na classe “Formado” são aqueles que conseguiram integralizar a carga horária prevista pelo curso. Por sua vez, o rótulo “Evadido” remete-se aos alunos que, por decisão própria ou processo de prescrição previsto em regimento da instituição, abandonaram a graduação. Os estudantes com matrículas ativas, porém que ultrapassaram um ano do prazo de conclusão estabelecido no currículo do curso foram classificados como “Retido”. Há na base de dados 65.758 (66,63%) amostras referentes a classe dos alunos formados; 25.581 (25,92%) dos registros, pertencem aqueles que desistiram dos estudos; e por fim, os alunos em retenção são menos representativos, 7.359 (7,46%).

3.3. Aplicação Proposta

Durante a etapa de *Data Mining*, foram testados algoritmos classificadores, a partir disso analisou-se a precisão global (acurácia) de cada um deles, para finalmente selecionar aquele que obteve uma taxa de acerto aceitável. Considerou-se ainda à seleção do algoritmo dois critérios: a representação dos resultados e o quanto esta informação pode ser interpretada por especialistas e usuários inseridos no domínio. Para estas finalidades, a Rede Bayesiana se mostra uma importante ferramenta, pelos seguintes aspectos: representação gráfica da relação entre estados; a rede expressa o conhecimento especialista acerca do domínio; e os resultados numéricos (probabilidades) podem ser visualizados através de gráficos.

A estratégia utilizada para segmentar a base de dados em conjuntos de treinamento e testes, destinados a estimar precisão e confiabilidade do modelo construído pelo classificador, foi a validação cruzada com k-conjuntos estratificada (*stratified k-fold cross-validation*), por ser uma das mais empregadas em mineração de dados [Han et al. 2012].

Os algoritmos de aprendizado supervisionado [Rezende 2005] empregados nesta pesquisa estão disponíveis na ferramenta de código aberto (*open source*) Weka [Weka 2017], divididos de acordo com as seguintes abordagens: árvores de decisão, probabilísticos, baseados em instâncias, baseados em funções e redes neurais artificiais. Os classificadores probabilísticos foram *Naive Bayes* e Redes Bayesianas (*Bayesian Networks*). Os métodos baseados em instâncias e funções foram representados pelos indutores *K-Nearest Neighbor* (KNN) e *Support Vector Machine* (SVM), respectivamente. O método *Multilayer Perceptron* foi empregado segundo a abordagem Redes Neurais Artificiais (RNA). Finalmente, testaram-se os algoritmos, de árvores de decisão, *Random Tree*, *Random Forest* e *Classification And Regression Trees* (CART).

4. Resultados

4.1. Análise de desempenho dos algoritmos

A Tabela 3 apresenta os 9 algoritmos e as métricas usadas: tempos para treinar e testar o modelo, acurácia e coeficiente Kappa. Os resultados mostram que a melhor solução foi conseguida através do indutor *Random Forest* cuja acurácia superou 87%. Não obstante o algoritmo *Bayesian Network* revelou precisão global próxima de 86% e tempos aceitáveis para construção e testes do modelo, além disso este algoritmo obteve valor de estatística Kappa igual a 0,6961, considerado um nível substancial de concordância interobservador [Viera e Garrett 2005]. Destaca-se que aplicações nas quais o tempo de processamento é considerado requisito crucial ao domínio, soluções como *Multilayer Perceptron* e SVM são consideradas inviáveis, embora apresentem boas taxas de acerto.

4.2. Análise da evasão e retenção via Redes Bayesianas

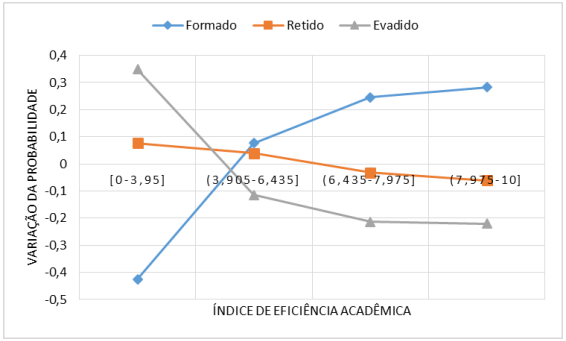
Foram selecionados os 14 atributos mais relevantes, de acordo com ganho de informação [Han et al. 2012]. Após a redução no número de variáveis, aferiu-se novamente a acurácia do algoritmo *Bayesian Network*, apresentando precisão de 83,5%, ratificando a sua robustez. O algoritmo de busca gulosa (*greedy search*) K2 [Cooper e Herskovits 1992] foi empregado para construção da topologia da rede, atingindo-se maior precisão global com o parâmetro de número esperado de pais por nó definido a 5.

Tabela 3. Métricas de desempenho geral dos classificadores

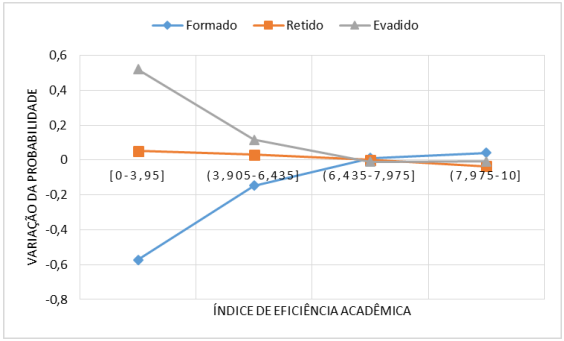
Algoritmos	Tempo para treino (s)	Tempo para teste (s)	Acurácia (%)	Kappa
Naive Bayes	0.31	2.14	78.7736	0.5688
Bayesian Network	6.77	1.26	85.865	0.6961
KNN	0.29	1153.27	83.8041	0.6483
SVM	1418.69	1288.21	86.6938	0.6999
Multilayer Perceptron	4739.78	3.56	86.2054	0.7048
C4.5	1.62	1.44	86.2449	0.6984
Random Tree	0.55	1.48	80.5021	0.5924
Random Forest	12.31	8.24	87.1102	0.7118
CART	236.39	0.63	86.5104	0.7023

Para a inferência Bayesiana, escolheram-se os atributos IEA (iea) e número de trancamentos (numero_trancamentos). Os estados, de todas as variáveis, foram conseguidos por intermédio da discretização com distribuição uniforme de frequência. Os atributos numero_trancamento e iea tiveram, respectivamente, três e quatro intervalos para conversão de dados contínuos em discretos, sendo estas quantidades determinadas após análise dos dados.

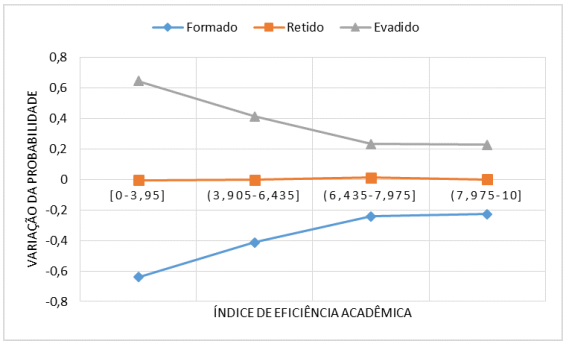
As Figuras 1(a) a 1(c) representam os gráficos das variações de probabilidade, comparativamente por classes de discentes, número de trancamento e evolução do IEA.



(a) Alunos sem trancamento



(b) Alunos que trancaram apenas uma vez a matrícula



(c) Alunos que trancaram mais de uma vez

Figura 1. Variação da probabilidade de acordo com a classe de alunos, número de trancamentos e Índice de Eficiência Acadêmica (IEA).

Depreende-se do gráfico da Figura 1(a) que, alunos sem trancamento, possuem

mais chances de graduar-se a mesma proporção do crescimento do IEA. O intervalo do índice de eficiência acadêmica entre 3,905 e 6,435, revela um importante diagnóstico – nesta faixa começa o decrescimento das chances de o aluno evadir-se e acentuam-se as possibilidades de diplomação em tempo previsto pela instituição de ensino. O aluno que conseguiu IEA entre 0 e 3,95, incrementa em mais de 30% as possibilidades de abandonar os estudos. O cenário desejável ocorre a um IEA superior a 7,975, pois neste caso há acréscimo de quase 30% de aluno formar-se e a redução que supera 20% para evasão.

Alunos com exatamente 1 trancamento têm suas possibilidades de formatura reduzidas em até 57,2%, caso possuam IEA próximo a 3,95, conforme ilustra o gráfico Figura 1(b) e crescem até 4,2%, caso tenham IEA excedente a 7,975%. Naquele mesmo cenário, intensificam-se os indícios de retenção (5,1%) e, principalmente, da evasão (52,1%). Constata-se um importante comportamento nos dados – apenas 1 trancamento do curso não implica a propensão à desistência do curso, se o índice de eficiência acadêmica inferior a 6,435, todavia é um alerta à retenção.

Após 1 trancamento, há uma variação positiva de 64,5% para evasão, caso o aluno possua até 3,95% de índice de eficiência acadêmica e atenua-se, ao passo que este indicador cresce, até 22,8%, de acordo com a Figura 1(c). A variação da probabilidade à formatura não atingiu em nenhum momento valor positivo, isto é, na prática a interrupção da matrícula por mais de uma vez diminuem as chances de o aluno conseguir a diplomação, elevando consideravelmente as chances da desistência.

5. Considerações Finais

Esta pesquisa utilizou mineração de dados sobre uma base de dados com quase cem mil registros acadêmicos dos discentes de graduação para entender as causas associadas ao abandono dos estudos e a permanência além do prazo estipulado para conclusão do curso. Neste trabalho foram testados nove algoritmos classificadores, sendo que o método *Random Forest*, apresentou a melhor acurácia, superior a 87%. Contudo, priorizou-se a escolha de um classificador capaz de possuir fácil representação de resultados e que esta possa expressar o conhecimento do especialista sobre o domínio estudado. Nessa perspectiva, o classificador *Bayesian Network* foi elegido e, ratificou sua escolha por também obter desempenho satisfatório, visto que sua precisão global ultrapassou 85%.

A Rede Bayesiana construída mediante o uso do método de buscas K2 viabilizou a extração de importantes conhecimentos a respeito dos problemas analisados, permitindo a sua vinculação ao índice de eficiência acadêmica e a interrupção da matrícula em período letivo. Os resultados alcançados não são exaustivos, dessa forma outras pesquisas são necessárias para consolidar as respostas acerca dos principais fatores ligados à evasão e retenção em âmbito universitário. Como trabalhos futuros, serão realizadas novas investigações com atributos adicionais de carácter socioeconômicos, fato que permitirá relacionar o desempenho acadêmico a situações de vulnerabilidades sociais e, principalmente, quantificar o impacto dessa dependência, o que propiciará aos gestores a criação de mecanismos eficazes de combate a evasão e retenção.

Referências

Cooper, G. F. e Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Mach. Learn.*, 9(4):309–347.

Cortez, P. e Silva, A. (2008). Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference*, pages 5–12, Porto, Portugal.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., e Uthurusamy, R., editors (1996). *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, Menlo Park, CA, USA.

Frawley, W., Piatetsky-Shapiro, G., e Matheus, C. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine*, 13(3):57–70.

Han, J., Kamber, M., e Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3 edition.

INEP (2014). Censo da Educação Superior 2014 - Notas Estatísticas. http://download.inep.gov.br/educacao_superior/censo_superior/documentos/2015/notas_sobre_o_censo_da_educacao_superior_2014.pdf. [Online; Acessado em 16/04/2016].

JEDM (2016). Journal of Educational Data Mining. http://www.educationaldatamining.org/JEDM. [Online; Acessado em 26/01/2016].

Manhães, L. M. B., da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., e Zimbrão, G. (2012). Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação Através de Sistemas Baseados em Mineração de Dados: Uma Abordagem Quantitativa. In *Simpósio Brasileiro de Sistemas de Informação*, pages 468–479, São Paulo.

Manhães, L. M. B., da Cruz, S. M. S., Zavaleta, J., e Zimbrão, G. (2014). The Impact of High Dropout Rates in a Large Public Brazilian University. In *CSEDU – 6th International Conference on Computer Supported Education*, pages 126–129, Barcelona, Spain.

Rich, E. e Knight, K. (1993). *Inteligência Artificial*. Makron Books.

Romero, C. e Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618.

Sachin, R. e Vijay, M. (2012). A Survey and Future Vision of Data Mining in Educational Field. In *Advanced Computing Communication Technologies (ACCT), 2012 Second International Conference on*, pages 96–100, Rohtak, Haryana, India.

Viera, A. e Garrett, J. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5):360–363.

Weka (2017). Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/. [Online; Acessado em 06/02/2017].