INSTITUTO FEDERAL GOIANO - CAMPUS CERES
BACHARELADO EM SISTEMAS DE INFORMAÇÃO
GUSTTAVO NUNES GOMES

ESTUDO DA EVASÃO EM ALUNOS DE GRADUAÇÃO POR MEIO DE MINERAÇÃO DE DADOS

CERES - GO
2019

GUSTTAVO NUNES GOMES

ESTUDO DA EVASÃO EM ALUNOS DE GRADUAÇÃO POR MEIO DE MINERAÇÃO DE DADOS

Trabalho de curso apresentado ao curso de Sistemas de Informação do Instituto Federal Goiano – Campus Ceres, como requisito parcial para a obtenção do título de bacharel em Sistemas de Informação, sob orientação do Prof. Dr. Marcos de Moraes Sousa.

CERES - GO

2019

INSTITUTO FEDERAL
Goiano

Repositório Institucional do IF Goiano - RIIF Goiano
Sistema Integrado de Bibliotecas

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610/98, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano, a disponibilizar gratuitamente o documento no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

Identificação da Produção Técnico-Científica

[ ] Tese	[ ] Artigo Científico
[ ] Dissertação	[ ] Capítulo de Livro
[ ] Monografia – Especialização	[ ] Livro
[X] TCC - Graduação	[ ] Trabalho Apresentado em Evento
[ ] Produto Técnico e Educacional - Tipo: _____	

Nome Completo do Autor: Gusttavo Nunes Gomes
Matrícula: 2016103202030044
Título do Trabalho: Estudo da evasão em alunos de graduação por meio de mineração de dados

Restrições de Acesso ao Documento

Documento confidencial: [ ] Não [ ] Sim, justifique: _____

Informe a data que poderá ser disponibilizado no RIIF Goiano: __/__/__
O documento está sujeito a registro de patente? [ ] Sim [X] Não
O documento pode vir a ser publicado como livro? [ ] Sim [X] Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a autor/a declara que:

1. o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
2. obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
3. cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

_____, ____/____/____.
Local Data

Gusttavo Nunes Gomes
Assinatura do Autor e/ou Detentor dos Direitos Autorais

Ciente e de acordo:

Assinatura do(a) orientador(a)

Sistema desenvolvido pelo ICMC/USP
Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas - Instituto Federal Goiano

Gomes, Gusttavo Nunes
G633e ESTUDO DA EVASÃO EM ALUNOS DE GRADUAÇÃO POR MEIO DE MINERAÇÃO DE DADOS / Gusttavo Nunes Gomes;orientador Marcos de Moraes Sousa. -- Ceres, 2019.
33 p.

Monografia (em Bacharelado em Sistemas de Informação) -- Instituto Federal Goiano, Campus Ceres, 2019.

1. Algoritmos de Classificação. 2. Plataforma Nilo Peçanha. 3. J48. 4. Weka. 5. Teste-T para Amostra Pareada. I. Sousa, Marcos de Moraes, orient. II. Título.

Responsável: Johnathan Pereira Alves Diniz - Bibliotecário-Documentalista CRB-1 n°2376

ANEXO IV - ATA DE DEFESA DE TRABALHO DE CURSO

Ao(s) 22 dia(s) do mês de OUTUBRO do ano de dois mil e 19, realizou-se a defesa de Trabalho de Curso do(a) acadêmico(a) GUSTTAVO NUNES GOMES, do Curso de SISTEMAS DE INFORMAÇÃO, matrícula 20161032020300 44, cujo título é "ESTUDO DA EVASÃO EM ALUNOS DE GRADUAÇÃO POR MEIO DE MINERAÇÃO DE DADOS". A defesa iniciou-se às 14 horas e 10 minutos, finalizando-se às 14 horas e 45 minutos. A banca examinadora considerou o trabalho APROVADO com média 8,8 no trabalho escrito, média 9,1 no trabalho oral, apresentando assim média aritmética final 9,0 de **pontos,** estando o(a) estudante APTO para fins de conclusão do Trabalho de Curso.

Após atender às considerações da banca e respeitando o prazo disposto em calendário acadêmico, o(a) estudante deverá fazer a submissão da versão corrigida em formato digital (.pdf) no Repositório Institucional do IF Goiano – RIIF, acompanhado do Termo Ciência e Autorização Eletrônico (TCAE), devidamente assinado pelo autor e orientador.

Os integrantes da banca examinadora assinam a presente.

Assinatura Presidente da Banca
MARCOS DE MORAES SOUSA

Assinatura Membro 1 Banca Examinadora
Adriano Honorato Braga

Assinatura Membro 2 Banca Examinadora
Leonardo Paulo Arantes.

"The world moves as fast as Instagram scrolls."

Virgil Abloh

RESUMO

A evasão universitária é um grande problema para os gestores educacionais. A maneira mais eficiente de se abordar a evasão é predizer e com isso tomar medidas para evitar que tal ação ocorra e para que essa antecipação ocorra, é utilizado análise de padrões dos dados de alunos que evadem. Porém para uma análise mais precisa, é necessário ferramentas que auxiliem os gestores de maneira mais precisa e eficiente e por isso que é utilizado a mineração de dados. Esse trabalho realizou uma revisão bibliográfica analisando os métodos usados na avaliação dessa predição, algoritmos usados, *softwares* usados e taxas de acerto de diversos trabalhos dos últimos dez anos. Posteriormente foi aplicado testes na base de dados dos alunos de Institutos Federais do ano de 2017 disponíveis na Plataforma Nilo Peçanha, usando 13 algoritmos de classificação no *software* livre *Weka* para avaliar qual algoritmo teria melhor desempenho e qual seria o mais recomendado para ser usado em uma aplicação que predissesse tal evasão. Foi verificado que o J48 foi o mais eficiente, com outros com desempenho bem próximo e que o Teste-T de Amostra Pareada é fundamental para uma comparação mais precisa dos algoritmos, pois testes individuais podem afetados por fatores externos e influenciar nos resultados, além do fato de ser um teste com maior valor estatístico .

Palavras-chave: Algoritmos de Classificação. Plataforma Nilo Peçanha. J48. *Weka*. Teste-T para Amostra Pareada.

ABSTRACT

University evasion is a major problem for the educational management.The most efficient way to address evasion is to predict and thereby take steps to prevent such action from occurring and for such anticipation to occur, analysis of student data patterns that evade is used. But for a more accurate analysis, tools are needed to help managers more accurately and efficiently and that is why data mining is used. This paper performed a bibliographical review analyzing the methods used in the evaluation of this prediction, algorithms used, software used and performance rates of several works of the last ten years. Subsequently, tests were applied to the database of students from Federal Institutes of the year 2017 available on the Nilo Peçanha Platform, using 13 Weka open source classification algorithms to evaluate which algorithm would perform best and which would be the most recommended to be used in an application that predicted such evasion. It was found that J48 was the most efficient, with others with very close performance and that the Paired Sample T-Test was fundamental for a more accurate comparison of the algorithms, since individual tests can be affected by external factors and influence the results, besides the fact that it is a test with higher statistical value.

Keywords: *Classification Algorithms. Platform* Nilo Peçanha. *J48. Weka*. *Paired Sample T-Test* .

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O desenvolvimento de um país, está muito relacionado com a formação de profissionais altamente capacitados para o mercado de trabalho e grande parte dessa qualificação se obtêm dentro das universidades, sendo fundamental que o governo invista no setor de educação superior.

Entretanto nem todos os estudantes que ingressam em graduações terminam, sendo a evasão, algo bem comum o que prejudica e muito esse processo. Em todo mundo, esse fenômeno é amplamente estudado pela comunidade acadêmica com intuito de se conhecer as causas e auxiliar os gestores universitários a tomarem medidas para tentar evitar essas evasões.

Uma das linhas de pesquisas mais comuns em relação a evasão, é justamente o estudo dos dados de alunos que evadiram e com base neles, tentar encontrar padrões com maior relevância, que mais influenciaram esses alunos a evadirem dos cursos de graduação, geralmente usando técnicas matemáticas e estatísticas para dar maior relevância e importância para esse tipo de estudo.

O estudo desses padrões é englobado por Baker et al. (2011) na área de mineração de dados educacionais, que desenvolvem metodologias para analisar conjuntos de dados coletados em um meio educacional e através disso, compreender melhor o processo de aprendizagem, assim como maior entendimento sobre fenômenos relacionados a aprendizagem, como a evasão.

Ao se conhecer o que pode gerar a saída universitária com uma certa precisão matemática, possibilita que possa ser pensado medidas preventivas para evitar tais ações, otimizando assim o uso dos recursos públicos com educação. O objetivo desse trabalho é verificar em uma plataforma de dados de educação, a Plataforma Nilo Peçanha, o perfil dos estudantes que desistem de cursos graduação por meio de mineração de dados e verificar quais algoritmos de classificação teria melhor eficácia, caso fossem usados em ferramentas futuras de predição.

2 REVISÃO BIBLIOGRÁFICA

A evasão universitária não é um problema apenas no Brasil, tendo registro de pesquisas em várias partes do mundo como na Europa, Yukselturk et al. (2014) e na Ásia, Kaur et al. (2015) por exemplo, e é um fenômeno que traz grandes consequências que os envolvidos buscam formas de evitar a evasão e muito do que é feito, geralmente são estudos com base no perfil dos estudantes que evadiram dos cursos, entretanto com o volume de dados obtidos das instituições é tão grande que tais medidas não são eficientes e com isso administradores e pesquisadores procuram métodos efetivos para extrair conhecimento útil desse grande volume de dados, como informa Yukselturk et al. (2014) que aponta a mineração de dados como uma das ferramentas de apoio a decisão que esses profissionais precisam.

Percebe-se uma necessidade de ferramentas ou mecanismos que automatizem esse processo de detecção precoce dos alunos com risco de evasão para evitar que isso ocorra, visto que essa predição de um possível estudante que venha a evadir, geralmente é feito de maneira "manual, subjetivo, empírico e sujeito a falhas" como define Manhães et al. (2011, p. 151). Além disso depende do envolvimento dos professores, da experiência acadêmica deles o que dificulta o reconhecimento das necessidades e acompanhamento delas com o intuito de se evitar a evasão.

A mineração de dados é mostrada por Manhães et al. (2011, p. 151) e Pascal (2016) como sendo uma técnica que tem muito a agregar no contexto de predição da evasão, entretanto antes de ver as características de mineração de dados e aprofundar em suas aplicações na educação, precisamos diferenciar a pura Mineração de Dados, do termo Descoberta de Conhecimento em Base de Dados, que por alguns autores é vista como sinônimo e por outros como sendo conceitos diferentes, mas que se correlacionam.

Mineração de Dados (MD) - em inglês *Data Mining* (DM) - e Descoberta de Conhecimento em Base de Dados (DCBD) - em inglês *Knowlegde Discovery Databases* (KDD) – é indicado por Baker et al. (2011) como sendo sinônimos visto que ambos extraem conhecimento de base de dados para que seja gerado uma nova informação. Kaur et al. (2015) também aborda essa interpretação no significado desses dois termos, mas complementa como sendo a descoberta de novas informações de grandes base de dados, o que é relativo essa concepção de grande em base de dados, se é grande a quantidade de dados em si, ou se a variedade das fontes desses dados e ao tipo de estruturação dos dados, está sendo considerada ou não porque poderia se questionar sobre as semelhanças com o conceito de *Big Data* o que não faz parte

do escobo da pesquisa, mas abre margem para essa possível interpretação.

Entretanto, alguns os autores assim como Baker et al. (2011) e Kaur et al. (2015), não aprofundam esse aspecto de detalhar o que seriam essas "grandes bases de dados", fazendo apenas essa generalização dos termos e já introduzindo a temática a ser pesquisada logo em seguida como Kaur et al. (2015) que associa KDD e DM como sendo a mesma coisa e altera a discussão para Mineração de Dados Educacionais (MDE) – em inglês, *Educational Data Mining* (EDM) – algo que tem sim, relação com os termos anteriores, mas que não aprofunda no porquê desses termos serem sinônimos.

Outros autores possuem uma visão mais ampla sobre KDD e DM como Paz; Cazella (2017) que aborda a mineração de dados como sendo parte do processo de Descoberta do Conhecimento em Base de Dados, ideia também defendida por Oliveira (2015) que detalha a DM como sendo a etapa do KDD em que são aplicados algoritmos que produzem resultados que serão convertidos em conhecimento após a análise. A generalização de todo o processo de criação desse conhecimento, não só a parte que se executa os algoritmos e sim desde a coleta, preparo dos dados e modelagem e análise como sendo o KDD.

Dehning et al. (2016) definem o KDD como um processo de seis etapas sendo elas: Seleção dos Dados; Processamento; Análise dos Impactos; Análise Preliminar dos Dados; Mineração de Dados e Interpretação; que ao final de todo o processo gera propriamente o conhecimento.

Para Oliveira (2015), essa diferenciação nos conceitos acontece pois alguns autores abordam Mineração de Dados em um contexto mais ampliado, que acabam considerando a mineração de dados como sendo todo o processo e não a etapa em si de análise com os algoritmos, sendo englobado a preparação dos dados, que não apresenta propriamente técnicas de mineração de dados e sim técnicas de conversão, adaptação dos dados para que fiquem compatíveis aos *softwares* para que a mineração possa de fato ser realizada. Já outros autores como Paz; Cazella (2017), Dehning et al. (2016) e Oliveira (2015), separam muito bem cada etapa e definem métodos para cada parte com o intuito de deixar mais claro os passos desenvolvidos, sem perder a qualidade dos dados e do conhecimento gerado, ficando de maneira mais estruturada o KDD em si.

A extração dos padrões dos dados seria a mineração de dados, já o KDD seria o processo cíclico de interação entre as fases e seus resultados com intuito de se refinar cada vez mais para produzir resultados melhores, no sentido de estarem mais íntegros e confiáveis, ou seja,

dentro de um KDD, etapas podem se repetir, como mineração de dados usando mais de um algoritmo, como Kaur et al. (2015) que usou desde rede neural artificial até árvores de decisão.

Como a abordagem de KDD sendo algo mais amplo e mineração de dados ser uma etapa desse processo e essa visão ser a mais aceita na literatura da área atualmente, é o que será usado como base para a pesquisa.

Dentro da área de mineração de dados, existe uma linha de estudo que seria a Mineração de Dados Educacionais (MDE) que para Oliveira (2015) é basicamente o uso de dados educacionais para a realização de mineração de dados. Rigo et al. (2014) confirma essa visão, porém complementa mostrando que a natureza desse dados, é diferente das que normalmente são encontradas o que demanda tratamento, adaptações e aplicação de técnicas nesses dados para se realizar tais análises.

Além do KDD, uma outra metodologia foi bastante usada pelos trabalhos que verificam a evasão universitária, que seria a técnica CRISP-DM (acrônimo de *CRoss-Industry Standard Process for Data Mining*) ou geralmente uma adaptação dessa técnica, que possui várias características similares ao KDD, e no trabalho de Ramos et al. (2017) ele toma a liberdade e cria o termo CRISP-EDM que seria o CRISP normal, mas com mineração de dados educacionais. CRISP-DM realmente tem várias semelhanças com KDD, mas com uma abordagem mais empresarial e de acordo com Kantorsky et al. (2016) sem necessariamente, depender de uma ferramenta ou tecnologia em específico, como Fernandes et al. (2019) demonstra avaliando o desempenho de alunos para verificar quais variáveis afetariam a evasão que nesse estudo foram identificadas como sendo onde o estudante reside e onde se encontra as instituições de ensino que foram as mais impactantes, só depois que fatores como nota nas disciplinas e idade foram aparecer com menor grau de impacto na evasão.

Em resumo, as etapas do CRISP-DM são: o entendimento do projeto, com a definição das metas e requisitos do projeto de uma perspectiva de negócio; coleta inicial dos dados e análise inicial com estatística descritiva; preparação dos dados para o dataset a ser usado na ferramenta; escolha da melhor modelagem; avaliação dos resultados; apresentação em si dos resultados e modelos propostos.

O KDD em si, se mostra como a forma mais eficiente de se obter informações de base de dados educacionais, entretanto uma abordagem que misture alguns fatores da CRISP-DM, podem ser interessantes para se encontrar resultados mais ricos em informações visto que os processos são bem parecidos e com essa visão, o KDD, pode ficar ainda mais abrangente no

estudo e mais eficiente.

Com a mineração de dados de dois períodos diferentes com a mesma amostra, Fernandes et al. (2019) verificou que a localização e a distância da residência do aluno até o local de ensino, foi o maior fator na taxa de falha dos alunos e que esse baixo desempenho acadêmico aumenta a possibilidade do estudante evadir. Pascoal (2016) que avaliou a nota no exame de admissão no curso de computação e correlacionou dados acadêmicos e socieconômicos com intuito de se ranquear as mais impactantes, grau de instrução e profissão dos pais foram as mais impactantes.

No estudo de Manhães et al. (2011) vários algoritmos foram aplicados, desde aprendizado de regras, tabela de decisão, árvore de decisão, modelos de regressão logística, rede neural artificial e modelos probabilísticos. Entretanto nesse caso o *OneR*, foi o algoritmo usado por ter o menor custo operacional de bem preciso. No trabalho Kaur et al. (2015) vários algoritmos foram aplicados, mas nesse uma rede neural artificial, foi mais eficiente.

Já Pascoal (2016), *Naive Bayes* foi o método mais eficiente, mostrando que fatores socioeconômicos que mais tiveram impacto na evasão dos alunos foram o grau de instrução e profissão dos pais (destaque para os das mães que mais impactavam estatisticamente). Esses diversos resultados aplicando diferentes tipos de algoritmos só demonstra que vários fatores influenciam os resultados pois cada base de dados e a forma que foram modelados os dados e como foram tratados vai influenciar diretamente no desempenho. Portanto é recomendado a aplicação de vários algoritmos com o intuito de se obter o melhor para determinada modelagem e geralmente esses softwares de mineração de dados já possuem ferramentas para implementar vários algoritmos como o *Weka* e o *RStudio*, amplamente usado por pesquisadores da área.

Um aspecto importante dos algoritmos é que ele tenha uma taxa de acerto alta, mas que também tenha uma taxa de erro baixa como propõe Malhães et al. (2011), atribuindo pontuação a áreas críticas do problema com o intuito de refinar ainda mais, pois atribuir que uma pessoa com alto risco de evasão não corre risco de evasão, seria um erro grave, mas atribuir evasão a uma pessoa que não tem esse risco, ainda sim seria um erro, porém mais brando.

Uma métrica usada para verificar a consistência de dados não-balanciados é o Coeficiente de Correlação de Matthews (*Matthews Correlation Coefficient - MCC)*, como é apontado por Boughorbel et al. (2017) e Powers (2011), pois é uma métrica que relaciona diretamente as classificações corretas e erradas como é mostrado na Fórmula 2.1 e pode ser usada para

verificar possíveis inconsistência nos dados.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2.1}$$

Na fórmula mostrada, TP significa *True Positive* (verdadeiro positivo) e TN seria *True Negative* (verdadeiro negativo), sendo essas variáveis os acertos do algoritmo. Entretanto a fórmula também engloba as classificações erradas que seria o FP que representa *False Positive* (falso positivo) e o FN é *False Negative* (falso negativo).

3 MATERIAIS E MÉTODOS

Nesse capítulo será apresentado os materiais e métodos usados na pesquisa desse trabalho que é uma pesquisa explicativa, por envolver mineração de dados, quantitativa, já que lida diretamente com as variáveis de uma base de dados e é um estudo transversal, já que é analisado dados de 2017.

Em 2018, foi disponibilizado na Plataforma Nilo Peçanha, resultados de algumas análises de dados referentes Institutos Federais do ano de 2017 que envolvem dados desde financeiros, dados dos servidores, até dados de eficiência acadêmica e das matrículas de alunos vinculados as instituições de ensino no ano estudado. Além dos resultados dos dados, a plataforma também disponibiliza os microdados brutos para realização de outras análises usando esses dados e a base de dados escolhida para a pesquisa, foi a referente as matrículas dos alunos que foi filtrada apenas com alunos de graduação.

Dentro dessa base escolhida, encontra-se variáveis referentes as matrículas e são elas: carga horária; carga horária mínima; código de matrícula no Sistec (Sistema Nacional de Informações da Educação Profissional e Tecnológica); raça do estudante; data do fim do previsto da matrícula; data de início do ciclo da matrícula; data de ocorrência da matrícula; eixo tecnológico do curso; fator esforço do curso; mês de ocorrência da mudança do estado da matrícula; modalidade de ensino do curso; nome do curso; fonte de financiamento da matrícula; renda familiar; sigla da instituição; sexo do aluno; situação da matrícula; sub eixo tecnológico; tipo do curso; tipo de oferta do curso; total de inscritos no processo seletivo; turno do curso; unidade de ensino; e quantidade de vagas ofertadas.

De todas as variáveis disponíveis, algumas foram consideradas menos relevantes em relação a evasão universitária e com isso foram desconsideradas com base no estudo de Fernandes et al. (2019), Kaur et al. (2015) e Kantorsky et al. (2016), sendo as variáveis que de fato foram usadas no estudo: carga horária; raça do estudante; eixo tecnológico do curso; mês de ocorrência da mudança do estado da matrícula; modalidade de ensino do curso; nome do curso; fonte de financiamento da matrícula; renda familiar; sigla da instituição; sexo do aluno; situação da matrícula; tipo do curso (apenas bacharelado e licenciatura); turno do curso; e quantidade de vagas ofertadas.

Esses microdados da plataforma, foram convertidos e inseridos em um Sistema de Gerenciamento de Banco de Dados (SGBD), no caso usado a ferramenta *MySQL Workbench* 8.0 para o tratamento dos dados e filtragem das variáveis escolhidas para análise e posteriormente

foram extraídos os dados e inseridos no *software* livre *Weka* versão 3.8.3 para a aplicação de algoritmos de classificação. A base final consistia de 169.326 estudantes, sendo desses 142.706 considereados alunos não-evadidos, e 26.620 de alunos que evadiram (aproximadamente 15.72% da base).

Os algoritmos de classificação envolvendo árvores de decisão usarão parte dos dados para treinamento, escolhidos aleatoriamente pelo próprio *Weka* sendo os algoritmos usados: *Decision Stump; J48; Random Forest; Random Tree; REPTree*; e *Hoeffding Tree*. Alguns tipos de redes bayesianas também serão aplicadas (*Naive Bayes; Naive Bayes Multinomial; Naive Bayes Updateable; BayesNet*).

Com base no teste individual de cada algoritmo e as comparações posteriores, foi realizado o *Paired Sample T-Test*, em inglês para Teste-T para Amostra Pareada, um teste estatístico realizado no próprio *Weka* que compara todos os algoritmos entre si de forma estatísica e se a diferença entre eles tem relevância ou não, estatisticamente falando.

Para esse teste, o *Weka* usou uma parte da base de dados para treinamento (66%) e o restante para validação. Executou todos os algoritmos 10 vezes e com base nos resultados, gerou uma média que foi usada no desempenho entre os algoritmos, com nível de 0.05 de significância e todos com as configurações padrões do *Weka*.

4 RESULTADOS

Dentro do ambiente do *Weka* foi aplicado os algoritmos 13 algoritmos de classificação propostos em uma mesma máquina, em execuções individuais e foram extraídos os resultados de desempenho, tempo em segundos e a métrica MCC , como é demonstrado na Tabela 1.

Algoritmo	Desempenho	MCC	Tempo (s)
RandomTree	86.9193%	0.383	2.22
PART	86.1085%	0.325	49.87
REPTree	85.9437%	0.305	3.9
DecisionTable	85.6313%	0.278	15.01
J48	85.5433%	0.261	2.34
OneR	84.5298%	0.179	0.54
DecisionStump	84.5298%	0.179	1.2
HoeffdingTree	84.4956%	0.200	2.25
ZeroR	84.2788%	-	0.85
NaiveBayesMultinomialText	84.2788%	-	1.49
NaiveBayes	83.8082%	0.210	1.59
NaiveBayesUpdateable	83.8082%	0.210	1.34
BayesNet	83.8076%	0.210	2.01

MCC: *Matthews Correlation Coefficient*

Tabela 1 – Desempenho individual dos algoritmos do *Weka* na classificação de evasão

A Tabela 1 foi ordenada de forma descrescente em relação ao desempenho dos algoritmos, valor esse que é gerado pelo próprio *Weka* e leva em consideração a quantidade de classificações corretas pelos erros e com isso é gerado esse desempenho.

O MCC assim como o desempenho, também leva em consideração as taxas de acerto e erro nas classificações, por isso que algoritmos com o MCC maior tendem a ter um desempenho maior, mas isso não é necessariamente uma regra, como é o caso do *OneR* e o *Decision Stump* que possuem MCC menor que outros algoritmos, mas o desempenho é melhor, como é comparado ao *Hoeffding Tree*. Quanto mais próximo de 1, mais forte é a correlação e quanto mais próximo de 0, mais fraca são as relações.

Nem todos os algoritmos dentro do *Weka* permitem o cálculo do MCC, como é o caso do *ZeroR* e o *Naive Bayes Multinomial Text* que pela forma que são codificados esses algoritmos , algumas variáveis usadas nesse cálculo adquirem valor de zero, que é uma das restrições para o cálculo dessa métrica.

Random Tree foi o algoritmo com maior desempenho, próximo de 87% e também com o maior MCC, próximo de 0.4, o que demonstra a boa estruturação do algoritmo, que permitiu

vários acertos com o mínimo de classificações erradas, já que esse fator tem grande impacto nos valores.

Apesar de não ser o mais rápido, o algoritmo teve sim um desempenho bom em relação ao tempo, que levando em consideração o tamanho da base de dados, ter um desempenho alto em relação aos acertos em 2.22 segundos como mostra a Tabela 1, realmente demonstra o quão eficiente é o *Random Tree*.

Outro algoritmo com desempenho muito bom, foi o *PART* com desempenho e MCC bem próximos ao do *Random Tree*, 86% e 0.3 respectivamente, o que demonstra ser uma alternativa para o outro algoritmo em aplicações diferentes, mas a grande desvantagem, pelo menos na base de dados estudada, foi o tempo de resposta, quase 50 segundos, um tempo considerável se considerar que a média de tempo é de aproximadamente 6.50 segundos, valor que é diretamente influenciado pelo *outlier* (valor discrepante em um conjunto de dados).

Já a mediana que é um valor que não sofre de *outliers*, tem o valor de 2.01, valor muito mais condizente com a realidade dos outros 12 algoritmos em comparação com o tempo do *PART*. O cálculo da média de tempo apenas desses 12, seria de 2.895, valor que descreve melhor a distribuição dos valores, mesmo com um valor mais alto que seria o tempo do *Decision Table*.

Isso demonstra o quão discrepante é o valor do tempo do *PART* que faz a média aritmética saltar de 2.895 para aproximadamente 6.50 e a mediana que com os 12 valores é de 1.8 ao inserir o *outlier* passa para 2.01, variação mínima mostrando a eficiência da mediana em não ser influenciada por valores discrepantes em relação ao restante do conjunto de dados.

Como o *Random Tree* foi o algoritmo com melhor desempenho foi realizado o *Paired Sample T-Test* dos algoritmos comparandos com ele, como é mostrado na Tabela 2 que está ordenada de forma decrescente em relação ao desempenho individual da Tabela 1.

O *Weka* ao fazer esse teste compara diretamente todos os algoritmos em relação ao primeiro algoritmo, no caso o *RandonTree* na Tabela 2, e apresenta três tipos de saídas (ou inglês, *outputs*) no *Weka*: v,* ou em branco. O *output* do sinal de (+), significa que o algoritmo é melhor estatisiticamente, que o algoritmo principal sendo comparado. O *output* do sinal de (-), significa que é pior estatisiticamente, em relação ao algoritmo principal. E o resultado em branco é justamente quando não existe diferença entre os algoritmos comparados.

Na Tabela 2, a maioria dos algoritmos foram considerados pelo *Weka* como sendo piores que o algoritmo principal (*Random Tree*) com excessão do *PART*, o *REPTree* e o *J48*.

Algoritmo	Desempenho	+	-
RandomTree	87.43		
PART	88.06		
REPTree	88.08	✓	
DecisionTable	86.72		✓
J48	88.20	✓	
OneR	61.48		✓
DecisionStump	49.00		✓
HoeffdingTree	87.09		✓
ZeroR	40.04		✓
NaiveBayesMultinomialText	72.41		✓
NaiveBayes	72.40		✓
NaiveBayesUpdateable	72.39		✓
BayesNet	72.49		✓

+ = Melhor resultado estatístico, se comparado com o *RandonTree*
- = Pior resultado estatístico, se comparado com o *RandonTree*

Tabela 2 – Teste-T pareado dos algoritmos comparados com o *RandonTree*

Uma diferença clara nesse teste da Tabela 2 e o da Tabela 1, é que o desempenho oscila muito mais, ficando mais distribuido os resultados que antes ficavam entre 83% e 87%, agora já ficam entre 40% e 88% aproximadamente, o que demontra a importância desse teste que avalia todos os algoritmos juntos.

O algoritmo *PART* nos testes da Tabela 2, apresentou desempenho igual ao algoritmo (*Random Tree*), não numericamente e sim estatisticamente e por isso que ficou com as duas colunas com nenhum valor preenchido.

Algoritmo	Desempenho	+	-
J48	88.20		
RandomTree	87.43		✓
REPTree	88.08		
PART	88.06		
DecisionTable	86.72		✓

+ : Melhor resultado estatístico, se comparado com o *J48*
- : Pior resultado estatístico, se comparado com o *J48*

Tabela 3 – Teste-T pareado dos 5 melhores algoritmos comparados com o *J48*

Resolveu-se realizar um novo teste-t, apenas com os 5 melhores algoritmos e os ordenando de acordo com o desempenho obtido na Tabela 2 para verificar se essa tendência se mantinha, os resultados podem ser vistos na Tabela 3.

O *J48* se manteve com o melhor desempenho que o *RandomTree* e o *Decision Table*. *REPTree PART* manteve seu desempenho muito próximo em relação ao *J48* e por isso ficou

com seus resultados, visto que de forma estatística, não tem diferença entre eles.

Com o embasamento estatístico do *Weka*, então esses três algoritmos (*J48*, *REPTree* e *PART*) podem ser considerados como os mais eficientes para abordar classificações nessa base de dados, visto que tiveram os melhores desempenhos, como foi demonstrado na Tabela 3.

Outra diferença no desempenho entre as Tabelas 1 e 2, foi a piora no desempenho de alguns algoritmos, destaque para o *ZeroR* que de 84%, que era considerado um algoritmo mediano para o problema, passou para 40% se tornando o pior algoritmo para a realização dos estudos nessa base de dados.

Na pesquisa realizada por Manhães et al. (2011), a eficiência ficou entre 75 e 80%, e apenas 39% da base eram de estudantes que evadiram, média de eficiência próxima as encontradas pelo nosso estudo, com vários desempenhos superiores aos 80% em uma base que 15% eram de alunos que evadiram.

Manhães et al. (2011) também cita que o desempenho dos algoritmos foram muito próximos, o que se confirma nos testes individuais realizados neste estudo, mas ao se aplicar o teste-t de amostra pareada, o desempenho mudou bastante e com significância estatística, mostrando a importância dessa análise combinada dos algoritmos através desse teste.

Nos estudo de Paz (2017) e de Lanes (2018), o desempenho do J48 foi um próximo a 91%, mas era bases pequenas com aproximadamente 3000 e 4000 registros respectivamente, nosso estudo teve desempenho próximo do J48 com 88.20 % com uma base próxima de 179.000 registros.

Já o estudo de Oliveira (2015), J48 teve desempenho de aproximadamente 85 e 87%, visto que pegou a base de dados e dividiu em diversos *dataset* de acordo com o filtro escolhido e com isso mudava os resultados, mas a eficiência fica bem próxima.

5 CONCLUSÃO

Com base nos resultados dos testes dos algoritmos de classificação, notou-se um desempenho consideravelmente similar já que todos estão na faixa entre 83% e 87% de eficiência do desempenho dos algoritmos, o que demonstra o quão bem estruturada e os dados extraídos da Plataforma Nilo Peçanha, que permitiu uma comparação com maior precisão, já que alguma inconsistência na estruturação dos dados pode afetar o desempenho de alguns algoritmos, que com base nos resultados, isso não aconteceu.

O uso do MCC (*Matthews Correlation Coefficient*) como métrica de verificação de desempenho, se demonstrou bastante eficiente pois ele quanto mais alto, também resultava em desempenhos melhores, claro que isso não é regra, mas é um bom indicador.

A Plataforma Nilo Peçanha disponibiliza os dados de maneira muito bem estruturada e permite uma mineração de dados mais rápida pois acelera etapas como o pré-processamento de dados sem precisar fazer nenhuma formatação ou generalização, mas no nosso estudo, algumas generalizações foram feitas e algumas variáveis desconsideradas.

Ao se considerar os resultados do Teste-T para Amostra Pareada, os algoritmos mais eficientes foram *J48*, *REPTree* e *PART*, respectivamente, e apesar de não serem os três melhores nos testes individuais, mas ficando entre os 5 melhores por diferença mínima, nos testes que comparavam os algoritmos entre si com várias repetições, eles demonstraram sim uma melhora considerável de eficiência.

Essa diferença entre as performaces, pode ser devido a forma que o *Weka* aborda os algoritmos em suas diferentes aplicações, a *Explorer* e *Experimenter* onde diferentes funcionalidades são disponibilizadas aos algoritmos que foram executados em ambos os ambientes com as configurações padrões e com várias repetições, mas o nível de acerto oscilou bastante, seja aumentando ou diminuindo os valores de eficiência.

Apesar dessa consideração em relação aos desempenhos dos algoritmos, outros testes podem ser realizados utilizando outras variáveis de saída ou até mesmo expandindo a base de dados como integrando os dados de outros anos da Plataforma Nilo Peçanha que pode ser proposto trabalhos futuros.

Todos os algoritmos foram executados em apenas um ambiente, no *Weka* que é codificado em *Java* e o fato de se usar essa linguagem de programação, pode ter influenciado o desempenho e é algo que se pretende analisar em trabalhos futuros, seja por meio da integração do *Weka* com o *Python*, o uso da biblioteca do *Weka* de forma nativa diretamente no *Java*

sem usar a interface gráfica do *Weka*, ou o teste desses mesmo algoritmos codificados em outras linguagens de programação, são fatores que pretendemos comparar com os resultados atuais dessa pesquisa.

A aplicação de algoritmos de associação como o *Apriori*, são planos para trabalhos futuros, assim como expandir a base de dados, além do ano de 2017, pegar dados de outros anos e analisar se teve mudança nos padrões de evasão e se o desempenho dos algoritmos de classificação, se mantem mesmo expandindo a quantidade de dados.

O trabalho conseguiu fazer um bom levantamento de aspectos relevantes em relação a produção científica da área com base na revisão bibliográfica e os testes com os algoritmos de classificação tiveram resultados muito interessantes e podem servir de base para novos estudos e para o desenvolvimento de ferramentas que abordem essa problemática.

REFERÊNCIAS

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 3–13, aug 2011. ISSN 1414-5685. Disponível em: ⟨http://www.br-ie.org/pub/index.php/rbie/article/view/1301⟩.

BOUGHORBEL, S.; JARRAY, F.; EL-ANBARI, M. Optimal classifier for imbalanced data using matthews correlation coefficient metric. **PloS one**, Public Library of Science, v. 12, n. 6, p. e0177678, 2017.

DEHNING, P. et al. Achieving Environmental Performance Goals - Evaluation of Impact Factors Using a Knowledge Discovery in Databases Approach. **Procedia CIRP**, v. 48, p. 230–235, 2016. ISSN 22128271. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S2212827116300786⟩.

FERNANDES, E. et al. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. **Journal of Business Research**, v. 94, p. 335–343, 2019.

KANTORSKI, G. et al. Predição da Evasão em Cursos de Graduação em Instituições Públicas. **Anais do XXVII Simpósio Brasileiro de Informática na Educação (SBIE 2016)**, v. 1, n. CBIE, p. 906–915, 2016.

KAUR, P.; SINGH, M.; JOSAN, G. S. Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. **Procedia Computer Science**, v. 57, p. 500–508, 2015. ISSN 18770509. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S1877050915019018⟩.

LANES, M.; ALCÂNTARA, C. Predição de Alunos com Risco de Evasão: estudo de caso usando mineração de dados. **Anais do XXIX Simpósio Brasileiro de Informática na Educação (SBIE 2018)**, v. 1, n. CBIE, p. 1921–1925, 2018.

MANHÃES, L. M. B. et al. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. **Anais do XXII SBIE - XVII WIE**, p. 150–159, 2011. ISSN 2316-6533.

OLIVEIRA, J. G. de. **IDENTIFICAÇÃO DE PADRÕES PARA A ANÁLISE DA EVASÃO EM CURSOS DE GRADUAÇÃO USANDO MINERAÇÃO DE DADOS EDUCACIONAIS**. 86 p. Tese (Dissertação (Mestrado)) — Programa de Pós-graduação em Computação Aplicada, Universidade Tecnológica Federal do Paraná. Curitiba,, 2015.

PASCOAL, T. et al. Evasão de estudantes universitários: diagnóstico a partir de dados acadêmicos e socieconômicos. **Anais do XXVII Simpósio Brasileiro de Informática na Educação (SBIE 2016)**, v. 1, n. CBIE, p. 926–935, 2016.

PAZ, F.; CAZELLA, S. Identificando o perfil de evasão de alunos de graduação através da Mineração de dados Educacionais: um estudo de caso de uma Universidade Comunitária. **Anais dos Workshops do VI Congresso Brasileiro de Informática na Educação (CBIE 2017)**, v. 1, n. CBIE, p. 624–633, 2017.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. **Bioinfo Publications**, 2011.

RAMOS, J. L. C. et al. Um Modelo Preditivo da Evasão dos Alunos na EAD a Partir dos Construtos da Teoria da Distância Transacional. **Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017)**, v. 1, n. Cbie, p. 1236, 2017.

RIGO, S. J. et al. Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. **Revista Brasileira de Informática na Educação**, v. 22, n. 01, p. 168–177, 2014. ISSN 1414-5685. Disponível em: ⟨http://www.br-ie.org/pub/index.php/rbie/article/view/2423/2478⟩.

YUKSELTURK, E.; OZEKES, S.; TÜREL, Y. K. Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. **European Journal of Open, Distance and E-Learning**, v. 17, n. 1, p. 118–133, jul 2014. ISSN 1027-5207. Disponível em: ⟨http://content.sciendo.com/view/journals/eurodl/17/1/article-p118.xml⟩.

APÊNDICE A – MODELO PARA SUBMISSÃO NA REVISTA BRASILEIRA DE SISTEMAS DE INFORMAÇÃO - ISYS

Submission date: dd/month/year
Resubmission date: dd/month/year
Camera ready submission: dd/month/year

1st round notification: dd/month/year
2nd round notification: dd/month/year
Available online: dd/month/year
Publication date: dd/month/year

Section: [editorial | regular article | thematic issue | survey | innovation case | application | extended version]
[If thematic version, then add the title of the thematic issue]

Instructions for Authors of iSys Articles

Title: If the article is written in Portuguese, then place the title in English here

Author 1 Full Name[1], Author 2 Full Name[2], …, Author N Full Name[1]

[1]Department/Institute name – University name (Acronym)
City, State/District – Country

[2] Department/Institute name – University name (Acronym)
City, State/District – Country

[N] Department/Institute name – University name (Acronym)
City, State/District – Country

{local-part-1, local-part-n}@domain, local-part-2@domain, e-mailx

Abstract. *This template describes the style to be used in articles for iSys – Brazilian Journal of Information Systems. For papers in English, you should add just an abstract while for the papers in Portuguese, we also ask for an abstract in Portuguese ("resumo"). In both cases, abstracts should not have more than 10 lines and must be in the first page of the paper.*

Keywords*. Add from 3 to 7 title-cased keywords, separated by semi-colon.*

Resumo. *Este modelo descreve o estilo a ser usado na confecção de artigos para publicação na iSys – Revista Brasileira de Sistemas de Informação. É solicitada a escrita de resumo e abstract apenas para os artigos escritos em português. Artigos em inglês deverão apresentar apenas abstract. Nos dois casos, o autor deve tomar cuidado para que o resumo (e o abstract) não ultrapasse(m) 10 linhas (cada), sendo que ambos devem estar na primeira página do artigo.*

Palavras-Chave. *Se o artigo for escrito em português, acrescente de 3 a 7 palavras-chave com a inicial em maiúsculo, separadas por ponto-e-vírgula.*

1. General Information

To avoid unnecessary errors, you are strongly advised to use the 'spell-check' and 'grammar-check'. We ask that authors follow the guidelines explained in this template, to achieve the highest quality possible and a standard presentation of the manuscripts of the journal. Be advised that manuscripts in a technically unsuitable form may be rejected anytime by the editors or reviewers.

All articles submitted to iSys should be written in English or in Portuguese. The format paper should be A4 with single column, 3.5 cm for upper margin, 2.5 cm for bottom margin and 3.0 cm for lateral margins, without headers or footers. The main font must be Times, 12 point nominal size, with 6 points of space before each paragraph.

As it is an Information Systems (IS) journal, it is expected the work is supported by IS theories, deals with innovative technology artifacts and correctly follows a scientific methodology, presenting novel contributions to the state-of-the-art of the area. We also welcome systematic reviews that include clear research questions, a critical analysis, and directions for new challenges in the area. There is no limit of number of pages for the manuscript, although usually it ranges from 15-30 pages.

2. First Page

The first page must display the paper title (if the article is written in Portuguese, then the "Title:" followed by the title in English is also mandatory), the name and address of the authors, the abstract and keywords in English, and "resumo" and "palavras-chave" in Portuguese ("resumos" and "palavras-chave" are required only for papers written in Portuguese). The title must be centered over the whole page, in 16 point boldface font and with 12 points of space before itself. Author names must be centered in 12 point font, bold, all of them disposed in the same line, separated by commas and with 12 points of space after the title. Addresses must be centered in 12 point font, also with 12 points of space after the authors' names. E-mail addresses should be written using font Courier New, 10 point nominal size, with 6 points of space before and 6 points of space after.

The abstract and "resumo" (if it is the case) must be in 12 point Times font, indented 0.8cm on both sides. The word **Abstract** and **Resumo**, should be written in boldface and must precede the text. The keywords and "palavras-chave" (if it is the case) must follow the abstract and "resumo", respectively, be in 10 point Times font, indented 0,8 on both sides, with 6 points of space before and 12 points of space after. The word **Keyword** and **Palavras-chave**, should be written in boldface and must precede the text.

3. Sections and Paragraphs

Section titles must be in boldface, 13pt, flush left. There should be an extra 12 pt of space before each title. Attention to the section numbering. The first paragraph of each section should not be indented, while the first lines of subsequent paragraphs should be indented by 1.27 cm.

3.1. Subsections

The subsection titles must be in boldface, 12pt, flush left.

4. Figures and Captions

Figure and table captions should be centered if less than one line (Figure 1), otherwise justified and indented by 0.8cm on both margins, as shown in Figure 2. The caption font must be Helvetica, 10 point, boldface, with 6 points of space before and after each caption.

Figure 1. A typical figure

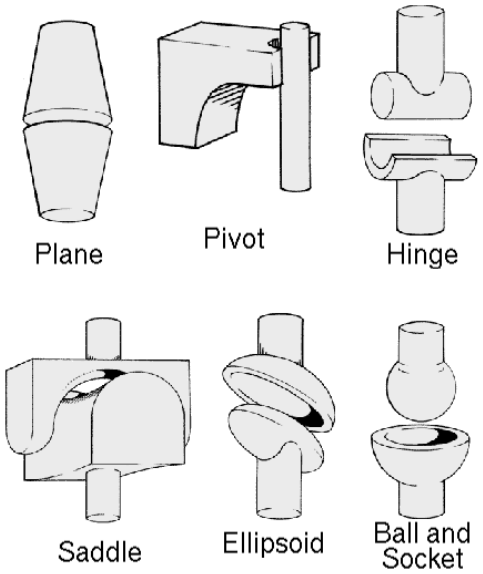

Figure 2. This figure is an example of a figure caption taking more than one line and justified considering margins mentioned in Section 5.

In tables, try to avoid the use of colored or shaded backgrounds, and avoid thick, doubled, or unnecessary framing lines. When reporting empirical data, do not use more decimal digits than warranted by their precision and reproducibility. Table caption must be placed before the table (see Table 1) and the font used must also be Helvetica, 10 point, boldface, with 6 points of space before and after each caption.

Table 1. Variables to be considered on the evaluation of interaction techniques

	Chessboard top view	Chessboard perspective view
Selection with side movements	6.02 ± 5.22	7.01±6.84
Selection with in-depth movements	6.29±4.99	12.22±11.33
Manipulation with side movements	4.66± 4.94	3.47±2.20
Manipulation with in-depth movements	5.71 ±4.55	5.37 ±3.28

5. Images

Images and illustrations may be colorful, but attention to the possibility of black-and-white, or gray tones, print. The image resolution on paper should be about 600 dpi. Do not include images with excessive resolution, as they may take hours to print, without any visible difference in the result.

6. Submission instructions

Fill up the metadata information with the most complete data possible as some indexing entities require detailed information and also for helping us to organize the journal data. Notice that if the paper is written in Portuguese, it is necessary to fill up the metadata in both languages (English and Portuguese).

Start with information for each author. Please provide the information according to the order of the authors. Although it is possible to reorganize the order, sometimes the system does not work well on performing reordering. Avoid shortening middle names (as it may make harder to identify other publications of the same author). In the URL field, choose to fill in with the curriculum URL (for those who have a Lattes curriculum, please fill in with the address to access the CV). Add the current affiliation of each author and fill in the country where the author works. Click on the "Add Author" button for adding other author, if that's the case.

Add the Title and Abstract of the manuscript. Attention to the "Form language" on the top of the form. If the manuscript is in English, then provide the title and abstract information in English. If the manuscript is written in Portuguese, then it is necessary to fill the title and abstract also in English. In this case, just change the form language on the top of the form and fill in the corresponding information. Notice the "Form language" is related to the information that is filled in the metadata. It does not change the language of the page.

For "Academic discipline", please consider the system suggestions, filling in the ones are related to the manuscript. For "Keywords", the authors are free to choose all the terms they consider important to indexing the manuscript. If the work is related to a specific geo-spatial area, please, provide the corresponding information in "Geo-spatial

coverage". Information about the research sample characteristics must be provided in the "Research sample characteristics" field. The research type, method and approach must be filled in the "Type, method or approach" field. The language field is related to the manuscript language. Therefore, this field is the same regardless of the form language. Attention to use semi-colon ";" to separate terms in each field of the indexing information.

Information about the contributors and supporting agencies must also be filled in the submission metadata. In case of approval, the authors may also add this information on the final version of the manuscript (usually it is added at the acknowledgement section). Again, attention to the "Form language". If the manuscript is in Portuguese, the authors are also required to fill in the metadata in English as in the title and abstract.

Provide a formatted list of references for works cited in the manuscript and separate individual references with a blank line. This field is the same regardless of the form language. Notice we follow the active link procedure, which is highly recommended to also include on the article. Then, when generating the PDF file, make sure the hyperlinks are active.

Acknowledgement

This template is based on the Brazilian Computer Society (SBC) template for papers.

References

Bibliographic references must be unambiguous and uniform. We recommend giving the author names references in brackets, e.g. [Knuth 1984], [Boulic and Renault 1991]; or dates in parentheses, e.g. Knuth (1984), Smith and Jones (1999).

The references must be listed using 12 point font size, with 6 points of space before each reference. The first line of each reference should not be indented, while the subsequent should be indented by 0.5 cm.

Whenever possible, include the links to the references (URL, DOI and Google Scholar link):

A. When the reference has a Link

- Make a clickable link on the respective URL (if you are using MS-Word, use the tool Insert Hyperlink, informing the URL).

B. Allow readers to search for the reference on Google Scholar

- Copy the title of the reference and put in between "%22", including the "+" character between each word: http://scholar.google.com/scholar?q= %22PASTE+TITLE+HERE%22&hl=en&lr=&btnG=Search
- If it is a common title, you may add the author, such as in: http://scholar.google.com/scholar?q=PASTE+AUTHOR+HERE+ %22PASTE+TITLE+HERE%22&hl=en&lr=&btnG=Search
- Or you may use the publication year (YEAR) to restrict the results, such as in: http://scholar.google.com/scholar?q=PASTE+AUTHOR+HERE+ %22PASTE+TITLE+HERE%22&hl=en&lr=&btnG=Search &as_ylo=YEAR&as_yhi=YEAR

- It is highly advisable to confirm if the link is correct (and if Google Scholar presents a correct result).
- Include the term "[GS SEARCH]" at the end of each reference and make "GS SEARCH" a hyperlink with the URL just created.

C. Allow readers to access references with DOI

- Add "doi:" followed by the DOI number. Then, make the DOI number a hyperlink using the corresponding URL (the URL can be created adding "http://doi.org/ in front of the DOI number).

D. Allow readers to access references with DOI

- Add "doi:" followed by the DOI number. Then, make the DOI number a hyperlink using the corresponding URL (the URL can be created adding "http://doi.org/ in front of the DOI number).

Boulic, R. and Renault, O. (1991) "3D Hierarchies for Animation", In: New Trends in Animation and Visualization, Edited by Nadia Magnenat-Thalmann and Daniel Thalmann, John Wiley & Sons ltd., England. [GS Search]

Dyer, S., Martin, J. and Zulauf, J. (1995) "Motion Capture White Paper", ftp://ftp.sgi.com/sgi/A%7CW/jam/mocap/MoCapWP_v2.0.html, December.

Holton, M. and Alexander, S. (1995) "Soft Cellular Modeling: A Technique for the Simulation of Non-rigid Materials", Computer Graphics: Developments in Virtual Environments, R. A. Earnshaw and J. A. Vince, England, Academic Press Ltd., p. 449-460. [GS Search]

Knuth, D. E. (1984), The TeXbook, Addison Wesley, 15th edition.

Pawlak, Z. (1981). Information systems theoretical foundations. Information Systems, 6(3), 205-218. doi: 10.1016/0306-4379(81)90023-5 [GS Search]

Smith, A. and Jones, B. (1999). On the complexity of computing. In *Advances in Computer Science*, pages 555–566. Publishing Press. [GS Search]