Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 31/10/2020; 1st round notif.: 18/12/2020; New version: 18/01/2021; 2nd round notif.: 02/02/2021;
Camera ready: 12/02/2021; Edition review: 19/02/2021; Available online: 12/03/2021; Published: 12/03/2021;

# Predição de Evasão Escolar na Licenciatura em Computação

***Title: School Dropout Prediction in Computer Science Degree***

*Hiago Oliveira de Jesus*
*Universidade do Estado do Amazonas*
*hodj.lic@uea.edu.br*

*Luis Cuevas Rodriguez*
*Universidade do Estado do Amazonas*
*lrodriguez@uea.edu.br*

*Almir de Oliveira Costa Junior*
*Universidade do Estado do Amazonas*
*adjunior@uea.edu.br*

## Resumo

*No primeiro ano de graduação e ao longo do curso de Licenciatura em Computação, os alunos expressam grandes dificuldades nas disciplinas de programação, seja pela ausência de conhecimento prévio, dificuldades na resolução de problemas, raciocínio lógico-matemático, abstração, entre outros fatores desconhecidos. Os dados dos históricos acadêmicos dos alunos, representam dados relevantes para prever o risco de evasão na Licenciatura em Computação da Universidade do Estado do Amazonas. Diante dos elevados índices de reprovações nas disciplinas do curso, foi levantada a seguinte hipótese "É possível prever os alunos evadidos na Licenciatura em Computação?". Este artigo apresenta uma mineração de dados educacionais, cujo objetivo é a previsão de alunos com risco de evasão. Esta pesquisa seguiu a metodologia de descoberta de conhecimento em base de dados, que consistiu em selecionar e preparar os dados para o treinamento do modelo preditivo de rede neural de múltiplas camadas. Os resultados obtidos com modelo preditivo foram avaliados, por meio de métricas de avaliação de desempenho, identificou-se com 98% de precisão os alunos com risco de evadir do curso.*

***Palavras-chave:*** *Evasão Escolar; Mineração de Dados Educacionais; Rede Neural Artificial.*

## Abstract

*In the first year of the degree and throughout the Degree in Informatics, students manifest great difficulties in the programming disciplines, whether due to the lack of prior knowledge, difficulties in solving problems, logical-mathematical reasoning, abstraction, among other unknown factors. The data of the students' school records represent relevant data to predict the risk of dropout in the Computer Science Degree Course of the educational institution of this study. In view of the high failure rates in the course subjects, the following hypothesis was raised "It is possible to predict dropout in the Computer Science Degree". This article presents an educational data mining, whose objective is to predict students at risk of dropout. This research followed the knowledge discovery methodology in the database, which consisted of selecting and preparing the data for training the predictive model of multilayer neural network. The results obtained with the predictive model were evaluated using performance evaluation metrics, with 98 % of correct answers identified students at risk of dropping out of the course.*

***Keywords:*** *School dropout; Educational Data Mining; Artificial Neural Network.*

*Cite as: de Jesus, H. O., Rodriguez, L. C., & Costa Junior, A. de O. (2021). School Dropout Prediction in Computer Science Degree (Predição de Evasão Escolar na Licenciatura em Computação). Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação – RBIE), 29, 255-272. DOI: 10.5753/RBIE.2021.29.0.255.*

# 1 Introdução

No ano de 1997, foi a primeira oferta do curso de Licenciatura em Computação na Universidade de Brasília (UnB), com o objetivo de formar um profissional habilitado para trabalhar com as mudanças tecnológicas, mediando o aprendizado através do uso do Computador (Luciano & Santos, 2013). Em 1999, o curso foi ofertado na Universidade de Santa Cruz do Sul (UNISC) e na Universidade Federal Rural de Pernambuco (UFRPE) (Falcão, Araújo, França, Andrade, & César, 2018) (Santos, 2017). Em 2002, o curso passou a compor as Diretrizes Curriculares Nacionais (DCN), referente aos cursos de graduação em Computação da Sociedade Brasileira de Computação (SBC, 2002).

O Licenciado em Computação é um profissional multidisciplinar, que integra outras áreas do conhecimento aos conceitos da Computação. No entanto, a carreira desse profissional não tem estabilidade e locais de trabalho plenamente definidos, em instituições de ensino e empresas (Castro & Vilarim, 2013). Segundo dados do Censo da Educação Superior do Ministério da Educação (Brasil, 2014), divulgados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, a evasão do curso de Ciência da Computação atingiu o índice de 60,2% em 2017.

Este artigo através do aprendizado supervisionado e por meio de redes neurais artificiais, criou um modelo preditivo para prever os alunos com risco de evadir do curso de Licenciatura em Computação da Universidade do Estado do Amazonas. O estudo foi dividido da seguinte maneira: a seção 2 discorre sobre o embasamento teórico deste trabalho. A seção 3 discute a metodologia empregada na mineração de dados educacionais e a realização do experimento. Por fim, são apresentados os resultados obtidos e as discussões dos resultados.

# 2 Trabalhos Relacionados

Os cenários de estudo de previsão de rendimento acadêmico estão relacionados ao contexto de abandono e reprovação. Desse modo, as pesquisas de mineração de dados educacionais, são incentivadas a elaborar modelos preditivos que proporcionem uma descoberta de conhecimento e auxiliem na tomada de decisão antecipada (Barber & Sharkey, 2012).

Em uma disciplina preliminar de programação, foi realizada uma análise de experiências com cinco cursos de Computação. Levando em conta, a estatística descritiva, foi analisado o rendimento dos alunos. Os autores afirmaram com base nos resultados do estudo, que a didática e metodologia empregada, era um ponto a ser melhorado (Coutinho, Lima, & Santos, 2017).

# 3 Referencial Teórico

## 3.1 Licenciatura em Computação

O uso das novas tecnologias no contexto educacional são ferramentas poderosas quando utilizadas em sala de aula. Apesar dos recursos investidos em sala de aula, pelo Plano Nacional de Educação na implantação de laboratórios de informática em escolas públicas (Brasil, 2014), existe uma

carência quanto à metodologia empregada no uso dos computadores em sala de aula. A carência de um profissional com as habilidades para o ensino dos conceitos de Computação nas escolas, nos níveis da educação básica. (Priecht & Pazeto, 2009).

A pioneira na oferta do curso de Licenciatura em Computação, foi a Universidade de Brasília no ano de 1997, seguido pela Universidade de Santa Cruz do Sul em 1999. No ano de 2002, o curso de Licenciatura em Computação, passou a compor as Diretrizes Curriculares Nacionais dos cursos de graduação em Computação da Sociedade Brasileira de Computação (SBC, 2002). O Licenciado em Computação é o profissional habilitado para integrar as mudanças tecnológicas na educação básica e atuar como mediador do aprendizado, através do desenvolvimento de novas tecnologias. No entanto, o curso de graduação encontra dificuldades no mercado de trabalho, já que o profissional de Computação não possui uma consistência em escolas e empresas, talvez pela falta de reconhecimento (Da Cruz, Becker, & Hinterholz, 2016).

O currículo de Licenciatura em Computação é referenciado pelos documentos das DCNs para cursos de Computação, definidos pelo Conselho Nacional de Educação/Conselho Pleno (CNE/CP, 2015), que trata da formação superior em cursos de Licenciatura, nos documentos do Conselho Nacional de Educação/Câmara de Educação Superior (CNE/CES, 2016) e referenciais de formação para cursos de graduação em Computação da Sociedade Brasileira de Computação (SBC, 2017). Os componentes curriculares do curso são divididos em formação básica, formação tecnológica e formação humanística. Na formação básica são abordados os fundamentos da Ciência da Computação, Matemática e áreas pedagógicas. Na formação tecnológica são discutidos os conceitos básicos de tecnologias de suporte. Na formação humanística são trabalhados os processos educacionais na prática do educador, o modo como um professor de Computação atua de modo multidisciplinar na Educação Básica (Matos, 2012).

### 3.2 Descoberta de Conhecimento em Base de Dados

A descoberta de conhecimento em bases de dados é um método de identificação de padrões em conjuntos de dados. Este conjunto de dados quando tratados e analisados, produzem informações relevantes para uma tomada de decisão (Feyyad, Shapiro, Smyth, & Uthurusamy, 1996).

O processo de descoberta de conhecimento utiliza dados estruturados e não estruturados (Silva & Silva, 2014). Primeiramente os dados são selecionados, organizados e tratados, antes de serem aplicados aos algoritmos de aprendizado de máquina (Han, Pei, & Kamber, 2011). Tendo como etapas, a preparação dos dados, lidando com dados ausentes, dados inconsistentes, dados redundantes e dados discrepantes. Em seguida, são definidos as técnicas e os algoritmos a serem aplicados em um determinado problema. Posteriormente é desempenhada uma busca por conhecimentos relevantes. Nas etapas da metodologia de descoberta de conhecimento em base de dados são realizados os processos de análise, interpretação e visualização dos dados.

### 3.3 Aprendizado de Máquina

O aprendizado de máquina é um processo que gera novos conhecimentos, a partir das instâncias contidas em uma base de dados (Alpaydin, 2010). A mineração de dados utiliza o aprendizado de máquina na obtenção de padrões presentes nos dados (Witten & Frank, 2005). As tarefas de aprendizado de máquina são separadas em tarefas descritivas. Nesta tarefa, os dados não

possuem rótulos e o algoritmo de aprendizado agrupa os dados, considerando as similaridades das instâncias do conjunto de dados. Nas tarefas preditivas, os dados possuem rótulos e o algoritmo de aprendizado tenta prever o rótulo das instâncias, através da avaliação de dados de treinamentos com dados de testes. Os algoritmos de aprendizado supervisionado realizam tarefas preditivas e os algoritmos de aprendizado não-supervisionado executam tarefas descritivas (Feyyad et al., 1996).


Figura 1: Os tipos de tarefas dos algoritmos de aprendizado de máquina.

### *3.3.1 Redes Neurais Artificiais*

As redes neurais artificiais consistem em neurônios artificiais com processamento interconectado, trabalhando em conjunto para produção de uma função de saída. A partir de uma função de ativação que limita a saída do neurônio, dentro de um intervalo de valores. As redes neurais artificiais do tipo *Feed Forward* são as redes neurais mais utilizadas, pois, possuem uma arquitetura dividida em camadas, que consiste em uma camada de entrada, ao menos uma camada oculta e uma camada de saída. Os neurônios das camadas ocultas são conectados as camadas anteriores e a camada seguinte, as arestas que interconectam as camadas associam-se a pesos individuais (Hamalainen & Vinni, 2010).

Uma particularidade da rede neural artificial do tipo *Feed Forward*, é a ramificação do sinal de cada neurônio, que é passado para a camada da frente. Quando a rede neural artificial apresenta mais de uma camada, essa rede é denominada de rede neural de múltiplas camadas. Essas redes recebem como entrada em cada camada, os valores de saída das camadas anteriores, delimitados pela função de ativação (Faceli, Lorena, Gama, & Carvalho, 2011). As redes neurais artificiais podem ser utilizadas em tarefas de classificação, regressão e agrupamento (Haykin, 2009).

## 3.4 Mineração de Dados

A mineração de dados realiza uma extração de conhecimento implícito em bases de dados para predição de padrões. (Witten & Frank, 2005). Nas tarefas de mineração de dados, quando utilizado um modelo do tipo descritivo, são retornados padrões descritivos por meio da avaliação do comportamento dos dados. Por outro lado, o modelo do tipo preditivo utiliza as instâncias da base de dados na predição de classes desconhecidas (De Amo, 2004).

A mineração de dados é definida como um processo de descoberta de novos conhecimentos em uma base de dados. (Zaki & Meira Junior, 2014). A descoberta de conhecimento está diretamente associada a mineração de dados (Bramer, 2007). A mineração de dados apresenta duas abordagens: a primeira abordagem é um processo de descoberta de conhecimento para encontrar novos padrões. A segunda abordagem é um processo de verificação de hipótese (Feyyad et al., 1996).

A análise exploratória de dados utilizada na mineração de dados, expõe uma base de dados compreensível, mantendo o máximo de instâncias e a identificação das características relevantes dos atributos pertencentes a base de dados (Kaski & Kohonen, 1996).

### *3.4.1 Mineração de Dados Educacionais*

A mineração de dados educacionais aplica a estatística e a mineração de dados no contexto educacional (C. Romero & Ventura, 2010). Uma área que abrange a análise e visualização de dados, desenvolvendo métodos para exploração de dados, provenientes de diferentes contextos educacionais (S. Romero C. Ventura, 2013). A mineração de dados educacionais aplica técnicas de aprendizado em dados oriundos de diferentes cenários (Cohen, Koedinger, & Matsuda, 2011).

### *3.4.2 Análise de Dados Educacionais*

A análise de dados educacionais é uma área da Informática em Educação, que desenvolve métodos de análise exploratória de dados (Daniel, 2016). Essa área de análise de dados, divide-se em três linhas de pesquisa: *Educational Data Mining*, *Learning Analytics* e *Academic Analytics*. As linhas de pesquisa diferem-se em relação ao modo como os dados são analisados.

A *Educational Data Mining* trabalha com a análise de dados derivados de ambientes de aprendizagem, realizando tarefas de predição, associação ou agrupamento. A *Learning Analytics*, realiza a análise de hipóteses, relacionadas as tarefas de aprendizagem do aluno, por meio da estatística. A *Academic Analytics* fundamenta-se nas abordagens de *Educational Data Mining* e *Learning Analytics*, analisando dados armazenados em sistemas educacionais administrativos e sistemas de gestão acadêmica (Silva, Silveira, Silva, Ramos, & Rodrigues, 2017).

A *Educational Data Mining* e *Learning Analytics* priorizam a análise de aprendizagem do aluno. Por outro lado, a *Academic Analytics*, concentra-se em dados acadêmicos provenientes de sistemas de instituições de ensino (Campbell & Oblinger, 2007; Baepler & Murdoch, 2007).

## 4 Metodologia

O processo descoberta de conhecimento em bases de dados seguiu as etapas da metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM). Uma metodologia fundamentada em seis processos, que viabilizam o desenvolvimento e a produtividade do modelo preditivo de mineração de dados (Chapman et al., 2000). Sendo assim, o processo de mineração de dados seguiu as etapas descritas na Figura 2.


Figura 2: Etapas da metodologia de mineração de dados.

As etapas da metodologia CRISP-DM relacionam-se e o resultado obtido em cada etapa, determina a próxima etapa a ser executada. Um metodologia com ciclo de vida recursivo, enquanto o modelo de mineração de dados não produz os resultados esperados ou aceitáveis (Chapman et al., 2000). Este trabalho seguiu a metodologia de mineração de dados da seguinte maneira:

- **Entendimento do Problema:** foram definidas as hipóteses a serem comprovadas.
- **Compreensão dos Dados:** foram descritas a origem da base de dados, bem como, o processo de coleta dos dados, a criação da base de dados e uma descrição das informações da base de dados.
- **Preparação dos Dados:** foram realizadas a limpeza da base de dados, tratando os dados ausentes, inconsistentes, redundantes e discrepantes e criado uma base de dados derivada.
- **Modelagem:** foram selecionados modelos de classificação. Em seguida, foram definidos as bases de dados, os parâmetros do modelo preditivo. Logo depois, foram escolhidas as métricas de avaliação de desempenho do modelo de aprendizado de máquina.
- **Avaliação:** foi avaliado com métricas de avaliação de desempenho o modelo preditivo.
- **Aplicação:** discutiu-se, uma proposta de utilização dos resultados alcançados.

### 4.1 Entendimento do Problema

Tendo em vista, os resultados da mineração de dados foram levantadas as seguintes hipóteses.

- Hipótese 1: "*A frequência do aluno em uma disciplina possui correlação com sua média?*"
- Hipótese 2: "*As disciplinas do primeiro ano de curso mais reprovam do que aprovam?*"
- Hipótese 3: "*As disciplinas que exigem uma boa prática de programação possuem muitas reprovações, faltas e trancamentos?*"
- Hipótese 4: "*É possível prever os alunos evadidos na Licenciatura em Computação?*"

### 4.2 Compreensão dos Dados

Os dados utilizados neste estudo foram coletados do sistema *Lyceum*[6] da Universidade do Estado do Amazonas. Coletou-se 136 arquivos, referentes aos 136 históricos acadêmicos dos alunos de Licenciatura em Computação, com a matricula ativa até 2019/1. Os arquivos foram tratados e agrupados, produzindo uma base de dados composta por 3396 instâncias e 8 atributos.

Os atributos *Matrícula*, *Ano/Período*, *Código*, *Disciplina* e *Situação* representam dados qualitativos. Os atributos *Carga Horária*, *Média* e *Frequência* retratam dados quantitativos. As escalas nominais e racionais predominam em comparação à escala ordinal (Tabela 1).

[6]https://www.lyceum.com.br

Tabela 1: Caracterização dos atributos da base de dados.

| Atributo | Tipo | Escala |
|---|---|---|
| Matrícula | Qualitativo | Nominal |
| Ano/Período | Qualitativo | Ordinal |
| Código | Qualitativo | Nominal |
| Disciplina | Qualitativo | Nominal |
| Carga Horária | Quantitativo discreto | Racional |
| Média | Quantitativo contínuo | Racional |
| Frequência | Quantitativo contínuo | Racional |
| Situação | Qualitativo | Nominal |

### 4.3 Preparação dos Dados

Os problemas mais comuns encontrados em base de dados são a presença de dados ausentes, dados discrepantes, dados inconsistentes e dados redundantes. Muitas vezes, é preciso detectar antecipadamente esses problemas e realizar o devido tratamento. Em alguns casos, dependendo do problema é necessário a exclusão dos dados problemáticos (Faceli et al., 2011). Nesse caso, para manter-se a integridade e qualidade dos dados, realizou-se a limpeza da base de dados.

O conjunto de dados apresenta 3397 instâncias, das quais 3396 instâncias caracterizam cada instância do conjunto de dados e 1 instância que identifica os rótulos dos atributos. Em relação aos tipos de atributos, o conjunto de dados dispõe de 3 atributos numéricos e 4 atributos categóricos.

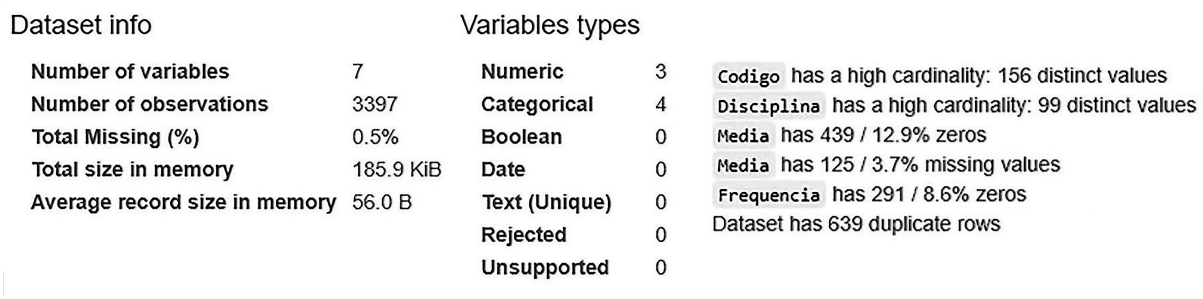


Figura 3: Relatório de características da base de dados.

Uma inconsistência foi identificada no relatório de características da Figura 3. A base de dados apresenta 99 disciplinas e 156 códigos, não necessariamente uma inconsistência. Devido a presença de diversos currículos acadêmicos na base de dados, uma determinada disciplina pode conter mais de um código de identificação. Em razão do *Código* apenas identificar uma disciplina, realizou-se a sua remoção da base de dados.

No relatório foi observada a presença de 125 instâncias do atributo *Média* contendo dados ausentes. Para verificar a importância dessas instâncias, realizou-se uma exploração dos dados mais detalhada. Na qual foi identificado 117 instâncias que correspondiam as ocorrências de disciplinas trancadas e 8 instâncias de disciplinas canceladas. Manteve-se as 125 instâncias e foi realizado o tratamento dos dados ausentes. Dentre as abordagens no tratamento de dados ausentes, a solução encontrada foi preenchimento dos dados faltantes com o valor (0.0).

A matriz curricular do curso de Licenciatura em Computação da Universidade do Estado do Amazonas, passou por atualizações ao longo dos anos. Algumas nomenclaturas de disciplinas

foram atualizadas. No entanto, mesmo com as alterações do currículo do curso, as nomenclaturas das disciplinas continuavam registradas na base de dados com nomenclaturas antigas.

O tratamento das redundâncias de nomenclaturas de 99 disciplinas, consistiu na remoção de acentuações. As acentuações incoerentes faziam com que uma disciplina fosse considerada como uma disciplina distinta. Em seguida, foi elaborado um arquivo CSV contendo as disciplinas do curso de Licenciatura em Computação da universidade e suas equivalências. Para identificação das disciplinas equivalentes, foi consultado os códigos das disciplinas dos currículos anteriores do curso. Mesmo com o tratamento das disciplinas equivalentes, foi verificado que as nomenclaturas continham caracteres não alfanuméricos, tornando as disciplinas redundantes. Esse problema foi tratado através de expressões regulares, haviam 99 disciplinas e restaram 63 disciplinas sem redundâncias.

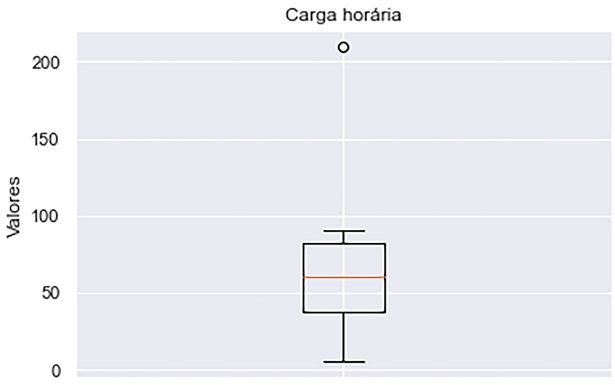


Figura 4: Identificação de Outlier no atributo carga horária.

Um gráfico do tipo *Boxplot* dos atributos numéricos foi elaborado, identificou-se um possível *outlier* no atributo carga horária (Figura 4). Em seguida, foi calculado os valores mínimo (e.g., Equação 1) (*min*=5) e máximo (e.g., Equação 2) (*max*=210) da carga horária. Esses valores apresentavam um afastamento relevante, em comparação aos demais valores. Os demais atributos foram mantidos e o atributo *Carga Horária* foi removido da base de dados.

$$Limite_{inferior} = Media - (2 \times Desvio.Padrao) \tag{1}$$

$$Limite_{superior} = Media + (2 \times Desvio.Padrao) \tag{2}$$

Os valores do atributo *Ano/Período* no formato (Fev/18 a Jun/18, Ago/18 a Dez/18) foram transformados em (2018.1, 2018.2). O atributo *Matricula* foi convertido em números inteiros ordinais no intervalo [0-136]. A base de dados percorreu diversos tratamentos, resultando em um conjunto de dados sem ruídos, com 3396 instâncias e 6 atributos.

Os modelos preditivos utilizam atributos preditores e um atributo alvo para a tarefa de predição. Nesse sentido, foram criados atributos derivados. Essa criação dos atributos derivados consistiu em produzir novas instâncias, para cada período cursado pelos 136 alunos registrados na base de dados. Diante disso, foram criados novos atributos (*Tempo, Aprovações, Reprovações,*

| | ID_aluno | ano_periodo | disciplina | media | frequencia | situacao |
|---|---|---|---|---|---|---|
| **2861** | 79 | 2018.1 | Introducao a Computacao | 4.8 | 83.3 | Rep Nota |
| **1759** | 36 | 2018.2 | Fundamentos de Software Educacional | 0.0 | 0.0 | Trancado |
| **2149** | 49 | 2017.1 | Metodologia de Pesquisa e Pratica Pedagogica | 8.1 | 100.0 | Dispensado |
| **2308** | 55 | 2017.2 | Matematica Discreta | 0.3 | 53.3 | Rep Freq |
| **101** | 1 | 2012.1 | Avaliacao de Aprendizagem | 9.5 | 100.0 | Aprovado |

Figura 5: Exemplos de atributos e instâncias da base de dados, ao final do processo de preparação dos dados.

*Trancamentos, Cancelamentos, Cursadas e Evasão*), o tempo de curso do aluno, a quantidade de aprovações, reprovações, trancamentos, cancelamentos e o total de disciplinas cursadas, incluindo todas as situações (Aprovado, Cancelado, Dispensado, Reprovado e Trancado). Em seguido, baseando-se nos atributos derivados, criou-se o atributo alvo (*Evasão*, que classifica os alunos em não evadido ou evadido. Os alunos enquadrados em ao menos um dos critérios (Tabela 2), foram considerados como evadido (1), caso contrário, não evadido (0).

Tabela 2: Critérios considerados na classificação de evasão.

| Critério | Condição | Situação |
|---|---|---|
| 1 | Reprovações e/ou trancamentos superior a 60% no período. | Evadido |
| 2 | Tempo de curso excedido, superior a 6 anos. | Evadido |
| 3 | Trancamento de todas as disciplinas matriculadas no período. | Evadido |

A criação da base de dados derivada, resultou em 733 instâncias e 9 atributos, sendo 2 atributos da base de dados original e 7 atributos derivados (Figura 6).

| | ID_aluno | Periodo | Tempo | Aprovacoes | Reprovacoes | Trancamentos | Cancelamentos | Cursadas | Evasao |
|---|---|---|---|---|---|---|---|---|---|
| **230** | 16 | 2015.1 | 1 | 2 | 3 | 0 | 0 | 5 | 0 |
| **490** | 48 | 2019.1 | 2 | 5 | 0 | 1 | 0 | 6 | 0 |
| **486** | 48 | 2017.2 | 0 | 6 | 1 | 0 | 0 | 7 | 0 |
| **129** | 8 | 2014.1 | 1 | 4 | 1 | 0 | 0 | 5 | 0 |
| **510** | 52 | 2018.2 | 1 | 0 | 4 | 0 | 0 | 4 | 1 |

Figura 6: Base de dados para predição de evasão.

## 4.4 Modelagem

Na etapa de preparação de preparação dos dados foi obtido duas bases de dados: a base original utilizada nas hipóteses (1, 2 e 3) e a base de dados derivada empregada na hipótese (4).

Os modelos preditivos (Decision Tree, Random Forest, Multilayer Perceptron, Support Vector Machine, Adaboost, Stochastic Gradient Descent, Naive Bayes) foram selecionados para teste de desempenho. Em seguida, foi realizada uma busca em grade para seleção dos melhores parâmetros do modelo preditivo com o melhor desempenho.

### 4.5 Avaliação

Existem quatro possibilidades de classificação de uma instância, como não evadido ou evadido:

- **Verdadeiro Positivo - VP**: a instância é rotulada como evadido e a instância é predita corretamente como evadido, conta-se como um verdadeiro positivo.
- **Falso Negativo - FN**: a instância é rotulada como não evadido mas a instância é predita incorretamente como evadido, conta-se como um falso negativo.
- **Verdadeiro Negativo - VN**: a instância é rotulada como não evadido e a instância é predita corretamente como não evadido, conta-se como um verdadeiro negativo.
- **Falso Positivo - FP**: a instância é rotulada como não evadido mas a instância é predita incorretamente como evadido, conta-se como um falso positivo.

O classificador pode ser avaliado por meio de métricas de avaliação de desempenho:

- **Acurácia**: a taxa de acertos é obtida pela razão entre a soma dos acertos das classes (não evadido, evadido) e o número total de exemplos de instâncias classificadas.
- **Precisão**: o percentual de acertos de verdadeiros positivos, ou seja, o número total de exemplos classificados como verdadeiros positivos e falsos positivos.
- **Revocação**: a taxa de verdadeiros positivos, ou seja, o percentual de verdadeiros positivos corretamente classificados.
- **F-Score**: a qualidade do modelo, indicada para conjuntos de dados que possuem classes desbalanceadas.

$$Accuracy = \frac{VP + VN}{Total.Exemplos} \tag{3}$$

$$Precision = \frac{VP}{VP + FP} \tag{4}$$

$$Recall = \frac{VP}{VP + FN} \tag{5}$$

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

### 4.6 Aplicação

Diante da ausência de apoio a interpretação dos dados acadêmicos de Licenciatura em Computação da Universidade do Estado do Amazonas. Para trabalhos futuros propõe-se o desenvolvimento de sistema de gerenciamento educacional, tendo como público alvo professores do curso.

O processo de desenvolvimento do sistema seguirá a metodologia *ICONIX*, dividida nas etapas de análise de requisitos, projeto preliminar e projeto. Primeiramente, será realizado o levantamento de requisitos e elaborado o protótipo de interface. Em seguida, será produzido os diagramas de caso de uso, diagramas de classe e os diagramas de sequência. Por fim, serão definidos as tecnologias e ferramentas a serem utilizadas na implementação do produto final.

## 5 Resultados

### 5.1 A frequência do aluno em uma disciplina possui correlação com sua média?

A correlação entre a frequência e a média do aluno foi calculada pelo coeficiente de correlação de Pearson (e.g., Equação 7), uma equação que expressa o grau de correlação entre duas variáveis, através de valores no intervalo de [-1,1] (Silvestre, 2007).

$$Pearson = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{7}$$

O valor obtido com o cálculo do coeficiente de correlação (0.654298) entre as variáveis média e frequência, reflete uma relação linear positiva, ou seja, a medida que cresce a assiduidade, aumenta a média e quando decresce a assiduidade, diminui a média.

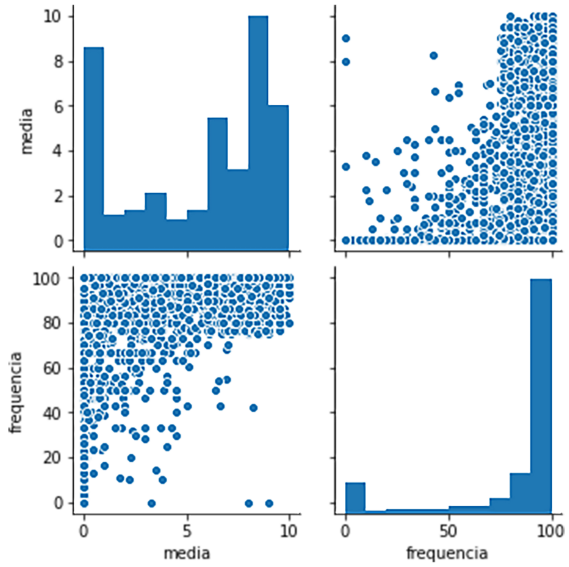


Figura 7: Matriz de correlação entre as variáveis.

Uma matriz de correlação foi construída contendo as distribuições (2º e 4º quadrante) dos

valores das variáveis média e frequência e as correlações (1º e 3º quadrante) entre as variáveis (Figura 7). As variáveis crescem ou diminuem ao mesmo tempo, pode-se afirmar a hipótese.

### 5.2 As disciplinas do primeiro ano de curso mais reprovam do que aprovam?

Tabela 3: Disciplinas do primeiro ano de curso.

| ID | Período | Disciplina |
|---|---|---|
| 01 | 1º | Filosofia da Educação |
| 02 | 1º | Introdução a Programação de Computadores |
| 03 | 1º | Introdução a Computação |
| 04 | 1º | Cálculo I |
| 05 | 1º | Legislação e Organização da Educação Brasileira |
| 06 | 2º | Álgebra Linear I |
| 07 | 2º | Matemática Discreta |
| 08 | 2º | Programação de Computadores e Algoritmos |
| 09 | 2º | Probabilidade e Estatística |
| 10 | 2º | Psicologia da Educação |
| 11 | 2º | Português Instrumental |

As disciplinas (ID 02, ID 04, ID 06, ID 07, ID 08) são as disciplinas que mais reprovam os alunos. O primeiro ano de curso contempla 11 disciplinas, foi observado que 6 disciplinas possuem um número de aprovados maior que o número de reprovados. Levando em conta, a quantidade de disciplinas com mais reprovações, não é possível afirmar a hipótese. No entanto, as disciplinas de formação básica são as disciplinas com mais reprovados no primeiro ano de curso.

| | Disciplina | Aprovados | Reprovados |
|---|---|---|---|
| 0 | Filosofia da Educacao | 109 | 29 |
| 1 | Introducao a Programacao de Computadores | 85 | 101 |
| 2 | Introducao a Computacao | 95 | 60 |
| 3 | Calculo I | 66 | 113 |
| 4 | Legislacao e Organizacao da Educacao Brasileira | 110 | 22 |
| 5 | Algebra Linear I | 40 | 80 |
| 6 | Matematica Discreta | 37 | 125 |
| 7 | Programacao de Computadores e Algoritmos | 54 | 65 |
| 8 | Probabilidade e Estatistica | 44 | 37 |
| 9 | Psicologia da Educacao | 92 | 10 |
| 10 | Portugues Instrumental | 83 | 20 |

Figura 8: As disciplinas do primeiro ano de curso e o número de aprovados e reprovados por disciplina.

### 5.3 As disciplinas que exigem uma boa prática de programação possuem muitas reprovações, faltas e trancamentos?

As disciplinas que abordam os conceitos e práticas de programação, consideradas nesta hipótese foram 5 disciplinas. Para cada disciplina foi calculado a quantidade de alunos que trancaram a disciplina, a quantidade de aprovados e reprovados nas disciplinas (Figura 9). Para 3 disciplinas foi observado um número de reprovados por nota, frequência e trancamentos superior ao número de aprovados. Desse modo, não é possível afirmar a hipótese. No entanto, foi percebido que essas disciplinas são ministradas no 1º, 2º e 3º período do curso.

| | Disciplina | Aprovado | Rep Nota | Trancado | Rep Freq |
|---|---|---|---|---|---|
| 0 | Introducao a Programacao de Computadores | 85 | 80 | 2 | 21 |
| 2 | Programacao de Computadores e Algoritmos | 54 | 52 | 2 | 13 |
| 3 | Algoritmos e Estrutura de Dados I | 35 | 32 | 2 | 10 |
| 1 | Projeto de Programas | 40 | 16 | 0 | 9 |
| 4 | Algoritmos e Estrutura de Dados II | 20 | 6 | 3 | 4 |

Figura 9: Situações das disciplinas de programação.

### 5.4 É possível prever os alunos evadidos na Licenciatura em Computação?

Nesta etapa foi utilizada a base de dados derivada, contendo 733 instâncias e 9 atributos. Os atributos preditores e o atributo alvo foram delimitados. Os modelos preditivos (Decision Tree, Random Forest, Multilayer Perceptron, Support Vector Machine, Adaboost, Stochastic Gradient Descent, Naive Bayes) foram treinados com a base de dados derivada e avaliados. O modelo (Multilayer Perceptron) com o melhor desempenho de acurácia (0.937778) foi selecionado.

| | acuracia | precision | recall | f-score |
|---|---|---|---|---|
| DT | 0.801111 | 0.470588 | 0.941176 | 0.627451 |
| RF | 0.877778 | 0.447368 | 1.000000 | 0.618182 |
| MLP | 0.937778 | 0.447368 | 1.000000 | 0.618182 |
| SVM | 0.816667 | 0.414634 | 1.000000 | 0.586207 |
| ADA | 0.885556 | 0.414634 | 1.000000 | 0.586207 |
| SGD | 0.751111 | 0.000000 | 0.000000 | 0.000000 |
| NB | 0.361111 | 0.000000 | 0.000000 | 0.000000 |

Figura 10: Métricas de desempenho dos modelos preditivos..

A técnica de *Grid Search*[8] foi utilizada na seleção dos melhores parâmetros e hiper-parâmetros do algoritmo de rede neural artificial de múltiplas camadas *MLPClassifier*[9] (Tabela 4).

[8]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[9]https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

Tabela 4: Definição dos parâmetros da rede neural artificial.

| **Parâmetro** | **Inicialização** | *Descrição* |
|---|---|---|
| activation | 'relu' | Função de ativação das camadas ocultas |
| alpha | 0.0001 | Parâmetro de regularização de penalidade |
| hidden_layer_sizes | (19, 77, 115) | O número de neurônios nas camadas |
| learning_rate | 'constant' | A taxa de aprendizado constante |
| learning_rate_init | 0.001 | A taxa de aprendizagem para atualizações de peso |
| max_iter | 1000 | O número máximo de épocas |
| max_fun | 15000 | O número máximo de chamadas da função de perda |
| n_iter_no_change | 10 | O número máximo de épocas sem otimização |
| shuffle | True | Condição para embaralhar as amostras em cada iteração |
| solver | 'adam' | O otimização de pesos |
| verbose | False | Condição para imprimir mensagens de progresso |

O modelo preditivo foi treinado por 100 repetições, para eliminar um possível viés. A cada repetição de treinamento foi aplicada uma *Validação Cruzada* do tipo *KFold*[10], com *10 folds*. Em cada *k-fold* a base de dados foi dividida em 10 subconjuntos de dados, sendo 9 subconjuntos para treinamento e 1 subconjunto para avaliação do modelo. A cada subconjunto os dados foram escalonados no intervalo [0,1]. Por fim, a capacidade de generalização do modelo preditivo foi avaliada 1000 vezes (100 repetições x 10 folds). As métricas de avaliação de desempenho (*Acurácia*[10], *Precisão*[11], *Revocação*[12] e *F1-Score*[13]) foram utilizadas na avaliação do modelo preditivo. As métricas foram calculados, obtendo-se 100 valores de cada métrica. Em seguida, foi calculada a média dos valores de cada métrica (Tabela 5).

Tabela 5: Resultados das médias de avaliação desempenho do modelo *Multilayer Perceptron*.

| **Acurácia** | **Precisão** | **Revocação** | **F1-Score** |
|---|---|---|---|
| 0.987595520177712 | 0.9863270666819447 | 0.9873175612491307 | 0.9863634973440952 |

O comportamento do modelo foi observado, por meio de uma partição *holdout* com 70% dos dados para treinamento e 30% para teste. Os dados da partição foram escalonados no intervalo [0,1], aplicado ao modelo treinado, criando-se uma matriz de confusão (Figura 11).

A matriz de confusão expõe a quantidade de classificação para cada classe do modelo. A partição *holdout*, dividiu as 727 instâncias da base de dados derivada, em 508 instâncias para dados de treinamento e 219 instâncias para dados de teste. Das quais, 157 instâncias foram classificadas como não evadido e 62 instâncias como evadido. O modelo treinado classificou corretamente todas as 157 instâncias com o rótulo não evadido e todas 62 instâncias com o rótulo evadido. Desse modo, um modelo preditivo com métricas em torno de 98%, é capaz de prever os alunos com risco de evasão, no curso de Licenciatura em Computação da Universidade do Estado do Amazonas. Logo, pode-se afirmar a hipótese para o cenário deste estudo.

[10]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html
[10]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
[11]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html
[12]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html
[13]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

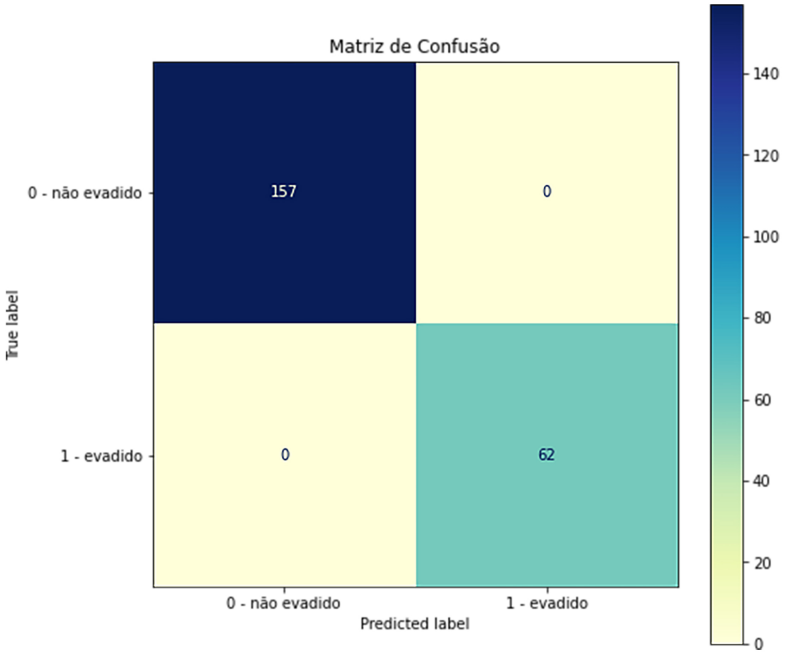


Figura 11: Matriz de Confusão.

## 6 Discussões dos Resultados

A partir dos resultados obtidos foram identificadas que as disciplinas (Introdução a Programação de Computadores, Cálculo I, Matemática Discreta) são as disciplinas do primeiro e segundo período de curso que mais possuem reprovações. As disciplinas de programação (Introdução a Programação de Computadores, Programação de Computadores e Algoritmos, Algoritmos e Estrutura de Dados I) são as disciplinas que mais possuem reprovações e trancamentos. Através da correlação de Pearson foi identificado, que a quantidade de reprovações tem relação com a assiduidade dos alunos na disciplina. A relação entre a média e a frequência implica que os alunos com baixa assiduidade, tendem a ter médias baixas e os alunos com médias altas possuem alta assiduidade.

Este estudo utilizou como base de dados, os históricos acadêmicos de 136 alunos do curso de Licenciatura em Computação, com matriculas ativas na instituição até 2019/1. Diante dos resultados, não é possível afirmar com toda a certeza, que um aluno evadiu do curso. Levando em conta, o desconhecimento de fatores externos, por exemplo, se o aluno trancou o curso e ingressou ou não em outra instituição. Tendo em vista, o contexto da base de dados foram definidos alguns critérios, como uma possibilidade de classificação de evasão. Desse modo, seguindo os critérios estabelecidos, para cada período os alunos foram classificados como evadidos e não evadidos. Para identificar os alunos com risco de evasão ao longo do curso, foi realizado um experimento com a rede neural artificial de múltiplas camadas e uma base de dados originada dos históricos acadêmicos. Os resultados obtidos, a partir da aplicação de métricas de avaliação de desempenho do modelo de aprendizado, mostraram que é possível prever os alunos com risco de evadir do curso de Licenciatura em Computação da Universidade do Estado do Amazonas.

## References


Alpaydin, E. (2010). Introduction to Machine Learning. *The MIT Press*. [GS Search]

Baepler, P., & Murdoch, C. J. (2007). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning*, *4*(2). doi: 10.20429/ijsotl.2010.040217 [GS Search]

Barber, R., & Sharkey, M. (2012). Course correction: Using analytics to predict course success. *International Conference on Learning Analytics and Knowledge*, 259-262. doi: 10.1145/2330601.2330664 [GS Search]

Bramer, M. (2007). *Principles of data mining*. New York: Cambridge University Press. Springer. [GS Search]

Brasil (2014). Planejando a próxima década: conhecendo as 20 metas do plano nacional de educação. *Ministério da Educação, MEC*. [Disponível em: http://portal.mec.gov.br/docman/junho-2013-pdf/13309-20metas-pne-lima/file]

Campbell, P., J. P. DeBlois, & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE review*, *42*(4), 40-57. [GS Search]

Castro, C. S., & Vilarim, G. O. (2013). Licenciatura em computação no cenário nacional: embates, institucionalização e o nascimento de um novo curso. *Revista Espaço Acadêmico*, *13*(148), 18-25. [GS Search]

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 step-by-step data mine guide*. CRISP-DM Consortium. [Disponível em: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2015-16/kdd/files/CRISPWP-0800.pdf]

CNE/CES (2016). Conselho Nacional de Educação/Câmara de Educação Superior. Resolução nº 5/2016. institui as diretrizes curriculares nacionais para os Cursos de Graduação em Computação. *Ministério da Educação. Processo 23001.000026/2012-95. Parecer CNE/CES número 136/2012. Homologação em Diário Oficial em 27 de outubro de 2016.*. [Disponível em: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=52101-rces005-16-pdf&category_slug=novembro-2016-pdf&Itemid=30192]

CNE/CP (2015). Conselho Nacional de Educação/Conselho Pleno. Resolução nº2, de 1º de julho de 2015, que define a Diretrizes Curriculares Nacionais para a formação inicial em nível superior (cursos de licenciatura, cursos de formação pedagógica para graduados e cursos de segunda licenciatura) e para a formação continuada. [Disponível em: http://portal.mec.gov.br/docman/agosto-2017-pdf/70431-res-cne-cp-002-03072015-pdf/file]

Cohen, N. L., Koedinger, W. K., & Matsuda, N. (2011). A machine learning approach for automatic student model discovery. *International Conference on Educational Data Mining, Eindhoven*, 44-53. [GS Search]

Coutinho, E. F., Lima, E. T. d., & Santos, C. C. (2017). Um panorama sobre o desempenho de uma disciplina inicial de programação em um curso de graduação. *Revista Tecnologias na Educação*. [GS Search]

Da Cruz, M. K., Becker, F., & Hinterholz, L. (2016). Carga horária prática na formação de professores de computação e informática educativa. *Anais do Workshop de Informática na Escola*. doi: 10.5753/cbie.wie.2016.698 [GS Search]

Daniel, B. K. (2016). *Big data and learning analytics in higher education: Current theory and practice*. Springer. doi: 10.1007/978-3-319-06520-5 [GS Search]

De Amo, S. (2004). Técnicas de mineração de dados. *Jornada de Atualização em Informática*. [GS Search]

Faceli, K., Lorena, A., Gama, J., & Carvalho, A. (2011). *Inteligência artificial: Uma abordagem de aprendizado de máquina*. LTC, Rio de Janeiro. [Disponível em: https://www.docsity.com/pt/inteligencia-artificial-uma-abordagem-de-aprendizado-de-maquina/4916888/]

Falcão, T. P., Araújo, D. R. B. d., França, R., Andrade, E. C. d., & César, F. (2018). Currículo da licenciatura em computaçao: uma proposta alinhada as novas diretrizes e demandas contemporaneas. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. doi: 10.5753/cbie.wcbie.2018.1108 [GS Search]

Feyyad, U. M., Shapiro, G. P., Smyth, P., & Uthurusamy, R. (1996). Data mining and knowledge discovery: making sense out of data. *IEEE Expert*, *11*(5), 20-25. doi: 10.1109/64.539013 [GS Search]

Hamalainen, W., & Vinni, M. (2010). *Classifiers for educational data mining*. Handbook of Educational Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. [GS Search]

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier Editora Ltda. [GS Search]

Haykin, S. (2009). *Neural networks and learning machines*. [GS Search]

Kaski, S., & Kohonen, T. (1996). Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. *In: CITESEER. Neural networks in financial engineering. Proceedings of the third international conference on neural networks in the capital markets*. [GS Search]

Luciano, A. P. C., & Santos, A. A. (2013). Caminhos do licenciado em computação no brasil: Estudo de mercado a partir de uma pesquisa com egressos. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. doi: 10.5753/cbie.sbie.2013.517 [GS Search]

Matos, G. F. B., E. e Silva (2012). Currículo de licenciatura em computação: uma reflexão sobre perfil de formação à luz dos referenciais curriculares da SBC. *Congresso da Sociedade Brasileira de Computação*. [GS Search]

Priecht, S. S., & Pazeto, T. A. (2009). Análise, sugestões e perspectivas de um curso de licenciatura em informática. *Anais do Workshop sobre Educação em Informática*. [GS Search]

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618. doi: 10.1109/TSMCC.2010.2053532 [GS Search]

Romero, S., C. Ventura (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12-27. doi: 10.1002/widm.1075 [GS Search]

Santos, W. O. (2017). Mulheres na computação uma análise da participação feminina nos cursos de licenciatura em computação. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. doi: 10.5753/cbie.wcbie.2017.814 [GS Search]

SBC (2002). Sociedade Brasileira de Computação. Currículo de Referência para Cursos de Licenciatura em Computação. Versão homologada na assembleia da SBC em julho de 2002

no Congresso da Sociedade Brasileira de Computação.
[Disponível em: https://www.sbc.org.br/documentos-da-sbc/summary/131-curriculos-de-referencia/763-curriculo-de-referencia-lic-versao-2002]

SBC (2017). Sociedade Brasileira de Computação. Referenciais de Formação para os Cursos de Graduação em Computação. Comissão de Educação da SBC.
[Disponível em: https://www.sbc.org.br/documentos-da-sbc/summary/131-curriculos-de-referencia/1165-referenciais-de-formacao-para-cursos-de-graduacao-em-computacao-outubro-2017]

Silva, L. A., & Silva, L. (2014). Fundamentos de mineração de dados educacionais. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. doi: 10.5753/cbie.wcbie.2014.568 [GS Search]

Silva, L. A., Silveira, I. F., Silva, L., Ramos, J. L. C., & Rodrigues, R. L. (2017). Ciência de dados educacionais: definições e convergências entre as áreas de pesquisa. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. doi: 10.5753/cbie.wcbie.2017.764 [GS Search]

Silvestre, A. L. (2007). *Análise de dados e estatística descritiva*. Escolar Editora. [GS Search]

Witten, I. H., & Frank, E. (2005). Practical machine learning tools and techniques. *Morgan Kaufmann Series in Data Management Systems*. [GS Search]

Zaki, M. J., & Meira Junior, W. (2014). *Data mining and analysis: Fundamental concepts and algorithms*. New York: Cambridge University Press. [GS Search]