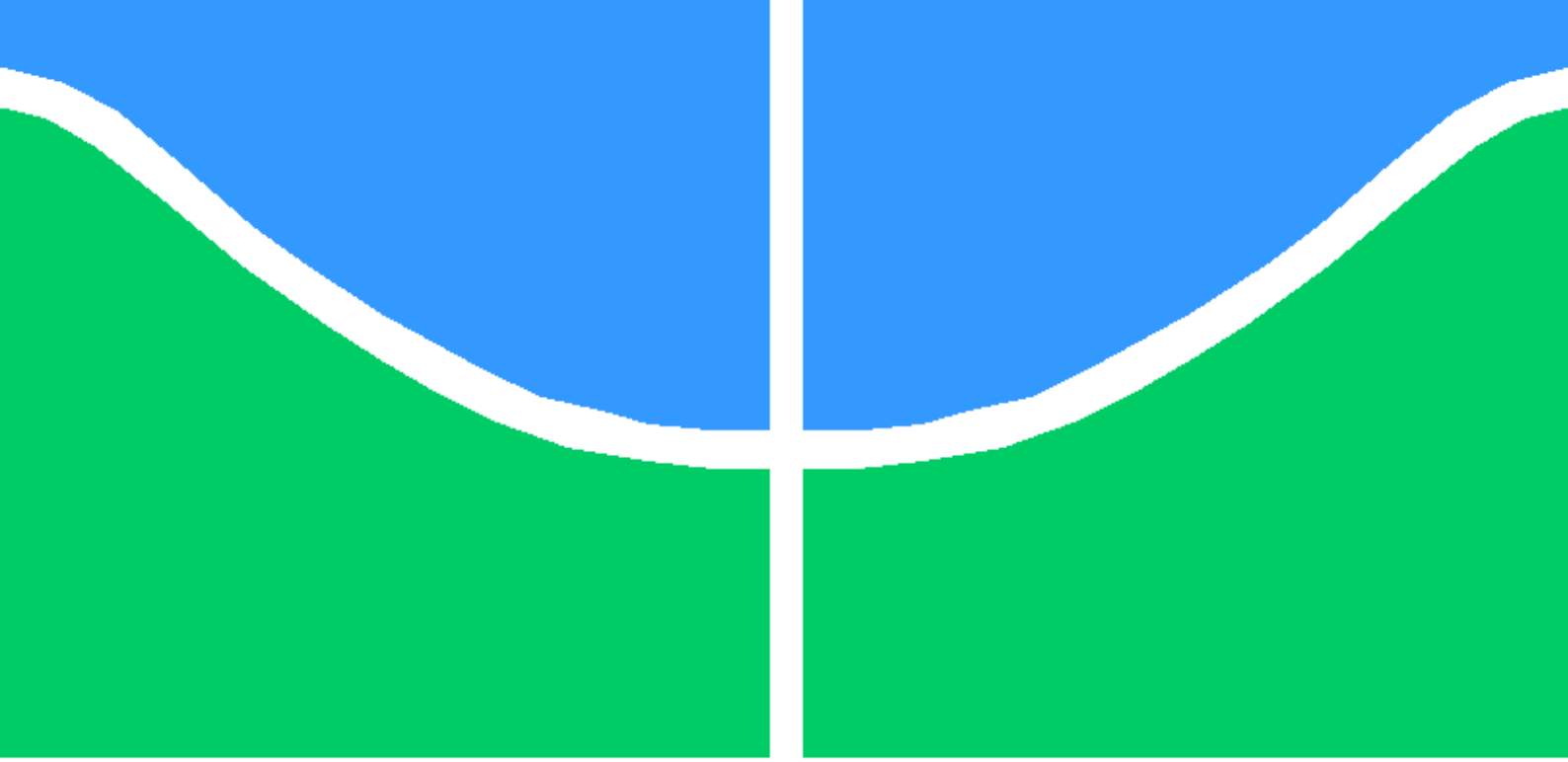


TRABALHO DE GRADUAÇÃO

Mineração de dados educacionais: um estudo de caso nos departamentos de Engenharia Elétrica e Ciência da Computação

Fernanda Amaral Melo

Luiz Fernando Neves de Araújo

Brasília, Novembro de 2019

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASILIA
Faculdade de Tecnologia

TRABALHO DE GRADUAÇÃO

Mineração de dados educacionais: um estudo de caso nos departamentos de Engenharia Elétrica e Ciência da Computação

Fernanda Amaral Melo

Luiz Fernando Neves de Araújo

Relatório submetido ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Engenheiro Eletricista

Banca Examinadora

Prof. Daniel Guerreiro e Silva, ENE/UnB _____
Orientador

Prof. José Edil G. de Medeiros, ENE/UnB _____
Co-Orientador

Prof. Guilherme N. Ramos, CIC/UnB _____
Examinador Externo

Prof. Levy Boccato, FEEC/UNICAMP _____
Examinador Externo

Those who stand for nothing fall for anything.

Alexander Hamilton

Agradecimentos

Agradeço a Deus por ter iluminado o meu caminho durante todos esses anos, por ter me oferecido a oportunidade de viver, evoluir, crescer e conhecer todas as pessoas que citarei abaixo.

Agradeço à minha família, o alicerce da minha história, que sempre esteve ao meu lado, em especial aos meus pais Suely de O. A. Melo e Edilson Fernando Melo que renunciaram às suas próprias conquistas para permitir que eu concluísse meus estudos. Agradeço também ao meu grande amor e companheiro de vida, André A. G. Perez, por todo o apoio fornecido durante a realização deste curso.

Agradeço a todos cujos caminhos já se encontraram com o meu, a quem me ajudou direta ou indiretamente nesta pesquisa, em especial aos meus amigos cuja presença definitivamente tornou a jornada dentro e fora da UnB mais afável, aos quais gostaria de citar nominalmente os meus sócios, que seguirão comigo nesta nova etapa da minha vida profissional, Caio Rondon B. de Carvalho, João Tribouillet M. de Menezes e Pedro Henrique L. da Costa.

Agradeço à Universidade de Brasília e ao Decanato de Planejamento, Orçamento e Avaliação Institucional, sem os quais não seria possível concluir este trabalho. Aos nossos professores orientadores, Daniel Guerreiro e Silva e José Edil G. de Medeiros, pelos ensinamentos e apoio dados. E, por fim, agradeço ao meu grande amigo e companheiro de curso e pesquisa, Luiz Fernando N. Araújo, por todo o empenho e dedicação ao presente trabalho.

Fernanda Amaral Melo

Agradeço a todos e todas que passaram e passarão por minha vida, principalmente aos que estão aqui agora, a quem ajudou de forma direta ou indireta na produção dessa pesquisa, seja provendo conhecimento para que ela fosse realizada, seja apoiando emocionalmente e permitindo que eu a realizasse. Agradeço a você leitor que se interessou pelo conhecimento aqui produzido, espero que essa contribuição que cá coloco possa de alguma forma melhorar minimamente os objetos de estudo. De forma alguma essa pesquisa é uma produção individual, ela é fruto de tudo, todos e todas, os quais me colocaram na posição que estou hoje, a esses, gostaria de tecer mais alguns agradecimentos sobre essas pessoas. Agradeço a Universidade de Brasília (UnB) por ter possibilitado essa pesquisa, bem como ao Decanato de Planejamento, Orçamento e Avaliação Institucional (DPO). Agradeço aos professores Daniel Guerreiro e Silva e José Edil G. de Medeiros pela oportunidade de trabalhar com um tema tão interessante e por todo apoio e orientação dados. Agradeço aos colegas que colaboram engrandecendo de forma significativa meus conhecimentos que usei nessa pesquisa, alguns que gostaria de destacar são Caio R. B. de Carvalho, João T. M. de Menezes e Thiago C. Pita. Agradeço a todos meus amigos que por todo esse período de pesquisa me confortaram, apoiaram e proveram suporte psicológico, entre os quais gostaria de destacar alguns como Bárbara Maria G. de Souza, Mariana R. Coutinho, Cristiele G. dos Santos, Marcus Vinicius B. Barbosa, Guilherme Henrique F. Ribeiro, Hugo P. G. Carvalho, Raphael C. N. Redigolo, Matheus Felipe N. Leal, Pedro Henrique L. da Costa, Gustavo V. Gianini, Gabriel T. de B. Tavares, Igor Moro C. S. Lousada, Camila de F. Dias e gostaria de agradecer novamente aos colegas, que aqui também coloco-os como amigos, Caio R. B. de Carvalho, João T. M. de Menezes e Thiago C. Pita. Agradeço a minha família por todas oportunidades que tive ao longo da minha vida bem como todo suporte que eles me proveram, nominalmente, gostaria de destacar meus pais Vanessa N. do N. de Araújo e Marcelo R. de Araújo, agradeço também a minha madrinha Paloma N do Nascimento e ao meu primo Kauwai L. T. Nobre. Agradeço também a todos os meus colegas que irão comemorar essa conquista e colaborarão, logo após a apresentação dessa pesquisa, no consumo de 1L de cerveja para cada página desse trabalho. Por último, gostaria de agradecer a Fernanda A. Melo por ser minha amiga, colega de curso e colega de pesquisa, sem a qual não seria possível concluir a produção dessa pesquisa dessa forma, a qual me orgulho muito de ter produzido e feito parte.

Luiz Fernando Neves de Araújo

RESUMO

A crescente preocupação com a eficiência do ensino, principalmente no nível superior, tem motivado a busca por formas de se compreender e aplicar melhor conceitos pedagógicos e fenômenos educacionais. Com isso, diversas áreas começaram a colaborar na busca de soluções para o processo educacional. Nesse contexto, a solidificação da área *Educational Data Mining* ao longo da década dos anos 2000 se mostra como uma ferramenta poderosa.

Neste trabalho, são apresentadas métricas estatísticas que permitem analisar informações sobre os estudantes do curso de Engenharia da Computação, sobre as disciplinas cursadas por esses estudantes e todas as disciplinas ofertadas pelos departamentos de Engenharia Elétrica e Ciência da Computação da Universidade de Brasília. Essas métricas buscam correlacionar informações, a fim de descobrir relações mais complexas e assim possibilitar a tomada de decisões em diversos âmbitos com base neste estudo. Além disso, foi desenvolvido um modelo preditivo da desistência dos estudantes que, de forma automática, estima a probabilidade de um determinado estudante concluir ou não a sua graduação.

O modelo preditivo desenvolvido, baseado em árvores de decisão, busca não só oferecer uma forma de se identificar estudantes com altas chances de sair do curso, mas também prover informações qualitativas sobre a correlação do desempenho dos estudantes ao longo dos semestres analisados e a sua desistência.

ABSTRACT

The rising concern about efficiency in education, especially regarding higher education, has stimulated searches for methods to better comprehend and apply pedagogical concepts and other phenomena related to the educational field. Accordingly, many academical fields began collaborating in order to reach solutions for the educational process. Thus, the consolidation of the *Educational Data Mining* field along the decade of the 2000s has presented itself as a powerful tool.

In this paper, statistical metrics are presented which facilitate the analysis of information regarding students of the computer engineering course, with respect to subjects these students participated and all other disciplines which are offered by both the Electric Engineering and the Computer Science departments of the Universidade de Brasilia. These metrics attempt to correlate information, in order to find out complex relations so that it is possible to make decisions in many ways that are influenced by this study. Also, a model was developed to predict student's dropping out, so in an automatic way, it works to figure out if a certain student would graduate or if he would drop out.

The predictive model created, is based on decision trees, not only intends to conceive a way to identify students with a high potential to drop out but it also provides qualitative information regarding the correlation between students' performance along the semesters analyzed and they effectively dropping out.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

Acrônimos

API	Application Programming Interface
BD	Banco de Dados
CAO	Comissão de Acompanhamento e Orientação
CEP	Código de Endereçamento Postal
CIC	Departamento de Ciência da Computação
CPF	Certificado de Pessoa Física
CRISP-DM	CRoss-Industry Standard Process for Data Mining
DM	Data Mining
DF	Distrito Federal
DPO	Decanato de Planejamento, Orçamento e Avaliação Institucional
EDM	*Educational Data Mining*
ENE	Departamento de Engenharia Elétrica
ENEM	Exame Nacional do Ensino Médio
GPA	Grade Point Average
KDD	Knowledge Discovery from Data
MOOCs	Massive Open Online Courses
PAS	Programa de Avaliação Seriada
PEC-G	Programa de Estudantes-Convênio de Graduação
SAA	Secretaria de Administração Acadêmica
SDSU	San Diego State University
SEMMA	Sample, Explore, Modify, Model, and Assess
Sisu	Sistema de Seleção Unificada
UnB	Universidade de Brasília

Capítulo 1

Introdução

A preocupação sobre como a educação é feita e os impactos dela na sociedade sempre foram de extremo interesse por governantes e por todos aqueles que se beneficiam direta e indiretamente do desenvolvimento educacional; Compreender é sempre o primeiro passo para se propor, então, melhorar qualquer tipo de processo. Nesse caso, a área que estuda os processos de aprendizado, a pedagogia, busca melhorar a relação entre como o conhecimento é transmitido e entender melhor as interações que acontecem durante esse processo.

Fora os fatores pedagógicos e científicos que visam o desenvolvimento educacional, existem razões financeiras que impactam diretamente no esforço coletivo pela compreensão dos fenômenos educacionais da sociedade. Nesse ponto, a educação como um todo vem se tornando um área que movimenta muito dinheiro: de acordo com o relatório (UNESCO, 2019), a estimativa para gastos anuais na área da educação é de $\$4,7$ trilhões de dólares no mundo, sendo majoritariamente percebido em países com um nível alto de renda. Esse valor é referente ao gasto de todos os níveis educacionais combinados, sendo que o nível universitário é responsável por cerca de 20% desse valor. Dessa forma, é perceptível que o ensino superior movimenta um mercado muito grande e, a partir disso, vários estudos tentam melhorar a efetividade da educação superior.

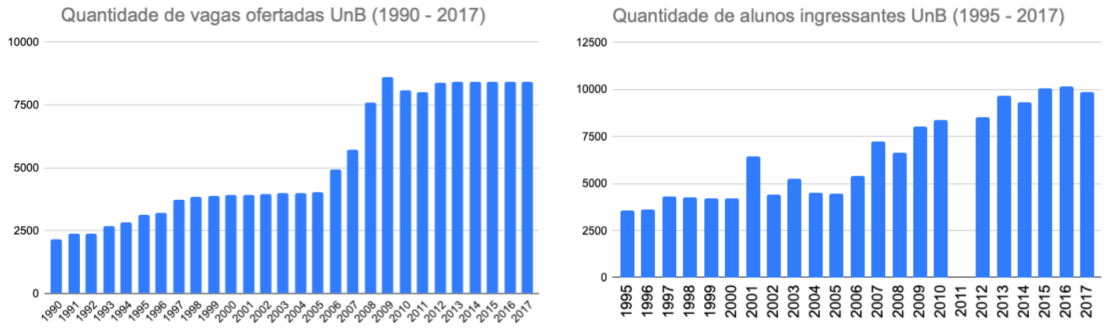
A produção e armazenamento de dados vêm crescendo ao longo do tempo. Estamos produzindo, armazenando e processando cada vez mais, ao mesmo tempo em que o conhecimento obtido a partir da análise de grandes bancos de dados (BD) vem quebrando diversas fronteiras interdisciplinares. A tarefa de produzir conhecimento a partir de um BD carrega consigo a necessidade de se entender as informações e o que pode ser feito com elas. Assim, é necessário cada vez mais pessoas capacitadas para interagir com ferramentas de ciência de dados nas áreas mais diversas e, dessa forma, conseguir trazer inferências interessantes e valorosas para esse campo.

Ciência de dados é uma área que permite o desenvolvimento de pesquisas avançadas para validação de fenômenos complexos. Uma das possibilidades de aplicação interdisciplinar dessa área é o melhor entendimento dos fenômenos na educação superior. Como se vê nos trabalhos de (YANG et al., 2015; FREEMAN et al., 2014; AUD; WILKINSON-FLICKER, 2013) a relação entre essas áreas pode trazer informações relevantes para o desenvolvimento da educação.

Analisando o caso da Universidade de Brasília (UnB), o crescimento da população universitária

pode ser explicado por uma política ativa de expansão de vagas. Utilizando-se as informações disponibilizadas pelo Decanato de Planejamento, Orçamento e Avaliação Institucional (DPO) nos relatórios (DECANATO DE PLANEJAMENTO, 2018b, 2013, 2008, 2002, 1998) é possível traçar algumas análises sobre esse crescimento, como é feito a seguir.

A Figura 1.1 mostra esse fenômeno, na qual podemos ver o crescimento do número de vagas ofertadas ao longo dos anos, bem como do número de estudantes que ingressaram na universidade. Percebe-se que houve um aumento significativo nas vagas ofertadas nos vestibulares, no Programa de Avaliação Seriada (PAS) da UnB e no Exame Nacional do Ensino Médio (ENEM), bem como na quantidade de ingressantes, alunos aprovados nos vestibulares, PAS, ENEM e em outras formas de ingresso. O número de vagas ofertadas subiu de 3.126 em 1995 para 8.439 em 2017, um aumento de aproximadamente 3 vezes $(269,96\%)$. Já o número de alunos ingressantes foi de 3.575 em 1995 para 9.878 em 2017, o que também define um aumento de aproximadamente 3 vezes $(276,63\%)$. Sem explicitarem o motivo, os relatórios em questão não apresentam valores para o número de ingressantes do ano de 2011.



(a) Vagas ofertadas de 1995 a 2017 pela UnB. (b) Alunos ingressantes na UnB de 1995 a 2017.

Figura 1.1: Distribuição ao longo dos anos das vagas ofertadas e alunos ingressantes.

Analisando os relatórios dos anos de 1998 e 2018 (DECANATO DE PLANEJAMENTO, 1998, 2018b), podemos ver que em 1995 existiam 53 opções de cursos e em 2017 um total de 153 opções, o que nos mostra, um aumento de cerca de 2,8 vezes $(288,67\%)$ na quantidade de opções de cursos na UnB. Percebe-se na Figura 1.2 que o aumento da população universitária ao longo dos anos, foi de 13.729 estudantes de graduação em 1995 para 39.624 em 2017, o que naturalmente condiz com o crescimento do número de vagas ofertadas e ingressantes ao longo dos anos, visto previamente na Figura 1.1. Um dado que pode ser usado para compreender melhor o crescimento da população universitária é o de número de formandos da graduação ao longo dos anos, que pode ser visto em 1.3a. Apesar do crescimento de quase 3 vezes $(276,63\%)$ ao longo dos anos, a relação entre o número de alunos que ingressam na faculdade e os formados é sempre menor do que um, conforme mostra a Figura 1.3b. Pode-se compreender que existem duas possíveis consequências desses dados, que seriam: a retenção de alunos na universidade, ou a saída de estudantes de outra forma, que não pela formatura, que no caso seria a desistência dos estudantes. Ambas as possibilidades não são desejáveis e reduzir tais índices seria uma ação interessante para a universidade.

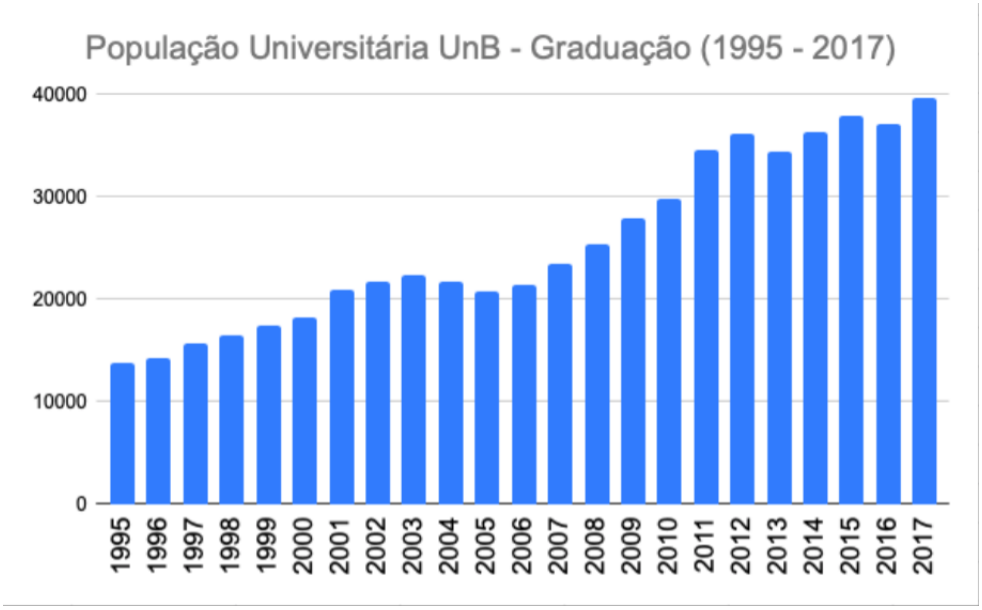
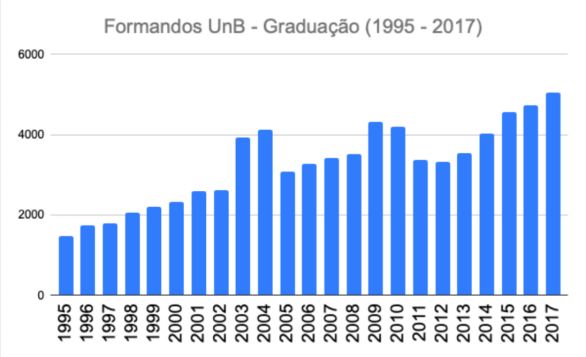
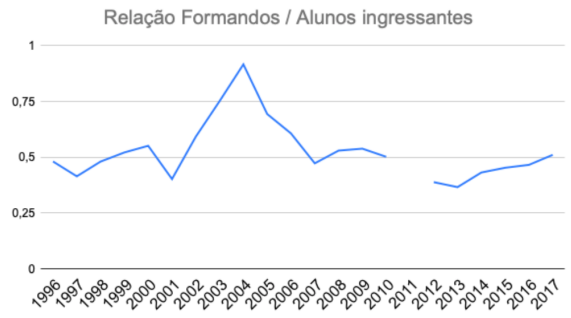


Figura 1.2: População universitária UnB - Graduação (1995 - 2017).



(a) Alunos ingressantes na UnB de 1995 a 2017.



(b) Formandos/Ingressantes UnB de 1995 a 2017.

Figura 1.3: Gráficos de formandos e relação formandos estudantes ingressantes UnB - Graduação (1995 - 2017)

1.1 Estrutura pedagógica da Universidade de Brasília

Como os objetos de estudo são relacionados à Universidade de Brasília (UnB), nesta seção serão disponibilizadas informações acerca do funcionamento de um curso, as quais são necessárias para a compreensão da pesquisa realizada e seu contexto.

Todos os cursos da Universidade de Brasília possuem em seu currículo créditos que deverão ser compostos por disciplinas obrigatórias e disciplinas optativas e podem incluir disciplinas de módulo livre. As disciplinas obrigatórias são as disciplinas que o aluno imprescindivelmente deve cursar, sem as quais ele não poderá se formar. Por sua vez, as disciplinas optativas são as disciplinas que fazem parte de uma lista, diferente para cada curso, em que os alunos podem escolher aquelas matérias que possuem maior afinidade para cursar, não sendo necessário que se curse alguma

em específico. No entanto o aluno deve cursar disciplinas optativas até concluir a quantidade de créditos mínimos para se formar em seu curso, ainda que o mesmo possa escolher quais disciplinas da lista pré-definida ele irá cursar. As disciplinas categorizadas como módulo livre são todas as outras disciplinas que não se encaixam nessas duas definições anteriores e o estudante pode integralizar até 24 créditos do total necessário com disciplinas desse tipo. É permitido que um aluno se forme sem cursar qualquer disciplina de módulo livre, obtendo os créditos exigidos apenas com os conjuntos de disciplinas obrigatórias e optativas.

Dentro de uma disciplina o estudante recebe uma menção que indica o seu desempenho ao longo do semestre. Mais especificamente, na Universidade de Brasília são utilizados códigos para representar a menção dos estudantes, esses códigos serão amplamente utilizados ao longo do trabalho e seus respectivos significados se encontram na Tabela 1.1.

Tabela 1.1: Menção e código associado

Menção	Significado
SS	Nota final entre 9 e 10
MS	Nota final entre 7 e 9
MM	Nota final entre 5 e 7
MI	Nota final entre 3 e 5
II	Nota final entre 1 e 3
SR	Nota final entre 0 e 1
CC	Crédito Concedido
TR	Trancamento
TJ	Trancamento justificado

O Índice de Rendimento Acadêmico (IRA) é a forma pela qual o desempenho do aluno é calculado de forma a sintetizar seu rendimento, o qual é expresso na Equação 1.1, cujo resultado pertence ao intervalo de 0 (mínimo rendimento) a 5 (máximo rendimento). Essas informações fora retiradas do guia do calouro da Universidade de Brasília (BRASÍLIA - UNB, 2018).

$$IRA = \left[1 - \frac{(0,6 * DT_b + 0,4 * DT_p)}{DC}\right] * \frac{\sum_i P_i * CR_i * P_{ei}}{\sum_i CR_i * P_{ei}} \tag{1.1}$$

Em que:

- DT_b = número de disciplinas obrigatórias trancadas
- DT_p = número de disciplinas optativas trancadas
- DC = número de disciplinas matriculadas (incluindo as trancadas)
- P_i = peso da menção (SS=5, MS=4, MM=3, MI=2, II=1, SR=0)
- P_{ei} = período em que uma dada disciplina foi cursada, obedecendo à seguinte limitação: min (6, período), que pega o menor valor entre o período do estudante e 6.
- CR_i = número de créditos de uma dada disciplina

Dentro da Universidade existem algumas condições para que o estudante possa permanecer matriculado em seu curso, as quais são apresentadas abaixo. Caso o aluno desrespeite alguma destas condições, ele entra em situação de risco de desligamento.

1. Ter duas reprovações na mesma disciplina obrigatória;
2. Não ter sido aprovado, em pelo menos, quatro disciplinas do seu curso, em dois períodos letivos regulares consecutivos;
3. Chegar ao último período letivo permitido no projeto pedagógico do seu curso sem possibilidades de concluí-lo.

O estudante em risco de desligamento deverá ser acompanhado por orientador acadêmico e deverá cumprir uma das seguintes condições, respectivas às situações descritas anteriormente:

1. Obter aprovação na(s) disciplina(s) obrigatória(s) anteriormente cursada(s) com duas reprovações;
2. Obter aprovação em disciplinas que somem o mínimo de créditos por semestre definido para o curso, em cada um dos dois períodos letivos subsequentes;
3. Cumprir plano de estudo aprovado pela Comissão de Acompanhamento e Orientação (CAO).

O plano de estudo consiste de um planejamento mínimo de um ano com as disciplinas a serem cursadas e que tenham probabilidade de êxito pelo discente. O plano de estudo, após ter sido elaborado com o professor orientador acadêmico, deve ser aprovado pelo colegiado do curso e encaminhado à Secretaria de Administração Acadêmica (SAA).

Caso o estudante entre em risco de desligamento não obedecendo uma das condições de permanência e o mesmo não cumpra a condição relativa ao risco de desligamento em que se encontra, então ele será desligado do curso.

As informações acima que descrevem o processo de desligamento do estudante foram retiradas do guia do calouro da Universidade de Brasília (BRASÍLIA - UNB, 2018).

Parte da proposta pedagógica da Universidade de Brasília é a flexibilização da organização curricular. Uma das formas mais simples de contribuir para isso consiste em dar aos estudantes a possibilidade de escolherem as disciplinas em que vão se matricular em cada semestre. Dessa maneira, os estudantes escolhem todas as matérias que vão cursar e quando irão cursar tais disciplinas. Toda a proposta pedagógica da UnB pode ser encontrada no documento (DECANATO DE PLANEJAMENTO, 2018a)

Sendo o curso de Engenharia da Computação um dos focos desta pesquisa, uma informação que pode colaborar com a compreensão dos próximos capítulos é o contexto da grade curricular do curso. Por isso, apresentamos no Anexo I o fluxograma completo da grade curricular do referido curso.

1.2 Justificativa

Dado que os valores e dados explicitados indicam problemas como o crescimento da população universitária e situações de desistência do curso, por parte dos estudantes, uma análise mais profunda permitiria entender de forma mais adequada as causas e as possíveis formas de evitar, ou pelo menos, de amenizar as consequências desses fatores. Para tal, o que se sugere nesse trabalho é a aplicação de mineração de dados para tentar entender e remediar as consequências do crescimento da população universitária e desistência dos estudantes. E durante esse processo buscar compreender melhor os fenômenos educacionais e pedagógicos que influenciam nesse processo, e a partir dessa compreensão, desenvolver um sistema de suporte para tomada de decisões.

Existem diversos desafios a se superar para se alcançar um ensino superior com alto rendimento e alta qualidade. A alta taxa de desistência de estudantes é certamente um destes desafios, que deve ser superado quando se busca padrões de ensino mais elevados. Uma forma usada para enfrentá-lo é uma análise e tomada de decisão baseada em dados dos estudantes. Esse problema pode ser visto tanto em instituições internacionais, como no estudo realizado por (AUD; WILKINSON-FLICKER, 2013), o qual mostra os dados de universidades nos Estados Unidos e a dimensão desse problema para elas, como no Brasil, como o estudo de (SILVA FILHO et al., 2007), o qual mostra que alta evasão também é um desafio para o ensino superior brasileiro.

Em diversas áreas do conhecimento já se faz o uso de ciência de dados, aprendizado de máquina e outras técnicas mais avançadas para tomada de decisões, desde áreas como economia, *marketing*, fisiologia, logística e diversos outros. O uso dessas técnicas na educação também já possui um largo histórico de décadas e vem crescendo cada vez mais.

Para tanto a produção de um trabalho interdisciplinar que visa simplificar o entendimento de problemas enfrentados em diversos cursos e universidades pelo mundo, com foco em entender e compreender a importância dos dados coletados e usados nesse processo de estudo é uma forma de engrandecer o entendimento da área *Educational Data Mining* (EDM), no sentido de estabelecer metodologias mais específicas e efetivas que possam cada vez mais melhorar o processo de aprendizado.

1.3 Objetivo Geral

Desenvolver um modelo preditivo de desistência de estudantes e usar estatísticas para compreender melhor os fatores que levam um estudante a sair de seu curso, com foco nos estudantes de Engenharia da Computação da UnB, de forma a tentar utilizar esse modelo como ferramenta de acompanhamento pedagógico por parte dos coordenadores do curso, dentro de uma futura estratégia de mitigação do problema.

1.3.1 Objetivos Específicos

- Desenvolver um modelo robusto com capacidade preditiva dentro dos valores esperados para a área de estudo (*Educational Data Mining*);
- Compreender melhor quais aspectos afetam o desempenho do estudante durante o curso;
- Explicitar dados problemáticos para o curso, ou departamento em questão;
- Apresentar um método de análise de dados educacionais, disponibilizados pela Universidade de Brasília

1.4 Metodologia

Neste trabalho a metodologia adotada para avaliação dos dados foi a CRISP-DM (*Cross Industry Standard Process for Data Mining*), a qual foi definida por (WIRTH; HIPP, 2000). Como pode se ver na Figura 1.4, existem fases do processo que se deve seguir de forma a se obter melhores resultados.

Como a área de EDM é relativamente recente, a mesma não possui uma metodologia de referência definida. Por isso, foi feito o uso da CRISP-DM que foi desenvolvida para a indústria de ciência de dados no geral. Apesar disso, é necessário fazer uma adaptação da metodologia em questão para adequação às especificidades da área, conforme descrito na pesquisa feita por (HAND, 2006), para tanto os aspectos pedagógicos dos dados foram levados em consideração para as análises realizadas.

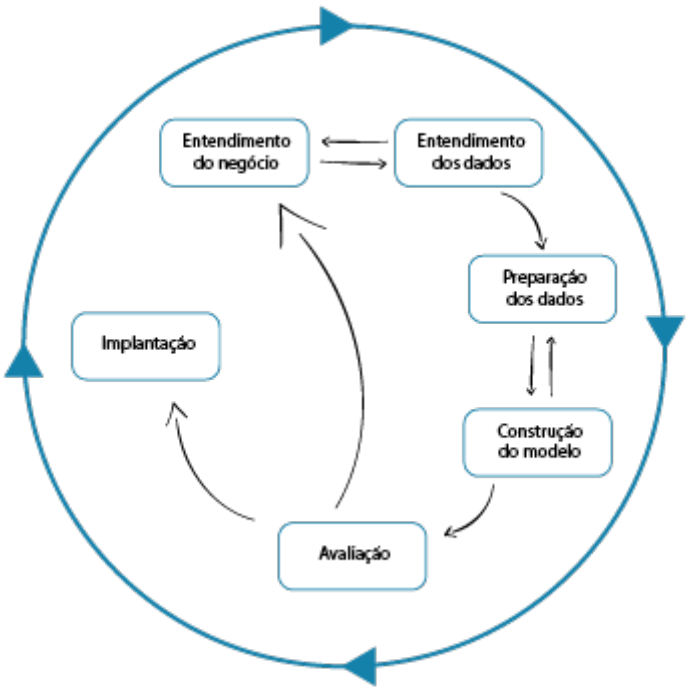


Figura 1.4: Fases do processo CRISP-DM.

Os processos seguidos foram:

1. Entendimento do problema e entendimento dos dados, no qual foi estudado o problema da evasão de estudantes e analisado os dados obtidos;
2. Preparação dos dados, , com a adoção de formatos que facilitassem sua análise e uso;
3. Modelagem, em que foram criados modelos de análise e predição baseados nos dados utilizados;
4. Avaliação, onde foram analisados os resultados obtidos a fim de verificar se correspondiam às expectativas e aos resultados encontrados na literatura de referência;
5. Implantação, onde foram utilizados os modelos criados para construir um sistema de visualização e predição.

As informações utilizadas nessa pesquisa foram retiradas de um recorte feito do ano de 2009 ao ano de 2018. O ano de 2009 foi escolhido por ter sido o primeiro ano do curso de Engenharia de Computação da Universidade de Brasília, e o ano de 2018 foi escolhido como forma de determinar uma janela de análise fechada, já que o presente estudo foi desenvolvido ao longo do ano de 2019. Os dados, já previamente anonimizados, foram disponibilizados pelo coordenador do curso de Engenharia da Computação, José Edil Guimarães de Medeiros. Dois conjuntos foram disponibilizados pelo mesmo, um contendo a relação de menções das turmas do departamento de Engenharia Elétrica (ENE) e do departamento de Ciência da Computação (CIC), os quais ofertam o maior número de disciplinas cursadas por estudantes do curso de Engenharia da Computação; o outro contém informações sobre os estudantes como Índice de Rendimento Acadêmico (IRA), região administrativa, desempenho em disciplinas, período de ingresso, período de saída, forma de ingresso e forma de saída. Nenhuma das informações disponibilizadas possuía qualquer valor que permitisse identificar isoladamente um estudante, sendo a utilização desses dados feita simplesmente para a criação de modelos estatísticos de forma semi-automatizada.

1.5 Apresentação do Trabalho

1. Capítulo 1: Apresentação da pesquisa realizada, incluindo problemática e apontamento de objetivos. Apresentação da abordagem escolhida para enfrentar os problemas apontados;
2. Capítulo 2: Apresentação e discussão dos conceitos e desafios envolvidos na elaboração da problemática da pesquisa;
3. Capítulo 3: Exposição da análise realizada acerca do impacto das disciplinas na problemática da pesquisa;
4. Capítulo 4: Exposição da análise realizada acerca do impacto socioeconômico na problemática da pesquisa;

5. Capítulo 5: Exposição e detalhamento da modelagem realizada;

6. Capítulo 6: Síntese e discussão dos resultados obtidos e abertura para trabalhos de melhoria nos temas e ferramentas apresentadas.

Capítulo 2

Fundamentação Teórica

Mineração de dados, também conhecida como *knowledge discovery from data* (KDD) é uma área que visa extrair informações de grandes bancos de dados para diversas aplicações. Uma definição, de acordo com (HAND, 2006) é:

> Mineração de dados é uma disciplina nova que surgiu em razão da convergência de muitas outras disciplinas, conduzida principalmente pela expansão de grandes bancos de dados. O elemento que propulsiona a importância da mineração de dados é o fato de que, bancos de dados extensos possuem informações valiosas para os respectivos proprietários; contudo, essas informações podem estar camufladas em uma massa de dados irrelevantes, de modo que precisarão ser exploradas. Ou seja, procura-se informações surpreendentes, inéditas, inesperadas ou valiosas, e o objetivo é extrair essas informações. Assim, é possível inferir que o referido campo de estudo está intimamente relacionado com a análise exploratória de dados. Entretanto, problemas relativos ao tamanho dos bancos de dados, bem como a existência de ideias e ferramentas importadas de outras áreas, fazem com que a mineração de dados seja uma área mais ampla do que simplesmente a análise exploratória de dados.[1](HAND, 2006, p.1) (tradução livre dos autores)

O surgimento da área de mineração de dados abriu uma ampla perspectiva para estudos estatísticos, visto que a partir de uma massa de dados é possível fazer uma inferência estatística sobre tais. Apesar das possibilidades, existem diversas dificuldades e problemáticas em como realizar o processo de descoberta de informações.

O processo de descoberta de informações a partir de base de dados é descrito por (HAN; PEI; KAMBER, 2011) como tendo as seguintes etapas:

[1]Data mining is a new discipline which has sprung up at the confluence of several other disciplines, driven chiefly by the growth of large databases. The basic motivating stimulus behind data mining is that these large databases contain information which is of value to the database owners, but this information is concealed within the mass of uninteresting data and has to be discovered. That is, one is seeking surprising, novel, unexpected, or valuable information, and the aim is to extract this information. This means that the subject is closely allied to exploratory data analysis. However, issues arising from the sizes of the databases, as well as ideas and tools imported from other areas, mean that there is more to data mining than merely exploratory data analysis.

1. Limpeza dos dados - remoção de dados inconsistentes;
2. Integração de dados - onde dados de diversas fontes podem ser combinados;
3. Seleção dos dados - consiste em selecionar os dados interessantes para a análise realizada;
4. Transformação dos dados - transformação dos dados em formas mais apropriadas para a mineração;
5. Mineração de dados - aplicação de métodos inteligentes para extração de padrões potencialmente úteis, a partir dos dados;
6. Avaliação - avaliação dos dados extraídos que representam informações interessantes;
7. Representação da informação - onde técnicas de visualização e representação são usadas para mostrar as informações para o usuário.

Uma discussão abrangente na área é sobre a sintetização de metodologias de mineração de dados. Como sugere a pesquisa de (MARISCAL; MARBAN; FERNANDEZ, 2010), vemos que as mais usadas são: CRISP-DM (*Cross Industry Standard Process for Data Mining*), SEMMA (*Sample, Explore, Modify, Model, and Assess*). Podemos ver que a discussão sobre o uso de metodologias de desenvolvimento na área de mineração de dados busca explicitar a importância do uso para a redução de erros ao longo do processo.

Concomitante ao desenvolvimento das técnicas de mineração de dados, diversas áreas correlatas usam dos processos de exploração estatística e extração de dados para o seu desenvolvimento. O uso de técnicas de aprendizado de máquina (*Machine Learning*) é um deles, as possibilidades que se desenvolveram a partir do uso dessas duas áreas juntas é extremamente relevante para diversos campos, como pode-se ver na descrição de (FRIEDMAN; HASTIE; TIBSHIRANI, 2001)

> A ciência da aprendizagem possui um papel importante nos campos da estatística, da mineração de dados e da inteligência artificial, convergindo com as áreas de engenharia e outras disciplinas.[2] (FRIEDMAN; HASTIE; TIBSHIRANI, 2001, p.1)(tradução livre dos autores).

As possibilidades geradas pela extração de dados em grandes sistemas e análise de aprendizado estatístico podem revolucionar diversas áreas, como: visão computacional, com a criação de sistemas inteligentes que conseguem identificar objetos; genética, na qual se busca elementos regulatórios que possuam impacto significativo em um fenótipo de interesse; e educação, em que sistemas preditivos de desempenho permitem melhorias na metodologia de ensino, dentre outras aplicações.

[2]The science of learning plays a key role in the fields of statistics, data mining and artificial intelligence, intersecting with areas of engineering and other disciplines.

2.1 *Educational Data Mining* (EDM)

Educational Data Mining (EDM) é uma área recente que visa aplicar recursos computacionais e técnicas de mineração de dados para tentar compreender fenômenos educacionais, como definido por (ROMERO; VENTURA, 2007):

> [...] A mineração de dados educacional é um novo campo relacionado a várias áreas bem estabelecidas de pesquisa, incluindo e-learning (ensino eletrônico), hipermídia adaptativa, sistemas de tutorias inteligentes, mineração da Web, mineração de dados etc. A aplicação da mineração de dados em sistemas educacionais possui requisitos próprios, não presentes em outras áreas, podendo destacar-se, principalmente, a necessidade de levar-se em consideração aspectos pedagógicos referentes ao estudante e ao sistema.[3] [ROMERO; VENTURA, 2007, p.144] (tradução livre dos autores)

O impacto da educação de nível superior na sociedade é incalculável e, assim, a busca por entender os fenômenos educacionais em universidades é crescente nos últimos anos. Uma forma de se tentar entender esse fenômenos é buscar a compreensão da relação dos estudantes com suas respectivas instituições de ensino.

As pesquisas e o interesse na área remonta de décadas atrás, podemos ver que desde o começo dos anos 90 já existiam estudos que se inseriam no campo de estudo dessa área, na pesquisa realizado por (KNOX; LINDSAY; KOLB, 1992), pode-se ver a preocupação com a performance de estudantes, a fim de se encontrar respostas elaboradas para compreender as demandas estudantis e, dessa forma, melhorar o ambiente nos quais esses estão inseridos e, a partir da melhora do ambiente, receber um impacto positivo no resultado acadêmico dos alunos.

Uma abordagem interessante é a vista no estudo realizado por (BEAN; BRADLEY, 1986) que mostra uma modelagem que visa compreender a relação do desempenho acadêmico com a satisfação dos estudantes. Avaliando essas relações a partir do *Grade Point Average* (GPA) dos alunos, tentando relacioná-lo a variáveis como performance no ensino-médio, integração acadêmica e outras. Na pesquisa de (BENJAMIN; HOLLINGS, 1997) é possível verificar uma busca maior pelas relações sociais dos estudantes e seu desempenho acadêmico; com efeito, segundo o estudo, existem vários fatores que podem afetar o desempenho acadêmico de um estudante, como sua relação com a instituição de ensino, relações pessoais, experiência prévias, dentre outros.

Com o desenvolvimento de sistemas computacionais e a popularização de seu uso em ambientes educacionais, foi possível desenvolver grandes base de dados sobre os estudantes e até mesmo sobre as instituições, o que permitiu que as pesquisas nessa área pudessem ser mais eficazes no sentido de obter dados estatísticos mais significativos. Partindo desse pressuposto, podemos ver na pesquisa de (ROMERO; VENTURA, 2007) que o uso de técnicas de mineração de dados no período analisado

[3]Educational data mining is an upcoming field related to several well-established areas of research including e-learning, adaptive hypermedia, intelligent tutoring systems, web mining, data mining, etc. The application of data mining in educational systems has specific requirements not present in other domains, mainly the need to take into account pedagogical aspects of the learner and the system.

(1995 - 2005), auxiliaram na compreensão e formação de relações complexas a partir de dados educacionais.

Os resultados que podem ser obtidos a partir de pesquisas de EDM são inúmeros e, dessa forma, a pesquisa feita por (ROMERO; VENTURA, 2010) busca categorizar as áreas de estudo e quais aspectos acadêmicos podem ser beneficiados por pesquisas em EDM, tais como: análise e visualização de dados, prover *feedback* para tutores, recomendações para estudantes, predição da performance de estudantes, modelagem dos estudantes, detecção de comportamento não desejado dos estudantes, agrupamento do comportamento de estudantes, análise das relações sociais, desenvolvimento de mapas de conceitos, construção de material didático, planejamento de cursos.

Dentro dessas subáreas, as que tem mais relação com este trabalho são as áreas de predição, modelagem, extração de dados para tomada de decisões e mineração de dados relacionais, de acordo com as definições de (BAKER; YACEF, 2009).

2.1.1 Extração de dados para tomada de decisões

A área de EDM também permite a análise de diferentes metodologias de ensino e sua eficiência, como por exemplo a diferença de turmas que seguem uma metodologia ativa e outra com uma metodologia tradicional, em (FREEMAN et al., 2014) compara-se turmas seguindo esses dois tipos de metodologias citados em cursos como ciências, engenharia e matemática, o estudo mostra que efetivamente existe uma melhora no desempenho dos estudantes mostrando que na média estudantes sob o primeiro modelo tem um índice de reprovação de $33,8\%$, enquanto no segundo modelo esse índice cai para $21,8\%$, esse tipo de estatística representa um modo de se compreender melhor metodologias educacionais e assim facilitar a tomada de decisões relacionadas.

Outro estudo que busca fazer esse comparativo de turmas é a pesquisa realizada por (MASON; SHUMAN; COOK, 2013), onde os autores comparam turmas com aprendizado tradicional e utilizando aprendizagem baseada em investigação, que consiste em uma metodologia ativa e, dessa forma, foi percebida uma melhora dos estudantes no aprendizado e desempenho do curso, o próprio estudo discute que essa melhora não se deve ao tempo extra gasto pelos estudantes e que o uso da metodologia teve grande impacto no resultado. Apesar dessa discussão, é também verificável na pesquisa de (KIM; AHN, 2017), uma redução na taxa de desistência do curso, bem como uma melhora da performance dos estudantes nos exames de meio e final de período, onde de fato foi aplicado uma metodologia de aprendizagem baseada em investigação.

2.1.2 Mineração de dados Web

Uma área relacionada a EDM que vem crescendo é a área que analisa e discute a qualidade e eficiência de cursos online. Nos chamados *Massive Open Online Courses* (MOOCs) é possível ver que menos de 10% dos estudantes inscritos chegam a concluir o curso, como demonstrado no estudo feito por (JORDAN, 2014). A depender do curso e da plataforma responsável, esses valores podem variar.

Em trabalhos mais recentes podemos ver a preocupação em tentar melhorar os índices de desempenho nesses cursos em diversos aspectos, como melhorar o desempenho dos estudantes, na pesquisa realizada por (YANG et al., 2015) é possível verificar que um fator que afeta significativamente o desempenho de estudantes nos chamados MOOCs é a confusão dos estudantes ao decidir o que fazer durante o curso, identificar esse fenômeno é a primeira etapa para se melhorar a qualidade dos MOOCs.

Apesar do escopo deste trabalho não lidar com MOOCs, devido à natureza dos dados utilizados, as concepções e estudos desenvolvidos sobre desistência de cursos em MOOCs é abrangente e demonstra vária abordagens para esse tipo de estatística e formas de análise para os dados.

2.1.3 Modelagem

Com o desenvolvimento de técnicas de aprendizado de máquina, concomitantes com a mineração de dados, a aplicação de algoritmos de aprendizado de máquinas na área foi acompanhando esse fenômeno, dado que com dados suficientes o entendimento de fenômenos e processos educacionais se torna mais fácil e simplifica o processo de tomada de ações e melhorias em todo o escopo. No trabalho de (GEIGLE; ZHAI, 2017) podemos ver a modelagem baseada em aprendizado profundo (*deep learning*), que busca descrever os padrões de comportamento em MOOCs. Um ponto extremamente interessante em modelagens é o entendimento de fenômenos que encaixam os indivíduos naquele padrão de comportamento, ao ponto que, ao analisá-los, é possível compreender melhor as problemáticas do método de ensino e aprendizado e, assim, melhorá-los.

O estudo feito por (AGRAWAL; GOLSHAN; PAPALEXAKIS, 2016), o qual usa bancos de dados para tentar criar planos de estudos mais efetivos para estudantes, mostra o potencial da área no auxílio e na tentativa de melhorar os sistemas de ensino; apesar de não chegar a uma conclusão estatisticamente relevante sobre a qualidade desses planos de estudos gerados, ele abre possibilidades de aperfeiçoamento e se mostra como uma estratégia interessante a ser explorada.

2.1.4 Predição

Predizer comportamentos de risco em estudantes é uma possibilidade que surge junto à área EDM, que demonstra uma forma de se atuar e melhorar os sistemas educacionais. Podemos ver na pesquisa de (PELAEZ et al., 2019) uma tentativa de identificar estudantes com risco de se desligarem da universidade, o estudo é feito sobre a *San Diego State University* (SDSU) para se entender melhor fatores de riscos e verificar se instruções suplementares podem auxiliar nesse processo.

Outro trabalho que demonstra o uso de análise de conjunto de dados para tomada de decisão é o realizado por (SAARELA; KÄRKKÄINEN, 2015), o qual mostra a relação entre as disciplinas cursadas pelo estudantes e o seu desempenho, demonstrando um caminho ótimo para o estudante durante sua graduação. Essa abordagem visa facilitar a vida do estudante, ao passo que melhora o desempenho acadêmico.

A partir da modelagem de sistemas, a capacidade de prever os resultados desse sistema se torna uma possibilidade. Predizer o desempenho do estudante se mostra como uma grande oportunidade para se perceber fenômenos educacionais, nesse aspecto, a pesquisa realizada por (SWEENEY et al., 2016), busca utilizar modelos de regressão para predizer a performance do estudante. Sistemas como esse demonstram a possibilidade de auxiliar os estudantes em disciplinas que eles potencialmente teriam um resultado não satisfatório. De fato, a partir deste tipo de sistema, seria possível criar uma rede de apoio para os estudantes.

Uma abordagem menos comum na área de EDM, como pode ser visto na pesquisa feita por (BAKER; YACEF, 2009), é o entendimento de fatores sociais no desempenho do estudante, na pesquisa feita por (BAYER et al., 2012) podemos ver um sistema de predição de desistência de estudantes baseado em aspectos sociais, tais como: gênero, ano de nascimento, ano de entrada na faculdade, quantidade de créditos feitos, quantidade de créditos que faltam e várias outras informações; esse estudo foi de grande importância como parâmetro para busca de informações a se utilizar na modelagem realizada neste trabalho.

2.2 Seção Final

Como se poderá ver adiante, os estudos desenvolvidos neste trabalho se utilizam de boa parte do referencial teórico em EDM, discorrido nas seções anteriores. Nos capítulos 3 e 4 busca-se analisar o impacto das disciplinas e de condições socioeconômicas no desempenho dos estudantes e na probabilidade de desistência. Assim como no trabalho de (YANG et al., 2015), a busca por compreender os fenômenos da desistência dos estudantes foi o objetivo dessa parte da pesquisa. Apesar das diferenças entre a origem dos dados de MOOCs e dados provenientes de um curso regular de uma universidade, existe, em ambos os casos, a preocupação em encontrar os fatores e características principais da desistência de estudantes.

Em suma, este trabalho tem como objetivo minerar dados educacionais, criar modelos estatísticos e produzir um sistema de predição de desistência de estudantes, a fim de auxiliar tomadas de decisões. O referencial teórico abordado neste capítulo serve, neste sentido, como ferramenta para auxiliar na escolha de características importantes e melhores formas de se analisar os problemas e desafios enfrentados.

Capítulo 3

Análise dos dados de disciplinas

Neste capítulo é realizada uma análise sobre as disciplinas oferecidas pelos departamentos CIC e ENE, com o propósito de se traçar a relação entre a reprovação em disciplinas e a desistência do estudante, para o conjunto de dados de estudantes do curso de Engenharia da Computação. O objetivo desse capítulo é levantar as estatísticas sobre as disciplinas, sobre o desempenho dos estudantes e explicitar dados relevantes para se indicar possíveis correlações entre desempenho e desistência.

Os principais aspectos a serem analisados nesse capítulo serão:

1. Distribuição de menções: compreender o aspecto geral do desempenho dos estudantes nas disciplinas;
2. Destaque das disciplinas com maiores índices de aprovação, reprovação e trancamento: destacar as disciplinas a fim de se localizar possíveis disciplinas problemáticas;
3. Relação entre desempenho nas disciplinas e desistência: verificar qual a correlação do desempenho do estudante em determinada disciplina e sua probabilidade de desistir do curso

3.1 Metodologia

Neste capítulo foram utilizadas duas bases de dados diferentes para realização das análises, ambas considerando o período de 2009 a 2018.

1. Coleta dos dados: A extração da primeira base de dados da análise foi feita por meio do Sistema de Informações Acadêmicas de Graduação (SIGRA), da Universidade de Brasília, pelo coordenador do curso de Engenharia de computação, José Edil Guimarães de Medeiros. Os dados coletados apresentam as disciplinas de um dado semestre e a distribuição de menções entre as turmas, para o departamento de Engenharia Elétrica e para o departamento de Ciência da Computação, essas informações foram utilizadas para realizar a análise da Seção 3.2. A extração da segunda base de dados da análise foi feita por meio do sistema de *Business*

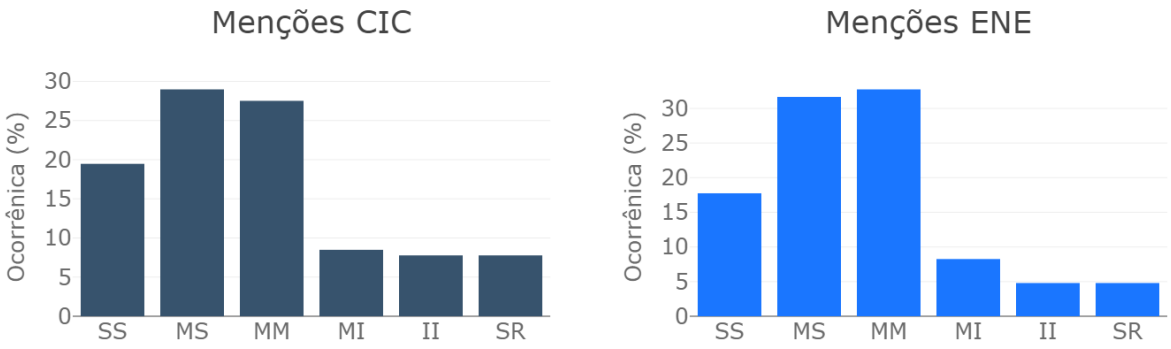
Intelligence, BI, da Universidade de Brasília, cujo acesso foi fornecido pelo *Decanato de Planejamento, Orçamento e Avaliação*, DPO, ao coordenador do curso de Engenharia de Computação, José Edil Guimarães de Medeiros. Os dados coletados relacionam, para cada aluno, a menção nas disciplinas cursadas em um semestre e a forma de saída do mesmo, para 700 alunos do curso.

2. Anonimização: A fim de manter o sigilo dos alunos analisados, para a segunda base de dados, o coordenador do curso substituiu o CPF de cada indivíduo por um código identificador, de modo que o número de matrícula dos estudantes não foi disponibilizado para os autores deste trabalho. A Figura 3.1 ilustra o formato dos dados após a etapa de anonimização, na qual o conteúdo da coluna *id_ aluno* foi substituído por um código identificador. A primeira base de dados apresenta apenas informações sobre as disciplinas, não sendo necessário nenhum tipo de anonimização dos dados.

id	id_aluno	ano_ingresso	semestre_ingresso	cod_disciplina	mencao	ano_referencia	semestre_referencia	Status
0	aluno88	2008	2	113093	SR	2009	2	Formatura
1	aluno88	2008	2	118001	MI	2009	2	Formatura

Figura 3.1: Modelo de dados utilizados com informações de ingresso, menção e forma de saída dos alunos.

3. Limpeza: Algumas análises iniciais dos dados nos permitem ver que a distribuição de menções ilustrada nos histogramas da Figura 3.2 parece ser satisfatória, de forma a se perceber que existe uma concentração de turmas com altos índices de aprovações e predominância das menções MM (Média) e MS (Média Superior), que representam notas de 5 a 8,9 na média final, com um baixo índice de alunos com aprovação recebendo a menção SS (Superior), que representa notas de 9,0 a 10,0.



(a) Histograma de distribuição de menções (CIC) (b) Histograma de distribuição de menções (ENE)

Figura 3.2: Distribuição de menções departamentos ENE e CIC.

No entanto, ao se aprofundar na base de dados é possível se extrair algumas informações que indicam um viés do estudo, como por exemplo, as turmas de disciplinas de trabalhos de conclusão de curso são ofertadas, quase que exclusivamente para cada aluno, isto é, cada

aluno teria uma turma para essa matéria, dessa forma, existe uma quantidade significativa de alunos dessa disciplina, a qual não reflete tão bem possíveis fenômenos educacionais que seriam observáveis na maioria das disciplinas convencionais do departamento. Ademais, é possível perceber também que este tipo de disciplina possui uma grande quantidade de menções SS. Sendo assim, os dados das seguintes disciplinas foram retiradas das análises:

(a) Projeto Final de Graduação

(b) Trabalho Conclusão de Curso

(c) Trabalho de Graduação 1

(d) Trabalho de Graduação 2

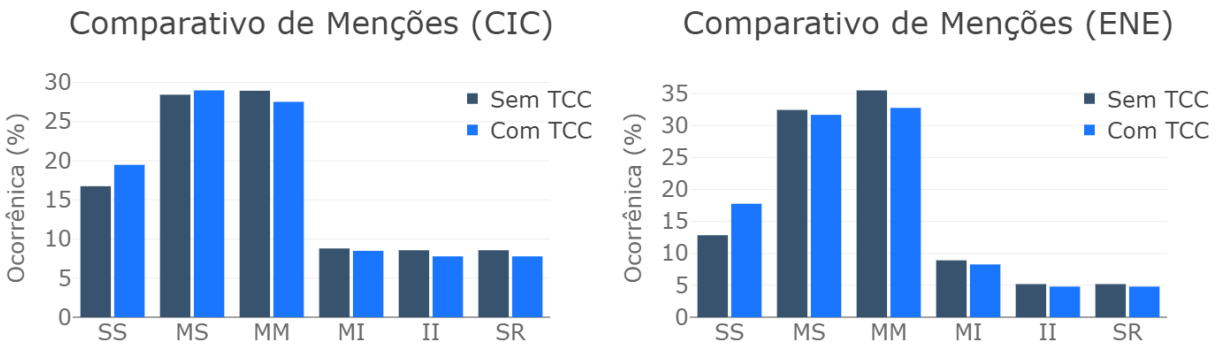
(e) Projeto de Licenciatura

(f) Projeto de Licenciatura 1

(g) Projeto de Licenciatura 2

(h) Projeto Final Engenharia

Consequentemente, um total de 20.915 turmas foram retiradas das análises, oriundas destas turmas de disciplinas de TCC ou análogas, dos alunos formandos entre 2009 e 2018. A Figura 3.3 mostra o histograma de menções antes e depois do processo de remoção. É possível observar a redução de menções de aprovação: SS, MS, MM, e o aumento percentual de menções de reprovação: MI (Média Inferior), II (Inferior) e SR (Sem Rendimento). Esse fenômeno pode ser percebido nos dois departamentos após a limpeza dos dados.



(a) Comparativo retirada de disciplinas (CIC)

(b) Comparativo retirada de disciplinas (ENE)

Figura 3.3: Histograma de distribuição de menções (ENE/CIC) antes e depois da retirada das disciplinas de TCC.

3.2 Exploração dos Dados de Disciplinas

A partir dos dados supracitados é possível tentar extrair relações que possam representar fenômenos educacionais relevantes. Busca-se as disciplinas que historicamente possuam altos índices de reprovação, de forma a tentar entender e, possivelmente, remediar esse efeito. As disciplinas foram

ordenadas por sua respectiva taxa de reprovação e a Figura 3.5 mostra as 10 primeiras, com índice de reprovação médio em torno de 40%, salientando que essa taxa é relativa apenas aos alunos que cursaram integralmente a disciplina, descontando assim aqueles que realizaram o trancamento[1] da matéria. Além disso, foram desconsideradas para esta seleção as disciplinas cujo número de alunos fosse inferior a 300, dessa forma, apenas disciplinas que foram regularmente ofertadas no período analisado (10 anos), entrem nas estatísticas e os resultados reflitam o processo pedagógico geral.

Aliado a essa métrica de disciplinas com maiores médias de reprovações ao longo do período analisando, segue as disciplinas com maiores índices de trancamento, um fator que é problemático são disciplinas com alto nível de reprovação e alto nível de trancamentos, como se pode ver nos gráficos das Figuras 3.5 e 3.7 as disciplinas: Sistemas de Programação, Sistemas Digitais 1, Princípios de Comunicação, Computação para Engenharia, possuem um índice de estudantes não aprovados indo de 40% a 50% e também estão entre as disciplinas com maiores índices de trancamento.

Ao se analisar essas mesmas relações para o CIC, observando os gráficos das Figuras 3.4 e 3.6 vemos as disciplinas: Sinais e Sistemas, Org arq de computadores (Organização e arquitetura de computadores), Automatos e (Autômatos e Computabilidade), Lógica computacional 1, possuem um índice de alunos não aprovados variando de cerca de 40% a 60%, os quais também se encontram na lista de disciplinas com maiores índices de trancamento.

Esse valores explicitam algumas disciplinas, as quais podem ocasionar problemas logísticos para os departamentos, tendo em vista que algumas dessas disciplinas compõe o corpo de matérias obrigatórias dos cursos, a demanda por elas durante a matrícula pode se tornar bem alta, aliado ao fato de 50% dos estudantes que cursam elas não serem aprovados, pode gera um acumulo de estudantes e assim dificultar a matrícula deles.

[1]Recurso da universidade para que o aluno possa se desvincular da disciplina desde que transcorrido até 50% do semestre, ainda que fique registrado em seu histórico com uma menção de trancamento.

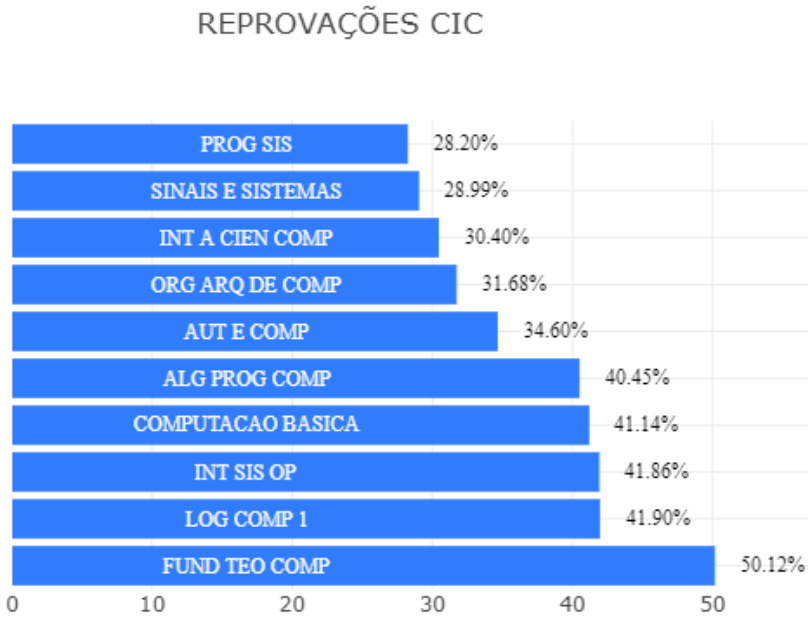


Figura 3.4: Maiores índices de reprovação do CIC.

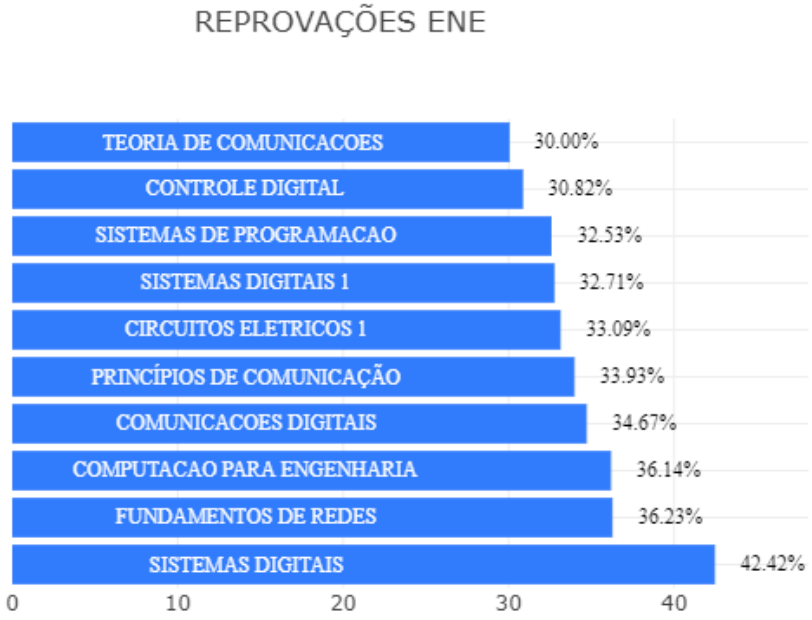


Figura 3.5: Maiores índices de reprovação do ENE.

As disciplinas listadas para o gráfico de reprovações do CIC (Figura 3.4) e seus respectivos códigos, em ordem crescente, são:

- PROG SIS - Programação Sistemática
- SINAIS E SISTEMAS - Sinais e Sistemas
- INT A CIEN COMP - Introdução à Ciência da Computação
- ORG ARQ DE COMP - Organização e Arquitetura de Computadores
- AUT E COMP - Autômatos e Computabilidade
- ALG PROG COMP - Algoritmos e Programação de Computadores
- COMPUTACAO BASICA - Computação Básica
- INT SIS OP - Introdução a Sistemas Operacionais
- LOG COMP 1 - Lógica Computacional 1
- FUND TEO COMP - Fundamentos Teóricos da Computação

As disciplinas listadas para o gráfico de reprovações do ENE (Figura 3.5) e seus respectivos códigos, em ordem crescente, são:

- TEORIA DE COMUNICACOES - Teoria de Comunicações
- CONTROLE DIGITAL - Controle Digital
- SISTEMAS DE PROGRAMACAO - Sistemas de Programação
- SISTEMAS DIGITAIS 1 - Sistemas Digitais 1
- CIRCUITOS ELETRICOS 1 - Circuitos Elétricos 1
- PRINCÍPIOS DE COMUNICAÇÃO - Princípios de Comunicação
- COMUNICAÇÕES DIGITAIS - Comunicações Digitais
- COMPUTACAO PARA ENGENHARIA - Computação para Engenharia
- FUNDAMENTOS DE REDES - Fundamentos de Redes
- SISTEMAS DIGITAIS - Sistemas Digitais

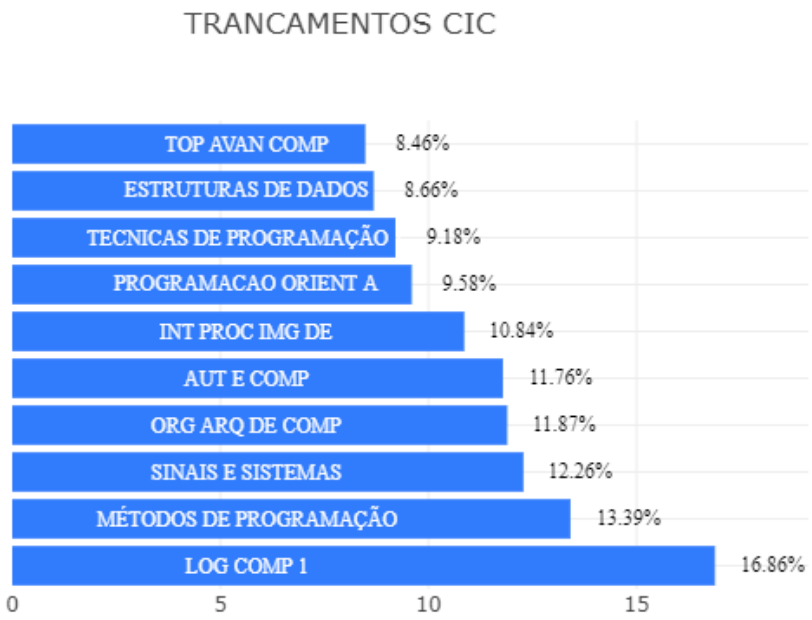


Figura 3.6: Maiores índices de trancamento do CIC.

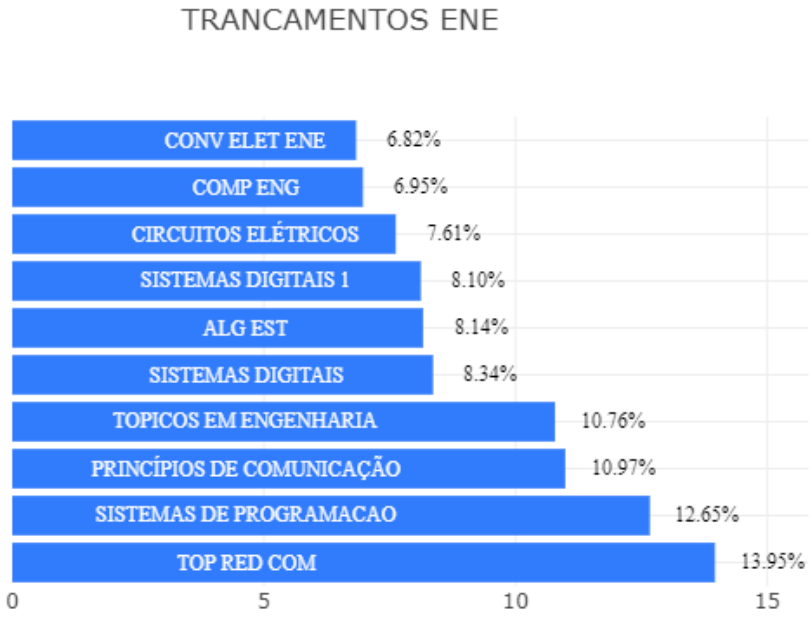


Figura 3.7: Maiores índices de trancamento do ENE.

As disciplinas listadas para o gráfico de trancamento do CIC (Figura 3.6) e seus respectivos códigos, em ordem crescente, são:

- TOP AVAN COMP - Tópicos Avançados em Computação
- ESTRUTURA DE DADOS - Estrutura de Dados
- TECNICAS DE PROGRAMAÇÃO - Técnicas de Programação
- PROGRAMACAO ORIENT A - Programação Orientada a Objetos
- INT PROC IMG DE - Introdução ao Processamento de Imagens
- AUT E COMP - Autômatos e Computabilidade
- ORG ARQ DE COMP - Organização e Arquitetura de Computadores
- SINAIS E SISTEMAS - Sinais e Sistemas
- MÉTODOS DE PROGRAMAÇÃO - Métodos de Programação
- LOG COMP 1 - Lógica Computacional 1

As disciplinas listadas para o gráfico de trancamento do ENE (Figura 3.7) e seus respectivos códigos, em ordem crescente, são:

- CONV ELET ENE - Conversão Eletromecânica de Energia
- COMP ENG - Computação para Engenharia
- CIRCUITOS ELETRICOS - Circuitos Elétricos
- SISTEMAS DIGITAIS 1 - Sistemas Digitais 1
- ALG EST - Algoritmos e Estrutura de Dados
- SISTEMAS DIGITAIS - Sistemas Digitais
- TOPICOS EM ENGENHARIA - Tópicos em Engenharia
- PRINCÍPIOS DE COMUNICAÇÃO - Princípios de Comunicação
- TOP RED COM - Tópicos em Redes de Comunicações

Usando os dados de aprovação e reprovação das disciplinas, tentou-se visualizar essas informações em diagramas de caixa (*boxplots*). Observando as Figuras 3.8 e 3.9, percebe-se que os índices para o ENE são melhores que os do CIC, dado que os índices de aprovação do ENE são superiores em todos os quartis e a variação dos valores superiores e inferiores para os quartis também é melhor. Por outro lado, os índices de trancamento associados ao ENE são inferiores aos do CIC, mesmo que a distância do menor valor para o primeiro quartil seja ligeiramente maior. Sendo assim, é possível considerar que os valores do ENE são mais satisfatórios. Ademais, é preciso enfatizar que embora ofereçam disciplinas diferentes, os dois departamentos lidam com estudantes de engenharia com formação semelhante, de modo que a comparação entre os dois é válida em certa medida.

No entanto, percebe-se que o diagrama de trancamentos apresenta dados discrepantes para ambos os departamentos, a detecção de valores discrepantes, pode ser usada como forma de se identificar pontos de atuação, disciplinas que possuem parâmetros atípicos seriam as ideais para uma atuação. Em diagramas de caixa, valores discrepantes são aqueles são pelo menos 1,5 vezes maior do que o valor do terceiro quartil (Q3). As disciplinas destacadas são:

- Tópicos em Redes de Comunicações (ENE)
- Sistemas de Programação (ENE)
- Princípios de Comunicação (ENE)
- Tópicos em Engenharia (ENE)
- Lógica Computacional 1 (CIC)

Como citado anteriormente, as disciplinas Lógica Computacional 1 e Princípios de Comunicação, além de possuírem uma alta taxa de reprovação, também possuem uma alta taxa de trancamento e perceptivelmente essa taxa se mostra como um valor atípico, dessa forma, os valores encontrados reforçam a argumentação realizada.

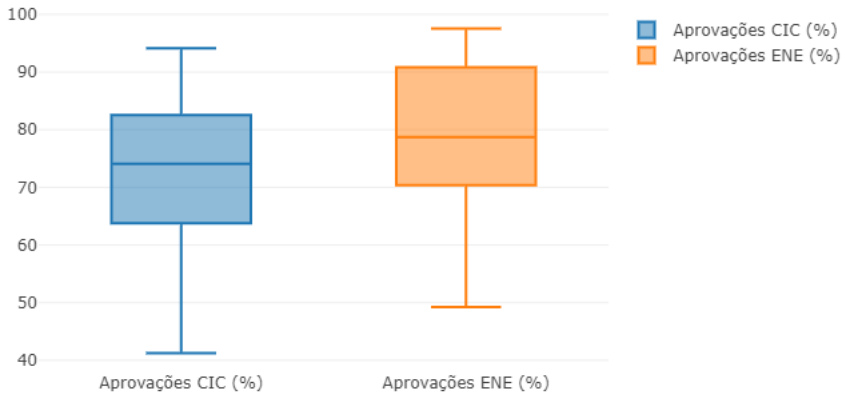


Figura 3.8: Diagramas de caixa aprovações ENE e CIC.

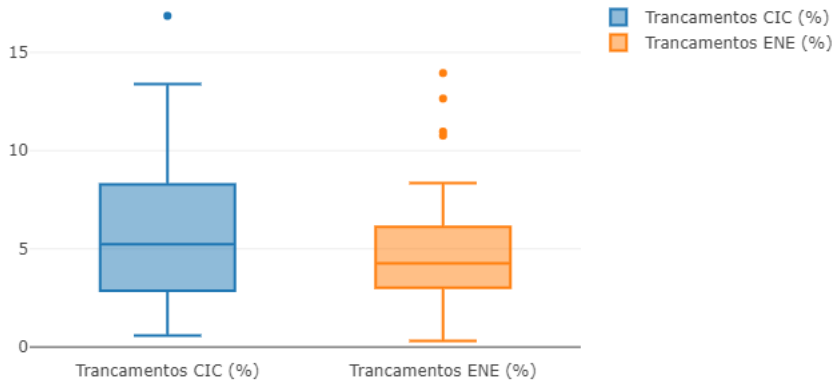


Figura 3.9: Diagramas de caixa trancamentos ENE e CIC.

3.3 Disciplinas e Desistência

Um fenômeno que se pode identificar no curso de Engenharia da Computação é a desistência por parte dos estudantes do curso, isto é, os alunos que de alguma forma são desligados do curso e passam a não compor mais o corpo discente do mesmo. Percebe-se que nos 2 primeiros anos tem-se as maiores taxas de desistência de estudantes, sendo o primeiro ano o maior valor, como se pode verificar na Figura 3.10. Devido a esse fator, uma análise focada no impacto das disciplinas e do desempenho dos estudantes nos dois primeiros semestres do curso foi realizada para se compreender melhor esse fenômeno, considerando os dados de 700 estudantes do curso dentro do recorte temporal

fornecido. .

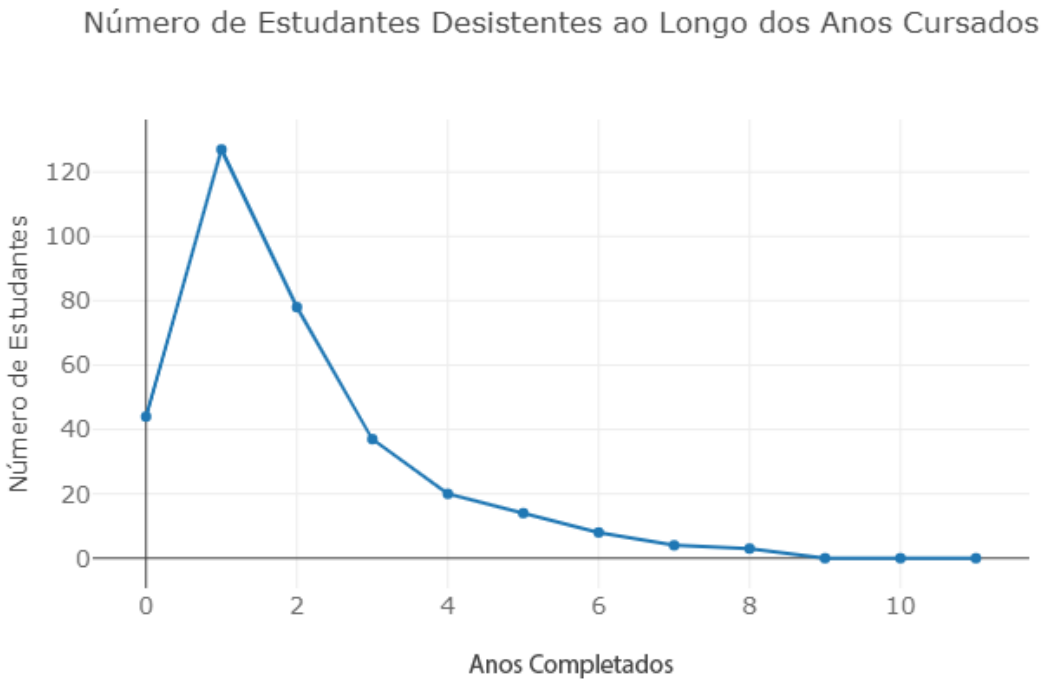


Figura 3.10: Desistência de Estudantes do curso de Engenharia da Computação

A partir dos dados desses estudantes, foram realizadas análise da taxa de desistência dos estudantes relacionadas com sua reprovação, utilizando o conceito de probabilidade condicional, como pode ser visto na Equação (3.1), em que A representa o evento de desistência do curso e B representa a reprovação em uma determinada disciplina. Dessa maneira, estima-se a probabilidade de desistência dada a reprovação em uma disciplina.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{3.1}$$

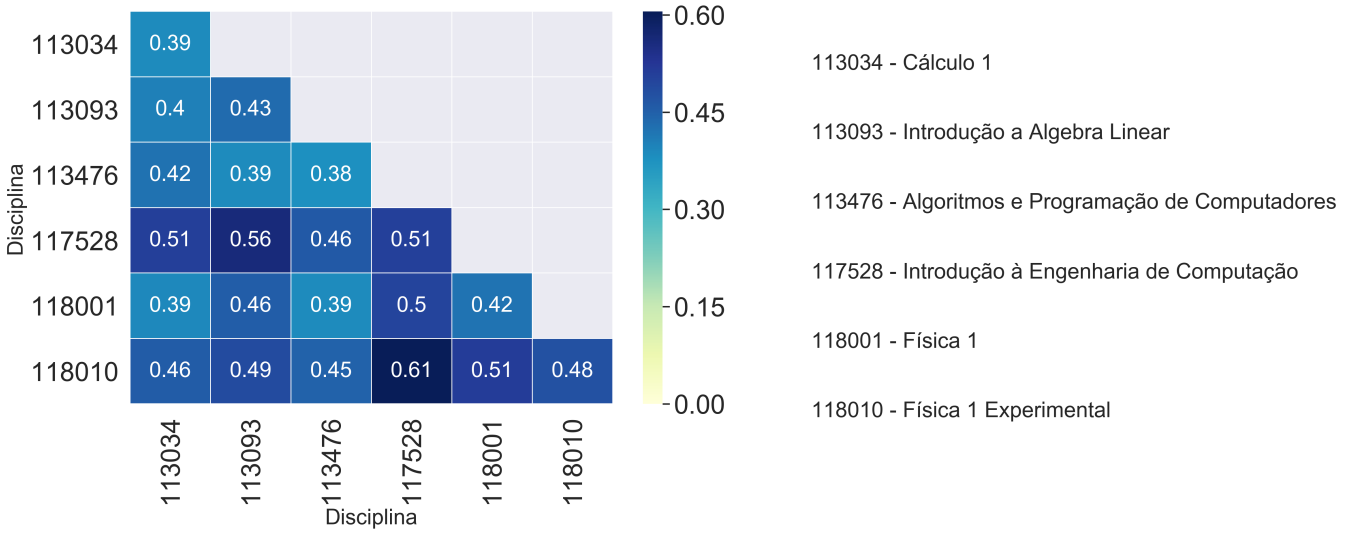


Figura 3.11: Taxa de desistência baseada nas disciplinas reprovadas no mesmo semestre.

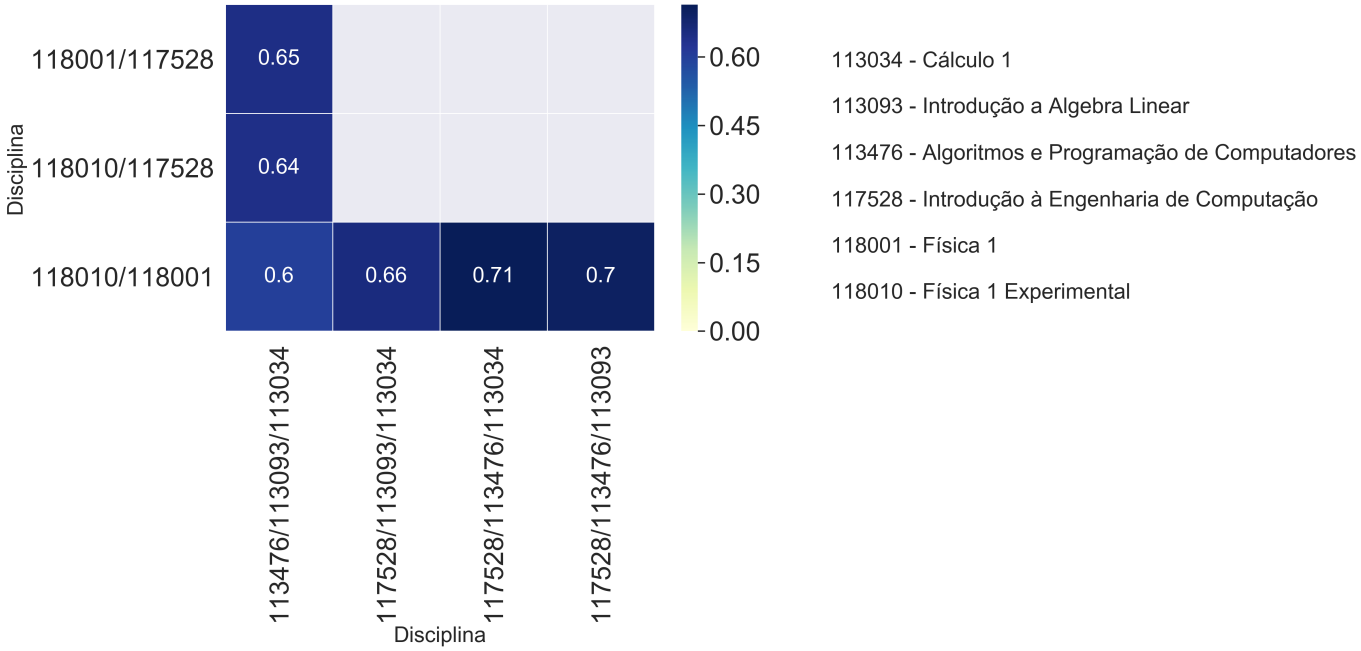


Figura 3.12: Taxa de desistência baseada nas disciplinas reprovadas no mesmo semestre.

As disciplinas em questão são as disciplinas indicadas para o primeiro semestre e as taxas de desistência representam a relação dos alunos que cursam essas disciplinas no mesmo semestre. As Figuras 3.11 e 3.12 representam essa taxa dado o conjunto de disciplinas reprovadas. Cada disciplina é indicada por seu respectivo código.

A disciplina de Computação Básica era colocada como pertencente ao fluxo do primeiro semestre, essa disciplina saiu do fluxo do curso e em seu lugar foi colocada a disciplina de Algoritmos e Programação de Computadores, para tanto, os códigos das duas disciplinas foram igualados nessa análise realizada.

As taxas são obtidas a partir dos alunos que reprovaram nas dadas disciplinas e desistiram no curso, para a Figura 3.11 a menor quantidade de estudantes em relação as taxas é para Física 1 (118001) e Física 1 Experimental (118010) em que 179 estudantes reprovaram ambas as disciplinas e 91 desses estudantes desistiram do curso, o que corresponde a cerca de 13% dos dados analisados, a maior quantidade de estudantes é relacionada a taxa de Cálculo 1 (113034) e Introdução a Engenharia da Computação (117528), em que 407 estudantes reprovaram as disciplinas e 208 desses estudantes desistiram do curso, correspondendo a cerca de 30% da quantidade total analisada. Enquanto que, para a Figura 3.12, a menor quantidade de estudantes em relação as taxas é para Física 1 (118001), Introdução a Engenharia da Computação (117528), Cálculo 1 (113034), Introdução a Algebra Linear (113093) e Algoritmos e Programação de Computadores (113476) em que 99 estudantes reprovaram ambas as disciplinas e 65 desses estudantes desistiram do curso, o que corresponde a cerca de 9% dos dados analisados, a maior quantidade de estudantes é relacionada a taxa de Física 1 Experimental (118010), Física 1 (118001), Cálculo 1 (113034), Introdução a Algebra linear (113093) Algoritmo e Programação de Computadores (113476), em que 121 estudantes reprovaram as disciplinas e 73 desses estudantes desistiram do curso, correspondendo a cerca de 10% da quantidade total analisada.

Verifica-se que o maior índice de desistência condicional é dado para a reprovação da disciplina Introdução à Engenharia da Computação (117528), com uma taxa de 0.51 (51%) de desistência, como pode ser verificado na Figura 3.12. Percebe-se também que a reprovação dessa disciplina associada a outras carrega sempre as maiores taxas de desistência, dadas as disciplinas analisadas, por outro lado, as disciplinas como Algoritmos e Programação de Computadores (113476) e Cálculo 1 (113034) apresentam os menores índices de desistência individualmente, apesar das taxas ainda serem altas, são menores do que outras disciplinas.

As associações que contêm as disciplinas Cálculo 1 (113034) e Algoritmos e Programação de computadores (113476) apresentam as menores taxas, enquanto as associações que possuem a disciplina Introdução à Engenharia da Computação (117528) seguem com taxas maiores. Observando a Figura 3.12, percebe-se que a reprovação das disciplinas: Física 1 (118001), Física 1 Experimental (118010), Introdução a Engenharia da Computação (117528), Algoritmos e Programação de Computadores (113476) e Cálculo 1 (113034) indica uma probabilidade condicional de 0,71 (71%) de desistência dos estudantes, dessa forma é possível verificar que estudantes que reprovem essas disciplinas no mesmo semestre possuem uma alta chance de saírem o curso.

A partir desses valores encontrados, é possível verificar que a disciplina Algoritmos e Programação de Computadores (113476) é responsável pela desistência de 35% dos estudantes que a reprovam, um valor que não é tão alarmante quanto os outros, mas quando analisamos conjuntamente esse dado, com o dado da Figura 3.4, vemos que essa disciplina historicamente é responsável pela reprovação de 40% dos estudantes que a cursam, sendo assim um valor considerável e que deve ser analisado, possivelmente com o intuito de se reduzir essas taxas. Como colocado anteriormente essa disciplina representa os dados das disciplinas Computação Básica e Algoritmos e Programação de Computadores, ambas as disciplinas se encontram na lista de disciplinas com os maiores índices de reprovação.

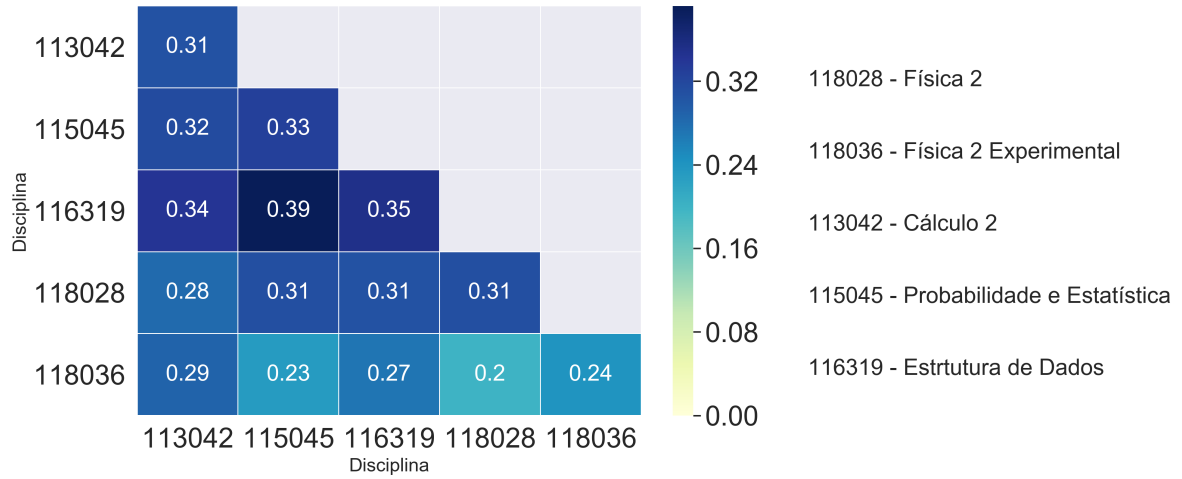


Figura 3.13: Taxa de desistência baseada nas disciplinas reprovadas no mesmo semestre

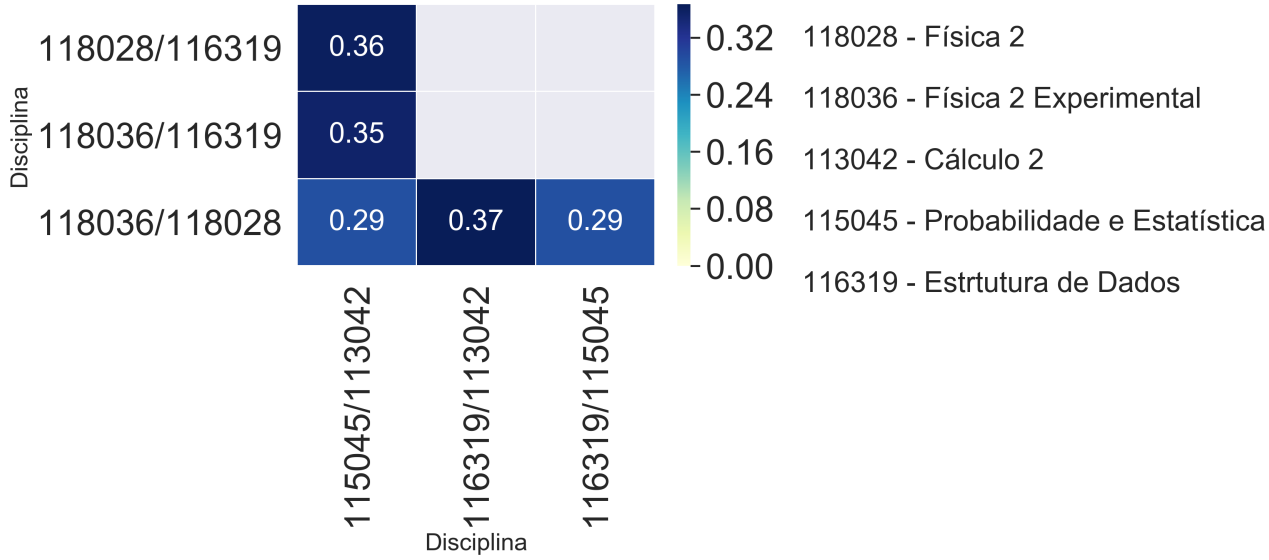


Figura 3.14: Taxa de desistência baseada nas disciplinas reprovadas no mesmo semestre

As Figuras 3.13 e 3.14 ilustram a taxa de reprovação para as disciplinas indicadas para o segundo semestre, nos mapas as disciplinas são descritas por seus códigos.

As taxas são obtidas a partir dos alunos que reprovaram nas dadas disciplinas e desistiram no curso, para a Figura 3.13 a menor quantidade de estudantes em relação as taxas é para Estrutura de Dados (116319) e Cálculo 2 (113042) em que 103 estudantes reprovaram ambas as disciplinas e 35 desses estudantes desistiram do curso, o que corresponde a cerca de 5% dos dados analisados, a maior quantidade de estudantes é relacionada a taxa de Probabilidade e Estatística (115045), em que 301 estudantes reprovaram as disciplinas e 99 desses estudantes desistiram do curso, correspondendo a cerca de 14% da quantidade total analisada. Enquanto que, para a Figura 3.14, a menor quantidade de estudantes em relação as taxas é para Física 2 Experimental (118036), Estrutura de Dados (116319), Cálculo 2 (113042) e Probabilidade e Estatística (115045), em que 107 estudantes reprovaram tais disciplinas e 38 desses estudantes desistiram do curso, o que corresponde a cerca

de 5% dos dados analisados, a maior quantidade de estudantes é relacionada à taxa de Física 2 Experimental (118036), Física 2 (118028), Probabilidade e Estatística (115045) e Estrtutura de Dados (116319), em que 118 estudantes reprovaram as disciplinas e 34 desses estudantes desistiram do curso, correspondendo a cerca de 5% da quantidade total analisada.

Percebe-se que as taxas das disciplinas indicadas para o segundo semestre são inferiores se comparadas às do primeiro semestre, cujos dados estão representados nas Figuras 3.11 e 3.13. De todo modo, os dados do segundo semestre ainda são significantes para entender a desistência dos estudantes, visto que esta se concentra no primeiro ano de curso, conforme ilustrado no Gráfico 3.10. Observar o conjunto de disciplinas com as maiores taxas de desistência associada a reprovação é uma forma de se entender esse fenômeno e, dessa forma, prover uma assistência aos estudantes.

3.4 Seção Final

Neste capítulo foram explicitadas as estatísticas sobre disciplinas dos departamentos ENE e CIC, onde é possível visualizar alguns dados sobre as disciplinas ofertadas que podem indicar complicações acerca do desempenho dos estudantes no curso. Outro ponto levantado foi a relação do desempenho dos estudantes com sua desistência, dessa forma é possível ver que existem correlações entre esses valores e utilizar esses dados como uma forma de foco de atuação é uma forma de se atuar no sentido de reduzi-los.

Ao longo deste capítulo foram evidenciados os índices das disciplinas Lógica Computacional 1 e Princípios de Comunicação como atípicos e potenciais causadores de problemas administrativos para a alocação de recursos e professores para ministrar a disciplina e atender as demandas dos estudantes. Verifica-se também que os índices de desistência dada a reprovação das disciplinas Introdução a Engenharia da Computação e Física 1 Experimental apresentam-se como os maiores. Utilizar esses dados como forma de amparar decisões e auxiliar coordenadores do curso é uma das aplicações possíveis para as informações geradas.

Capítulo 4

Análise de dados dos estudantes

Neste capítulo será discutida a relevância das informações socioeconômicas dos estudantes e o seu impacto no âmbito acadêmico. Como a análise da relação entre método de admissão, escola de origem e região administrativa com o desempenho escolar. Da mesma forma que o Capítulo 3 buscava encontrar relações entre as disciplinas e os estudantes, busca-se aqui encontrar relações sólidas que possam de alguma forma engrandecer a compreensão do desempenho dos estudantes no curso.

Os principais aspectos a serem analisados nesse capítulo serão:

1. Distribuição dos estudantes ativos, desistentes e formados: compreender como se dá o processo de desistência e formatura dos estudantes no curso;
2. Relação entre rendimento acadêmico e instituição prévia ao ingresso: compreender o impacto desse fator na relação dos estudantes e seu rendimento;
3. Relação entre o rendimento acadêmico e o método de admissão: verificar a relação entre a forma pelo qual o aluno ingressou e a relevância para o seu rendimento;
4. Relação entre o rendimento acadêmico e a região administrativa do estudante: relacionar a região administrativa do estudante ao ingressar na faculdade com seu rendimento acadêmico e identificar fatores relevantes nesse processo

4.1 Metodologia

Os dados que deram origem a esta análise provêm da mesma base utilizada na Seção 3.3, contendo o histórico escolar de 700 estudantes do ano de 2009 a 2018.

1. Coleta dos dados: A extração dos dados da análise foi feita por meio do sistema de *Business Intelligence*, BI, da Universidade de Brasília cujo acesso foi fornecido pelo *Decanato de Planejamento, Orçamento e Avaliação*, DPO, ao coordenador do curso de Engenharia

de Computação, José Edil Guimarães de Medeiros. Os dados coletados relacionam, o método de ingresso do aluno e seu rendimento, instituição prévia e seu rendimento e código de endereçamento postal (CEP) e seu respectivo rendimento.

2. Anonimização: Para os dados utilizados nesse capítulo não foi necessário anonimizá-los, visto que nenhuma informação permite identificar univocamente um estudante, há apenas a relação dos valores de rendimento acadêmico com outras informações, como pode-se ver nas Figuras 4.1, 4.2 e 4.3.

	Semestre/Ano	CEP	0	0.0333	0.0625	0.08	0.1473	0.1667	0.1778	0.1818	...	4.6969	4.7006	4.7425	4.75	4.7647	4.8244	4.8379	4.8621
0	1996/1	71699016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	1996/2	71699016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	1997/1	71699016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	1997/2	71699016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	1998/1	71699016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figura 4.1: Modelo de dados utilizados com CEP e IRA

Forma de ingresso	AcordoCultural-PEC-G	Convênio-Andifes	Convênio-Int	Enem	MatrículaCortesia	PIE	PortadorDiplomCursoSuperior	ProgramadeAvaliaçãoSeriada	Seleção
0.0000	1.0	1.0	0.0	2.0	1.0	0.0	0.0	18.0	16.0
0.0333	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0625	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
0.0800	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.1473	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Figura 4.2: Modelo de dados utilizados com forma de ingresso e IRA

	IRA	Não_informado	Particular	Pública
0	0.0000	17.0	37.0	33.0
1	0.0333	1.0	0.0	0.0
2	0.0625	0.0	0.0	1.0
3	0.0800	0.0	1.0	0.0
4	0.1473	1.0	0.0	0.0

Figura 4.3: Modelo de dados utilizados com instituição prévia e IRA

3. Limpeza: Não foi necessário remover nenhum dado das análises realizadas neste capítulo, entretanto algumas informações foram adicionadas, para construir a relação da região administrativa do estudante e seu desempenho foi convertido a informação de Código de Endereçamento Postal (CEP) que existiam nos dados para latitude e longitude, para tal foi utilizada a *Application Programming Interface* (API) dos Correios para, a partir do CEP do estudante, obter-se a região administrativa, após esse processo aglutinou-se as informações dos estudantes e então obteve-se os valores de latitude e longitude de cada bairro para desenvolver os mapas de distribuição dos estudantes.

4.2 Impacto do Método de Admissão e Instituição Prévia no Rendimento

A performance de um estudante ao longo do curso é significativa para a formação de um bom profissional, para tanto, a preocupação com os fatores que afetam o desempenho dos estudantes devem ser estudados e explicitados. No Capítulo 3 foram apresentados dados que relacionam as disciplinas cursadas com o desempenho dos estudantes e sua desistência, aqui busca-se apresentar outras relações que podem afetar esse desempenho, com um foco maior nos estudantes.

Para se compreender melhor o objeto de estudos, pode-se observar a Figura 4.4 em que vemos o que acontece com os estudantes de acordo com os anos completados de curso. Observando o gráfico, percebe-se que boa parte dos alunos desistem do curso nos dois primeiros anos e que boa parte deles se formam entre o quinto e sexto anos cursados.

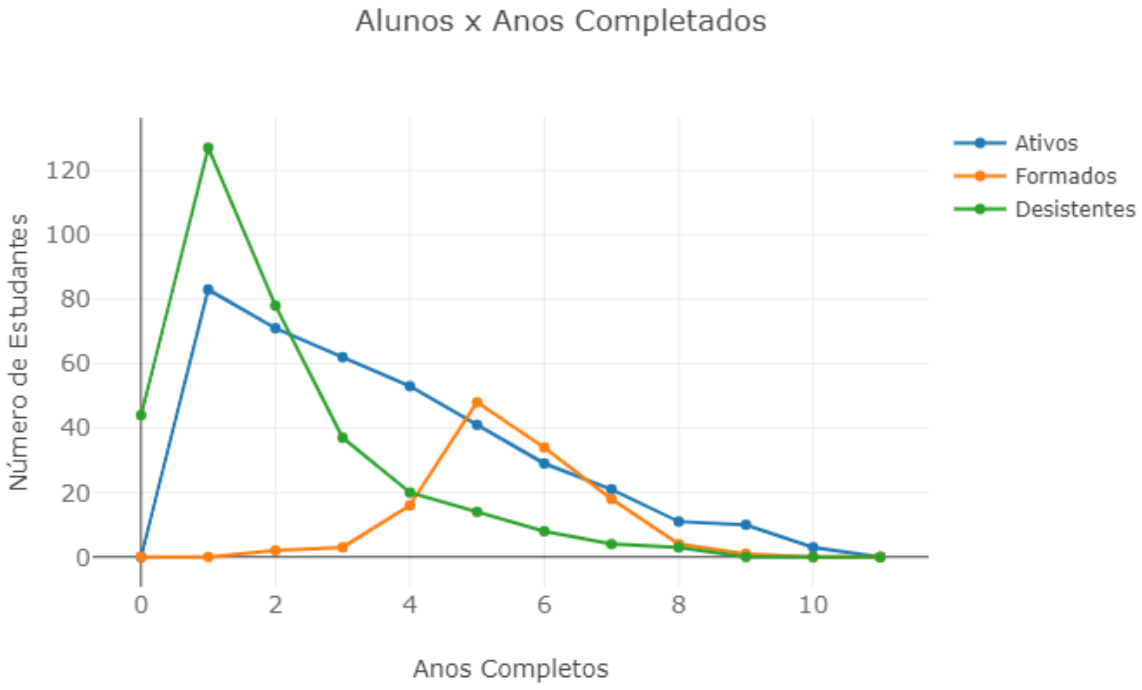


Figura 4.4: Comportamento do estudantes ao longo dos anos

A partir disso é possível ter uma compreensão maior do comportamento do objeto de estudo. A alta desistência dos estudantes na primeira metade do curso, e tendo em vista que durante esse período os estudantes cursam diversas disciplinas de base, no Capítulo 3 foi mostrado que existe um certo nível de correlação entre a reprovação das disciplinas de começo de curso e a desistência dos estudantes, foi verificado que essa correlação é mais acentuada nas disciplinas do primeiro semestre quando comparadas com as disciplinas do segundo semestre. Esse fato pode indicar que o problema da desistência vai além da relação direta dos mesmos com as disciplinas.

Observando a Figura 4.5, pode-se verificar que a distribuição dos Índices de Rendimento Aca-

dêmico (IRAs) dos estudantes é razoavelmente similar entre aqueles que ingressam partindo de uma escola pública e de uma escola privada. O que pode indicar que esse fator não é determinante para determinar o rendimento do estudante.

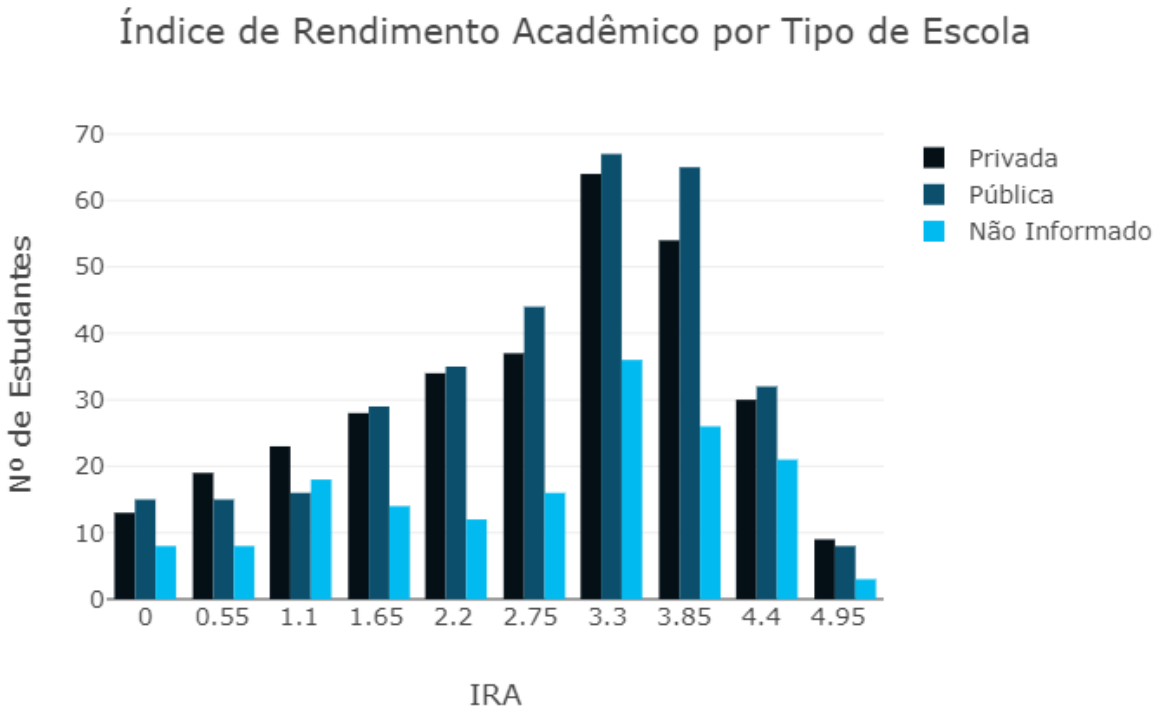
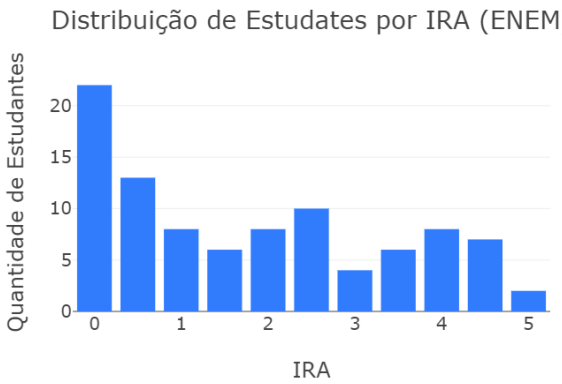


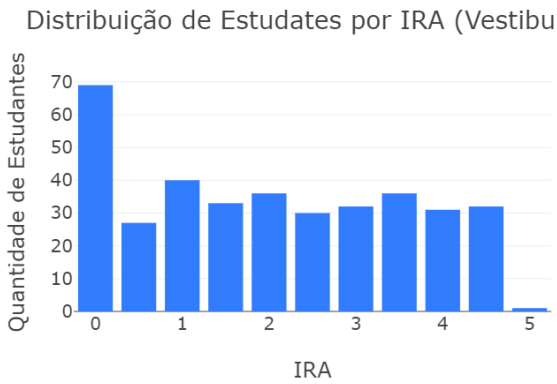
Figura 4.5: Distribuição IRA por tipo de instituição prévia

Uma outra análise possível é verificar o modo de acesso do estudante ao curso, para tal, realizou-se uma análise dos estudantes separando os três principais métodos de acesso ao curso, os quais são: vestibular tradicional, Programa de Avaliação Seriada (PAS) e Exame Nacional do Ensino Médio (ENEM), via Sistema de Seleção Unificada (Sisu). Esses três métodos de acesso são responsáveis pela forma de acesso de 620 do 700 estudantes analisados, os outros entraram por outros métodos como vestibular para portadores de diplomas, Programa de Estudantes-Convênio de Graduação (PEC-G) e outros. Esses 620 alunos são distribuídos como sendo 94 ingressantes pelo ENEM, 367 pelo vestibular e 159 pelo PAS.

A partir da Figura 4.6 é possível verificar distribuições razoavelmente parecidas entre os métodos de acesso ao curso, os estudantes que utilizam o ENEM como forma de entrada no curso aparentam ser os mais destoantes entre os três, uma possível justificativa é o fato dessa forma de entrada ser a que possui a menor quantidade de estudantes no dados analisados, porém ainda assim as distribuições se assemelham.



(a) Distribuição dos estudantes por IRA (ENEM)



(b) Distribuição dos estudantes por IRA (Vestibular)

(c) Distribuição dos estudantes por IRA (PAS)

Figura 4.6: Distribuição de menções departamentos ENE e CIC

Uma justificativa para a leve diferença entre as distribuições pode ser visualizada na Figura 4.7, dado que o vestibular e o PAS são os métodos mais antigos de admissão e existem desde o começo do curso, suas distribuições são mais parecidas, o que diferencia as duas é o fato de o vestibular ter reduzido o número de vagas para admissão em detrimento do ENEM. Como se pode analisar, apesar de terem percentuais diferentes, as proporções entre alunos formados e desistentes para o PAS e vestibular aparentam ser semelhantes, enquanto para o ENEM essa distribuição parece ser completamente atípica, visto que não houve tempo hábil para a formatura de muitas turmas após a inserção deste modo de ingresso na UnB, fator que pode justificar as diferenças nas distribuições de IRAs da Figura 4.6. Dada a semelhança entre as distribuições para os três métodos de ingresso, salvas as obervações acerca da proporção de formandos, se pode concluir que o método de ingresso não é um fator relevante para a desistência.

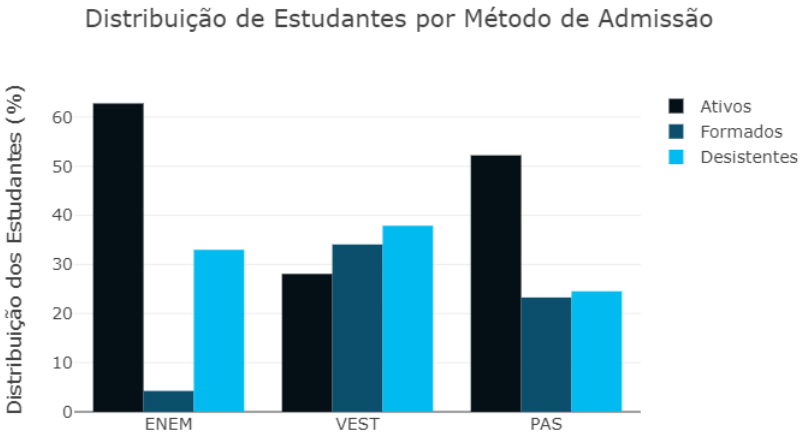


Figura 4.7: Distribuição de Estudantes por Método de Admissão

4.3 Rendimento Acadêmico e Distribuição Geográfica

Para além das relações da instituição de ensino do estudante prévias ao ingresso na universidade, uma outra análise feita foi a relação entre a região administrativa do estudante e o seu desempenho acadêmico.

Ao ingressar na universidade os estudantes fornecem informações de ondem residem, apesar de alguns dos estudantes poderem mudar, parte-se do pressuposto que alguns ainda se manterão em sua região administrativa por boa parte do curso e relacionar essa informação com o desempenho destes estudantes pode auxiliar no entendimento de alguns fenômenos que ocorrem no curso.

A partir dessas informações foram confeccionados mapas de distribuição dos estudantes com relação a sua informação de residência e seu desempenho acadêmico, foram agregados os estudantes que moravam nos mesmos bairros e a partir desse conjunto foi calculado o IRA médio de cada aglomerado desses, esse valor foi obtido para todos os alunos ativos no semestre dado e com informações sobre região administrativa no banco de dados utilizados. As Figuras 4.8, 4.9 e 4.10 ilustram a relação entre o IRA e a distribuição geográfica dos estudantes no DF, em que o tamanho da circunferência indica a quantidade de estudantes no conjunto e a cor representa o IRA médio daquele conjunto, variando de vermelho que indica IRA 0, ou bem próximo a esse valor, e azul indica IRA 5, ou bem próximo desse valor, e tons de verde indicam IRA próximos a 3.

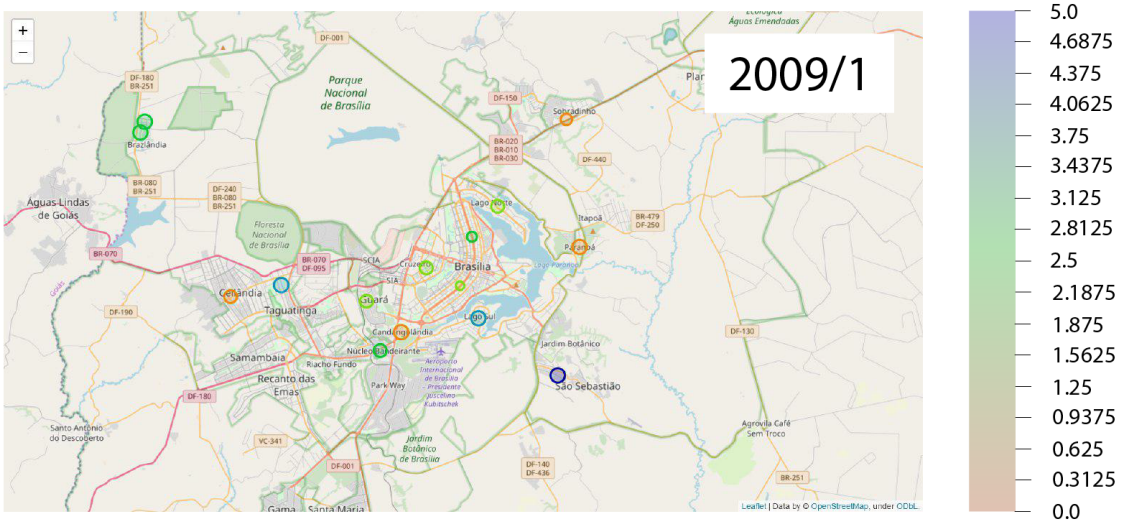


Figura 4.8: Mapa De Distribuição de Estudantes com Coloração Baseada no IRA (2009/01)

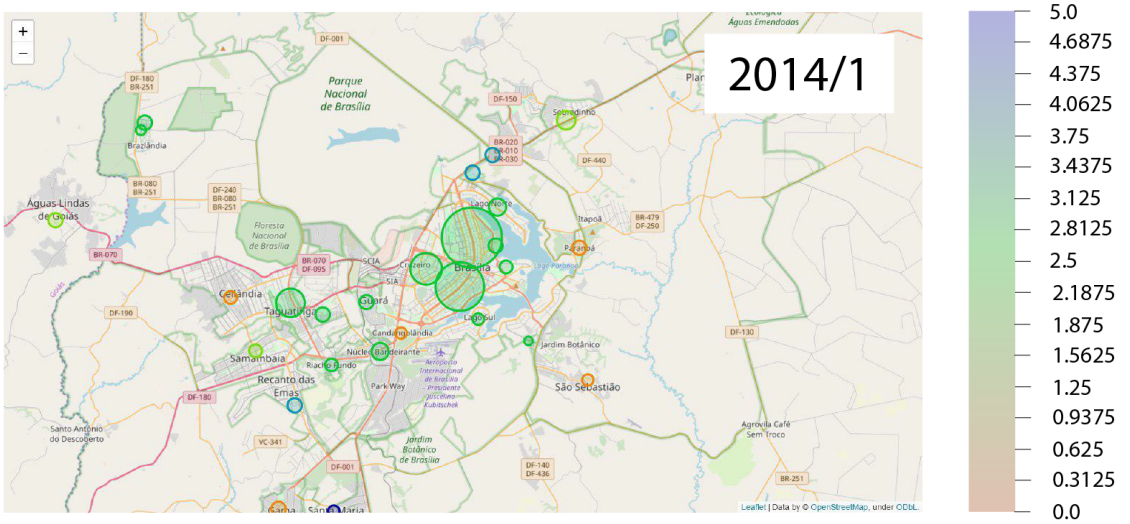


Figura 4.9: Mapa De Distribuição de Estudantes com Coloração Baseada no IRA (2014/01)

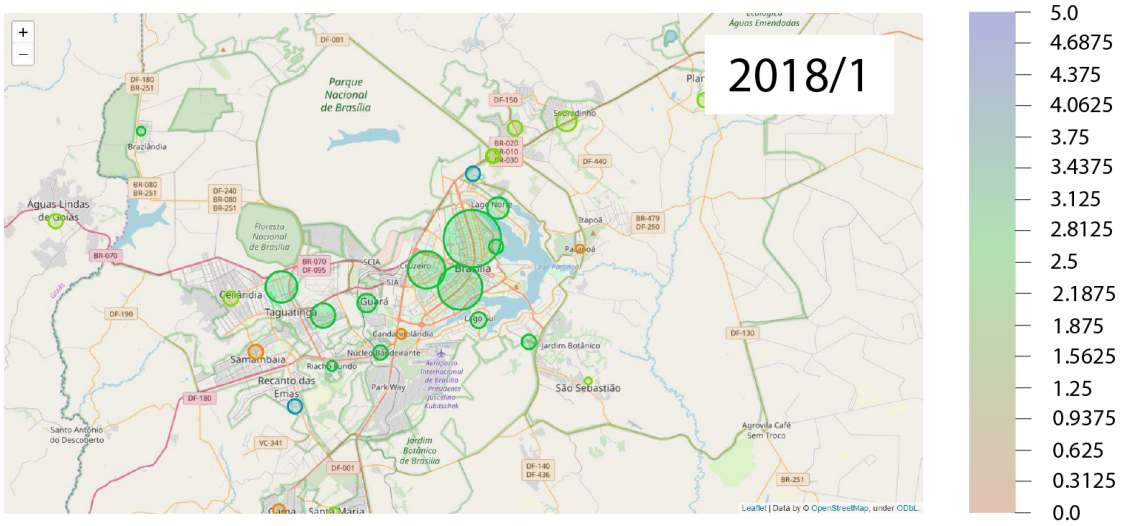


Figura 4.10: Mapa De Distribuição de Estudantes com Coloração Baseada no IRA (2018/01)

A partir das Figuras 4.8, 4.9 e 4.10 é possível verificar dois fenômenos interessantes, o primeiro é a distribuição da região de residência dos estudantes ao longo dos anos em que se percebe que o número de estudantes é concentrado nas regiões centrais de Brasília, como Asa Sul e Asa Norte e grupos ligeiramente menores, mas também predominantes em Taguatinga e a região que engloba Cruzeiro, Sudoeste e Octogonal, a partir do mapa do ano de 2014, Figura 4.9, podemos ver um crescimento de estudantes de outros bairros como Guará e Águas Claras. A partir dessa sequência de imagens é possível ver a progressão da região administrativa de ingresso dos estudantes do curso de Engenharia da Computação ao longo dos anos e suas concentrações, que por si só é um dado interessante para questões administrativas e desenvolvimento do perfil do estudante que ingressa no curso.

O segundo fenômeno que se pode analisar é a média do IRA de algumas regiões se mantendo em valores abaixo de 3 ao longo dos anos, mesmo com o crescimento do número de estudantes dessas regiões , como por exemplo ocorre com a região de Sobradinho que ao se observar nas Figuras 4.8, 4.9 e 4.10, apesar da evolução da sua média, sempre mantém seu valor ligeiramente abaixo de 3, esses dados podem ser visualizados de forma mais nítido nas Tabelas 4.1, 4.2 e 4.3.

Utilizando-se da pesquisa realizada por (PLANEJAMENTO DO DISTRITO FEDERAL – CODEPLAN, 2012), foi desenvolvida a Tabela 4.4 em que se vê a distribuição de renda *per capita* por região administrativa do Distrito Federal (DF), relacionando as tabelas com o IRA médio dos estudantes, com essa informação é possível realizar outras interpretações dos dados, como por exemplo: as regiões administrativas que possuem renda *per capita* superior a média do DF, como Asa Sul, Asa Norte, Núcleo Bandeirante, Guará, Cruzeiro, Sudoeste/Octogonal, Lago Sul, Águas Claras, Jardim Botânico e Vicente Pires, mostram-se como as que são origem da maior parte dos estudantes e a média de IRA desses estudantes se mostra como 3, ou acima desse valor, como pode-se ver nas regiões, enquanto regiões com renda *per capita* inferior a essa média como Gama, Planaltina, Paranoá, Ceilândia, Samambaia, Santa Maria, Candangolândia são origem de uma quantidade menor de estudantes e o desempenho acadêmico médio deles se mostra como abaixo

de 3.

Essa informações demonstram que o perfil do estudante que está no curso de engenharia da computação, apesar de estar mudando com o decorrer dos anos, se demonstra como majoritariamente de pessoas que provém de regiões administrativas com renda *per capita* superior a média do DF e essa informação aparenta estar correlacionada com o desempenho do estudante.

Tabela 4.1: Distribuição de IRA e Estudantes por bairro (2009/01)

Região Administrativa	Qtd. Estudantes	IRA (Médio)
Asa Sul	6	3.0
Cruzeiro / Sudoeste / Octogonal	2	2.3
Asa Norte	5	3.1
Guara	2	2.0
Lago Norte	2	2.3
Paranoá	1	2.0
Jardins Mangueiral	1	0.7
Candangolândia	1	1.6
Ceilândia	2	1.4
Sobradinho	3	1.5
Núcleo Bandeirante	2	3.0
Brazlândia	1	3.1
Lago Sul	1	4.0
Vicente Pires	1	3.9
Taguatinga	1	4.3

Tabela 4.2: Distribuição de IRA e Estudantes por bairro (2014/01)

Região Administrativa	Qtd. Estudantes	IRA(Médio)
Taguatinga	30	3.2
Asa Sul	50	3.2
Cruzeiro / Sudoeste / Octogonal	33	3.3
Asa Norte	62	3.3
Guara	14	3.2
Lago Norte	17	3.2
Brazlândia	5	3.1
Lago Sul	12	3.4
Jardim Botânico	9	3.2
São Sebastião	3	1.8
Candangolândia	3	1.6
Núcleo Bandeirante	17	3.0
Águas Claras	14	3.4
Ceilândia	13	1.9
Samambaia	2	2.8
Santa Maria	3	0.2
Sobradinho	19	2.9
Gama	1	1.9
Paranoá	1	1.8
Planaltina	5	2.6
Setor de Mansões Isoladas Norte	1	3.0
Vila Planalto	2	3.8
Riacho Fundo	2	3.8
Vicente Pires	1	3.9
Recanto das Emas	1	4.5
Varjão	1	4.1

Tabela 4.3: Distribuição de IRA e Estudantes por bairro (2018/01)

Região Administrativa	Qtd. Estudantes	IRA (Médio)
Asa Sul	46	3.4
Cruzeiro / Sudoeste / Octogonal	38	3.4
Asa Norte	59	3.0
Guara	19	3.0
Lago Norte	22	3.2
Paranoá	6	1.9
Lago Sul	16	3.7
Jardim Botânico	14	3.4
São Sebastião	7	2.5
Núcleo Bandeirante	15	3.3
Candangolândia	4	1.1
Águas Claras	25	3.7
Taguatinga	33	3.5
Ceilândia	15	2.7
Samambaia	1	1.9
Sobradinho	20	2.9
Gama	3	1.5
Santa Maria	4	2.5
Santo Antônio do Descoberto	1	2.0
Planaltina	7	2.7
Setor de Mansões Isoladas Norte	1	3.0
Brazlândia	6	3.0
Riacho Fundo	4	3.6
Recanto das Emas	1	4.5
Varjão	1	4.1

Tabela 4.4: Renda Per Capita em Salários Mínimos (2012)

Região Administrativa	Renda Mensal per Capita
Brasília	6,7
Gama	1,86
Taguatinga	2,41
Brazlândia	1,18
Sobradinho	2,67
Planaltina	1,16
Paranoá	0,89
Núcleo Bandeirante	2,55
Ceilândia	1,18
Guará	3,4
Cruzeiro	3,71
Samambaia	1,06
Santa Maria	1,21
São Sebastião	0,92
Recanto das Emas	0,9
Lago Sul	10,56
Riacho Fundo	1,56
Lago Norte	8,93
Candangolândia	1,5
Águas Claras	4,36
Riacho Fundo II	1,03
Sudoeste/Octogonal	8,67
Varjão	0,78
Park Way	6,71
SCIA (Estrutural)	0,56
Sobradinho II	2,44
Jardim Botânico	6,33
Itapoã	0,63
Setor de Ind. E Abastecimento	1,52
Vicente Pires	3,13
Distrito Federal	2,42

4.4 Seção Final

Neste capítulo foi discutido a relação entre o desempenho dos estudantes do curso de Engenharia da Computação e informações socio-econômicas para se compreender melhor fatores que podem influenciar no desempenho do estudante durante o curso.

Ao longo do capítulo foi mostrado que a diferença entre o desempenho dos estudantes que provêm de instituições públicas e privadas possuem diferença bem baixa. O desempenho dos estudantes também não parece ser afetado pela forma de ingresso quando se compara PAS e vestibular, enquanto os estudantes que ingressaram no curso por meio do ENEM aparentam possuir um desempenho diferente das outras duas formas trabalhadas, foi mostrado que como esse método de ingresso possui uma distribuição de estudantes ativos, formados e desistentes diferentes, essa diferença pode ser justificada, e ao passo que o tempo pode igualar esses perfis e o desempenho dos estudantes, visto que o ENEM entre os 3 métodos analisados foi o que começou a ser adotado por último.

Outra relação traçada foi entre a região administrativa do estudante ao ingressar no curso e seu desempenho acadêmico, verificou-se que regiões com renda *per capita* superior são da onde a maior parte dos estudantes provém e esse fator aparenta estar correlacionado com o desempenho dos estudantes. Essas informações são extremamente interessantes para se traçar o perfil do estudantes do curso e potencialmente compreender melhor fenômenos acadêmicos diversos.

Capítulo 5

Predição de Evasão

Neste capítulo será realizada uma análise acerca do histórico dos alunos de Engenharia de Computação a partir das menções obtidas nas disciplinas obrigatórias do curso em cada semestre da grade curricular. O objetivo deste capítulo é a construção de um modelo de predição de evasão que possa fundamentar a tomada de decisão dos coordenadores do curso acerca da construção de políticas de auxílio para os alunos em risco de desligamento.

O sistema de predição proposto baseia-se em Árvore de Decisão, um modelo estatístico que utiliza um método de treinamento supervisionado para classificação e previsão de dados. A escolha do algoritmo utilizado se deu por uma série de motivos, explicitados a seguir.

1. Facilidade de interpretação: Dado que o objetivo da construção deste sistema é auxiliar o processo de tomada de decisão dos coordenadores, a utilização de um modelo inteligível e compreensível é fundamental, por isso, uma estrutura gráfica como a apresentada por uma árvore de decisão se mostra como uma alternativa adequada.

2. Capacidade em lidar com diferentes tipos de dados: Foi explorada na Seção 4 a influência de aspectos sociais no desempenho dos estudantes, revelando a importância de informações como método de ingresso, local de moradia, tipo de escola de origem, dentre outros. Apesar de esta análise ter sido feita exclusivamente com os dados do histórico escolar dos alunos, a escolha de um modelo de predição que possibilite o uso de diferentes tipos de dados se mostra importante para a evolução futura do sistema.

3. Tempo de aprendizado: Quando comparado a outros algoritmos de classificação, a árvore de decisão apresenta um tempo de treinamento consideravelmente reduzido, o que permite uma ágil alteração do vetor de características (*features*) utilizadas devido a mudanças na grade curricular, a exemplo da substituição da disciplina Computação Básica pela nova Algoritmos e Programação de Computadores.

5.1 Metodologia

Os dados que deram origem a esta análise provêm da mesma base utilizada na Seção 3.3, contendo o histórico escolar de 700 estudantes do ano de 2009 a 2018. A Figura 5.1 ilustra todas as etapas do processo de extração, limpeza e modelagem dos dados utilizados na análise.

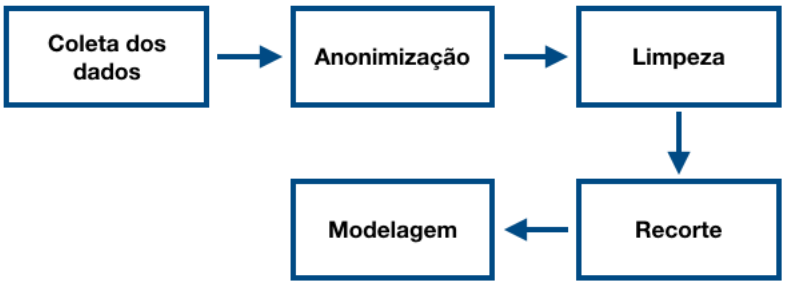


Figura 5.1: Fluxograma de extração, limpeza e modelagem dos dados.

1. Coleta dos dados: A extração dos dados da análise foi feita por meio do sistema de *Business Intelligence*, BI, da Universidade de Brasília cujo acesso foi fornecido pelo *Decanato de Planejamento, Orçamento e Avaliação*, DPO, ao coordenador do curso de Engenharia de Computação, José Edil Guimarães de Medeiros. Os dados coletados relacionam, para cada aluno, a menção nas disciplinas cursadas em um semestre e a forma de saída do mesmo.

2. Anonimização: A fim de manter o sigilo dos alunos analisados, o coordenador do curso substituiu o CPF de cada indivíduo por um código identificador, de modo que o número de matrícula dos estudantes não foi disponibilizado para os autores deste trabalho. A Figura 5.2 ilustra o formato dos dados após a etapa de anonimização, na qual o conteúdo da coluna *id_ aluno* foi substituído por um código identificador.

id	id_aluno	ano_ingresso	semestre_ingresso	cod_disciplina	mencao	ano_referencia	semestre_referencia	Status
0	aluno88	2008	2	113093	SR	2009	2	Formatura
1	aluno88	2008	2	118001	MI	2009	2	Formatura

Figura 5.2: Fluxograma de extração, limpeza e modelagem dos dados.

3. Limpeza: O conteúdo da coluna *status* de cada aluno foi determinado como o conjunto de classes do problema, podendo assumir os valores 'Formatura', 'Evasão', 'Falecimento' e 'Ativo'. Na análise, foi desconsiderada a classe 'Falecimento', por não estar relacionada de forma direta ao desempenho estudantil. Além disso, não é possível fazer a distinção do modo de saída dos alunos da classe 'Ativo' visto que os mesmos ainda estão no decorrer do curso. Como o objetivo deste trabalho é a predição de evasão, estes alunos podem ser utilizados para a validação futura do sistema, mas não para o seu treinamento. Portanto, os alunos da classe 'Ativo' também foram desconsiderados.

   A Figura 5.3 ilustra a proporção entre a quantidade de alunos por classe da base de dados utilizada. Do total de 700 alunos, foram retirados 284 alunos ativos e 2 falecidos, restando

portanto um total de 414 alunos divididos entre as classes evasão, com 289 alunos, e formatura, com 125 alunos. Além disso, os dados foram divididos em um conjunto para treinamento do modelo, contendo 70% dos dados, e outro para a fase de testes, com 30%, resultando em um total de 290 elementos para o treino e 124 para o teste.

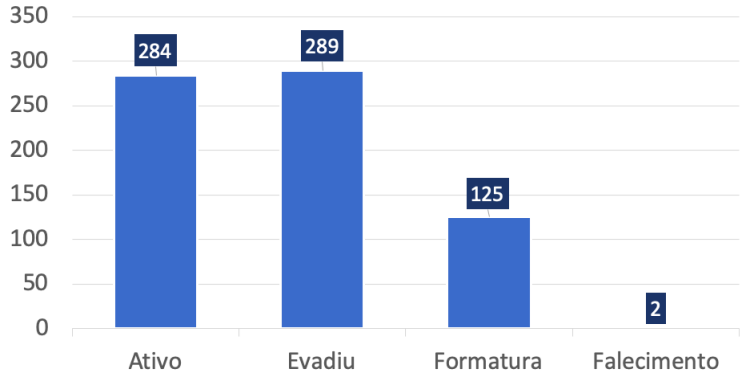


Figura 5.3: Quantidade de alunos por classe.

4. Recorte: A análise feita na Seção 4 aponta que historicamente a evasão estudantil se concentra no primeiro ano do curso. Isso possibilitou um direcionamento acerca do grupo de alunos a ser analisado. Para tal, a nota dos alunos nas matérias obrigatórias dos dois primeiros semestres do fluxo, listadas abaixo, foram elencadas como os atributos da análise.

   - Cálculo 1
   - Introdução à Álgebra Linear
   - Algoritmos e Programação de Computadores
   - Introdução à Engenharia de Computação
   - Física 1
   - Física 1 Experimental
   - Física 2
   - Física 2 Experimental
   - Cálculo 2
   - Probabilidade e Estatística
   - Estruturas de Dados

5. Modelagem: As menções obtidas em cada disciplina foram utilizadas como os atributos do modelo. Para tal, cada menção recebeu um valor numérico que representasse o seu desempenho associado, de modo que valores positivos indicam uma aprovação na disciplina, valores negativos indicam reprovação ou trancamento, e um valor nulo indica a não realização da disciplina. Os pesos associados a cada menção estão relacionados na Tabela 5.1.

   Como o cálculo do Índice de Rendimento Acadêmico da UnB considera um valor menor para reprovação quando comparada a um trancamento, essa característica foi mantida na

Tabela 5.1: Peso definido para cada menção.

Menção	Significado	Peso definido
SS	Nota final entre 9 e 10	3
MS	Nota final entre 7 e 9	2
MM	Nota final entre 5 e 7	1
MI	Nota final entre 3 e 5	-1
II	Nota final entre 1 e 3	-2
SR	Nota final entre 0 e 1	-3
CC	Crédito Concedido	0
TR	Trancamento	-0.5
TJ	Trancamento justificado	0
–	Disciplina não cursada	0

análise em questão, associando à menção TR um valor intermediário entre -1, que representa uma reprovação com menção MI, e 0, que representa uma disciplina não cursada. De modo similar, o peso definido para as menções CC e TJ foi nulo porque estas não são consideradas no cálculo do Índice de Rendimento Acadêmico.

Para os alunos que cursam a mesma disciplina diversas vezes o valor correspondente ao atributo foi calculado como a soma dos pesos das menções obtidas a cada tentativa, a fim de introduzir uma informação temporal na análise. Desse modo, a pontuação de um aluno se torna mais negativa conforme o número de reprovações aumenta.

Como a disciplina de Computação Básica foi substituída por Algoritmos e Programação de Computadores, os códigos das duas foram igualados nesta análise.

5.2 Modelo de Predição

Os modelos de aprendizado podem ser classificados em caixa preta e orientado a conhecimento. Por um lado, os sistemas do tipo caixa preta possuem uma estrutura interna complexa e de difícil interpretação, limitando a análise do problema à relação entre entradas e saídas. Em contrapartida, os sistemas do tipo orientado a conhecimento, objetos de estudo desta seção, consistem na obtenção de modelos que sejam de fácil utilização e entendimento.

Neste capítulo foi proposta a utilização de um algoritmo CART *Classification And Regression Tree*, primeiramente introduzido por (BREIMAN et al., 1984). A árvore de decisão é um tipo de modelo orientado à conhecimento que utiliza a estratégia de dividir para conquistar, por meio da qual um problema complexo é decomposto em sub-problemas mais simples de modo recursivo, conforme definido por (GAMA, 2004). Este modelo consiste em regras 'se-então' de fácil compreensão através da divisão das amostras existentes em conjuntos menores até que este seja pequeno o suficiente para representar uma classe. A estrutura de dados de uma árvore é composta por:

1. Nó de decisão: Representa cada subconjunto de uma árvore.
2. Raiz: É o primeiro nó da árvore, que representa toda a amostra a ser dividida.

3. Regra: Teste de atributo que divide o conjunto em duas partes distintas

4. Folhas: Nós que não se dividem e que representam alguma classe.

A Figura 5.4 contém a representação de uma árvore de decisão. O nó raiz, elipse X1, contém todo o conjunto, que pode assumir valores distintos a1, a2, etc. Os nós de decisão, elipses X2, X3, X4, correspondem aos subconjuntos a serem divididos por meio das regras, que compreendem testes de atributo. Por fim, cada retângulo ilustra uma folha, contendo a representação de uma classe.

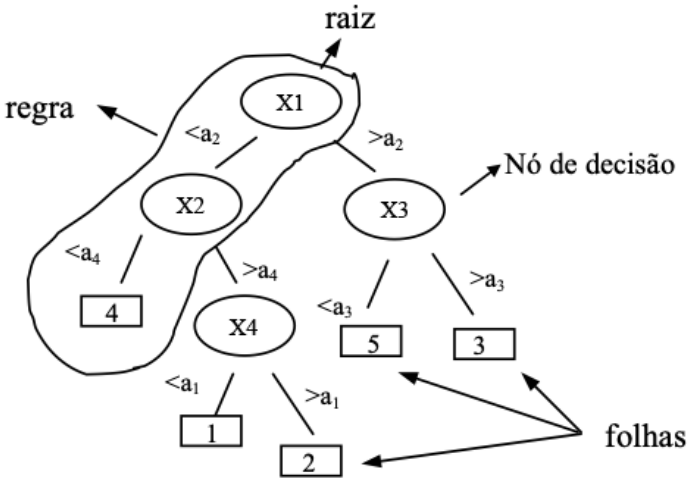


Figura 5.4: Representação da árvore de decisão

O critério utilizado na definição das regras que particionam o conjunto de exemplos foi o índice de Gini, desenvolvido por Conrado Gini em 1912, que mede o grau de heterogeneidade dos dados. Logo, pode ser utilizado para medir o grau de impureza de um nó (ONODA, 2001). A escolha deste, em detrimento de outros critérios para a partição, se deu por adequação aos dados na apresentação de melhores resultados para a predição de desistência.

A Equação (5.1) ilustra o cálculo de índice de Gini em um determinado nó, que se aproxima de zero conforme o nó se torna mais puro. No processo de indução (treinamento) de uma árvore busca-se isolar num ramo os exemplos do conjunto de treinamento que representam uma determinada classe.

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \tag{5.1}$$

Onde:

p_i é a frequência relativa de cada classe em cada nó

c é o número de classes

Desse modo, a regra que compõe cada nó de decisão é a que melhor consegue dividir os dados

em dois conjuntos com graus de impureza reduzidos. A Figura 5.5 ilustra a estrutura de cada nó de modo individual, na qual os elementos associados são:

1. *Rule*: Teste de atributo que divide o conjunto em duas partes distintas.
2. *Gini*: Indica o grau de impureza do conjunto em questão. Quando gini é nulo, o nó é totalmente puro, de modo que todos os seus elementos pertencem à mesma classe. Quando Gini é igual a 0.5, o nó é totalmente impuro, indicando que o conjunto em questão é composto pelo mesmo número de elementos de cada classe. A informação do grau de impureza foi transmitida para a representação gráfica do nó através de um sistema de cores a fim de facilitar a visualização da sua classe mais presente. Os tons mais fortes de azul indicam um conjunto muito puro de evasão, cores muito próximas de laranja indicam um conjunto muito puro de formatura, consequentemente a cor branca indica que o conjunto não pertence majoritariamente a nenhuma das duas classes, representando o índice de Gini igual a 0.5.
3. *Samples*: À medida que a árvore é percorrida, ocorrem divisões dos exemplos em um nó, fazendo com que a quantidade de elementos seja reduzida em cada folha, começando pelo total de amostras no nó raiz. Este atributo indica a quantidade de elementos presentes no nó específico.
4. *Value*: Vetor contendo a quantidade de elementos em cada classe. Para o nó em questão o primeiro valor corresponde aos elementos da classe 'Formatura' e o segundo aos elementos da classe 'Evasão'. A soma dos elementos do vetor é igual ao '*samples*'.
5. *Class*: Classe que representa a maioria dos elementos do nó. Se for uma folha, este atributo representa a classe da predição.

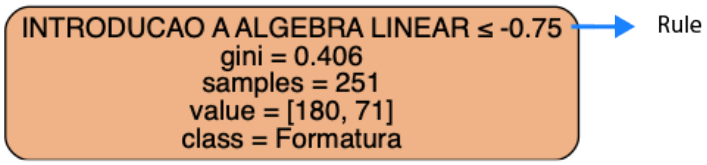


Figura 5.5: Estrutura de um nó.

No algoritmo de indução, a árvore estende a sua profundidade até o ponto de classificar perfeitamente os elementos do conjunto de treinamento. Quando este não possui ruído, o número de erros no treinamento pode ser zero. Quando este conjunto, entretanto, é ruidoso ou não representativo, este algoritmo pode produzir árvores com *overfitting*.

Uma forma de avaliar a árvore gerada para um determinado conjunto de teste é fazer o cálculo da acurácia:

$$\text{Acurácia} = \frac{\text{Número de elementos corretamente classificados}}{\text{Número total de elementos}} \tag{5.2}$$

O gráfico da Figura 5.6 ilustra a acurácia das árvores geradas por 50 rodadas independentes de treinamento. Seu comportamento instável reafirma a necessidade da determinação de um conjunto de treino adequado. Por um lado, árvores com acurácia muito baixa indicam que o *dataset* de treinamento gerou um modelo incapaz de prever o comportamento real dos dados. Por outro, árvores com acurácia muito acima da média podem indicar uma divisão enviesada entre validação e treinamento. A fim de evitar ambos os cenários, o conjunto escolhido, representado pelo marcador vermelho, foi aquele que obteve acurácia mediana dentre os avaliados.

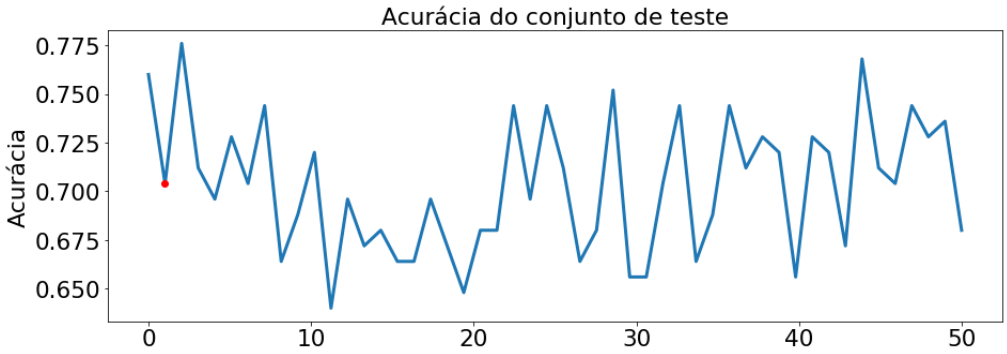


Figura 5.6: Acurácia de teste para diversos conjuntos de treinamento

Com auxílio da biblioteca *Scikit-learn* de código aberto para a linguagem de programação *Python*, foi gerada uma árvore de decisão para o *dataset* escolhido, que pode ser visa na Figura 5.7, e ser conferida com mais detalhes no Anexo III. Como os exemplos de treinamento são apenas uma amostra de todos os exemplos possíveis, foi possível adicionar arestas na árvore que melhoraram seu desempenho nos dados de treinamento, mas que pioram seu desempenho em um conjunto de teste, gerando *overfitting*. Para resolver este problema, é necessário realizar a *poda* da árvore, que pode ser feita de dois modos.

A pré-poda consiste em interromper o crescimento da árvore enquanto ela é induzida, baseado no ajuste de alguns parâmetros, como o número máximo de nós, profundidade máxima, grau de impureza do nó, dentre outros. Entretanto, segundo (MONARD; BARANAUSKAS, 2003), a pré-poda pode interromper o crescimento da árvore em um estágio muito prematuro, evitando assim, a obtenção do melhor modelo para os dados analisados.

De modo contrário pode-se fazer a pós-poda, na qual uma árvore é induzida até o seu tamanho máximo e, em seguida, é podada por meio de métodos recursivos gerando diversas sub-árvores e possibilitando a escolha daquela com o melhor desempenho. Este processo é computacionalmente mais ineficiente, mas pode gerar árvores mais adequadas quando comparado à pré-poda.

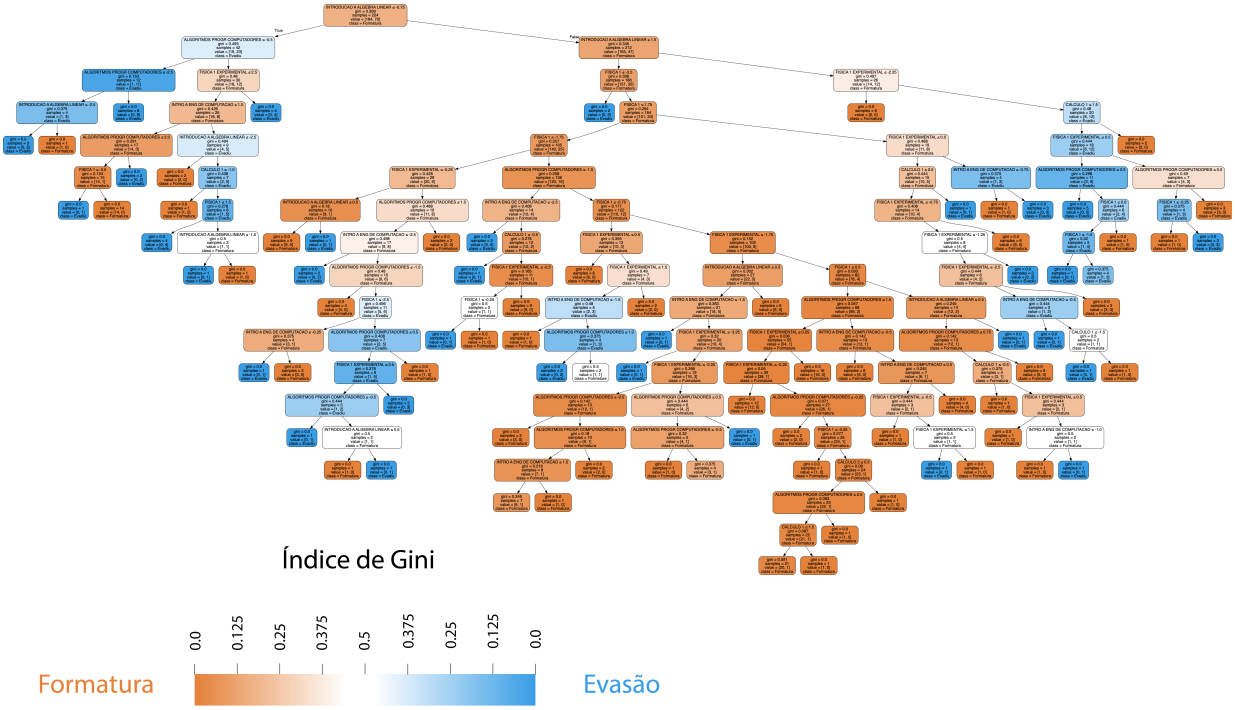


Figura 5.7: Árvore de decisão induzida.

Dadas as vantagens da pós-poda, esta foi a técnica escolhida para otimização da árvore por meio do método de complexidade do erro proposto por (BREIMAN et al., 1984), cujo objetivo é minimizar o custo, definido pela Equação (5.3), que relaciona o erro e a complexidade da árvore.

$$R_\alpha(T) = R(T) + \alpha \cdot |f(T)| \tag{5.3}$$

Onde:

$R(T)$ é o erro de treinamento

$f(T)$ é o número de folhas da árvore

α é um parâmetro de regularização

O número de folhas da árvore é utilizado como uma medida da complexidade da mesma, e o parâmetro α é um multiplicador que relaciona o peso da complexidade no custo. Quando $\alpha = 0$, apenas o erro é considerado na poda, e quando $\alpha \to \infty$, apenas a complexidade é considerada; neste caso, a árvore final é a própria raiz, minimizando o número de folhas.

Desse modo, é possível encontrar um conjunto de valores reais $-\infty < \alpha_1 = \alpha_{\min} < \alpha_2 < \cdots < \alpha_K < +\infty$ e uma sequência de sub-árvores associadas $T_0 > T_1 > \cdots > T_K > Raiz$ e, em seguida, escolher a sub-árvore que melhor represente o problema avaliado, sendo aquela que apresenta a maior acurácia média em rodadas independentes de treinamento.

Para encontrar o conjunto α foi implementado um algoritmo iterativo, iniciando em $\alpha_1 = 0$,

que calcula a cada iteração o valor de $g_i(t)$ definido pela Equação (5.4). O nó com o menor valor de $g_i(t)$ é podado e o α da iteração seguinte recebe o valor de $g_i(t)$, até que a árvore podada seja igual à raiz.

$$g_i(t) = \frac{R(t) - R\left(T_t^i\right)}{\left|f\left(T_t^i\right)\right| - 1} \tag{5.4}$$

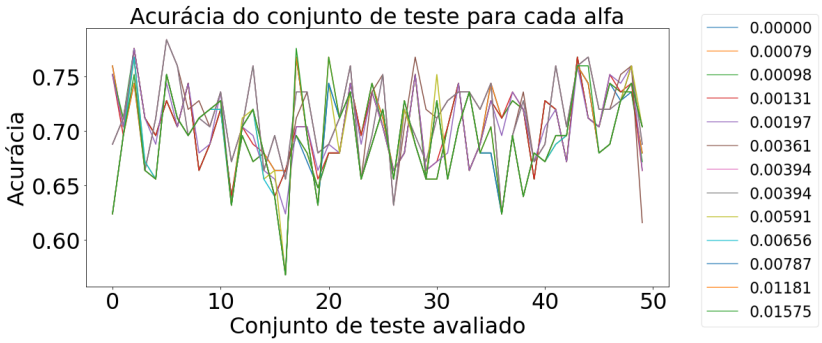
Onde:

$R(t)$ é o erro de treinamento no nó t

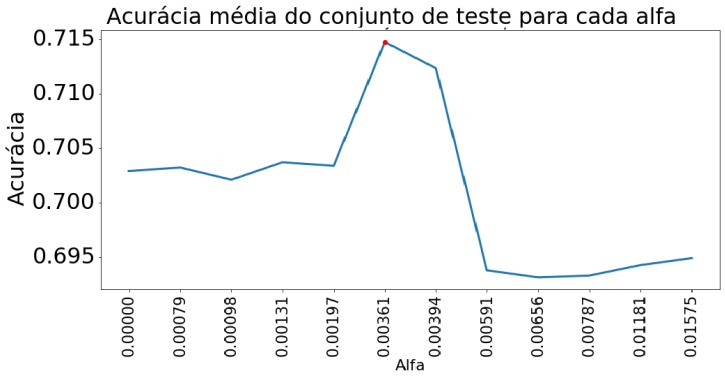
$R(T_t)$ é o erro de treinamento de uma sub-árvore T_t cuja raiz é o nó t

$f\left(T_t^i\right)$ é o número de folhas a serem podadas a partir do nó t

Ao final da implementação do método de complexidade do erro foram encontrados 13 valores de α e 13 sub-árvores associadas, resultados da poda da árvore da Figura 5.7. Para avaliar o comportamento de cada α na poda da árvore, e assim possibilitar a escolha do modelo com melhor desempenho, passou-se para um procedimento de validação cruzada, com 50 rodadas independentes de amostragem do conjunto de dados, divididos, conforme já mencionado, em 70% para treinamento e 30% para validação. A cada rodada foi induzida e podada uma nova árvore usando o algoritmo de complexidade do erro com o α especificado. A acurácia das árvores para cada um dos conjuntos de teste está ilustrada na Figura 5.8a e a acurácia média de cada α está representada no gráfico da Figura 5.8b. O parâmetro de regularização escolhido para o modelo final de predição foi $\alpha = 0.0036$ que corresponde ao ponto máximo da curva da Figura 5.8b, apresentando uma acurácia média de 71.5%.



(a) Acurácia do conjunto de testes para cada valor de α encontrado.



(b) Acurácia média para cada α.

Figura 5.8: Acurácia das árvores de decisão geradas para 50 conjuntos de teste diferentes com os 13 α encontrados pelo método de complexidade do erro.

Após isso, foi definido um limiar máximo de 3% do total de *samples* para que um nó pudesse se dividir, caso contrário, este se tornaria uma folha. Isso foi feito porque, mesmo após a aplicação do método de complexidade do erro, algumas folhas possuíam um número muito reduzido de amostras, sendo assim mais improvável a relação entre a forma de saída destes alunos e seus desempenhos acadêmicos, em detrimento de outras condições externas ao curso e a esta análise.

A árvore de decisão resultante do método de poda está ilustrada na Figura 5.9, que pode ser conferida com mais detalhes no Anexo **??**. É possível observar a redução expressiva da quantidade de folhas quando comparado ao modelo apresentado anteriormente na Figura 5.7. Além do ganho de acurácia, que saiu de 69.4% para 72.5%, decorrente da amenização do fenômeno de *overfitting*, a poda tornou o sistema consideravelmente mais inteligível, viabilizando o seu uso como uma ferramenta de auxílio para os coordenadores.

O fato de a maioria das folhas serem compostas por um conjunto de amostras "puras", com índices de Gini reduzidos, indica uma boa adequação da árvore aos dados, visto que estas folhas são capazes de representar muito bem a classe associada a elas, não havendo a necessidade de tantas outras ramificações para divisão dos dados, como as existentes na árvore induzida originalmente.

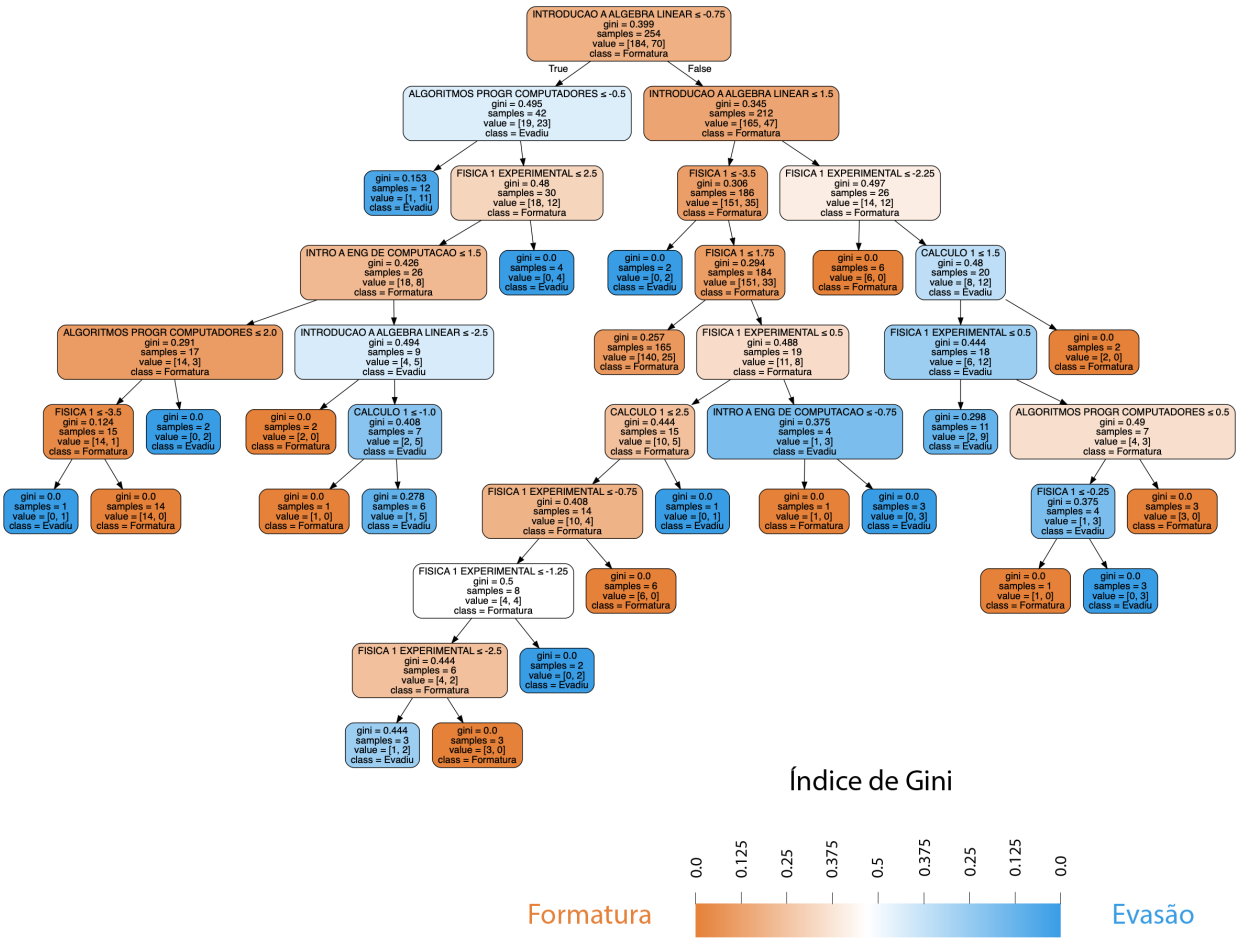


Figura 5.9: Árvore de decisão podada.

5.3 Análise dos Resultados

A árvore de decisão pode ser utilizada para classificar a forma de saída de um novo aluno como *Formatura* ou *Evasão* antes do mesmo efetivamente sair do curso. Para isso, basta partir do nó raiz da árvore e ir percorrendo-a, através das respostas aos testes dos nós internos, até chegar em um nó folha, o qual indica a classe correspondente do mesmo.

Analisando a estrutura da árvore encontrada, nota-se que a disciplina *Introdução à Álgebra Linear*, a raiz, representa o nó mais informativo da árvore, no sentido de ser responsável pelo *split* que divide os dados em dois conjuntos mais homogêneos, um de cada classe, sendo portanto estabelecida pelo modelo como a matéria mais importante na determinação da evasão. Baseado na modelagem do sistema de menções descrita na Tabela 5.1 temos que a regra que define este nó, $Nota <= -0.75$, divide os alunos que reprovaram a disciplina dos demais, dado que a menção MI recebeu o peso -1. Esse fenômeno pode ser devido ao fato de a disciplina em questão, apesar de ser oriunda do Departamento de Matemática, apresentar conceitos fundamentais para os sistemas de programação, tais como vetores e operações matriciais.

Nota-se também que, seguida desta, uma reprovação em *Algoritmos e programação de Computadores* é quase decisiva para a evasão, dando origem a uma folha com índice de Gini consideravelmente baixo. Esta matéria, por sua vez, é a disciplina do primeiro semestre que representa o primeiro contato da maioria dos alunos com o conceito de algoritmo, sendo extremamente representativa do resto do curso e da prática profissional de um Engenheiro de Computação, fato que pode fazer com que uma reprovação seja desestimulante o suficiente para acarretar na desistência. Este fator, em específico, deve ser alvo de preocupação visto que, conforme explicitado no gráfico 3.4, a disciplina de *Algoritmos e Programação de Computadores* se mostra como a quinta colocada em termos de nível de reprovação do departamento. Conforme mencionado anteriormente, houve uma alteração no fluxo do curso que colocou esta disciplina no lugar de *Computação Básica*. No entanto, a taxa de reprovação se manteve praticamente a mesma, apresentando uma melhora sutil de 0.69%.

Além do aspecto social e econômico, explicitado no Capítulo 4 como relevante para o desempenho estudantil, existem questões que influenciam a performance geral das turmas que não estão sendo levadas em consideração nesta análise, como greves, alterações na ementa ou professor, etc; fatores que podem gerar algumas inconsistências no modelo gerado.

Com auxílio dos dados relativos às disciplinas do CIC utilizados na Seção 3 foi traçado o gráfico da Figura 5.10, contendo a informação temporal das matérias do departamento. É possível notar a inconstância da taxa de reprovação da disciplina de *Computação Básica* e sua sucessora *Algoritmos e Programação de Computadores* ao longo dos semestres, que se apresentaram como folhas relevantes na determinação da evasão. Devido a este comportamento irregular, conclui-se que a pontuação do aluno em determinada disciplina não é suficiente, por si só, para determinar o seu desempenho acadêmico. Uma reprovação em um semestre cuja média geral foi baixa, por exemplo, pode não representar, na prática, uma nota similar à uma turma em que os demais alunos conseguiram aprovação; esse fator pode levar a inconstâncias no modelo.

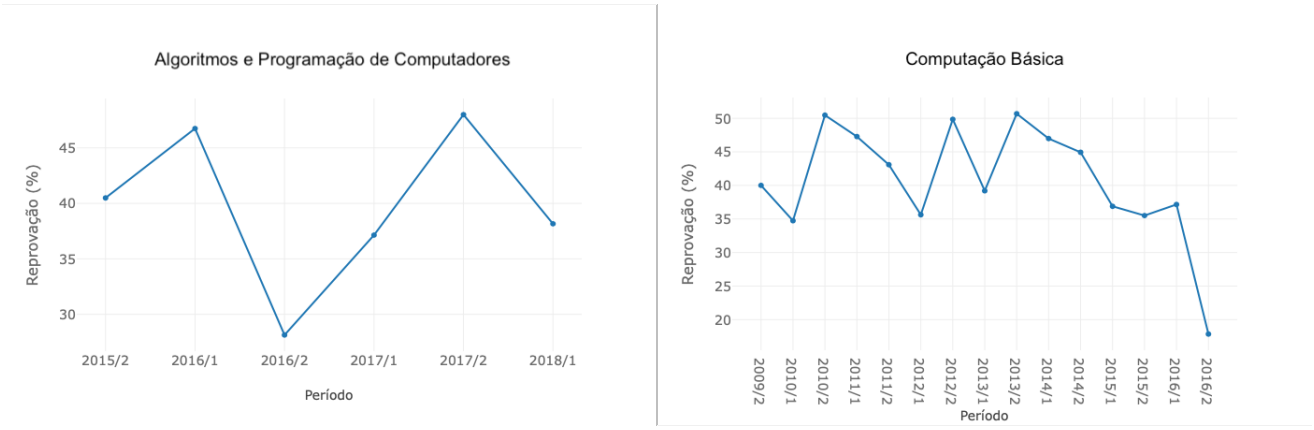


Figura 5.10: Taxa de reprovação de CB e APC ao longo do tempo.

Conforme supracitado, a modelagem do problema foi orientada para a predição de evasão. Uma vez que as informações não referentes ao primeiro ano foram desconsideradas, o modelo apresentou uma precisão menor na predição de formatura. O gráfico da Figura 4.4 aponta que a formatura

acontece, majoritariamente, entre o quinto e sexto anos do curso, sendo um evento temporalmente distante do recorte de tempo analisado. Esse fato é reiterado pela matriz de confusão da árvore para o conjunto de validação definido, que está ilustrada na Figura 5.11.

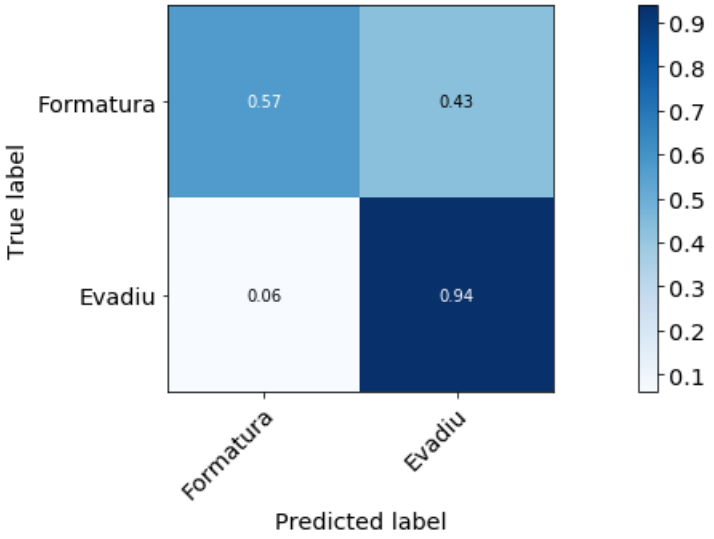


Figura 5.11: Matriz de confusão para o conjunto de validação.

As linhas da matriz de confusão representam, percentualmente, as classes dos alunos do conjunto de validação, enquanto as colunas representam a predição feita pela árvore. É possível observar que a determinação da classe formatura se deu de modo quase aleatório, com uma acurácia de apenas 57%, indicando que o primeiro ano, por si só, não é capaz de determinar se um aluno irá concluir o curso, sendo necessária a análise do seu desempenho nas disciplinas dos semestres seguintes, que são mais específicas do curso e do departamento.

Em contrapartida, a predição de evasão se mostrou bastante satisfatória, com uma acurácia de 94% para o conjunto de validação analisado. De acordo com os objetivos definidos para o trabalho, a existência de *Falsos Positivos*, situação na qual o modelo determina a formatura de um aluno que vai evadir é o pior cenário possível, pois este estudante não receberia o suporte que o coordenador do curso se propõe a fazer, e é exatamente nesta determinação que o modelo de predição apresentou bons resultados. Para este conjunto de dados específico, apenas 6%, 5 dos 78 alunos desistentes, estariam fora do grupo que o modelo indicaria ao coordenador para a realização de alguma ação preventiva.

5.4 Seção Final

Neste capítulo foram analisados os históricos escolares dos alunos de Engenharia da Computação, com foco nas disciplinas do primeiro ano do curso, tido como o que concentra o maior índice de evasão. A partir disso foi gerado um modelo de predição baseado no algoritmo da árvore de decisão com índice de Gini para cálculo de impureza e com uso do método de complexidade do

erro para poda.

Algumas considerações acerca da possível relação entre a forma de saída dos alunos e as menções obtidas em disciplinas específicas foram feitas com base na estrutura da árvore obtida ao final do processo de poda. Foi evidenciada, por exemplo, a relação entre evasão e a disciplina Algoritmos e Programação de computadores do primeiro semestre, correspondente a um dos maiores índices de reprovação do departamento, além da análise feita em cima de disciplinas que não são do CIC, mas que apresentaram uma contribuição expressiva para a determinação da desistência.

Ao final, obteve-se uma árvore de decisão com acurácia geral de 71.5%, que apresenta desempenho insatisfatório para a determinação da formatura, classe que não fazia parte do escopo deste capítulo para predição e cujo recorte temporal foi descartado para inserção no modelo. Não obstante, o modelo obtido atingiu uma acurácia de 94% para predição de evasão, o que coloca este modelo como uma nova ferramenta de auxílio para a coordenação nos anos que se seguem.

Capítulo 6

Conclusão

Nessa pesquisa foram levantadas diversas informações relacionadas ao desempenho dos estudantes para os departamentos de Engenharia Elétrica e Ciência da Computação, como os histogramas de menções dos departamentos, a lista de matérias com altos índices de reprovação, os mapas de distribuição demográfica por IRA, dentre outros. A partir dos dados fornecidos pelo coordenador de Engenharia da Computação foi proposto um modelo de predição de evasão de estudantes do curso.

Dentre os objetivos específicos deste trabalho, definidos no Capítulo 1, estava a compreensão dos aspectos que afetam o desempenho estudantil durante o curso. Para tal, o Capítulo 3 levantou dados relativos aos aspectos puramente educacionais, com o estudo das disciplinas de um departamento, enquanto no Capítulo 4 foi feito um estudo da relação entre o rendimento acadêmico e aspectos socioeconômicos de um estudante, como a forma de ingresso, região administrativa e tipo de escola de origem.

Além disso, também foi definido como objetivo a apresentação de dados problemáticos intrínsecos ao curso em questão, dentre os quais se pode citar os resultados da análise feita no Capítulo 3, explicitando índices atípicos para disciplinas obrigatórias que podem ser potenciais causadoras da alta taxa de desistência, como as disciplinas de Lógica Computacional 1 e Princípios de Comunicação, que apresentaram concomitantemente taxas altas de reprovação e trancamento. Além disso, foi calculada a probabilidade entre a reprovação simultânea de duas ou mais disciplinas com a desistência, constatando a existência de algumas disciplinas de início do curso altamente correlacionadas com a evasão, como Introdução à Engenharia da Computação e Física 1. Esse tipo de análise possibilita uma melhoria na estrutura do curso como um todo, como a realocação inteligente dos recursos humanos do departamento, aumento de vagas de monitoria para as disciplinas tidas como problemáticas no que tange ao índice de reprovação, dentre outros.

Por fim, foram definidos os objetivos de criação de um modelo de predição e a apresentação de um método de análise de dados educacionais. Mais especificamente, foi desenvolvida uma árvore de decisão focada na detecção prévia de desistência de um estudante do curso de Engenharia da Computação da UnB. O modelo desenvolvido atingiu uma acurácia geral de 71,5% e de 94% para predição de desistência. A título de comparação, na pesquisa realizada por (BAYER et al.,

2012), em que também se desenvolve um modelo preditivo para desistência de estudante, tendo como base apenas as informações dos dois primeiros semestres, sem considerar as relações sobre o comportamento social dos mesmos, a acurácia varia entre 50,21% a 71,77% com o uso de diferentes algoritmos. Sendo assim, pode-se dizer que o resultado obtido neste trabalho foi satisfatório.

Existem diversas formas de se utilizar os resultados obtidos nesta pesquisa para melhorar o desempenho dos estudantes e diminuir a evasão, como a realização de políticas de auxílio para alunos do primeiro ano cuja previsão seja a desistência, reestruturação das disciplinas com índices atípicos de reprovação e trancamento, dentre outros. No entanto, esses objetivos só poderão ser observados depois da publicação deste trabalho. Esperamos que a pesquisa tenha um impacto, perceptível, em todas essas métricas.

6.1 Trabalhos Futuros

Existem vários aspectos que podem ser aprimorados em trabalhos futuros, entre os quais destacamos:

1. Ampliação do escopo: expandir a análise realizada para períodos posteriores ao primeiro ano do curso;
2. Inserção de novas *features*: como foi visto no trabalho, existe uma correlação entre o desempenho do estudante e sua condição socio-econômica; adicionar esses dados ao modelo poderia melhorar significativamente seu desempenho;
3. Utilização de tipos de modelos diferentes: apesar das métricas alcançadas serem boas para o problema proposto, árvores de decisão não são conhecidas pelo seu alto desempenho e sim por sua capacidade interpretativa dos dados. Por isso deve-se testar outros tipos de modelagem que possam aumentar as taxas de predição;
4. Verificar as taxas de modelos em novos dados: utilizar o modelo criado para predizer a taxa de alunos do ano de 2019 e verificar se seu desempenho se mantêm;
5. Acompanhar o impacto da pesquisa nas estatísticas: acompanhar por um período a utilização dessa pesquisa como forma de diminuir as taxas de evasão dos estudantes e concluir o seu impacto no mundo real;
6. Ampliar a modelagem para outros cursos: utilizar bancos de dados de outros cursos e aplicar a metodologia utilizada nesse estudo para produzir mais informações.

Além de melhorar o modelo desenvolvido, o que se propõe também é desenvolver aplicações para visualização desses dados de forma a amparar sistemas de decisões baseados em dados, além de disponibilizar essas estatísticas a quem pode tomar decisões, para melhorar esses índices.

Referências

AGRAWAL, Rakesh; GOLSHAN, Behzad; PAPALEXAKIS, Evangelos. Toward data-driven design of educational courses: A feasibility study. **Journal of Educational Data Mining**, International Working Group on Educational Data Mining, v. 8, n. 1, p. 1–21, 2016.

AUD, Susan; WILKINSON-FLICKER, Sidney. **The condition of education 2013**. [S.l.]: Government Printing Office, 2013.

BAKER, Ryan SJD; YACEF, Kalina. The state of educational data mining in 2009: A review and future visions. **JEDM| Journal of Educational Data Mining**, v. 1, n. 1, p. 3–17, 2009.

BAYER, Jaroslav et al. Predicting Drop-Out from Social Behaviour of Students. **International Educational Data Mining Society**, ERIC, 2012.

BEAN, John P; BRADLEY, Russell K. Untangling the satisfaction-performance relationship for college students. **The Journal of Higher Education**, Taylor & Francis, v. 57, n. 4, p. 393–412, 1986.

BENJAMIN, Michael; HOLLINGS, Ann. Student Satisfaction: Test of an Ecological Model. **Journal of College Student Development**, ERIC, v. 38, n. 3, p. 213–28, 1997.

BRASÍLIA - UNB, Universidade de. **Universidade de Brasilia - Guia do Calouro**. Brasília, 2018.

BREIMAN, L. et al. **Classification and Regression Trees**. Monterey, CA: Wadsworth e Brooks, 1984.

DECANATO DE PLANEJAMENTO, Orçamento e Avaliação Institucional - DPO. **Projeto Político-Pedagógico Institucional da Universidade de Brasília**. Brasília, 2018.

_____. **Universidade de Brasilia - Anuário Estátistico**. Brasília, 1998.

_____._____. Brasília, 2002.

_____._____. Brasília, 2008.

_____._____. Brasília, 2013.

_____._____. Brasília, 2018.

FREEMAN, Scott et al. Active learning increases student performance in science, engineering, and mathematics. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 111, n. 23, p. 8410–8415, 2014.

FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Robert. **The elements of statistical learning**. [S.l.]: Springer series in statistics New York, 2001. v. 1.

GAMA, João. Functional Trees. In: v. 55, p. 59–73. DOI: <10.1007/3-540-45650-3_9>.

GEIGLE, Chase; ZHAI, ChengXiang. Modeling MOOC student behavior with two-layer hidden Markov models. In: ACM. PROCEEDINGS of the fourth (2017) ACM conference on learning@ scale. [S.l.: s.n.], 2017. p. 205–208.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011.

HAND, David J. Data Mining. **Encyclopedia of Environmetrics**, Wiley Online Library, v. 2, 2006.

JORDAN, Katy. Initial trends in enrolment and completion of massive open online courses. **The International Review of Research in Open and Distributed Learning**, v. 15, n. 1, 2014.

KIM, Yeon; AHN, Changsun. Effect of combined use of flipped learning and inquiry-based learning on a system modeling and control course. **IEEE Transactions on Education**, IEEE, v. 61, n. 2, p. 136–142, 2017.

KNOX, William E; LINDSAY, Paul; KOLB, Mary N. Higher education, college characteristics, and student experiences: Long-term effects on educational satisfactions and perceptions. **The Journal of Higher Education**, Taylor & Francis, v. 63, n. 3, p. 303–328, 1992.

MARISCAL, Gonzalo; MARBAN, Oscar; FERNANDEZ, Covadonga. A survey of data mining and knowledge discovery process models and methodologies. **The Knowledge Engineering Review**, Cambridge University Press, v. 25, n. 2, p. 137–166, 2010.

MASON, Gregory S; SHUMAN, Teodora Rutar; COOK, Kathleen E. Comparing the effectiveness of an inverted classroom to a traditional classroom in an upper-division engineering course. **IEEE Transactions on Education**, IEEE, v. 56, n. 4, p. 430–435, 2013.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos Sobre Aprendizado de Máquina. In: SISTEMAS Inteligentes Fundamentos e Aplicações. Barueri-SP: Manole Ltda, 2003. p. 89–114. ISBN 85-204-168.

ONODA, Maurício. Estudo sobre um Algoritmo de Árvore de Decisão Acoplado a um Sistema de Banco de Dados Relacional, 2001.

PELAEZ, Kevin et al. Using a Latent Class Forest to Identify At-Risk Students in Higher Education. **JEDM| Journal of Educational Data Mining**, v. 11, n. 1, p. 18–46, 2019.

PLANEJAMENTO DO DISTRITO FEDERAL – CODEPLAN, Companhia de. **Distrito Federal em Síntese Informações Socioeconômicas e Geográficas**. Brasília, 2012.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational data mining: a review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, Ieee, v. 40, n. 6, p. 601–618, 2010.

_____. Educational data mining: A survey from 1995 to 2005. **Expert systems with applications**, Elsevier, v. 33, n. 1, p. 135–146, 2007.

SAARELA, Mirka; KÄRKKÄINEN, Tommi. Analysing student performance using sparse data of core bachelor courses. **Journal of educational data mining**, International Working Group on Educational Data Mining, v. 7, n. 1, 2015.

SILVA FILHO, Roberto Leal Lobo et al. A evasão no ensino superior brasileiro. **Cadernos de pesquisa**, SciELO Brasil, v. 37, n. 132, p. 641–659, 2007.

SWEENEY, Mack et al. Next-term student performance prediction: A recommender systems approach. **arXiv preprint arXiv:1604.01840**, 2016.

UNESCO. **UNESCO. 2018. Global Education Monitoring Report 2019: Migration, Displacement and Education – Building Bridges, not Walls. Paris, UNESCO.** 2019. Disponível em: <<https://unesdoc.unesco.org/ark:/48223/pf0000265866>>.

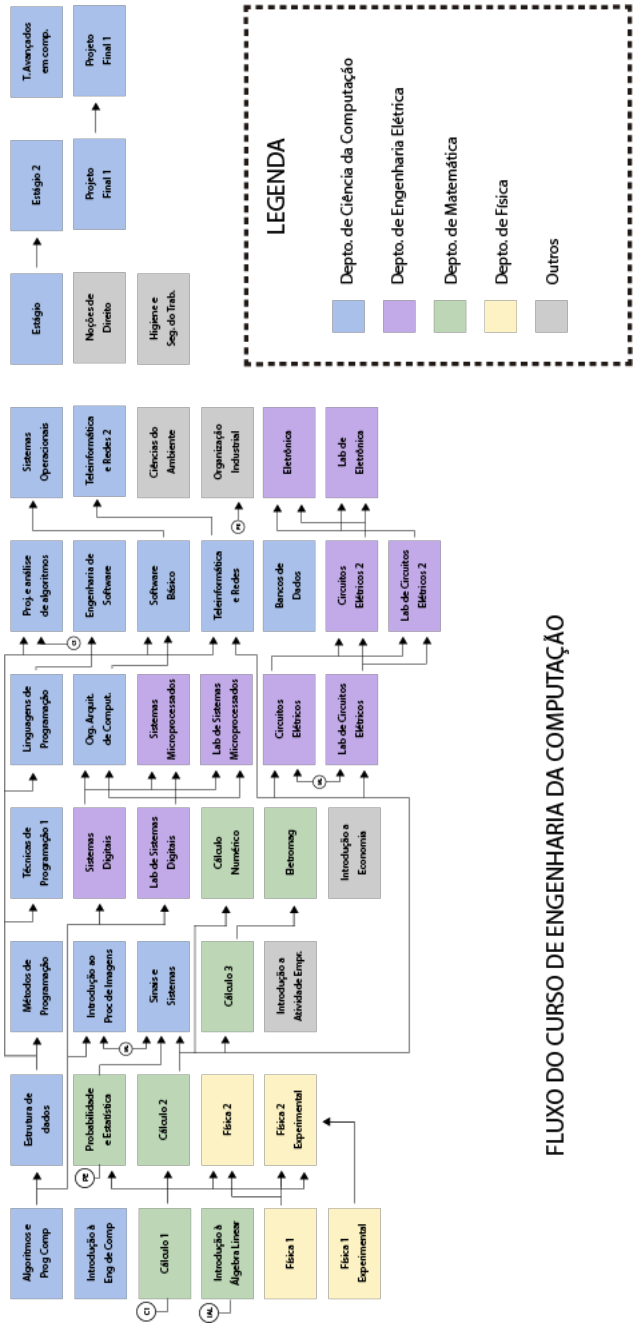
WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: Towards a standard process model for data mining. In: CITESEER. PROCEEDINGS of the 4th international conference on the practical applications of knowledge discovery and data mining. [S.l.: s.n.], 2000. p. 29–39.

YANG, Diyi et al. Exploring the effect of confusion in discussion forums of massive open online courses. In: ACM. PROCEEDINGS of the second (2015) ACM conference on learning@ scale. [S.l.: s.n.], 2015. p. 121–130.

ANEXOS

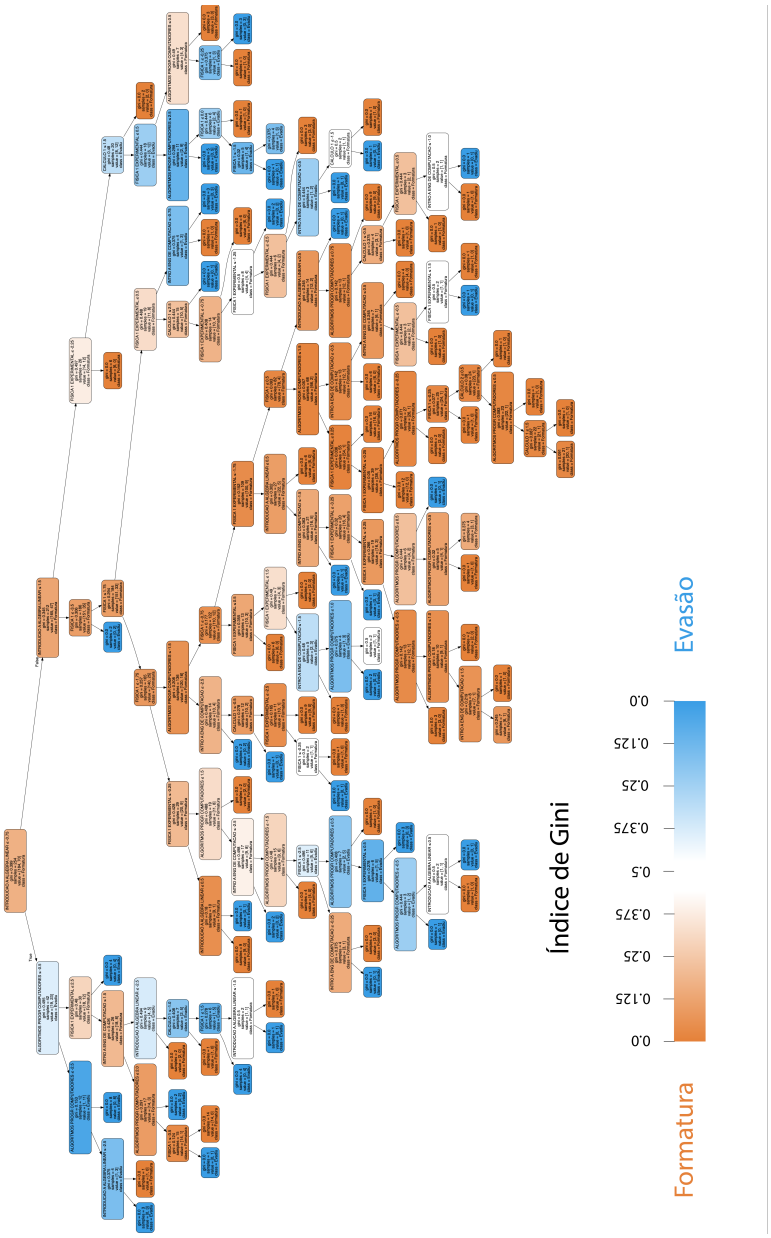
ANEXO I

Fluxo de Engenharia da Computação



ANEXO II

Árvore de Decisão antes da Poda



ANEXO III

Árvore de Decisão depois da Poda