UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANTÔNIO NILO DE ARAÚJO NETO

AVALIAÇÃO DA ESTRUTURA DO CURRÍCULO DO ENSINO SUPERIOR COM APRENDIZADO DE MÁQUINA

FORTALEZA

2018

ANTÔNIO NILO DE ARAÚJO NETO

AVALIAÇÃO DA ESTRUTURA DO CURRÍCULO DO ENSINO SUPERIOR COM APRENDIZADO DE MÁQUINA

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação

Orientador: Prof. Dr. João Paulo Pordeus Gomes

Coorientadora: Prof[a]. Dr[a]. Emanuele Marques dos Santos

FORTALEZA

2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

A687a Araújo Neto, Antônio Nilo de.
Avaliação da estrutura do currículo do ensino superior com aprendizagem de máquina / Antônio Nilo de Araújo Neto. – 2018.
58 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2018.
Orientação: Prof. Dr. João Paulo Pordeus Gomes.
Coorientação: Prof. Dr. Emanuele Marques dos Santos.

1. Aprendizado de máquina. 2. Estrutura curricular. 3. Análise pedagógica. 4. Método de controle sintético. 5. Estrutura de graduação. I. Título.

CDD 005

ANTÔNIO NILO DE ARAÚJO NETO

AVALIAÇÃO DA ESTRUTURA DO CURRÍCULO DO ENSINO SUPERIOR COM APRENDIZADO DE MÁQUINA

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. João Paulo Pordeus Gomes (Orientador)
Universidade Federal do Ceará (UFC)

Prof[a]. Dr[a]. Emanuele Marques dos Santos (Coorientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. João Fernando Lima Alcântara
Universidade Federal do Ceará (UFC)

Prof. Dr. João Paulo do Vale Madeiro
Universidade da Integração Internacional da Lusofonia Afro-Brasileira (UNILAB)

A um mundo que use a informação para decisão.

AGRADECIMENTOS

Agradeço à minha família pelo apoio, aos meus professores pela guia, ao meu orientador pela paciência.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“Knowledge is power”

(Francis Bacon)

RESUMO

A estrutura curricular de um curso de graduação é um assunto de extrema importância para a educação do ensino superior, pois é ela quem define não apenas o conteúdo programático que os alunos verão no curso, como também a ordem cronológica na qual os alunos verão esse conteúdo. Uma estrutura mal construída pode comprometer a formação do aluno, ou até mesmo produzir a evasão dos estudantes, um problema bastante frequente no ensino superior brasileiro. Essa estrutura tradicionalmente é construída a partir de critérios qualitativos, podendo ser suportados através de diretrizes, mas que acabam sendo elaborados a partir do bom senso dos educadores responsáveis pela concepção do currículo, ou seja, aquilo que eles creem como sendo necessário para a formação dos estudantes. Neste trabalho, utilizaremos o método de controle sintético, um modelo de regressão linear que pode fazer uso de conhecimento especialista, juntamente com os dados dos alunos do curso de Ciência da Computação da Universidade Federal do Ceará, para fornecer um ferramentário quantitativo para auxiliar na avaliação do desempenho dos alunos e da estrutura curricular a partir da relação entre as disciplinas do curso.

Palavras-chave: Aprendizado de Máquina. Estrutura curricular. Análise Pedagógica. Método de controle sintético. Estrutura de graduação

ABSTRACT

The curricular structure of a bachelor degree is a very important matter for education as a whole, and it's development usually revolves around qualitative factors. In the present work, we'll use the synthetic control method, alongside the data from the students of Computer Science from Universidade Federal do Ceará, to provide a quantitative framework to evaluate such structure The curricular structure of a bachelor degree is of utmost importance for higher education, as it defines not only the contents of the degree itself, but also the chronological order ofunder wihch the students will study this content. A poorly built structure might compromise the degree, or even increase the drop out rates of the students, a frequent problem in brazilian higher education. This structure traditionally is made upon qualitative criteria, which then relies on the good sense of the pedagogycal professionals involved in the production of this structure, which will reflect what the educators believe should be present in the curricula. In the present work, we'll use the Synthetic Control Method, a linear regression model which can make use of specialist domain knowledge, together with the data of students from Computer Science degree in Universidade Federal do Ceará, in order to providea quantitative toolkit to assist in the evaluation of the perfomance of the students and of the curricular structure based on the relationship between the disciplines of the course.

Keywords: Machine Learning. Curricular structure. Synthetic control method. Bachelor structure. Pedagogic analysis

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ALGEL	Álgebra Linear
APSIS	Análise de Projetos e Sistemas
ARQUI	Arquitetura de Computadores
AUTOM	Autômatos e Linguagens Formais
BANCO	Fundamentos de Bancos de Dados
CALC1	Cálculo Diferencial e Integral I
CALC2	Cálculo Diferencial e Integral II
CANAL	Construção e Análise de Algoritmos
CIRCU	Circuitos Digitais
COMPG	Computação Gráfica
COMPI	Construção de Compiladores
ENGEN	Engenharia de Software
ESTAT	Introdução a Probabilidade e Estatística
ESTRU	Estruturas de Dados
FISIC	Física
FUNDP	Fundamentos de Programação
GRAFO	Algoritmos em Grafos
INTAR	Inteligência Artificial
IRA	Índice de Rendimento Acadêmico
LINGP	Linguagens de Programação
LOGIC	Introdução a Lógica Matemática
MATED	Matemática Discreta
METO1	Métodos Numéricos I
METO2	Métodos Numéricos I
PROGR	Programação
REDES	Redes de Computadores
SCM	Synthetic Control Method
SIGBD	Sistema de Gerenciamento de Banco de Dados
SISOP	Sistemas Operacionais
TECNI	Técnicas de Programação
TEORI	Teoria da Computação

TRANS Transmissão de Dados

SUMÁRIO

1 INTRODUÇÃO

Nas últimas décadas, muito se tem discutido o papel da educação como propulsor da qualidade de vida em um país. A esse nível, não sobram dúvidas quanto à relevância do ensino superior para o desenvolvimento de uma nação ao longo prazo, e inúmeras são as discussões para o enfrentamento das dificuldades do estudante universitário.

Um estudo feito em **??**, por exemplo, aborda a problemática da perspectiva dos estudantes de classes economicamente vulneráveis e seu acesso à Universidade, e (FILHO *et al.*, 2007) aborda o problema da evasão dos alunos no ensino superior. Um tipo de abordagem que tem se tornado cada vez mais comum para tentar explorar esses problemas é definida em (SIEMENS GEORGE; LONG, 2011), onde o termo *Learning Analytics* é definido como sendo a mensuração, coleção, análise e divulgação da informação sobre estudantes e seu contexto, visando a compreensão e otimização do aprendizado e de seu ambiente. (HURN, 2013) por exemplo, lista aplicações de *Learning Analytics* em algumas universidades e seus resultados, como o *SIGNALS*, um sistema feito para acompanhar em tempo real o desempenho dos alunos da Universidade de Purdue, ou o *SSP*, desenvolvido em Sinclair Community College, para aconselhamento e melhora na retenção dos estudantes. Esses, entre outros, tentam fazer uso da grande massa de dados produzida dentro das unidades de ensino superior, para a melhora do sistema de ensino.

No entanto, não encontramos nenhum trabalho cujo objetivo fosse avaliar a estrutura do currículo do ensino superior. A estrutura curricular possui importante papel na definição da experiência de um aluno. (RODRIGUES Y. K. O.; PORTO, 2013) discute múltiplas definições para um currículo, mas aqui entenderemos o currículo como sendo o conjunto de disciplinas que o aluno precisa cursar para se formar, além da definição de quais disciplinas um aluno precisa para cursar uma outra disciplina. Sendo assim, o currículo define, primeiramente, que conteúdo aquele aluno precisa assimilar.

A estrutura curricular precisa, inicialmente, respeitar certas restrições pedagógicas, mas também ser atrativa para os alunos. No entanto, frequentemente se observa um distanciamento entre a estrutura e as habilidades e desejos dos alunos (LENNON; MAURER, 2003), o que acaba tornando os alunos dependentes de suas experiências passadas (MELIA; PAHL, 2007) (SAMPLES, 2002). Além disso, essa organização curricular também define implicitamente uma ordem sob a qual as disciplinas serão cursadas, visto que cada disciplina possui um conjunto de requerimentos que a antecedem, o que gera uma rede de dependências que limita as opções

disponíveis para um aluno e um determinado momento do curso.

Logo, destacamos a importância de uma boa estrutura curricular, que do contrário pode acarretar problemas para a formação do aluno. A carga horária é o primeiro deles. Como podemos avaliar, quantitativamente, a importância relativa de uma disciplina obrigatória para o aluno? Será que ela possui importância própria, ou é lecionada apenas como preâmbulo de outras disciplinas? Nesse caso, é possível eliminá-la? A seguir, podemos mencionar o desempenho e a retenção dos alunos. Se um aluno cursa uma disciplina porque é obrigado, e essa disciplina é apenas uma aparente requisição de uma outra, é possível que isso não só desestimule o aluno a dar o seu melhor, como também que o faça desistir do curso, especialmente se todas as disciplinas que ele cursa em um dado momento são dessa natureza. Logo, o questionamento que naturalmente surge é se é possível fazer uma avaliação objetiva da estrutura curricular.

1.1 Objetivos

É sobre esse problema que esse trabalho se debruça: uma abordagem quantitativa para avaliar a estrutura de um curso de ensino superior, a partir de uma abordagem de *Learning Analytics*. Intuitivamente, o que iremos fazer é buscar uma relação entre uma determinada disciplina e as demais disciplinas que a antecedem.

Mais explicitamente, iremos, a princípio, fazer uma análise exploratória sobre os dados disponíveis, e a partir deles tentar encontrar alguns fatores que possam estar relacionados às notas. A seguir, iremos fazer uso de um método de aprendizado de máquina relativamente recente, onde nos propomos a avaliar a estrutura curricular do curso de graduação em ciência da computação na Universidade Federal do Ceará. Para tanto, analisamos as notas dos alunos que concluíram o curso entre 2005 e 2016, e levando em conta a ordem temporal na qual os alunos cursam as disciplinas, tentamos traçar uma relação de dependência entre uma disciplina e aquelas que são cursadas anteriormente, tendo como objetivo obter uma combinação entre elas que represente ao máximo possível as notas obtidas pelos alunos e algumas características pré-determinadas das disciplinas .

Como resultado, encontraremos, graças ao método, uma relação entre cada uma das disciplinas obrigatórias e todas as demais disciplinas, também obrigatórias, que estejam num momento anterior no tempo. Dessa forma, iremos avaliar, através de critérios numéricos, quanto uma disciplina é descrita por seus pré-requisitos

Além disso, o método também fornece informação sobre quais disciplinas anteriores

são consideradas relevantes para as disciplinas sendo cursadas, e assim, indicar quais são as que, apesar de não serem consideradas pré-requisitos, demonstraram, numericamente, ter relevância para uma boa execução, por parte dos alunos, em termos de notas finais.

agens futuras que se pode ter com os resultados desse trabalho.

1.2 Trabalhos Relacionados

A avaliação do currículo do ensino superior possui certas dificuldades e algumas de suas abordagens recebem críticas (LEATHWOOD; PHILLIPS, 2000). Mesmo assim, a demanda por trabalhos quantitativos para esse problema aumenta consideravelmente ano após ano, em busca de uma maneira de tornar os resultados do ensino superior mais transparentes (JOHNES; TAYLOR, 1990). Infelizmente, uma grande parte dos estudos ainda se concentra basicamente em *Ambientes Virtuais de Ensino*, e os principais objetivos são a modelagem do comportamento dos estudantes, previsão de desempenho, e estudo de retenção de alunos (PAPAMITSIOU; ECONOMIDES, 2014). Por exemplo, (MACFADYEN; DAWSON, 2010) , utilizando dados de um *Ambiente Virtual*, tais como número de mensagens enviadas pelo sistema, tempo online, número de links visitados, etc, tenta prever a nota final obtida pelos alunos naquele curso. (KIZILCEC *et al.*, 2013), ainda em uma abordagem para um Ambiente Virtual, utiliza dados de como os alunos se engajam assistindo videoaulas e solucionando os problemas de um curso online para, através de algoritmos de classificação, encontrar fatores que determinam a retenção dos alunos. Alguns outros trabalhos relacionam a avaliação de currículo como sendo o conjunto de escolhas que o aluno fez ao longo do curso sobre quais disciplinas cursar a cada momento. Como exemplo, (WU; HAVENS, 2005) e (BALDONI *et al.*, 2011) auxiliam o estudante nessa tomada de decisão.

Um trabalhado mais focado em organização curricular se encontra em (PRIYAMBADA *et al.*, 2017). Nele, a sequência de disciplinas escolhida por cada aluno é comparada com a sequência sugerida pelo currículo, e a partir dos resultados encontrados, os alunos são agrupados. Como resultado, encontrou-se que um determinado tipo de caminho está associado a melhor desempenho, o que pode sugerir uma ordem entre as disciplinas.

De forma semelhante, (WANG; ZAïANE, 2015) e (PECHENIZKIY *et al.*, 2012) realizam uma mineração de processos, e a partir das múltiplas sequências encontradas, tenta associar às sequências o desempenho dos alunos que a realizaram, encontrando conjuntos de possíveis sequências que representariam uma maneira apropriada de dispor as disciplinas do

curso. Diferentemente dessas abordagens, iremos encontrar um único conjunto de relações entre as disciplinas, que não apenas possui uma interpretação semântica mais simples, como ainda permite a realização de previsão de desempenho.

1.3 Estrutura

Inicialmente, veremos a teoria básica sobre regressão linear no capítulo 2, além de alguns dos problemas comuns envolvidos com esse método. A seguir, veremos uma explicação sobre o Método de Controle Sintético, principal algoritmo que utilizaremos neste trabalho.

Uma vez tendo o arcabouço teórico, iremos nos debruçar sobre o o conjunto de dados no capítulo 3, que será utilizado como entrada pelo Método de Controle Sintético. Tentaremos entender a estrutura que esses dados possuem, e nele realizar uma análise exploratória, abordando alguns aspectos que influenciam as notas.

A seguir, iremos apresentar os resultados encontrados através do Método de Controle Sintético no capítulo 4, além de realizar algumas comparações com outros métodos e encontrar métricas para avaliar os resultados encontrados.

Finalmente, no capítulo 5, discutiremos algumas conclusões referentes aos dados encontrados e possíveis abord

2 MODELOS LINEARES PARA REGRESSÃO

De modo geral, uma regressão é um método cujo objetivo é encontrar a relação entre as variáveis de um conjunto de dados. De forma um pouco mais específica, uma regressão é um modelo que tenta encontrar uma função, que recebe um conjunto de dados de entrada, cujo objetivo é prever da melhor forma possível o valor de uma variável contínua de saída. Existem tipos diferentes de regressão, e aqui iremos apresentar o arcabouço do modelo que iremos utilizar.

2.1 Regressão Linear

Considere dois conjuntos de dados $\{x_i\}$ e $\{y_i\}$. Na regressão linear, iremos buscar uma relação entre esses dois conjuntos através de um modelo linear. A formulação deste modelo pode ser representada a partir da equação a seguir.

$$\overline{y_i} = w_0 + w_1 x_i \tag{2.1}$$

Onde x_i é a variável explanatória, e $\overline{y_i}$ é o valor obtido pelo modelo, que desejamos que seja o mais próximo possível de y_i. Estamos, portanto, buscando uma reta que represente a relação entre $\{x_i\}$ e $\{y_i\}$, cujo coeficiente linear é w_1 e coeficiente angular é w_0.

No caso mais geral, estaremos levando em consideração diversas variáveis explicativas x_{ij} influenciando $\overline{y_i}$ ao mesmo tempo. Na regressão multivariada, a estimação da i-ésima amostra é dada por:

$$\overline{y_i} = w_0 + x_{i1} w_1 + x_{i2} w_2 + ... + x_{im} w_m \tag{2.2}$$

em que x_{ij} e o j-ésimo atributo da i-ésima amostra. Ou seja, estamos representando a relação entre $\{x_i\}$ e $\{y_i\}$ como um hiperplano.

Iremos agora mostrar uma representação mais compacta desse problema a partir de formulação com matrizes. Seja a amostra x_i representada por $[1, x_{i1}, x_{i2}, ..., x_{im}]^T$, onde o 1 e adicionado ao vetor por conveniência. Construímos uma matriz X de atributos onde $X = [x_1, x_2, ..., x_n]^T$. Além disso, seja $w = [w_1, w_2, ..., w_n]^T$ o vetor de pesos. Dessa forma, podemos escrever a saída do modelo como

$$\bar{y_i} = w^T x_i \tag{2.3}$$

2.2 Métodos de Ajuste de Parâmetros

Como já vimos na equação 2.2, estamos buscando um hiperplano que represente uma relação linear entre $\{x_i\}$ e $\{y_i\}$. No caso em que x_i tem apenas uma dimensão, esse hiperplano se degenera na forma de uma reta definida pela equação 2.1. Iremos agora apresentar maneiras de ajustar os coeficientes w_i para tentar encontrar a melhor relação linear possível entre esses conjuntos. Precisaremos, para tanto, definir um critério de quão bom um modelo é, o que faremos a partir de uma definição de erro. Há diversas maneiras de mensurar o erro entre dois conjuntos numéricos, e aqui utilizaremos uma das mais difundidas: o erro quadrático médio, que pode ser definido a partir da função J a seguir:

$$J(w) = \frac{1}{2n}\sum_{i=1}^{n} e_i^2 \tag{2.4}$$

Onde $e_i = y_i - \overline{y_i} = y_i - w^T x_i$, ou seja, o quadrado da diferença entre o valor da variável de saída e a saída fornecida pelo modelo quando a entrada é x_i. Observe que o quadrado remove o sinal da diferença.

2.2.1 Gradiente descendente

O método do gradiente descendente baseia-se em alterar o vetor w levemente a cada iteração. Para tanto, decrementaremos w de uma pequena fração do gradiente da função J no ponto w. O valor α é um valor convenientemente pequeno (mais sobre esse valor na seção 2.3.1) que representa o tamanho da fração. Esse processo está representado na equação a seguir.

$$w = w - \alpha\vec{\nabla}J(w) \tag{2.5}$$

A equação 2.6 representa o gradiente da função J no ponto w, em que $\partial J/\partial w_i$ é a derivada parcial de J em w_i. Fazendo a derivada parcial de J e aplicando o resultado na equação 2.5, encontramos o resultado a equação 2.7.

$$\vec{\nabla}J(w) = (\frac{\partial J}{\partial w_0}, \frac{\partial J}{\partial w_1}, ..., \frac{\partial J}{\partial w_m}) \tag{2.6}$$

$$w_j = w_j - \alpha\frac{1}{n}\sum_{i=1}^{n}(y_i - w_j x_{ij})x_i \tag{2.7}$$

A cada decremento de w na direção contrária de seu gradiente, o novo valor de w normalmente fornece uma aproximação melhor para a regressão. Assim, repetimos esse processo um certo número de vezes. Chamamos essa quantidade de número de *épocas*. Fazemos isso até que a solução seja boa o suficiente, ou seja, o erro quadrático médio J do w encontrado esteja próximo de zero, ou que a mudança no valor de w não represente mais mudanças significativas em J. (Mais detalhes na seção 2.3.2)

Nem sempre é conveniente ter que iterar sobre todos os valores de $\{x_i\}$ para efetuar o decremento descrito na equação 2.7, seja porque nem todos os valores estão disponíveis a priori, seja porque o conjunto de dados é demasiadamente grande. Nesse caso, é mais conveniente utilizar o método do gradiente estocástico.

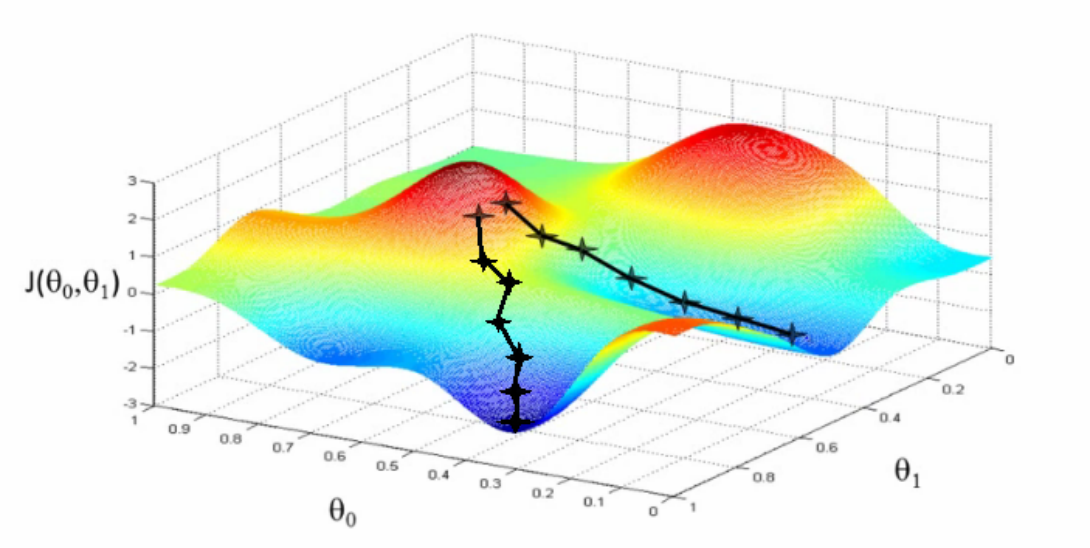

Figura 1 – Cada ponto no traçado é um valor obtido através da mudança de w. Regiões vermelhas representam picos e azuis, vales. Observe que o algoritmo está saindo de uma região de pico para chegar numa região de valor mínimo.

2.2.2 Gradiente descendente estocástico

O gradiente descendente estocástico traça uma abordagem um pouco diferente do gradiente descendente tradicional. Em vez de utilizar todos os valores do conjunto de dados para determinar o próximo valor de w, ele utiliza apenas um valor de x_i por iteração. A cada época, uma permutação dos valores de $e_i x_i$ é gerada, e essa sequência é utilizada para atualizar os valores de w. No caso de uma nova amostra sendo adicionada ao conjunto, o que caracteriza uma aprendizagem sequencial, ou aprendizagem *online*, o valor do novo x_i é utilizado para atualizar o peso do modelo já treinado com os dados antigos. Esse algoritmo é conhecido como *least mean*

squares, ou *LMS Algorithm* (BISHOP, 2006). A formula de atualização dos pesos passa a ser:

$$w = w - \alpha e_i x_i \tag{2.8}$$

2.2.3 Em lote - Batch

É possível encontrar uma solução analítica para o problema de minimização descrito pela equação 2.4. Essa abordagem é chamada de *ordinary least squares*, e pode ser utilizada quando o cálculo de matrizes inversas não for um problema. Primeiro, iremos transformar a equação de erro quadrático para uma forma vetorial. Sendo $Y = [y_1, y_2, ..., y_n]^T$ e $\overline{Y} = [\overline{y_1}, \overline{y_2}, ..., \overline{y_n}]^T$, nosso problema é reduzir o valor do erro quadrático, ou seja, minimizar a função:

$$J(w) = \frac{1}{2}(Y - \overline{Y})^T (Y - \overline{Y}) \tag{2.9}$$

Escrevendo $\overline{Y} = w^T X$ como uma consequência da equação 2.3 e fazendo $\partial J / \partial w = 0$, encontramos a equação 2.10, chamada de *equação normal*, cuja solução é dada pela equação 2.11 (BARBER, 2012). A porção $(X^T X)^{-1} X^T$ é chamada de *pseudo-inversa de Moore-Penrose*, que pode ser interpretada como a matriz inversa de matrizes não quadradas (RAO; MITRA, 1972).

$$\sum_{i=1}^{n} y_i x_i = \sum_{j=1}^{n} x_j x_j^T w \tag{2.10}$$

$$w = (X^T X)^{-1} X^T Y \tag{2.11}$$

2.3 Hiperparâmetros

Dois hiperparâmetros merecem atenção na regressão, o α e o número de épocas, cada um influenciando de certa maneira o algoritmo. Não existem valores ideais para esses hiperparâmetros, apesar de existir abordagens que tentam melhorar seus usos. Aqui, a melhor saída é a experimentação. Valores típicos de números de época giram em torno de 1000, enquanto que α fica em torno de 0.01

2.3.1 Alfa

O valor de α é o valor multiplicado ao valor que pretende ser incrementado ao vetor w. Ele determina o tamanho do salto que é dado na função que representa o erro (veja figura

1). Se α for grande, os saltos serão grandes e o algoritmo rapidamente chega ao valor mínimo. Todavia, ao se aproximar da solução ótima, o algoritmo não consegue chegar ao ponto exato, pois ele não é capaz de dar saltos pequenos o suficiente para chegar a um ponto preciso.

Para entender melhor o problema, suponha que você deseja chegar a um ponto que está a 5 metros de distância, mas você só é capaz de dar passos de 7 metros. Se, assim como o algoritmo, você está sempre caminhando em direção a solução ótima, então você nunca chegará ao ponto desejado. De fato, ao dar seu passo, você estará a 2 metros do ponto. Mas ao se mover em direção a ele novamente, você voltará a posição inicial do problema.

Naturalmente, se α for pequeno, esse problema é resolvido, mas isso pode tornar o algoritmo lento, ou até virtualmente parar, se o incremento tornar-se zero.

2.3.2 Número de épocas

O número de épocas é o número de vezes que se repete o algoritmo. Em geral, a cada iteração, fica-se mais próximo da solução ótima. O número necessário de épocas é desconhecido, assim, dois problemas podem surgir: se o número de épocas utilizado for menor que o necessário, talvez não se chegue próximo o suficiente da solução ótima; se por outro lado o número de épocas for maior que o necessário, o algoritmo irá perder tempo com iterações que não melhoram a solução encontrada. Isso pode ser resolvido com o término do algoritmo caso a iteração não melhore a solução, ou se a solução obtida já for considerada boa o suficiente.

2.4 Modelos não lineares

Os modelos lineares que vimos até aqui possuem fundamental importância teórica, porém são fortemente limitados na prática. Os dados no mundo real dificilmente são *linearmente separáveis*, ou seja, não podem ser classificados simplesmente através da regressão linear. É todavia possível utilizar a regressão com modelos não lineares. Podemos por exemplo adicionar à equação 2.2 a combinação quadrática de $\{x_i\}$:

$$\overline{y_i} = w_0 + x_{i1}w_1 + ... + x_{im}w_m + x_{i1}^2 w_1' + x_{i2}^2 w_2' + ... + x_{in}^2 w_m' \tag{2.12}$$

De fato, é possível complicar arbitrariamente o modelo, aumentando indefinidamente sua expressividade. Essa complicação, porém, nem sempre é desejada.

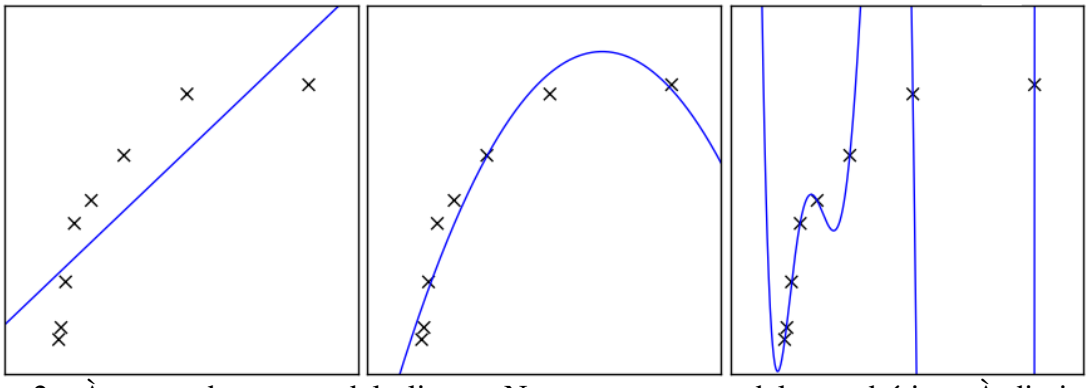

Figura 2 – À esquerda, um modelo linear. No centro, um modelo quadrático. À direita, um modelo polinomial de grau 6. Note como o modelo acerta cada vez melhor os dados fornecidos

2.4.1 Overfitting

O primeiro problema advindo do aumento de complexidade é um problema inerente a todos os algoritmos de aprendizagem de máquina: chama-se *overfitting*. O problema ocorre quando a saída do modelo se aproxima demais aos dados da amostra, mas não aos dados *fora dela*. Dizemos que o erro de aprendizado é quase zero, mas mesmo assim o modelo não generaliza bem (BISHOP, 2006).

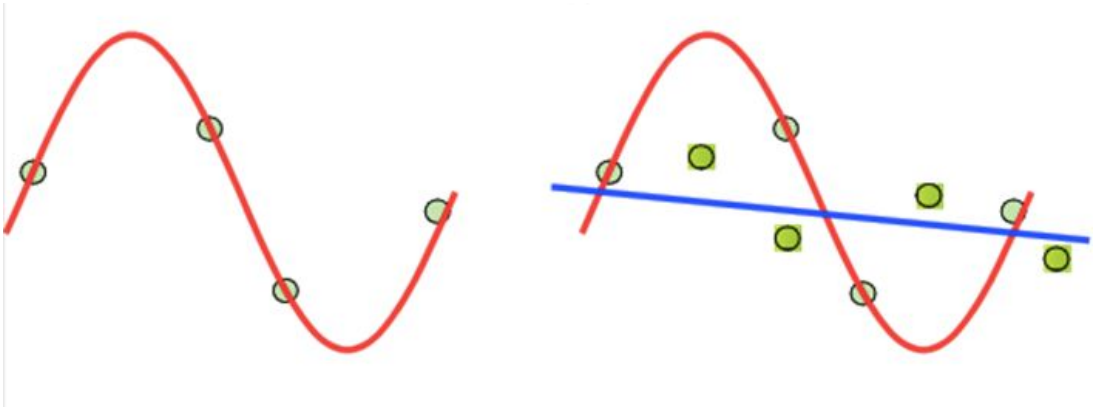

Figura 3 – Nos dados da amostra, o modelo quadrático se aproxima muito bem (a esquerda). Porém, quando se incluem dados fora dela, modelo linear possui um erro menor (a direita)

À esquerda da figura 3, temos o modelo obtido, em linha vermelha, quando o conjunto de treinamento é composto pelos quatro pontos dados. Mas ao final, o modelo não representa bem todos os dados da distribuição. De fato, um modelo linear, menos complexo, representaria melhor.

De modo semelhante, na figura 2, observando a amostra, o modelo quadrático responde melhor a generalização, e que o modelo polinomial de grau 6, apesar de ter um erro

quadrático menor, não representa bem a tendência dos dados.

Esse comportamento matemático em que o modelo se *esforça demais* para diminuir o erro, fazendo o gráfico dar muitas curvas para tocar os dados fornecidos é o que caracteriza o *overfitting*. Observe que o resultado do overfitting é um modelo que parece cada vez menos natural para explicar o fenômeno.

Assim, encontramos um problema dual na generalização: se o modelo é simples demais, o erro nos dados observados é grande demais. Se o modelo é complexo demais, as previsões para dados fora da amostra erram muito.

Uma abordagem para resolver esse problema é dividir o conjunto de amostras em duas partes. A primeira parte, chamada de *conjunto de treinamento*, é utilizada para treinar o modelo, melhorando os coeficientes através das iterações. É importante notar que o erro quadrático médio medido nesse grupo sempre tenderá a um determinado valor, que é o ponto de máximo overfitting do modelo (que depende de sua complexidade), um valor que pode chegar a zero. Entretanto, isso não significa que o modelo generaliza bem. Para saber se o modelo generaliza bem, mediremos o erro quadrático médio do segundo grupo, chamado de *conjunto de teste*, que para o modelo, funciona como um dado desconhecido e novo. Se o erro do segundo grupo também cair, então isso implica que o modelo está aprendendo a prever a amostra. Eventualmente, o overfitting nos dados de treinamento fará o erro aumentar no conjunto de testes, o que representa o momento adequado para parar as iterações.

2.4.2 Regularização

Uma outra abordagem para diminuir o overfitting é a regularização (BISHOP, 2006). Quando temos muitos atributos, o modelo passa a ter muita liberdade para tentar produzir um overfitting. Se formos capazes de dizer quais atributos são menos importantes para o treinamento, então podemos tirar do modelo parte de sua liberdade.

Faremos isso alterando a equação 2.4 do erro:

$$J(w) = \frac{1}{2n}\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{m} w_j^2 \tag{2.13}$$

Em que λ representa o coeficiente de regularização, que controla a importância relativa do termo de regularização. Como outros hiperparâmetros, o valor ideal de λ deve ser encontrado a partir de experimentação.

Nesse momento, essa equação não representa mais o erro, mas sim o *custo* que queremos minimizar. Esse custo é composto pelo erro, além do tamanho do vetor w, excetuando-se a componente w_0. A intenção e que tente-se reduzir os coeficientes ao máximo possível sem que isso aumente o erro. Naturalmente, alguns coeficientes diminuirão mais que outros, esses se caracterizando como menos vitais para a generalização.

Se derivarmos J com relação a w e igualarmos a zero, teremos a solução como sendo:

$$w = (\lambda I + X^T X)^{-1} X^T Y \tag{2.14}$$

2.5 O método de controle sintético

No âmbito das ciências sociais, as comparações entre diferentes classes é praticamente obrigatória, sejam elas países, estados, representantes políticos ou medidas socioeconômicas. Um dos maiores exemplos de comparações para pesquisa quantitativa foi feito em (CARD; KRUEGER, 1993). Nesse trabalho, Card estuda o efeito do aumento de salário mínimo sobre a taxa de desemprego em Nova Jersey em 1991. Para tanto, ele compara os estabelecimentos da Pensilvânia, onde não ocorreu aumento salarial, com os estabelecimentos de Nova Jersey, utilizando o método da diferença das diferenças. A suposição básica aqui é que se as trajetórias de desemprego dos dois estados possuíam comportamento semelhante antes do aumento, então é possível prever a taxa de desemprego de Nova Jersey após o aumento a partir da taxa da Pensilvânia.

Expandindo essa ideia, Abadie em (ABADIE; GARDEAZABAL, 2003), sugeriu o método de controle sintético (*SCM - Synthetic Control Method*), em que a previsão do PIB per capita de uma certa região da Itália poderia ser feita utilizando-se as características de um conjunto de outras regiões da Itália que tivessem características semelhantes. A motivação de Abadie era mensurar o impacto de certas medidas. Em (ABADIE *et al.*, 2015), ele estima o impacto econômico da queda do muro de Berlim na economia alemã. Para tanto, ele utiliza os dados de um grupo de países para prever o PIB per capita da Alemanha ocidental antes da queda do muro. A suposição é que se esses países têm características semelhantes, então é possível prever o PIB de um a partir dos demais. No entanto, se a partir do ano da queda do muro a previsão passa a falhar, isso indica que as características não são mais semelhantes, e que o erro na previsão evidencia o tamanho do impacto. Além disso, o método também emprega critérios quantitativos para selecionar quais países do grupo melhor representam a economia alemã antes

da queda do muro, eliminando um possível viés subjetivo que pode acompanhar a seleção dos países.

Essa metodologia se assemelha aos testes de eficácia de medicamentos. De fato, é desse ramo que Abadie toma emprestado alguns dos termos utilizados. Considere, por exemplo, que procuramos saber o impacto do uso de uma substância para o tratamento de uma doença. Para tanto, divide-se os pacientes em dois grupos: um grupo de controle e um grupo de tratamento. No grupo de tratamento, será administrada a substância em questão, e no grupo de controle, não será ministrado nada, se certificando, além disso, que nada de novo será tentado. Se no grupo de tratamento houver melhora da doença, então pode-se concluir que o fato do medicamento ter sido ministrado impactou o tratamento dos pacientes.

No caso econométrico, Abadie chama a Alemanha ocidental de grupo de tratamento, ou em particular, unidade de tratamento. Já os demais países são candidatos ao controle. O algoritmo então escolhe um subconjunto desses países que melhor representam as características da economia alemã, e atribui a cada um deles um peso. A esse conjunto, dá-se o nome de grupo de controle sintético. A soma das características de cada um desses países, ponderada pelos pesos, pode ser então chamada de Alemanha sintética. No escopo desse trabalho, iremos trabalhar com as notas dos alunos em cada disciplina: encontraremos um grupo de controle sintético para cada disciplina, utilizaremos o grupo sintético para tentar prever notas, e analisaremos como os pesos encontrados se relacionam com a estrutura de dependência das disciplinas dentro do currículo.

Considere novamente a noção de grupo de controle e grupo de tratamento. Dentro desses grupos, teremos o que chamaremos de unidades. A entidade sobre a qual iremos tentar realizar estimativas será chamada de unidade de tratamento, e as demais de unidades de controle. Cada uma dessas unidades possui uma variável de saída, em diferentes momentos. Por exemplo: um país tem uma medida de desemprego a cada ano, ou uma disciplina dá a um aluno uma nota, ou seja, o país e a disciplina são unidades, e um ano e um aluno são momentos.

Tendo $n+1$ unidades, seja a matriz $X \in \mathbb{R}^{n+1 \times T}$ em que X_{it} representa a variável de saída para a unidade i no momento t. Além disso, cada unidade i tem um conjunto de m preditores, que são variáveis que variam apenas em função da unidade, e não do momento, que descrevem características desta unidade. Seja $Z \in \mathbb{R}^{n+1 \times m}$ a matriz onde Z_{ij} é o preditor de tipo j da unidade i.

Além disso, considere que a unidade de tratamento esteja representada pela última linha de X , e que as k primeiras colunas serão utilizadas como treinamento. Assim, todas as k

primeiras colunas da matriz serão utilizadas no método, enquanto as demais servirão como teste.

Suponha que X_{it} é dado pela equação:

$$X_{it} = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \varepsilon_{it} \tag{2.15}$$

onde

- $\delta_t \in \mathbb{R}$ é um fator desconhecido em função do momento
- $Z_i \in \mathbb{R}^{m\times 1}$ contém os preditores da unidade i
- $\theta_t \in \mathbb{R}^{1\times m}$ contém os pesos dos preditores no momento t
- $\lambda_t \in \mathbb{R}^{1\times F}$ contém parâmetros desconhecidos em função do momento
- $\mu_i \in \mathbb{R}^{F\times 1}$ contém coeficientes
- $\varepsilon_{it} \in \mathbb{R}$ é um erro

Considere um vetor $W \in \mathbb{R}^{1\times n}$, onde $0 \leq w_i \leq 1$ e $\sum_{i=1}^{n} w_i = 1$. Cada valor de W representa um grupo sintético, ou seja, uma média ponderada de unidades de controle. O valor da variável de saída no momento t para um grupo sintético representado por W é:

$$\sum_{i=1}^{n} w_i X_{it} = \delta_t + \theta_t \sum_{i=1}^{n} w_i Z_i + \lambda_t \sum_{i=1}^{n} w_i \mu_i + \sum_{i=1}^{n} w_i \varepsilon_{it} \tag{2.16}$$

Suponha que exista um W^* tal que

$$\forall j \in \{1,2,...,k\}, \qquad \sum_{i=1}^{n} w_i^* X_{ij} = X_{(n+1)j} \tag{2.17}$$

e

$$\sum_{i=1}^{n} w_i^* Z_i = Z_{(n+1)} \tag{2.18}$$

Ou seja, W* é um grupo sintético que representa perfeitamente a unidade de tratamento até o momento k, tanto nos valores da variável de saída quanto nos valores dos preditores. Abadie demonstra em (ABADIE *et al.*, 2007) que, se $R = \sum_{t=1}^{k} \lambda_t^T \lambda_t$ é uma matriz invertível, então:

$$X_{(n+1)t} - \sum_{i=1}^{n} w_i^* X_{it} = \sum_{i=1}^{n} w_i^* \Big(\sum_{j=1}^{k} \lambda_t R^{-1} \lambda_j^T (\varepsilon_{ij} - \varepsilon_{(n+1)j})\Big) - \sum_{i=1}^{n} w_i^* (\varepsilon_{it} - \varepsilon_{(n+1)t}) \tag{2.19}$$

Além disso, ele também demonstra no apêndice que, se ε_{it} é uma variável independente de i e de t, e se k é grande com relação à escala dos valores de ε_{it}, então a média do lado direito da equação 2.19 tende a zero, o que sugere que o grupo sintético é uma boa

aproximação da unidade de tratamento mesmo quando $k \in \{k+1, k+2, ..., T\}$, pois a média de $X_{(n+1)t} - \sum_{i=1}^{n} w_i^* X_{it}$ também tende a zero.

O vetor W* das equações 2.17 e 2.18 representa uma combinação convexa de todas da variáveis de saída, de todos os momentos até k, e dos preditores. É irrealista supor que tal vetor exista, então uma busca é feita tal que essas equações sejam aproximadamente satisfeitas. É ainda possível que as unidades sejam tão diferentes entre si que não exista um vetor W para o qual a equação 2.19 seja aproximadamente verdade. O valor dessa diferença, portanto, deve ser avaliado para cada aplicação. O pesquisador pode restringir quais unidades podem ou não fazer parte do grupo de controle, a fim de tornar as unidades mais semelhantes entre si, mas se mesmo assim uma boa aproximação não puder ser encontrada, então o método provavelmente não é uma boa abordagem para o problema em questão.

Por fim, iremos introduzir o conceito de pesos para os preditores. A equação 2.18 indica que os preditores do grupo sintético devem representar bem os preditores da unidade de tratamento. Os valores de Z_i, no entanto, perpassam pela seleção do pesquisador. Em (ABADIE *et al.*, 2015), por exemplo, as médias das seguintes variáveis são escolhidas como variáveis preditoras:

- pib per capita - 0.442
- taxa de investimento - 0.245
- abertura a negócios - 0.134
- escolaridade - 0.107
- inflação - 0.072
- fração da insdústria no pib - 0.001

Podemos conceder a esses preditores pesos na forma de uma matriz diagonal V, em que o elemento da i-ésima posição representa o peso do i-ésimo preditor, de maneiras que o vetor que representa o preditor passa a ser $Z_i' = VZ_i$. Os valores dos pesos de V podem ser dados pelo próprio pesquisador, caso haja conhecimento prévio de como as unidades se assemelham entre si, ou podem ser encontrados usando um método sistemático, que elimina um possível viés subjetivo, como veremos adiante.

2.6 Algoritmo do método de controle sintético

Para encontrar um grupo sintético que melhor aproxime as variáveis de saída e os preditores da unidade de tratamento, utilizamos um método iterativo com duas otimizações,

descritas pelas equações:

$$V = \arg\min_{V} \sum_{t=1}^{k} |X_{(n+1)t} - \sum_{i=1}^{n} w_i X_{it}| \tag{2.20}$$

$$W = \arg\min_{W} (Z_{n+1} - \sum_{i=1}^{n} w_i Z_i)^T V (Z_{n+1} - \sum_{i=1}^{n} w_i Z_i) \tag{2.21}$$

Na equação 2.20, desejamos saber qual o grupo sintético que minimiza a diferença entre as variáveis de saída, encontrando um vetor W que é uma função de V. Como discutimos anteriormente, V é uma matriz diagonal que representa os pesos relativos de cada um dos preditores. Já a equação 2.21 minimiza a diferença ponderada entre os preditores da unidade de tratamento e do grupo sintético.

Em (SYNTH, 2018), é possível encontrar a implementação em MATLAB fornecida pelos autores, em que se utiliza uma otimização iterativa: utilizando um valor inicial de V composto pelo desvio padrão de Z, embora um valor aleatório também possa ser utilizado, um valor para W é encontrado, e com ele, um novo valor para V. A otimização continua até que o valor do erro da equação 2.20 seja baixo o suficiente.

3 APRESENTAÇÃO DOS DADOS

Neste capítulo, iremos introduzir os dados utilizados, sua estrutura, e realizar uma analise exploratória, investigando alguns conceitos básicos que impõem uma arquitetura do curso, como a estrutura curricular, as unidades curriculares e a integralização, e verificando o impacto desses fatores no desempenho dos alunos.

O conjunto de dados utilizado é composto pelas notas obtidas pelos alunos que se matricularam no curso de Ciência da Computação na Universidade Federal do Ceará entre 2005 e 2016, seguindo um currículo estabelecido em 2000. Esse currículo foi alterado em 2015.

Apesar de poder ser considerado um dos melhores cursos do Brasil (RANKING. . . , 2018) , o curso ainda sofre de uma alta taxa de evasão e reprovação. De 2005 até 2015, a taxa de evasão média foi superior a 45%. Além disso, a maioria dos estudantes leva mais tempo para se graduar do que o tempo sugerido de 8 semestres: o tempo médio é de 9 meses, com uma alta concentração de estudantes se formando entre 9 e 12 semestres.

Esses números podem, portanto, indicar uma falha na estrutura curricular. Assim, uma investigação se faz necessária, e uma primeira pergunta que se pode fazer a respeito da validade da estrutura curricular é sobre a relação entre as notas das disciplinas obrigatórias e a relação entre uma disciplina e seus pré-requisitos.

3.1 Estrutura dos dados

Todos os anos, cerca de 50 alunos se matriculam no curso de computação. Apesar das flutuações, em geral a quantidade total de alunos no curso permanece a mesma.A figura 4, ilustra a quantidade de matrículas por ano.

Estão disponíveis nos dados o período em que o aluno cursou uma dada disciplina, sua nota nessa disciplina, e o resultado obtido nessa disciplina, ou seja, aprovação ou reprovação, seja esta por nota ou por falta. Temos também a frequência do aluno na disciplina, e o IRA do aluno quando da matrícula naquela disciplina. O IRA, ou *Índice de Rendimento Acadêmico*, definido pela equação 4.1, é uma nota que indica quão bem o aluno está ao longo do curso.

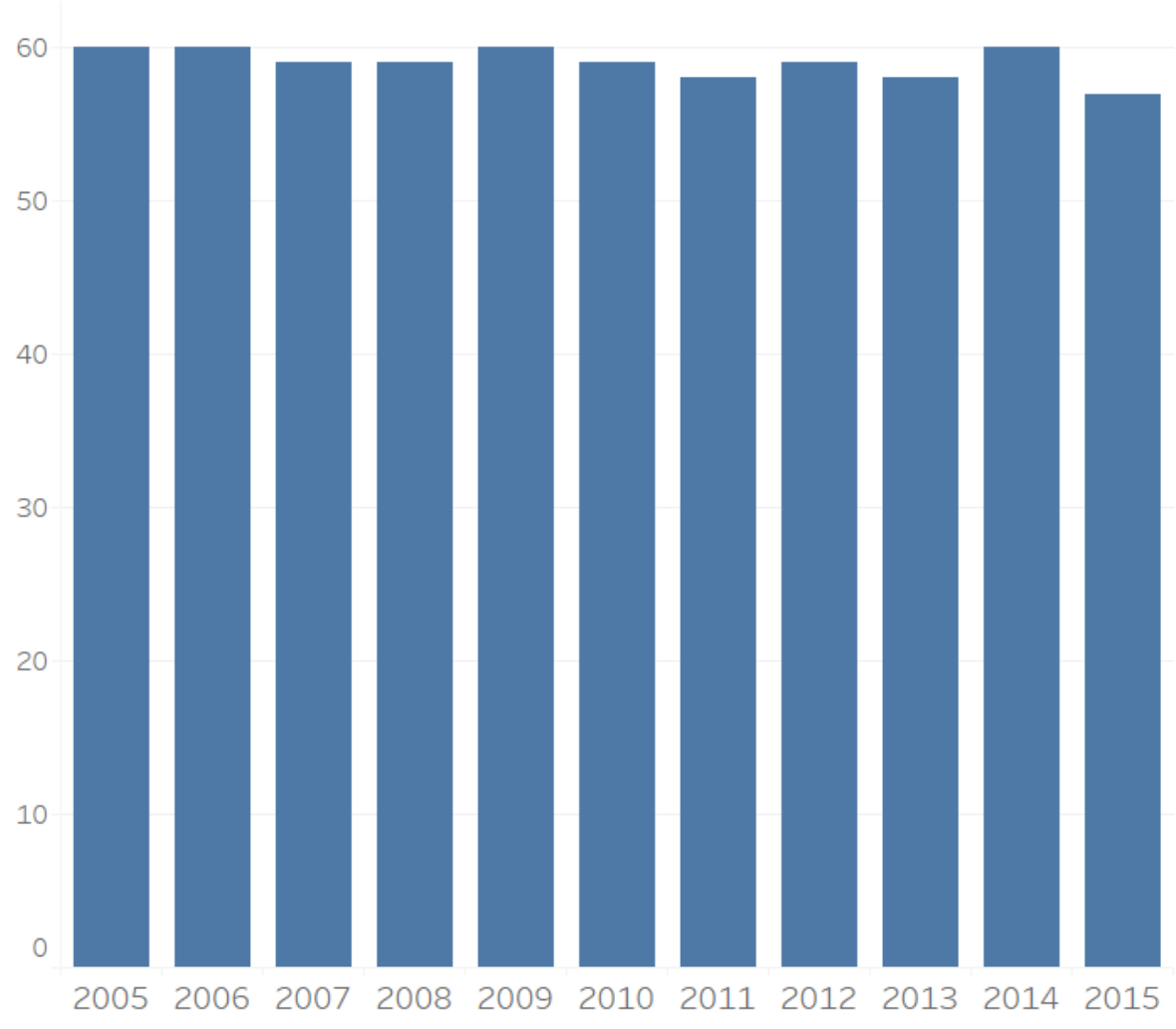

Figura 4 – Quantidade de matrículas realizadas por ano

3.2 Arquitetura do curso

3.2.1 Estrutura

Os alunos da computação precisam concluir um conjunto de 31 disciplinas obrigatórias para se formar. Cada uma delas possui um semestre, a contar a partir do ingresso do aluno, no qual ela deveria ser cursada. Uma disciplina do último semestre, *Informática e Sociedade*, no entanto, é desconsiderada, pois ela não possui pré-requisitos nem é pré-requisito de nenhuma disciplina. Algumas dessas disciplinas são pré-requisitos de outras, o que cria uma rede de dependência que impõe uma ordem sobre quais disciplinas um aluno pode cursar em um dado semestre para se formar em tempo mínimo. Assim, a maioria dos alunos que ingressam no mesmo período irá cursar as mesmas disciplinas, que são as disciplinas típicas daquele semestre. Na tabela a seguir podemos ver quais disciplinas pertencem a qual semestre.

1º semestre	2º semestre	3º semestre	4º semestre	5º semestre	6º semestre	7º semestre
CALC1	CALC2	LOGIC	CANAL	METO2	INTAR	COMPI
ALGEL	FISIC	ESTAT	METO2	REDES	SISOP	TEORI
MATED	ESTRU	GRAFO	COMPG	SIGBD	APSIS	
FUNDP	PROGR	TECNI	BANCO	ENGEN	AUTOM	
CIRCU	TRANS	ARQUI	LINGP			

No entanto, a média de notas por semestre de uma varia substancialmente. Na figura 5 encontramos a média das notas por semestre. Facilmente notamos que a menor nota média ocorre com as disciplinas do 4º semestre. Uma explicação alegórica que existe dentro do curso é a de que a partir do 4º semestre começam as disciplinas mais difíceis do curso, e que um aluno que não desiste ali irá até o fim do curso. Os dados não concordam totalmente com essa tese. Na figura 6, podemos observar as taxas de continuidade de cada semestre: no segundo semestre, temos 11% menos alunos do que havia no primeiro. A queda do quarto para o quinto semestre é notória: 14.91%, mas não é a maior. Os alunos ainda tem um último ponto destacado de desistência: o 6º semestre.

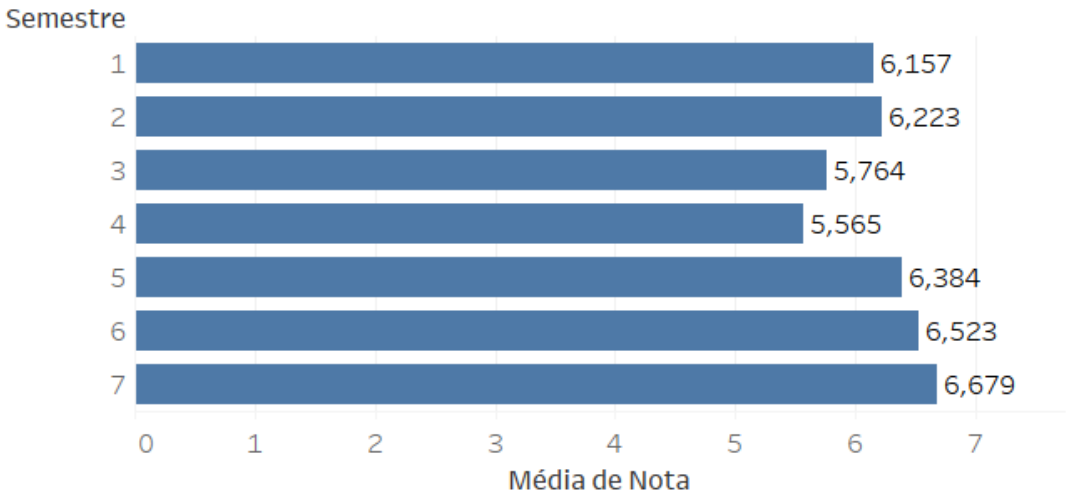

Figura 5 – Nota média de disciplinas de um mesmo semestre

As baixas notas obtidas no 4º semestre são também recorrentes ao longo dos anos. Na figura 7, notamos que a média de notas do 4º semestre permanece em baixa ao longo do tempo. É bastante natural, portanto, que a forma como essas disciplinas estão agregadas possam melhorar ou piorar o desempenho dos alunos. A subseção 3.2.2 ilustra outra característica do curso, que pode estar também relacionado a esse aspecto.

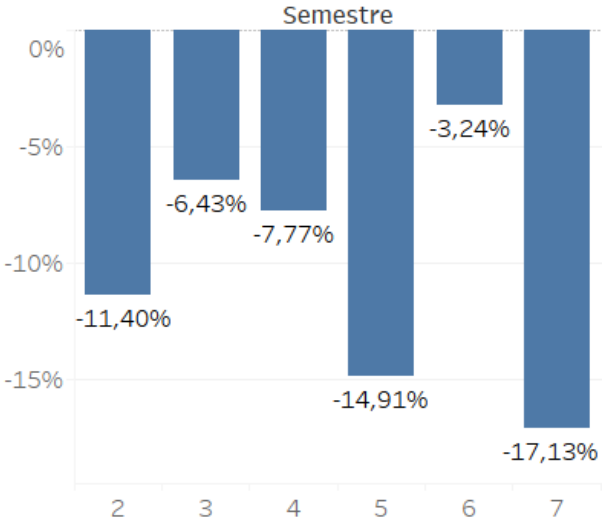

Figura 6 – Redução na quantidade de alunos por início de semestre. Do 6º para o 7º semestre, encontramos a maior taxa de desistência.

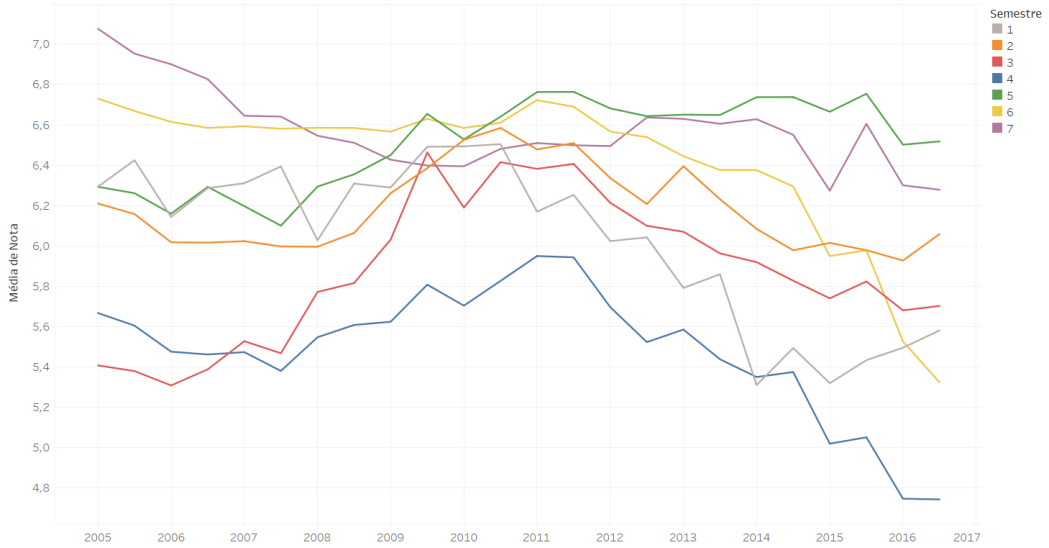

Figura 7 – Média de notas de cada semestre ao longo dos anos.

3.2.2 Unidades Curriculares

As unidades curriculares são grandes áreas sob as quais as disciplinas são agrupadas de acordo com seu papel na formação de cientista da computação, conforme definido no Projeto Politico Pedagógico do Bacharelado em Ciência da Computação (PROJETO, 2018), seguindo diretrizes sugeridas pelo MEC. As unidades curriculares são as seguintes: *Matemática*, *Programação*, *Teoria da computação*, *Sistemas de informação* e *Sistemas de computação*. Disciplinas de uma mesma unidade possuem características abstratas semelhantes, como foco maior em teoria do que prática, abstração maior ou menor de modelos e contato com ferramentas utilizadas no mercado.

A tabela a seguir mostra as unidades curriculares e suas disciplinas.

Matemática	Programação	Teoria da Computação	Sistemas de Informação	Sistemas de Computação
CALC1	FUNDP	GRAFO	APSIS	ARQUI
ALGEL	PROGR	AUTOM	COMPG	CIRCU
MATED	TECNI	COMPI	ENGEN	REDES
CALC2		CANAL	BANCO	SISOP
ESTAT		ESTRU	INTAR	TRANS
METO1		LINGP	SIGBD	
METO2		LOGIC		
ESTAT		TEORI		

Uma primeira suspeita é a de que a unidade curricular possa impactar nas notas dos alunos: podemos supor que disciplinas mais abstratas e mais teóricas tenham notas menores do que disciplinas mais práticas. Isso se confirma a partir da figura 8.

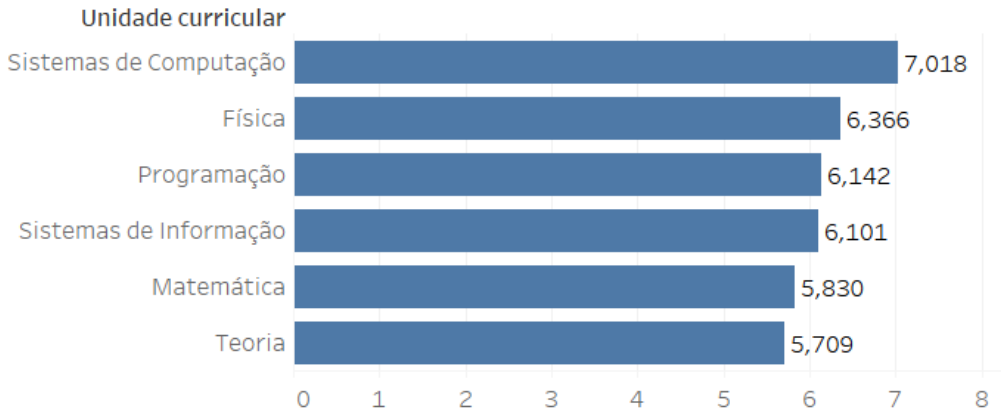

Figura 8 – Nota média por unidade curricular

Como mencionado, cada disciplina tem um semestre prático para ser cursado, que boa parte dos alunos segue. Na figura 9, podemos observar que cada semestre possui uma matiz diferente de composição de unidades curriculares, como exemplo, com *Matemática* mais presente no primeiro semestre, e *Teoria* sendo a única presente no último. Podemos observar a média de notas obtidas por semestre na figura 5.

Podemos observar também as notas médias obtidas por unidade curricular ao longo do tempo na figura 10 demonstrando que as tendências ao longo do tempo são basicamente as mesmas, aumentando a robustez da conclusão.

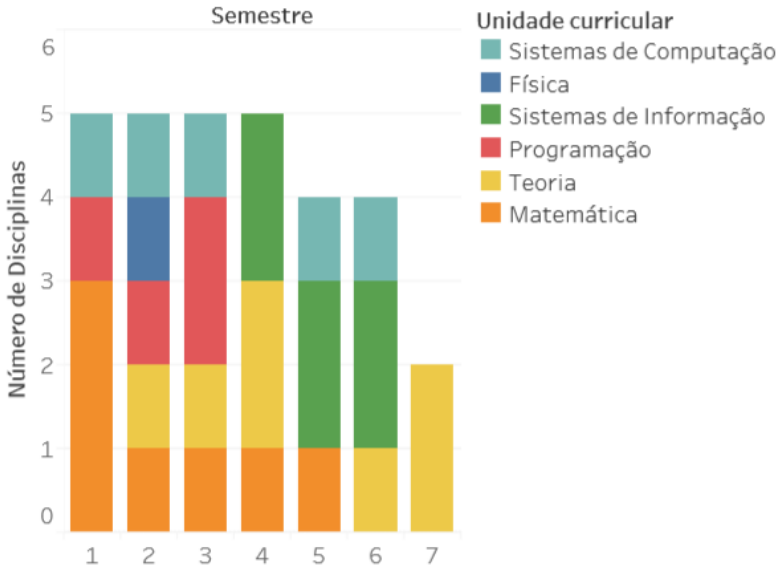

Figura 9 – Quantidade de disciplinas por semestre divididas por unidade curricular

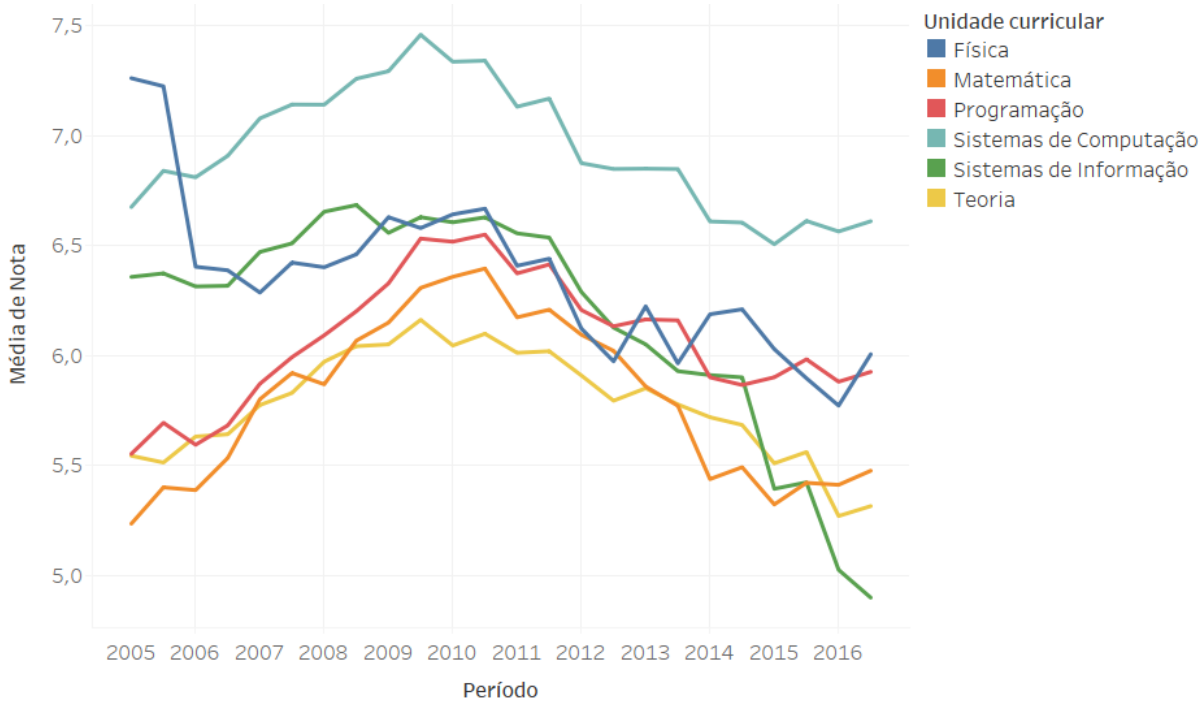

Figura 10 – Nota média por unidade curricular por período

3.2.3 Integralização

A integralização de uma disciplina será definida aqui como a quantidade de pré-requisitos diretos ou indiretos que serão exigidas. Suponha por exemplo que a disciplina A exige B como pré-requisito, e que B exige C e D como pré-requisitos, então B é um pré-requisito direto de A, e C e D são pré-requisitos indiretos de A. Disciplinas com maior integralização, tendo mais pré-requisitos, acabam sendo cursadas em semestres mais tardios. Logo no ingresso, espera-se que todos os novos alunos façam todas as disciplinas do primeiro semestre. A seguir,

os alunos que não reprovarem nenhuma disciplina do primeiro semestre provavelmente farão todas as disciplinas do segundo semestre, e assim por diante. Mesmo os alunos que reprovarem alguma disciplina do primeiro semestre provavelmente farão todas as disciplinas do segundo que puderem. Caso não o façam, é possível que o tempo de conclusão do curso seja afetado, o que acaba induzindo os alunos a seguirem a mesma ordem de disciplinas.

Como exemplo, no departamento de computação, a disciplina *Teoria da Computação* tem como pré-requisito *Autômatos e Linguagens Formais*, que por sua vez tem como pré-requisito *Introdução à Lógica Matemática*. No entanto, é possível não apenas que *Introdução à Lógica Matemática* seja profundamente diferente de *Teoria da Computação*, como também que *Teoria da Computação* esteja disponível para a matrícula, mas *Introdução à Lógica da Matemática* não. Uma consequência dessa estrutura é que, dependendo das afinidades do aluno, o acesso a determinados interesses fica dificultado. Não é possível, por exemplo, que um aluno curse *Banco de Dados* logo no segundo semestre. Esse engessamento do cronograma do curso pode se tornar especialmente problemático quando o aluno não pode cursar nenhuma das disciplinas que o interessam em um determinado momento da graduação, obrigando-o a cursar apenas disciplinas com as quais ele tem menos afinidade, causando, naturalmente, frustração.

As tabelas 1 a 6 mostram os pré-requisitos das disciplinas de cada semestre. Por exemplo, na tabela 2, vemos que os pré-requisitos de *Técnicas* são *Estruturas de Dados* e *Programação*.

Tabela 1 – Pré-requisitos para disciplinas do 2º semestre

Requisitos	CALC2	FISIC	ESTRU	PROGR	TRANS
CALC1	X	X	-	-	-
MATED	-	-	X	-	-
FUNDP	-	-	X	X	-
CIRCU	-	-	-	-	X

Tabela 2 – Pré-requisitos para disciplinas do 3º semestre

Unidade de controle	LOGIC	ESTAT	GRAFO	TECNI	ARQUI
MATED	X	-	-	-	-
CALC2	-	X	-	-	-
ESTRU	-	-	X	X	-
PROGR	-	-	-	X	-
TRANS	-	-	-	-	X

A integralização possui forte impacto para uma disciplina durante o curso, pois

Tabela 3 – Pré-requisitos para disciplinas do 4º semestre

Requisito	CANAL	METO1	COMPG	BANCO	LINGP
ALGEL	-	X	X	-	-
FUNDP	-	X	-	-	-
CALC2	-	X	-	-	-
ESTRU	-	-	-	X	X
ESTAT	X	-	-	-	-
GRAFO	X	-	-	-	-

Tabela 4 – Pré-requisitos para disciplinas do 5º semestre

Requisito	METO2	REDES	SIGBD	ENGEN
ESTAT	-	X	-	-
TECNI	-	-	-	X
ARQUI	-	X	-	-
METO1	X	-	-	-
BANCO	-	-	X	-

Tabela 5 – Pré-requisitos para disciplinas do 6º semestre

Requisito	INTAR	SISOP	APSIS	AUTOM
ESTRU	-	X	-	-
LOGIC	X	-	-	X
ARQUI	-	X	-	-
CANAL	X	-	-	-
ENGEN	-	-	X	-

Tabela 6 – Pré-requisitos para disciplinas do 7º semestre

Requisito	COMPI	TEORI
ARQUI	X	-
LINGP	X	-
AUTOM	X	X

quanto maior essa integralização for, mais difícil se torna realizar a matrícula nessa disciplina, dada a a maior quantidade de requisitos que serão necessários para se matricular nela. Podemos imaginar, a priori, que uma alta integralização pode aumentar o desempenho dos alunos, tendo em vista que temos uma garantia maior de que os alunos tem os conhecimentos necessários para aquela disciplina, algo que não se pode garantir para disciplinas introdutórias. Por outro lado, podemos supor que uma maior integralização pode prejudicar o desempenho dos alunos, uma vez que o aluno pode ser obrigado a dominar ainda mais áreas para poder ir bem naquela disciplina.

Na figura 11, observamos a nota média por número de requisitos diretos, que, supostamente, explicitam uma necessidade direta imediata para a compreensão da disciplina em questão.

Podemos verificar então na figura 12 a média de notas por quantidade de requisitos

Tabela 7 – Integralização de disciplinas

DISCIPLINA	Pré-requisitos diretos	Pré-requisitos indiretos	Total
CALC1	0	0	0
ALGEL	0	0	0
MATED	0	0	0
FUNDP	0	0	0
CIRCU	0	0	0
CALC2	1	0	1
FISIC	1	0	1
ESTRU	2	0	2
PROGR	1	0	1
TRANS	1	0	1
LOGIC	1	0	1
ESTAT	1	1	2
GRAFO	1	2	3
TECNI	2	3	5
ARQUI	1	1	2
CANAL	2	5	7
METO1	3	1	4
COMPG	1	0	1
BANCO	1	2	3
LINGP	1	2	3
METO2	1	4	5
REDES	2	4	6
SIGBD	1	3	4
ENGEN	1	5	6
INTAR	2	7	9
SISOP	2	4	6
APSIS	1	6	7
AUTOM	1	1	2
COMPI	3	6	9
TEORI	1	2	3

indiretos. Enquanto podemos entender requisitos diretos como sendo as disciplinas necessárias para sustentar uma base, podemos também entender os requisitos indiretos como sendo necessários para formar a *base dessa base*, ou seja, interpretando a relação entre as disciplinas como uma pilha, com uma em cima da outra.

Finalmente, na figura 13, observamos a média de notas por quantidade de requisitos diretos e indiretos, que podemos interpretar como o tamanho total da pilha necessária para se compreender uma determinada disciplina. Em todas as figuras, no entanto, temos uma correlação linear muito espúria, com um baixo valor para o coeficiente linear, baixos valores de R^2 e baixa significância estatística. Logo, notamos que a integralização parece ter baixo impacto na média

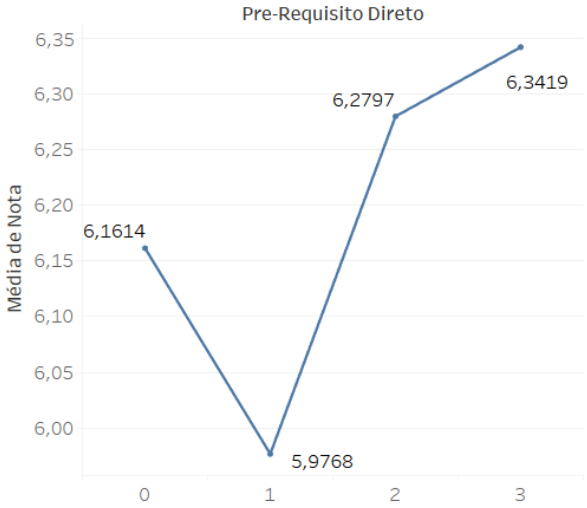

Figura 11 – Nota média por quantidade de pré-requisitos diretos

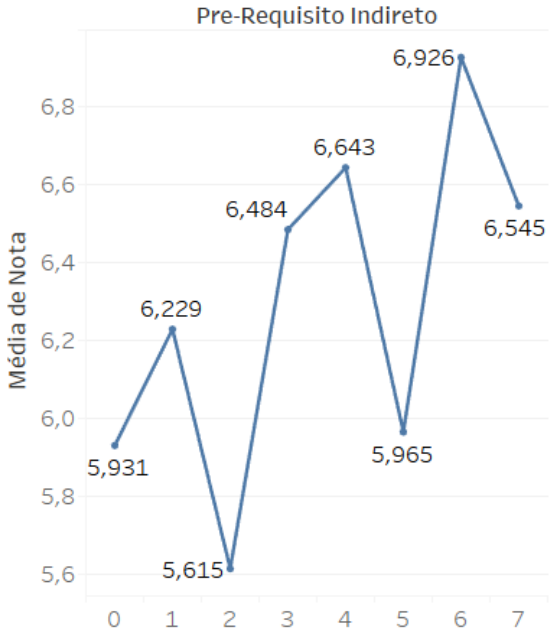

Figura 12 – Nota média por quantidade de pré-requisitos indiretos

de notas dos alunos.

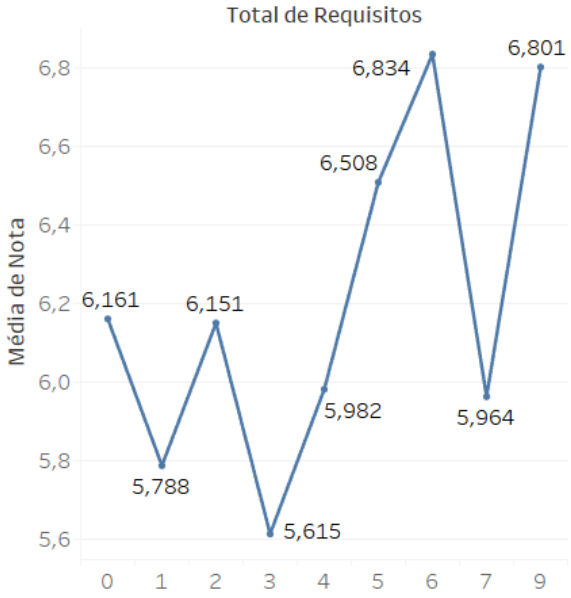

Figura 13 – Nota média por quantidade de pré-requisitos diretos e indiretos

4 ANÁLISE DE ESTRUTURA CURRICULAR USANDO O MÉTODO DE CONTROLE SINTÉTICO

Nesta seção. iremos descrever a problemática envolvida na análise curricular proposta. Após entender a estrutura do conjunto de dados, iremos derivar alguns atributos, e então explicar como o Método de Controle Sintético é utilizado. Em linhas gerais, o método utiliza um modelo linear que tenta prever as notas dos alunos numa dada disciplina baseando-se nas notas obtidas nas disciplinas anteriores. Os coeficientes encontrados no método linear explicitam quais disciplinas podem ser utilizadas para realizar a previsão. Usando esse resultado, podemos inferir uma possível relação de dependência entre as disciplinas.

4.1 Utilizando o método de controle sintético

Como vimos na seção 2.5, o método de controle sintético tenta encontrar um controle sintético composto por unidades de controle para melhor representar uma unidade de tratamento. A tese que *Introdução à Lógica Matemática* é um bom pré-requisito para *Teoria da Computação* se fortalece caso a nota da segunda seja capaz de prever a primeira, ou seja, que *Introdução à Lógica Matemática* esteja no controle sintético de *Teoria da Computação*. Se isto for verdade, então teremos evidência de que essas disciplinas possuem características comuns, e que faz sentido uma anteceder a outra.

No nosso caso, no entanto, desejamos conhecer a relação entre uma disciplina e as demais que a antecedem. O algoritmo do método não limita ou pondera determinadas unidades de maneira arbitrária, e assim sendo, devemos limitar os candidatos a unidade de controle manualmente para cada unidade de tratamento que estivermos estudando. Ou seja, para uma disciplina pertencente ao semestre i, apenas as disciplinas do semestre $i-1$ e anterior serão consideradas como candidatas a unidades de tratamento.

Para entender todos os dados utilizados pelo método, precisamos entender também alguns critérios relacionados às métricas utilizadas pela UFC na avaliação de seus alunos. O primeiro que veremos é o critério de aprovação. Para cada disciplina, o aluno obtém uma nota ao final do semestre, NS. Se $NS \geq 7$, o aluno é aprovado. Para esse caso, diremos que o aluno foi *aprovado com conceito A* . Se, no entanto, $NS < 4$, o aluno é considerado reprovado. Finalmente, se $4 \leq NS \leq 7$, ele deverá realizar uma avaliação final, de onde ele uma obterá uma nova nota AF. Se $(AF + NS)/2 \geq 5$, o aluno será aprovado. No entanto, esse aluno passou por um processo diferenciado, e por isso, aqui iremos considerar esse aluno como sendo *aprovado com conceito*

B. O conjunto de dados utilizado pelo *SCM* possui todas as notas dos com as quais os alunos foram aprovados em cada disciplina, seja com conceito A ou B. Caso o aluno tenha cursado a mesma disciplina mais de uma vez, apenas a vez em que ele foi aprovado é considerada.

O conjunto de dados possui também a frequência do aluno, ou seja, o percentual de aulas em que ele esteve presente. Finalmente, também está disponível o *IRA* do aluno quando ele ingressou na disciplina. O *IRA*, ou *Índice de Rendimento Acadêmico*, é uma medida utilizada pela UFC para avaliar o desempenho de um estudante dentro do curso. o *IRA* é calculado a partir da seguinte fórmula.

$$IRA = (1 - 0,5d)\left(\frac{\Sigma_i P_i C_i N_i}{\Sigma_i P_i C_i}\right) \tag{4.1}$$

- d é a fração entre o número de horas de disciplinas que o aluno desistiu e o número total de horas cursadas pelo aluno.
- N_i é a nota final da disciplina i.
- C_i é a carga horária da disciplina i.
- P_i é o mínimo entre o 6 e o semestre em que a disciplina foi cursada.

Assim, são utilizadas como preditores de uma disciplina os seguintes dados:

1. A nota média dos alunos daquela disciplina, se a nota for maior que ou igual a 7
2. A nota média dos alunos daquela disciplina, se a nota for menor que 7
3. A nota média de todos os alunos daquela disciplina
4. O IRA médio, ao ingressar, dos alunos cuja nota daquela disciplina for maior que ou igual a 7
5. O IRA médio, ao ingressar,dos alunos cuja nota daquela disciplina for menor que 7
6. A fração de alunos cuja nota é maior que ou igual a 7
7. A fração de alunos cuja nota é menor que 7
8. A frequência média de todos os alunos
9. A frequência média dos alunos cuja nota naquela disciplina é maior que ou igual a 7
10. A frequência média dos alunos cuja nota naquela disciplina é menor que 7

Precisamos também delimitar um conjunto de treino e um conjunto de teste. Em (ABADIE *et al.*, 2007), a escolha é simples: todos os momentos antes do projeto de lei serão do conjunto de treino, os demais serão conjunto de teste. No nosso caso, não temos um momento diferenciado, então selecionaremos aleatoriamente 75% dos alunos para servir de grupo de treino, os demais para grupo de teste.

Finalmente, nem todos os alunos do conjunto de dados concluíram o curso. De fato, apenas 186 são concludentes e com todas as notas necessárias para serem utilizadas pelo algoritmo. Alguns alunos entraram no curso através de transferência de curso, e não cursaram as mesmas disciplinas que os demais. Isto, aliado a uma grande evasão, faz com que o número de concludentes seja bastante baixo.

Neste trabalho não foi utilizado nenhuma abordagem para preenchimento dos valores *missing*. A primeira alternativa para lidar com esses valores é trabalhar apenas com os alunos que possuem nota em todas as disciplinas. A segunda é alterar o conjunto de dados para cada disciplina que estiver sendo analisada como unidade de tratamento. Como veremos na seção 4.1, o método de controle sintético utiliza apenas parte dos dados. Isso significa que se um certo aluno possui todas as notas até o 4º semestre, mas apenas algumas do 5º, então as notas desse aluno podem ser utilizadas no conjunto de dados para todas as disciplinas até o 4º semestre, e nas disciplinas do 5º semestre em que o aluno possui notas. No entanto, essa abordagem não representou melhora aos resultados encontrados no primeiro método, como veremos na seção 4.3.

4.2 Resultados dos Concludentes

O primeiro resultado encontrado diz respeito à execução do método nos 186 alunos concludentes do conjunto de dados. Nele, podemos encontrar o erro médio relativo do conjunto de teste para cada disciplina. Aqui, sendo y_i uma nota e $\hat{y}_i$ sua previsão de acordo com o algoritmo, definimos o erro relativo médio como sendo:

$$\Sigma_i \frac{|\hat{y}_i - y_i|}{y_i} \tag{4.2}$$

Utilizando o método *Lasso* com *alpha = 0* e coeficientes positivos *(positive=True)* a partir da biblioteca Scikit-learn (PEDREGOSA *et al.*, 2011), fazemos uma comparação dos resultados encontrados com uma regressão linear de coeficientes positivos. Os resultados, exibidos da tabela 8, têm um erro médio de 13%, um ótimo valor para um modelo linear. Isso é especialmente verdade para o método de controle sintético, que objetiva minimizar o erro entre a previsão e o valor real, assim como a regressão linear, mas também oferece significado a partir de seus coeficientes.

Analisamos também a relação estabelecida, a partir dos coeficientes, entre os pré-requisitos verdadeiros de uma disciplina e aqueles sugeridos pelos coeficientes do método de

controle sintético e da regressão linear com coeficientes positivos. Nesse caso, consideramos que uma disciplina *A* é selecionada pelos coeficientes para representar um pré-requisito de uma disciplina *B* se o coeficiente de *A* for maior que 0.1. A partir desse critério, interpretamos como um problema de classificação em que uma disciplina é ou não um pré-requisito, e avaliamos quão próximo estão as escolhas do verdadeiro conjunto de pré-requisitos a partir do F_1 *Score*, conforme (POWERS, 2011), definido por

$$F_1\,Score = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = \frac{2TP}{2TP + FN + FP} \tag{4.3}$$

TP representa a quantidade positivos verdadeiros, *FN* representa a quantidade de falsos negativos e *FP* representa a quantidade de falsos positivos. Os resultados de F_1 *Score* estão na tabela 10.

Tabela 8 – Erro relativo médio por disciplina para alunos concludentes

Disciplina	SCM	Regressão Restri.	Disciplina	SCM	Regressão Restri.
CALC2	0.16	0.15	BANCO	0.11	0.08
FISIC	0.15	0.15	LINGP	0.17	0.12
ESTRU	0.16	0.14	METO2	0.11	0.12
PROGR	0.13	0.14	REDES	0.12	0.11
TRANS	0.13	0.11	SIGBD	0.09	0.10
LOGIC	0.16	0.13	ENGEN	0.09	0.12
ESTAT	0.16	0.13	INTAR	0.16	0.17
GRAFO	0.16	0.16	SISOP	0.11	0.14
TECNI	0.10	0.12	APSIS	0.08	0.09
COMPG	0.10	0.11	AUTOM	0.14	0.16
CANAL	0.15	0.15	COMPI	0.11	0.18
METO2	0.14	0.14	TEORI	0.13	0.16
COMPG	0.19	0.12			

Tabela 9 – Erro relativo médio e F1-Score

	Método de Controle Sintético	Regressão Linear
Erro Médio	0.13	0.13
F1 Score	0.27	0.20
Precision	0.18	0.58
Recall	0.13	0.44

As tabelas a seguir contêm os os valores obtidos para os coeficientes do método de controle sintético para cada uma das disciplinas. Elas estão divididas em grupos por semestre.

Tabela 10 – F_1 Score por unidade

Disciplina	SCM	Regressão Linear
CALC2	0.5	0
FISIC	0.4	0.5
ESTRU	0.4	0.4
PROGR	0.5	0.66
TRANS	0.4	0
LOGIC	0.4	0.5
ESTAT	0.4	0.4
GRAFO	0.5	0.66
TECNI	0	0.8
ARQUI	0.4	0
CANAL	0.4	0.28
METO1	0	0
COMPG	0	0
BANCO	0	0
LINGP	0	0
METO2	0.4	0.25
REDES	0	0.28
SIGBD	0.5	0.5
ENGEN	0.4	0
INTAR	0.33	0
SISOP	0	0
APSIS	0.4	0
AUTOM	0	0
COMPI	0	0.57
TEORI	0.4	0

As disciplinas de um semestre tem o mesmo conjunto de unidades de controle, constituído de todas as disciplinas de semestres anteriores. Os valores foram aproximados para 2 casas decimais. Na tabela 12, por exemplo, temos que *Arquitetura de Computadores*, é representada pela combinação convexa das seguintes unidades de controle: *Fundamentos de Programação* (0.18), *Circuitos Digitais* (0.40), *Cálculo II* (0.02), *Programação* (0.12) e *Transmissão de Dados* (0.27). Essa disciplina sintética, que atinge uma semelhança baseada em F_1 Score de 0.4 com os verdadeiros pré-requisitos, aproxima as notas dos alunos do conjunto de teste com um erro médio de 10%.

Tabela 11 – Coeficientes do SCM para disciplinas do 2º semestre

Unidade de controle	CALC2	FISIC	ESTRU	PROGR	TRANS
CALC1	0.33	0.37	0.46	0.24	0.27
ALGEL	0.32	-	0.21	0.01	0.21
MATED	-	0.29	-	-	-
FUNDP	0.35	0.18	0.33	0.57	0.40
CIRCU	-	0.16	-	0.18	0.12

Tabela 12 – Coeficientes do SCM para disciplinas do 3º semestre

Unidade de controle	LOGIC	ESTAT	GRAFO	TECNI	ARQUI
CALC1	0.37	-	0.22	0.14	-
ALGEL	-	-	-	0.06	-
MATED	0.18	-	0.14	-	-
FUNDP	-	-	-	-	0.18
CIRCU	-	0.05	-	-	0.40
CALC2	0.16	0.28	0.05	0.06	0.02
FISIC	0.04	-	-	0.17	-
ESTRU	0.17	0.13	0.55	-	-
PROGR	-	0.14	-	-	0.12
TRANS	0.08	0.41	0.04	0.56	0.27

Tabela 13 – Coeficientes do SCM para disciplinas do 4º semestre

Unidade de controle	CANAL	METO1	COMPG	BANCO	LINGP
CALC1	0.26	-	-	0.26	-
CALC2	-	0.10	0.40	0.18	0.35
LOGIC	0.39	0.14	-	0.08	-
ESTAT	-	0.37	-	-	-
GRAFO	0.35	-	-	-	-
TECNI	-	0.39	0.60	0.48	0.65

Tabela 14 – Coeficientes do SCM para disciplinas do 5º semestre

Unidade de controle	METO2	REDES	SGBD	ENGEN
CALC1	-	-	0.06	-
ALGEL	-	0.03	0.08	0.18
FUNDP	0.19	0.16	0.09	0.09
TECNI	-	-	0.20	0.19
METO1	0.28	-	-	0.20
COMPG	0.23	0.23	-	-
BANCO	-	0.25	0.32	-
LINGP	0.30	0.32	0.25	0.35

Tabela 15 – Coeficientes do SCM para disciplinas do 6º semestre

Unidade de controle	INTAR	SISOP	APSIS	AUTOM
CALC1	-	-	-	0.33
ALGEL	-	-	0.06	-
MATED	0.02	-	-	-
FUNDP	-	0.06	0.21	-
CIRCU	-	0.17	-	-
CALC2	0.14	-	-	0.27
ESTRU	0.25	-	-	-
PROGR	-	0.33	-	-
TRANS	0.15	0.20	-	-
LOGIC	-	-	0.15	-
TECNI	-	-	0.29	-
ARQUI	-	0.03	-	-
CANAL	0.44	-	-	0.40
COMPG	-	-	-	0.01
METO2	-	0.21	-	-
ENGEN	-	-	0.29	-

Tabela 16 – Coeficientes do SCM para disciplinas do 7º semestre

Unidade de controle	COMPI	TEORI
COMPG	0.26	0.14
BANCO	0.19	-
LINGP	0.03	0.38
METO2	0.38	-
ENGEN	0.13	-
INTAR	-	0.22
AUTOM	-	0.27

4.3 Resultados com notas não nulas

O número de concludentes corresponde a uma fração relativamente pequena do conjunto de dados usado. Numa tentativa de aumentar o uso desse conjunto, foram considerados também alunos não-concludentes da seguinte maneira: se até um determinado semestre ele não possuir nenhuma nota zero, então essas notas podem ser utilizadas. Como resultado, o número de alunos disponíveis para análise varia por disciplina. Um mesmo aluno desistente pode ter as todas não-nulas até o 4º semestre, ter uma nota válida em Redes mas não em SGBD. Logo, ele pode ser utilizado em uma mas não em outra. Assim sendo, o número de notas disponíveis vai caindo ao longo do tempo, uma vez que mais alunos desistem com o passar dos semestres. Isso significa que a quantidade de dados para disciplinas do último semestre muda pouco, enquanto que haverá mais notas para as disciplinas de semestres iniciais, como demonstra a figura 14.

No entanto, houve perda significativa com esse procedimento. O erro médio no conjunto de teste que era de 13% passou para 18%. A figura 15 ilustra o valor de erro médio por disciplina.

Isso parece estar diretamente relacionado a quantidade de alunos desistentes na disciplina. A figura 16 relaciona o tamanho do erro por disciplina. Podemos observar a diferença na distribuição das notas de *Cálculo II* na figura 18. Ela é uma estimativa de densidade obtida através do método *kdeplot* da biblioteca *Seaborn* (SEABORN, 2018). Em azul, temos a distribuição das notas dos alunos concludentes. Já em vermelho, a distribuição das notas obtidas a partir do critério de nota não nula. Observe que nas notas não nulas, a curva possui uma calda com um pico em torno de 2, um comportamento inexistente nas notas dos concludentes. Como foi explicado em 4.1, um aluno concludente precisa ter todas as suas notas acima de pelo menos 5. Além disso, a curva em azul se aproxima mais facilmente de uma curva normal, enquanto que a curva vermelha tem um pico mais acentuado em volta de 5. É possível especular que os picos em torno de 5 e 7 sejam devidos a natureza de nota de corte entre os conceitos A e B

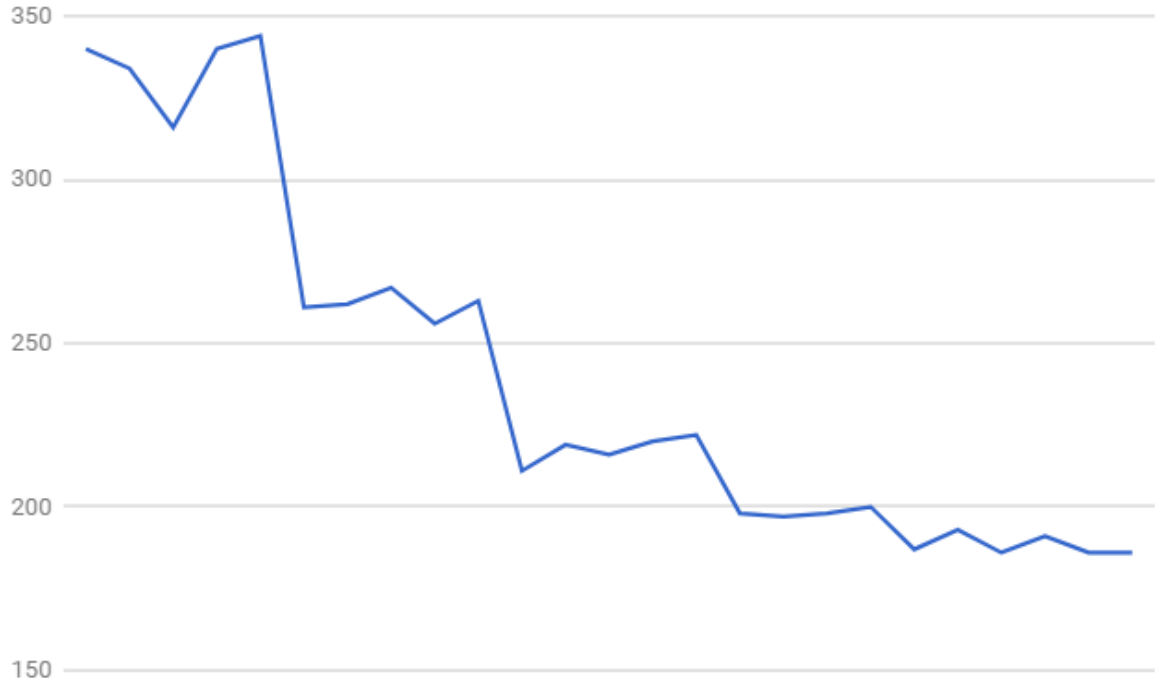

Figura 14 – Quantidade de alunos sem zero ao longo das disciplinas. Quanto mais a direita, maior o semestre

quando da aprovação de um aluno. Essa distinção causa um aumento de erro de 16% para 20% quando prevendo a nota de *Cálculo II*. Esses resultados sugerem que misturar notas de alunos concludentes com as demais notas pode atrapalhar a generalização do modelo.

4.4 Discussão

4.4.1 Pré-requisitos e disciplina sintética

Vimos na tabela 11 que a disciplina sintética que representa *Cálculo II* é composta de 33% de *Cálculo I*, 32% de *Álgebra Linear* e 35% de *Fundamentos de Programação*, e que essa aproximação tem em média 16% de erro no conjunto de teste. O método, a partir desse resultado, dá insumos que há fatores comuns entre as disciplinas componentes e a disciplina representada. Como era esperado, *Cálculo I* possui fator descritivo para *Cálculo II*. De um ponto de vista programático, no entanto, o pesquisador poderia se surpreender com o fato de que o coeficiente de *Cálculo I* não ser maior, afinal, de acordo com a estrutura curricular oficial utilizada pela *UFC*, apenas *Cálculo I* é pré-requisito de *Cálculo II*.

É possível acreditar, entretanto, que existam categorias de alunos que se dão melhor em alguns tipos de disciplina, mas não tão bem em outras. Um exemplo clássico a se imaginar é

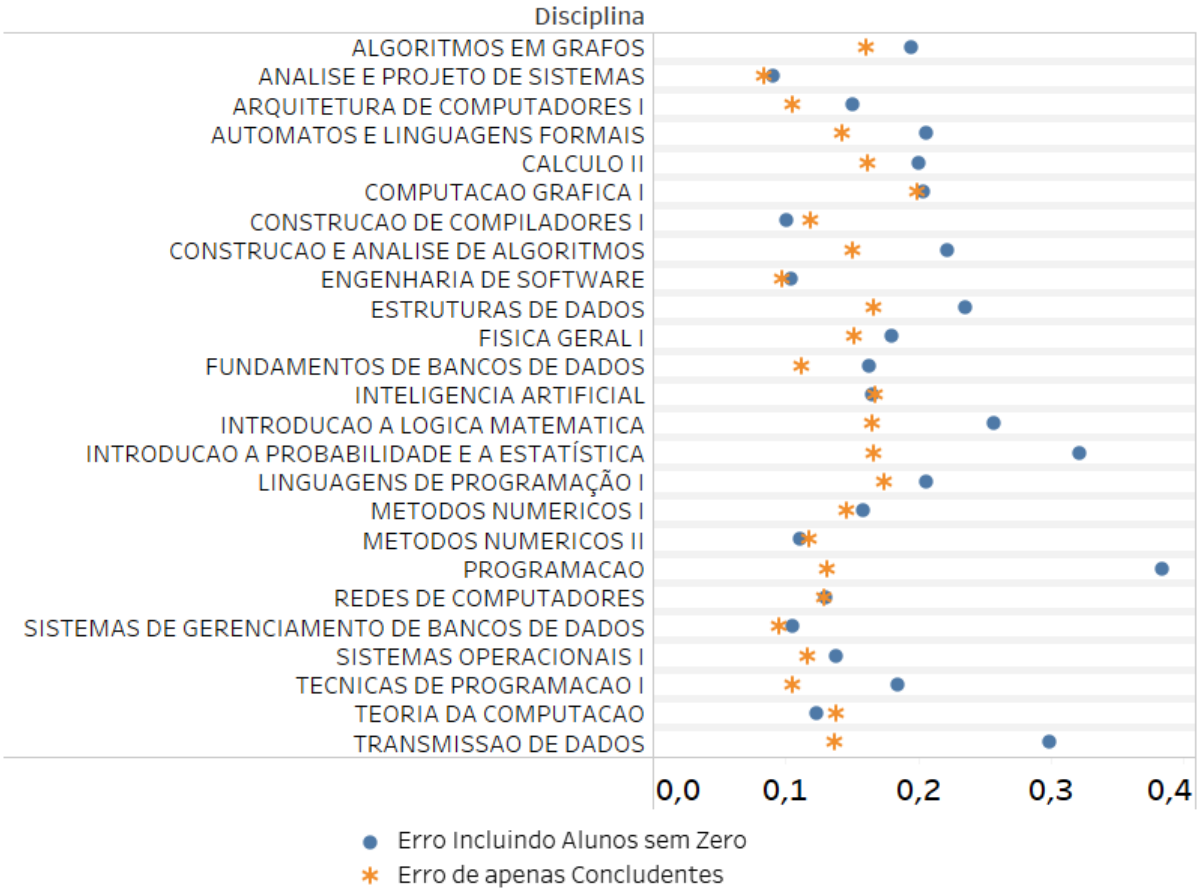

Figura 15 – Erro relativo médio por disciplina para concludentes e alunos sem zero. Os círculos representam o erro incluindo alunos sem zero, enquanto que os asteriscos representam os concludentes.

o tipo de aluno que prefere disciplinas teóricas a disciplinas práticas, ou vice-versa. Assim sendo, é possível conjecturar que existe uma malha invisível de características que afetam o desempenho de um aluno, mas que não estão relacionadas aquilo que é visto em disciplina, mas sim a como se vê ou como se pratica aquela disciplina. Isso torna possível que duas disciplinas, inicialmente distintas a uma análise rápida, estejam correlacionadas, mesmo que elas não compartilhem nada no que diz respeito ao conteúdo programático.

Com relação aos critério adotados no algoritmo, o erro relativo médio entre os preditores da disciplina sintética que representa *Cálculo II* e a disciplina original, por exemplo, é de 6%, o que demonstra não só que as características da disciplina original podem ser reproduzidas pela sua contraparte sintética, como também que o método logrou em aproximar com boa margem tanto as notas da disciplina como também suas características. A tabela 17 mostra os valores de erro relativo médio entre os preditores da unidade de tratamento e a unidade sintética.

Um interessante exemplo pode ser encontrado no resultado de *Construção e Análise de Algoritmos*. Sua disciplina sintética consiste de 26% de *Cálculo I*, 39% de *Introdução à*

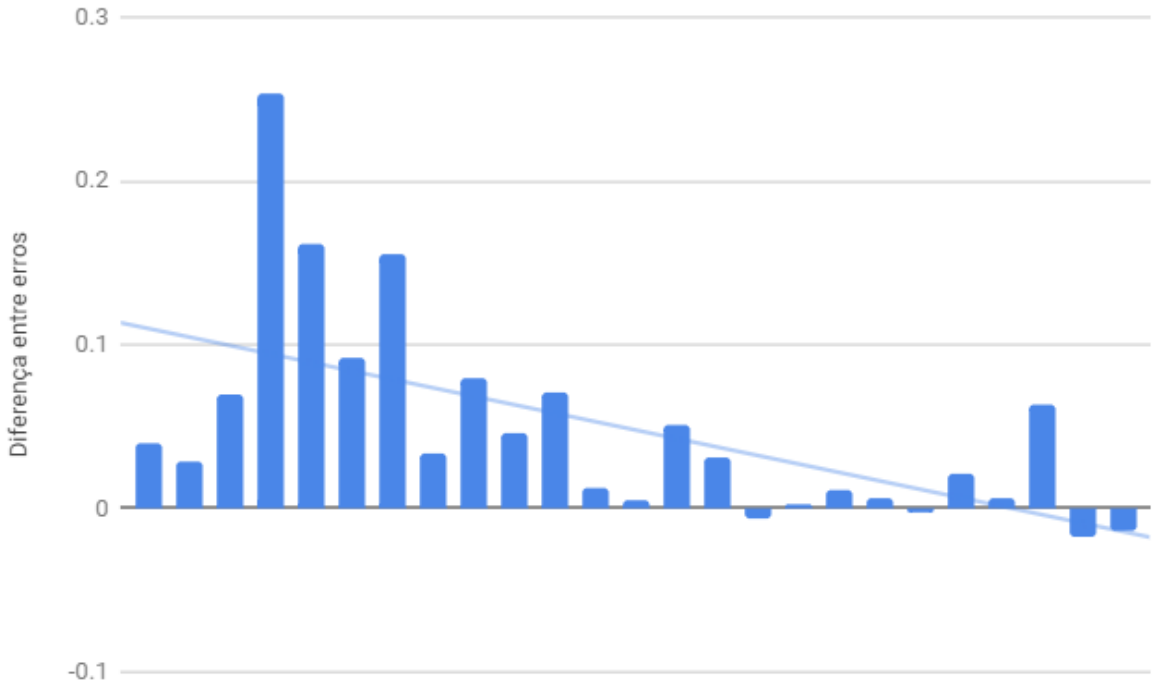

Figura 16 – Diferença entre erro utilizando concludentes e alunos sem zero. A linha clara mostra a tendência entre maior semestre e menor diferença.

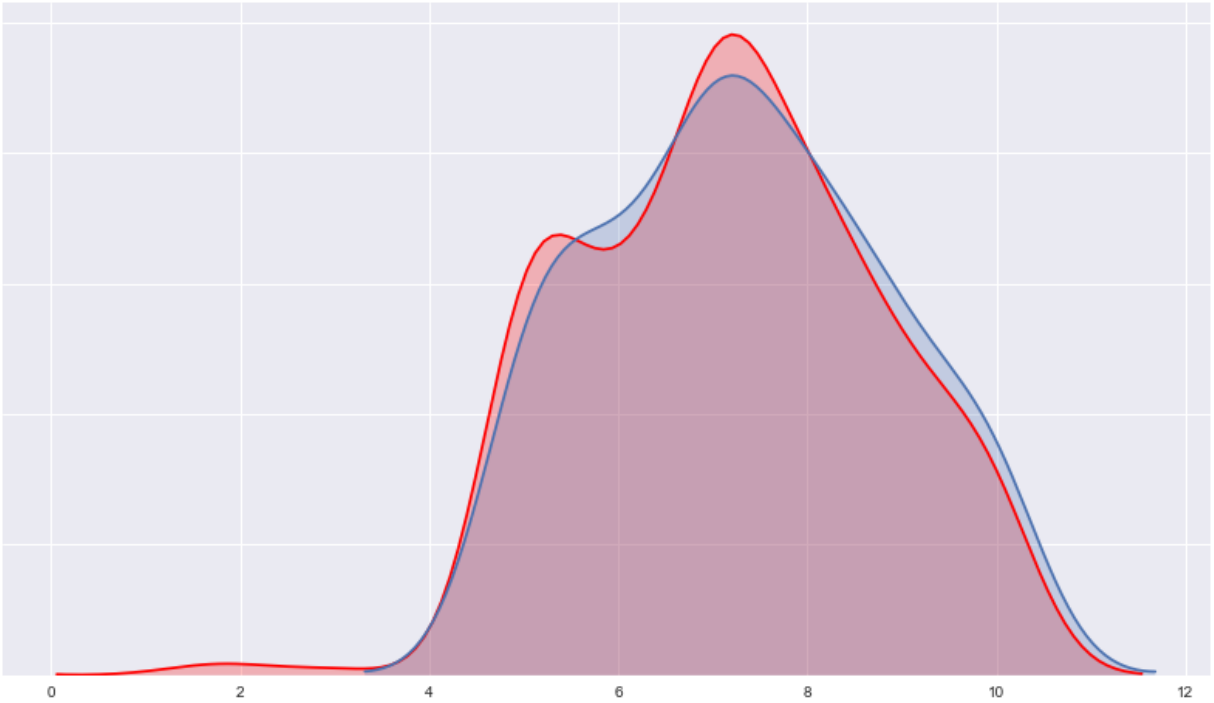

Figura 17 – Distribuição aproximada de kernel das notas obtidas em *Cálculo II*. A curva em vermelho representa a distribuição de notas não nulas, enquanto que a azul representa a distribuição das notas de alunos concludentes da disciplina

Lógica Matemática e 25% de *Algoritmos em Grafos*. Um de seus pré-requisitos obrigatórios, *Introdução à Probabilidade e Estatística*, não está presente. No entanto, já em 2016, ela deixou

Tabela 17 – Erro relativo médio dos preditores da unidade sintética

Disciplina	Erro Relativo Médio
Média	0.05
Calculo II	0.06
Fisica	0.02
Estrutura de Dados	0.05
Programacao	0.03
Transmissao	0.12
Logica	0.01
Estatistica	0.04
Grafos	0.01
Tecnicas	0.03
Arquitetura	0.18
Cana	0.04
Metodos I	0.02
CG	0.12
Bancos de Dados	0.03
LIP	0.09
Metodos II	0.05
Redes	0.01
SGBD	0.01
Engesoft	0.01
IA	0.01
SO	0.01
APS	0.14
Automatos	0.01
Compiladores	0.00
Teoria	0.01

de ser considerada um pré-requisito pelo departamento de Computação da *UFC*. Esse exemplo foi citado em (BARBOSA ARTUR; ARAUJO,), onde foi desenvolvida uma ferramenta de visualização cujo objetivo é facilitar a análise dos resultados encontrados no *SCM*.

Há diversas disciplinas que não contém algum ou todos os seus pré-requisitos. De fato, observando os resultados da tabela 10, pode-se concluir que na verdade os coeficientes encontrados concordam pouco com a estrutura curricular utilizada. *Computação Gráfica*, por exemplo, está costumeiramente muito atrelado a *Álgebra Linear*, que é um pré-requisito. Mesmo assim, o método não a tomou como unidade de tratamento. O mesmo ocorre com *Bancos de Dados*, que tem como pré-requisito *Estrutura de Dados*. Nesses casos, validar a estrutura curricular passa a não ser tão simples, uma vez que esses pré-requisitos tem uma relação programática muito forte com a disciplina posterior. Uma possível solução para esse problema é criar novos preditores, dessa vez binários, que representam a presença em uma categoria. Dessa forma, o algoritmo considerará disciplinas de uma mesma categoria mais semelhantes. Isso significa que se uma determinada disciplina pertence a uma categoria, então se torna mais provável que as disciplinas do grupo sintético sejam daquela categoria.

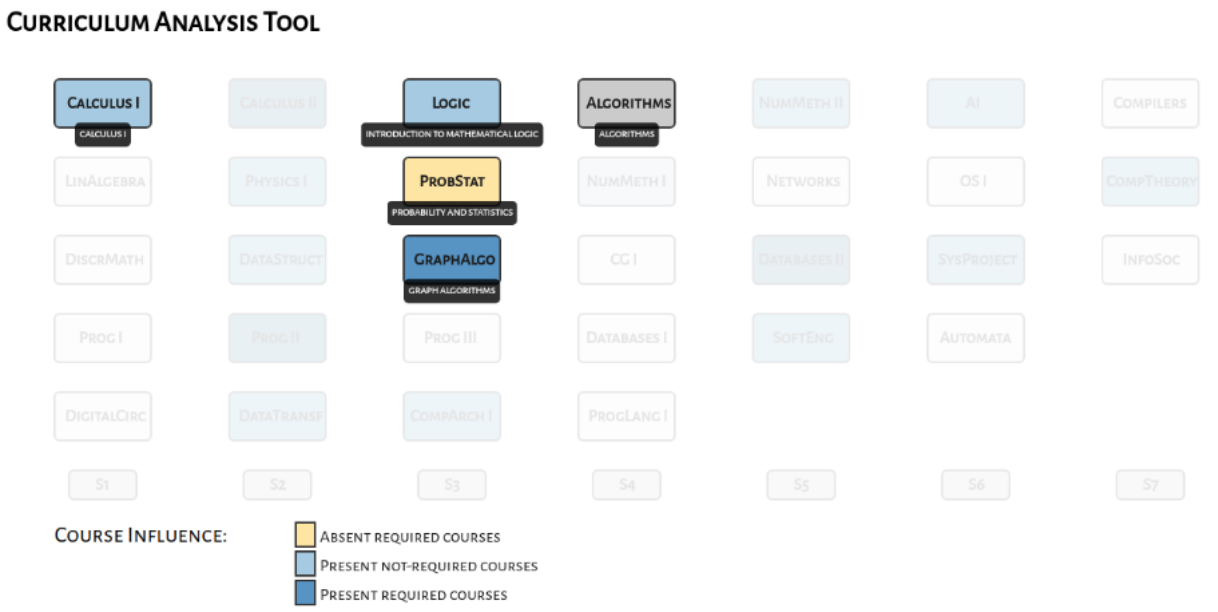

Figura 18 – Visualização construída para avaliar os resultados obtidos pelo *SCM*. A disciplina alvo, *CANAL*, está em cinza, a disciplina em azul escuro representa uma disciplina obrigatória que está presente no resultado do método, o azul claro representando disciplinas que são consideradas relevantes pelo método, e em amarelo, *ESTAT*, uma disciplina obrigatória, mas que não foi considerada relevante pelo método.

4.5 Quebra de requisito

O erro relativo médio obtido através do grupo de controle sintético é suficientemente bom para traçar estimativas que dizem respeito à generalização em termos de média da disciplina. No entanto, o desvio padrão do erro ainda é consideravelmente alto. Na figura 20, temos a distribuição do erro obtido com as notas previstas pelo grupo sintético de *Cálculo II*. Esse resultado levanta alguns questionamentos. Primeiro, se um educador, em posse dos resultados do algoritmo que demonstram que um determinado aluno não obterá nota suficiente baseando-se em suas notas passadas, deveria ou não interceder? Uma intervenção mal planejada pode ter piorar ainda mais a situação: se um aluno descobre que as estimativas apontam a sua reprovação, ele irá se esforçar mais para mudar as tendências ou irá se desestimular diante da possível fatalidade? Além disso, a partir de que margem de erro deveria o educador agir? Qual o intervalo de confiança mínimo que deveria ser tomado para tentar realizar mudanças estruturais?

Dentro deste contexto, o educador pode ou não realizar a quebra de pré-requisito: se um aluno não atende a um pré-requisito de uma certa disciplina, e o algoritmo aponta que aquele pré-requisito não tem relevância significativa para a estimação daquela nota, o educador pode optar por permitir que aquele aluno faça a disciplina mesmo sem atingir as exigências necessárias, desde que, utilizando as notas já obtidas, o algoritmo diga que ele tirará uma nota

suficientemente boa.

Nesse momento, uma característica interessante do resultado do método de controle sintético pode ser apontada. A relação de dependência entre as notas forma um grafo direcionado sem ciclos. De fato, a disciplina d_{ij}, a i-ésima disciplina do j-ésimo semestre, $j > 1$, sempre dependerá de um conjunto $\{d_{pq}\}$ de disciplinas, onde $q < j$. Caso uma dessas disciplinas de $\{d_{pq}\}$ não estiver disponível, e $q > 1$, então ela também pode ser obtida a partir de outras notas de um semestre anterior a q. Esse método transitivo pode ser utilizado para estimar uma vasta gama de notas assim que um nível básico esteja disponível.

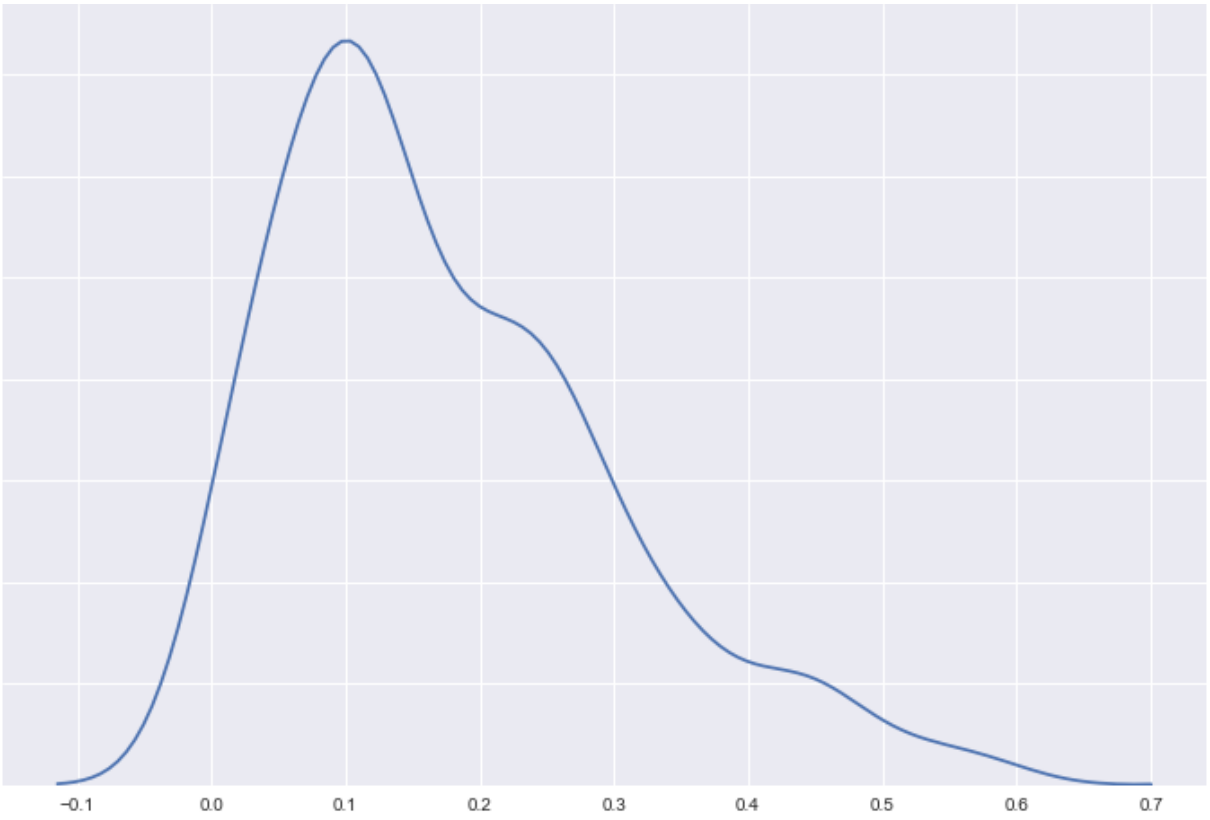

Figura 19 – Distribuição aproximada de kernel do erro relativo das notas obtidas em *Cálculo II* pelo controle sintético

5 CONCLUSÃO

Vimos nesse trabalho uma breve análise sobre alguns fatores que influenciam as notas dos alunos no curso de Ciência da Computação, como quantidade de pré-requisitos e unidade curricular. Vimos também o método de controle sintético, um algoritmo que visa encontrar uma unidade sintética construída a partir de unidades reais para ter uma contraparte cujo objetivo é realizar comparações que não poderiam ser feitas com facilidade entre uma unidade alvo e as demais que lhe são relativamente semelhantes. Para tanto, o algoritmo utiliza tanto uma grandeza que represente bem uma determinada dimensão estudada sobre as unidades, como também um conjunto de características que descrevem as unidades em estudo. Mostramos que nas ciências sociais e econômicas, o contexto em que o método está tradicionalmente inserido, o algoritmo pode ser utilizado para encontrar um grupo de comparação, baseando-se em indicadores econômicos e encontrando uma comparação a partir de critérios objetivos. No nosso estudo, utilizamos o método para encontrar um grupo de disciplinas que construa uma disciplina sintética com a qual é possível comparar uma disciplina alvo. A partir dessa comparação, tentamos encontrar uma relação entre uma disciplina e seus pré-requisitos, e a partir desses resultados, discutimos como podemos tentar traçar uma série de observações que podem ser úteis tanto para melhorar a estrutura curricular do curso, como também para auxiliar melhor alunos que tanto podem estar em dificuldades.

Como trabalhos futuros, podemos implementar um serviço que forneça os resultados do método utilizado para a administração de um departamento de educação, e entender como os profissionais envolvidos podem utilizar essa ferramenta como guia para construção curricular, ou permissão de quebra de requisito para alunos. Também podemos tentar utilizar mais atributos para melhorar o valor de erro encontrado.

REFERÊNCIAS

ABADIE, A.; DIAMOND, A.; HAINMUELLER, J. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. **Journal of the American Statistical Association**, v. 105, p. 493–505, 02 2007.

ABADIE, A.; DIAMOND, A.; HAINMUELLER, J. Comparative politics and the synthetic control method. **American Journal of Political Science**, v. 59, n. 2, p. 495–510, 2015. ISSN 1540-5907.

ABADIE, A.; GARDEAZABAL, J. The economic costs of conflict: A case study of the basque country. **American Economic Review**, v. 93, n. 1, p. 113–132, March 2003. Disponível em: http://www.aeaweb.org/articles?id=10.1257/000282803321455188.

BALDONI, M.; BAROGLIO, C.; BRUNKHORST, I.; HENZE, N.; MARENGO, E.; PATTI, V. Constraint modeling for curriculum planning and validation. **Interactive Learning Environments**, Routledge, v. 19, n. 1, p. 81–123, 2011. Disponível em: https://doi.org/10.1080/10494820.2011.528893.

BARBER, D. **Bayesian Reasoning and Machine Learning**. New York, NY, USA: Cambridge University Press, 2012. ISBN 0521518148, 9780521518147.

BARBOSA ARTUR; ARAUJO, N. S. E. G. J. Using learning analytics and visualization techniques to evaluate the structure of higher education curricula. **Proceedings of the XXVIII Brazilian Symposium on Computers in Education**.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738.

CARD, D.; KRUEGER, A. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. **American Economic Review**, v. 84, 04 1993.

FILHO, R. L. L. e. S.; MOTEJUNAS, P. R.; A, O. H.; LOBO, M. B. d. C. M. A evasÃno ensino superior brasileiro. **Cadernos de Pesquisa**, scielo, v. 37, p. 641 – 659, 12 2007. ISSN 0100-1574.

HURN, D.-U. . Using learning analytics to predict (and improve) student success: A faculty perspective. **Journal of Interactive Online Learning**, v. 12, p. 17–26, 2013.

JOHNES, J.; TAYLOR, T. Performance indicators in higher education. **Buckingham: SRHE and the Open University Pres**, 1990.

KIZILCEC, R. F.; PIECH, C.; SCHNEIDER, E. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In: **Proceedings of the Third International Conference on Learning Analytics and Knowledge**. New York, NY, USA: ACM, 2013. (LAK '13), p. 170–179. ISBN 978-1-4503-1785-6.

LEATHWOOD, C.; PHILLIPS, D. Developing curriculum evaluation research in higher education: Process, politics and practicalities. **Higher Education**, v. 40, n. 3, p. 313–330, Oct 2000. ISSN 1573-174X.

LENNON, J.; MAURER, H. Why it is difficult to introduce e-learning into schools and some new solutions. **Journal of Universal Computer Science**, v. 9, n. 10, p. 1244–1257, oct 2003.

MACFADYEN, L.; DAWSON, S. Mining lms data to develop an "early warning system"for educators: A proof of concept. **Computers Education**, v. 54, n. 2, p. 588–599, 2010.

MELIA, M.; PAHL, C. Pedagogical validation of courseware. In: **Proceedings of the Second European Conference on Technology Enhanced Learning: Creating New Learning Experiences on a Global Scale**. Berlin, Heidelberg: Springer-Verlag, 2007. (EC-TEL'07), p. 499–504. ISBN 3-540-75194-7, 978-3-540-75194-6.

PAPAMITSIOU, Z.; ECONOMIDES, A. A. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. **Journal of Educational Technology Society**, International Forum of Educational Technology Society, v. 17, n. 4, p. 49–64, 2014. ISSN 11763647, 14364522.

PECHENIZKIY, M.; TRCKA, N.; De Bra, P.; TOLEDO, P. Currim : Curriculum mining. International Educational Data Mining Society (IEDMS), p. 216–217, 2012. Editor(s): Yacef, K.; Zaïane, O.R.; Hershkovitz, A.; Yudelson, M.; Stamper, J.C. Proceedings of the 5th International Conference on Educational Data Mining (Chania, Greece, June 19-21, 2012); 5th International Conference on Educational Data Mining, EDM 2012 ; Conference date: 19-06-2012 Through 21-06-2012.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

POWERS, D. M. W. Evaluation: From precision, recall and f-measure to roc, informedness, markedness correlation. **Journal of Machine Learning Technologies**, v. 2, p. 37—-63, 2011.

PRIYAMBADA, S. A.; MAHENDRAWATHI, E.; YAHYA, B. N. Curriculum assessment of higher educational institution using aggregate profile clustering. **Procedia Computer Science**, v. 124, p. 264 – 273, 2017. ISSN 1877-0509. 4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia.

PROJETO. 2018. Disponível em: https://cc.ufc.br/curso/projeto-pedagogico/. Acesso em: 12 set. 2018.

RANKING de Cursos. 2018. Disponível em: http://ruf.folha.uol.com.br/2017/ranking-de-cursos/computacao/. Acesso em: 1 mar. 2018.

RAO, C. R.; MITRA, S. K. Generalized inverse of a matrix and its applications. In: **Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics**. Berkeley, Calif.: University of California Press, 1972. p. 601–620.

RODRIGUES Y. K. O.; PORTO, B. S. Texto de apoio pedagógico-curricular para (re)elaboração de projetos pedagógicos. 2013.

SAMPLES, J. The pedagogy of technology – our next frontier? **Connexions**, 2002.

SEABORN. 2018. Disponível em: https://seaborn.pydata.org. Acesso em: 1 mar. 2018.

SIEMENS GEORGE; LONG, P. Penetrating the fog: Analytics in learning and education. **EDUCAUSE Review**, v. 46, p. 30–32, 2011.

SYNTH. 2018. Disponível em: https://web.stanford.edu/~jhain/synthpage.html. Acesso em: 1 mar. 2018.

WANG, R.; ZAïANE, O. Discovering process in curriculum data to provide recommendation. **Proceedings of the 5th International Conference on Educational Data Mining**, p. 580–581, 2015. Cited By 2.

WU, K.; HAVENS, W. S. Modelling an academic curriculum plan as a mixed-initiative constraint satisfaction problem. In: KÉGL, B.; LAPALME, G. (Ed.). **Advances in Artificial Intelligence**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. p. 79–90. ISBN 978-3-540-31952-8.

APÊNDICEA – HISTOGRAMAS DAS NOTAS DAS DISCIPLINAS

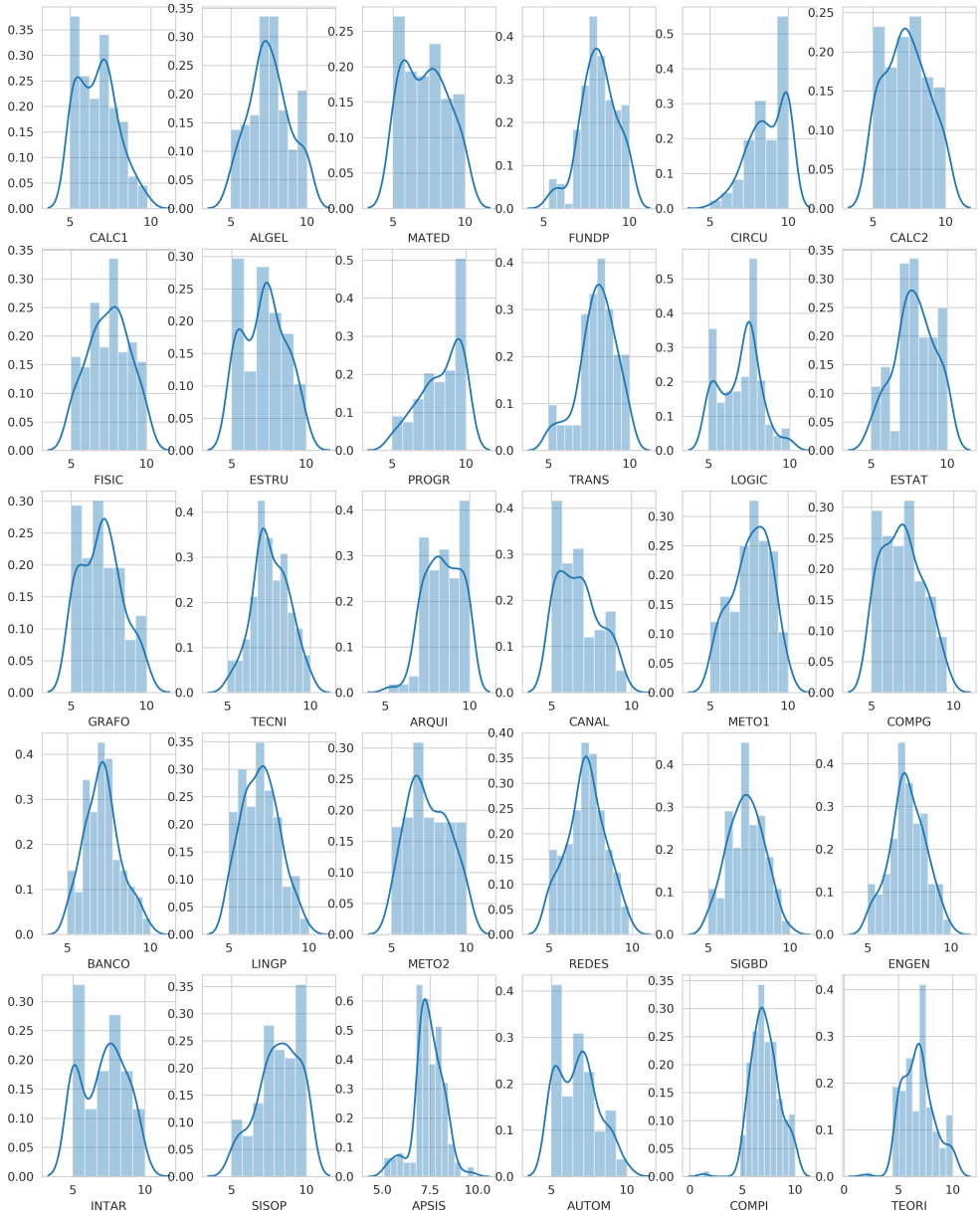

Figura 20 – Distribuição aproximada de kernel das notas de cada disciplina do conjunto de dados utilizado