Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Previsão automática de evasão estudantil: um estudo de caso na UFCG

Allan Sales da Costa Melo

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Mineração de dados educacionais

Leandro Balby
Adalberto Cajueiro

Campina Grande, Paraíba, Brasil

Resumo

A evasão estudantil é uma das maiores preocupações dos institutos de ensino superior brasileiros já que ela pode ser uma das causas de desperdício de recursos da Universidade. A previsão dos estudantes com alta probabilidade de evasão, assim como o entendimento das causas que os levaram a evadir, são fatores cruciais para a definição mais efetiva de ações preventivas para o problema. Nesta dissertação, o problema da detecção de evasão foi abordado como um problema de aprendizagem de máquina supervisionada. Utilizou-se uma amostra de registros acadêmicos de estudantes considerando-se todos os 76 cursos da Universidade Federal de Campina Grande com o objetivo de obter e selecionar atributos informativos para os modelos de classificação e foram criados dois tipos de modelos, um que separa os estudantes por cursos e outro que não faz distinção de cursos. Os dois modelos criados foram comparados e pôde-se concluir que não fazer distinção de alunos por curso resulta em melhores resultados que fazer distinção de alunos por curso.

Abstract

Students' dropout is a major concern of the Brazilian higher education institutions as it may cause waste of resources. The early detection of students with high probability of dropping out, as well as understanding the underlying causes, are crucial for defining more effective actions toward preventing this problem. In this paper, we cast the dropout detection problem as a supervised learning problem. We use a large sample of academic records of students across 76 courses from a public university in Brazil in order to derive and select informative features for the employed classifiers. We create two classification models that either consider the course to which the target student is formally committed or not consider it, respectively. We contrast both models and show that not considering the course leads to better results.

Agradecimentos

Primeiramente gostaria de agradecer a minha família, com destaque a minha mãe, Claudiram Sales de Melo, e irmã, Anne Karolyne Sales de Melo, por todo apoio e paciência comigo. Eu sei que é preciso ter muita paciência comigo. Aos meus primos e amigos Rômulo, Robson e Ronyel Roldão que foram um escape para as dores de cabeça e estresse de trabalho e do dia-a-dia e também ao meu primo, amigo e ex-companheiro de estudos e trabalho Luiz de Oliveira que esteve comigo em diversos momentos e foi determinante na escolha do meu tema de pesquisa.

Aos companheiros e ex-companheiros de laboratório e trabalho, em especial a Caio Santos, Iury Dewar, Ricardo Oliveira e principalmente a José Gildo pelas dezenas de discussões acadêmicas e não acadêmicas que me ajudaram a melhorar tanto como pesquisador quanto como pessoa.

Aos meus orientadores Leandro Balby, pelos conselhos e até mesmo por aceitar se tornar meu orientador com a pesquisa já em andamento e num momento de necessidade, e Adalberto Cajueiro, por me acolher desde o começo no processo de seleção.

E a amigos que se fizeram presente com alguma contribuição direta ou indireta. Dentre todos, não poderia deixar de destacar a participação de Thamyres Cardoso, Henrique Truta, Rommel Raphael, Valkercyo Feitosa e principalmente de Andryw Marques, que foi importante tanto em discussões sobre a pesquisa quanto em momentos de confraternização e amizade.

Conteúdo

Lista de Símbolos

UFCG - *Universidade Federal de Campina Grande*

MDE - *Mineração de Dados Educacionais*

OSS - *One-Sided Selection*

CNN - *Condensed Nearest Neighbor*

VP - *Verdadeiro Positivo*

VN - *Verdadeiro Negativo*

FP - *Falso Positivo*

FN - *Falso Negativo*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Contextualização

Com o surgimento de diversas políticas públicas com o objetivo de expandir o acesso da população às Universidades brasileiras, o número de matrículas cresceu consideravelmente nos últimos anos. Em 2013, por exemplo, foram registrados mais de 7 milhões de matrículas, e esse número tem crescido continuamente [7]. No entanto, estima-se que apenas 62.4% do total de alunos matriculados concluem os seus cursos [5], porcentagem que pode indicar a existência de um alto índice de evasão estudantil.

A evasão estudantil pode ser encontrada de forma abrangente em diversos níveis de educação formal ao redor do mundo. Notas baixas, aulas ruins, disciplinas mal estruturadas, trabalhar e estudar em paralelo, falta de perspectivas de trabalho após formado, problemas familiares e falta de aptidão na área [9; 4; 1; 2] são exemplos de razões que comumente levam alunos a abandonarem seus cursos. Diversos estudos também apontam que a ocorrência da evasão é maior no início dos cursos, devido às razões já mencionadas [8; 23], e causam um prejuízo, devido principalmente a desperdício de recursos, para a Universidade [6]. O gráfico superior da Figura 1.1, criada com base nos dados dos 76 cursos da Universidade Federal de Campina Grande (UFCG) - Brasil[1], mostra a quantidade de evasões por número de períodos de curso e confirma que a evasão ocorre principalmente no início do curso.

[1] `http://www.ufcg.edu.br`

1.2 Motivação

Apesar de o número de evasões decrescer fortemente ao longo de semestres subsequentes, o custo gerado para a Universidade decresce mais lentamente (ou chega a crescer), mostrando que quanto mais tempo a evasão leva para acontecer mais custosa ela se torna. Dessa forma, conclui-se que a evasão causa desperdício independente do momento em que ela acontece. A Tabela 1.1 e a Figura 1.1 mostram o custo da evasão por período, considerando alunos da UFCG do primeiro ao décimo período matriculados no ano de 2013. Para calcular o custo total da evasão por período, tomou-se como referência o investimento médio anual, em 2013, do governo [6] por aluno e a fórmula 1.1 (não-oficial, desde que não é oriundo de uma fonte do próprio governo ou Universidade):

$$C(p) = p \times C \times n \tag{1.1}$$

onde:

- $C = 21.383/2$ e representa o custo, em R$, referente a 6 meses de curso (equivalente ao tempo de 1 período) para um aluno;
- p representa o número de períodos que o aluno passou matriculado no curso; e
- n é o número de evasores no período.

A evasão tem impacto em diferentes âmbitos:

- Sociedade: que tenderá a carecer de profissionais qualificados;
- Estudantes: que levarão mais tempo para obter um diploma do ensino superior, assumindo que o evasor voltará a cursar um programa de ensino superior;
- Universidade: que investe em infraestrutura - como computadores, salas, professores, carteiras - que será subutilizada devido a falta de estudantes. Foi estimado que em 2013 cada estudante custava R$21.383,00 [6] para as Universidades brasileiras.

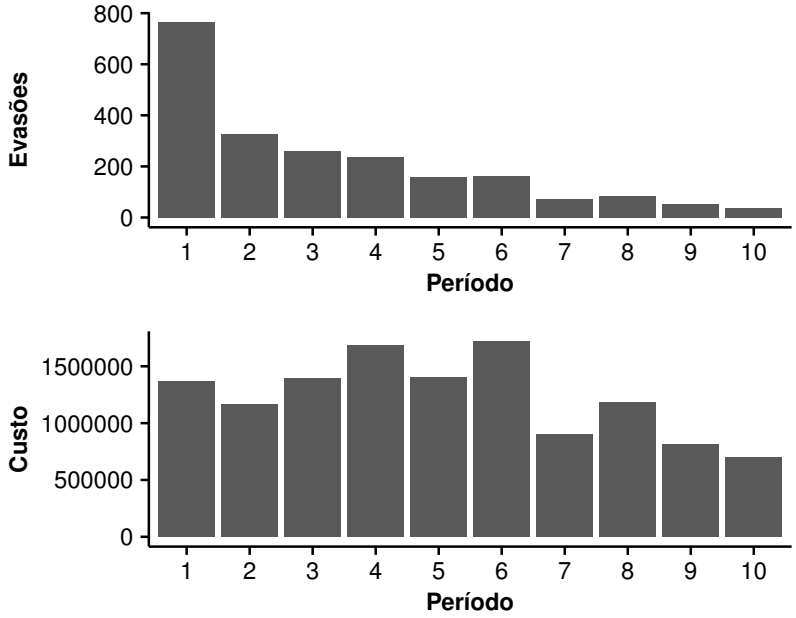

Figura 1.1: Número de evasões e custo (R$) por período na UFCG.

Períodos	Número de evasões	Custo (R$)
1	766	1.365.012,00
2	326	1.161.864,00
3	261	1.395.306,00
4	236	1.682.208,00
5	157	1.398.870,00
6	161	1.721.412,00
7	72	898.128,00
8	83	1.183.248,00
9	51	817.938,00
10	39	694.980,00

Tabela 1.1: Custo total (R$) por Período

1.3 Contribuições

A aplicação de um modelo de predição de evasão é importante já que pode ajudar administradores da universidade a, baseados na identificação dos evasores, tomarem providências a fim de prevenir que novos casos venham a se repetir no futuro.

Nesta dissertação, a detecção de evasão é abordado como um problema de aprendizagem

supervisionada, usando atributos extraídos do histórico acadêmico dos alunos. Com esse objetivo, foram aplicados modelos de classificação que categorizam os estudantes em uma de duas classes pré-definidas: "evasor"ou "não-evasor". Isto é, tem-se como objetivo identificar quais estudantes possivelmente não vão continuar na Universidade após o fim de um período. Foram criados dois modelos de classificação, onde o primeiro leva em consideração os alunos e seus respectivos cursos e o segundo que não faz distinção dos cursos que os alunos estão matriculados. A ideia é investigar se vale a pena criar um modelo especializado para cada par curso/semestre, em vez de apenas criar um modelo global (i.e. que considera todos cursos ao mesmo tempo) por semestre. Também realizou-se uma seleção de atributos e foram avaliadas várias configurações de aprendizado diferentes com uma variedade de atributos, e concluiu-se que não fazer distinção de alunos por curso resulta em melhores resultados do que fazer distinção de alunos por curso para qualquer métrica utilizada (i.e. diferença de 14.3% em relação ao *F-measure*).

Esta pesquisa está inserida em um campo conhecido como Mineração de Dados Educacionais (MDE) [25] que tem sido uma ferramenta de grande importância no suporte a instituições de educação para a definição de melhores ações preventivas e corretivas, tais como melhorar a alocação de recursos e de pessoal além da orientação de alunos identificados como potenciais desistentes. Alguns trabalhos surgiram recentemente propondo usar aprendizado de máquina para detectar potenciais evasores (ver capítulo 3). Esta pesquisa estende esses trabalhos com as seguintes contribuições:

- Foram utilizados dados de dados de registos acadêmicos de 32.342 alunos de todos os 76 cursos de uma Universidade pública brasileira. Essa base de dados é significativamente maior do que as bases de dados tipicamente usada pelos trabalhos relacionados;

- São propostos novos atributos discriminativos que não foram encontrados na literatura. Como por exemplo PORCENTAGEM.CURSO.COMPLETO, que indica quanto do curso já foi completo pelo aluno;

- É conduzida uma ampla análise de seleção de atributos, usando diferentes algoritmos de seleção de atributos, com o objetivo de descobrir que atributos tem maior impacto no desempenho dos classificadores nas duas perspectivas já descritas anteriormente;

- São propostas duas perspectivas diferentes de abordagem ao problema: criar um classificador por semestre ou um classificador por curso/semestre. Para cada perspectiva foram avaliadas diversas configurações de algoritmos de classificação considerando apenas o semestre ou o par curso/semestre. A melhor configuração de cada perspectiva foram escolhidas e comparadas.

Também é importante destacar que parte desta pesquisa - onde foi proposto um classificador para detectar estudantes com alta probabilidade de evasão no primeiro período de curso - já foi publicada no Symposium on Knowledge Discovery, Mining and Learning (KDMiLE) [27], convidada a ser estendida e atualmente está sendo avaliada com a possibilidade de ser publicada no Journal of Information and Data Management (JIDM).

1.4 Organização

O restante desta dissertação está organizado da seguinte forma. O capítulo 2 apresenta toda a fundamentação teórica para o entendimento desta pesquisa e a formulação do problema que foi proposto de se resolver. O capítulo 3 discorre a respeito dos trabalhos relacionados e inclui pesquisas que preveem a ocorrência da evasão universitária em 1) cursos específicos, 2) períodos de tempo acadêmico específico e 3) usando dados de contextos diferentes dos adotados no nosso trabalho. O capítulo 4 mostra como estão organizados os dados disponíveis, desde sua forma inicial até a transformação em atributos usados nos modelos. O capítulo 5 relata o processo de criação, escolha e avaliação dos modelos Global e Específico e, finalmente, o capítulo 6, apresenta as conclusões e direções para trabalhos futuros.

Capítulo 2

Fundamentação teórica

Neste capítulo é apresentada a fundamentação teórica necessária para o entendimento desta pesquisa. Abaixo estarão descritos alguns conceitos básicos relativos a classificação 2.1, com foco em Florestas Aleatórias, Seleção de atributos, Balanceamento de classes e métricas que são utilizadas para avaliar os experimentos e em seguida será explicado o processo de funcionamento da universidade e o contexto no qual o aluno é inserido (seção 2.2), a fim de definir todas as nomenclaturas que estão sendo adotadas no texto. Na seção 2.3 também é apresentado a formulação do problema que será tratado no restante da dissertação.

2.1 Classificação

Classificação é uma tarefa da aprendizagem de máquina que consiste em associar um objeto a uma classe pré-determinada. Essa técnica é comumente utilizada para prever a que classe pertencerá um novo objeto ou para separar objetos em classes. A Figura 2.1 retrata um exemplo de classificação: existe um conjunto de objetos e deseja-se separá-los utilizando algum critério. No contexto da evasão, pode-se pensar que existe um conjunto de alunos e é desejado separá-los em dois outros grupos: 1) o de possíveis evasores e 2) o de não evasores. Na prática, dentro de um contexto existem dados que são rotulados e utilizados para treinar um modelo de classificação que, por sua vez, será utilizado para classificar novas instâncias de dados dentro do mesmo contexto.

Dentro do contexto da evasão, queremos predizer a que classe o aluno pertencerá no fim do período: "evasor"ou "não evasor". Existem diversos algoritmos de classificação na litera-

tura, tais como: Support Vector Machine [14], Multilayer Perceptron [26], Naive Bayes [14], Árvore de Decisão C5.0 [14], Regressão Logística [14], dentre outros. Neste trabalho foi adotado a Floresta Aleatória por ser, dentre os métodos conhecidos pelos autores, o que apresenta maior poder preditivo e alguma capacidade de interpretação de seus componentes. Além disso, diversos outros aspectos dos dados devem ser considerados antes de definirmos um modelo de classificação, dentre os mais importantes estão a Seleção de Atributos e o Balanceamento das classes.

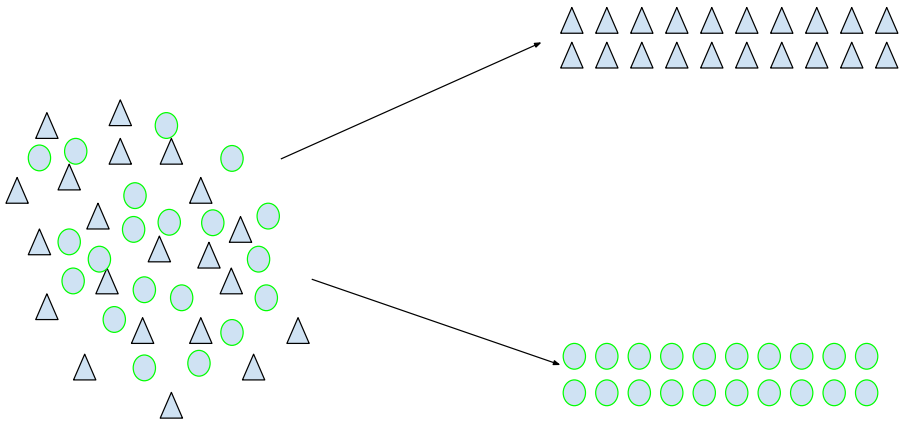

Figura 2.1: Exemplo de classificação.

2.1.1 O Processo de Classificação

Para classificar é necessária inicialmente a disponibilidade de dados que, seguindo algum critério, serão divididos em conjunto de treino - partição de dados rotulados que servirá para criar o modelo - e de teste - partição de dados não rotulados que servirá para fazer a avaliação

do modelo criado. Na Figura 2.2 é possível observar o processo de criação de um modelo, desde a sua criação até o seu teste.

Os passos indicados na Figura 2.2 são descritos a seguir:

- 1) cria-se o conjunto de treino com seus rótulos (ou classes) bem definidos;
- 2) realiza seleção de atributos a fim de selecionar atributos mais discriminativos;
- 3) submete-se o conjunto de treino a algum algoritmo de classificação a fim de criar o modelo;
- 4) utiliza-se o conjunto de teste, não rotulado, como entrada do modelo; e
- 5) realiza-se a classificação de cada uma das instâncias de teste.

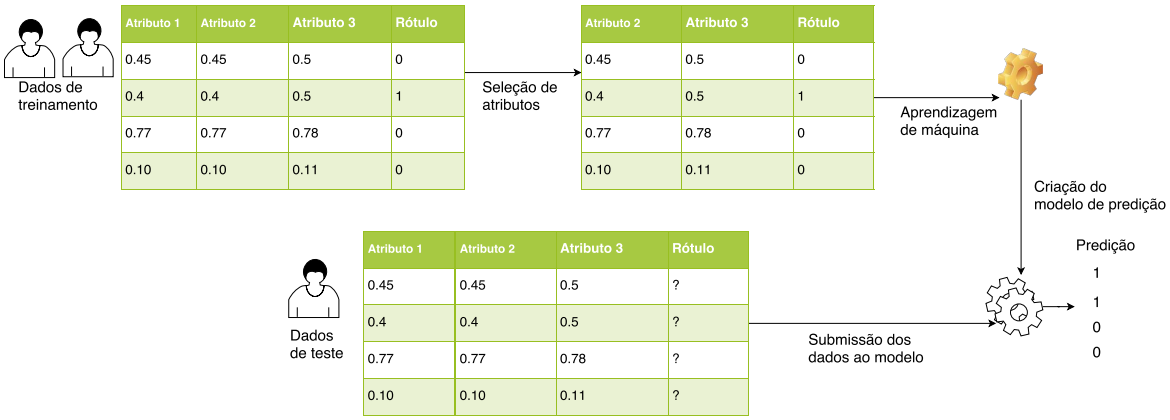

Figura 2.2: Arquitetura de uma classificação.

2.1.2 Seleção de Atributos

Seleção de Atributos é uma etapa do processo de aprendizagem, onde o objetivo é selecionar os atributos mais discriminativos com a intenção de, dentre outras, reduzir o número de dimensões da base de dados. Essa etapa é importante pelas seguintes razões:

- Pode ser computacionalmente caro utilizar todos os atributos disponíveis para criar um modelo;

- Na base de dados disponível podem haver atributos que não acrescentam nenhum tipo de informação ao modelo;

- Em diversas situações, menos atributos implica em uma maior facilidade de interpretação dos resultados.

Nesta pesquisa, foram utilizados e comparados três algoritmos de Seleção de Atributos: 1) *Information Gain* [16], 2) *Gain Ratio* [24], 3) *Symmetrical Uncertainty* [10] e o *Gini Index* [11]. Esses algoritmos foram escolhidos devido a serem comumente utilizados na literatura especializada, apresentam suas peculiaridades e são brevemente descritos a seguir.

- *Information Gain*. Faz uso do conceito de entropia e pode ser definido como: redução da entropia causada por dividir as classes alvo de acordo com um determinado atributo. Quanto maior o valor de *Information Gain*, maior o ganho ao utilizar um atributo para separar classes. Esse algoritmo também é conhecido por mostrar-se enviesado a respeito de tender a apresentar valores mais altos a medida que o atributo utilizado pode assumir mais valores [11];

- *Gain Ratio*. Métrica criada a fim de corrigir o viés do *Information Gain*. Para isso, no seu cálculo, é levado em consideração a quantidade de valores que o atributo pode assumir. Ao fazer essa consideração, o *Gain Ratio* acaba por priorizar atributos que assumem poucos valores;

- *Symmetrical Uncertainty*. Também propõe corrigir o viés do *Information Gain* dividindo o valor de *Information Gain* pela entropia do conjunto inicial somado a entropia do conjunto após a separação das classes por algum atributo. O *Symmetrical Uncertainty* é normalizado no intervalo $[0, 1]$.

- *Gini Index*. Comumente utilizado para medir o grau de desigualdade dentro de um contexto social - i.e. distribuição de renda familiar - em países, pode ser utilizado no contexto da classificação para medir o grau de heterogeneidade de uma amostra. O valor de Gini é 0 quando há apenas instâncias de uma classe na amostra e é máximo quando há a mesma quantidade de instâncias de cada classe.

2.1.3 Balanceamento de classes

Na mineração de dados, o problema de classes desbalanceadas é caracterizado pela dominância do número de instâncias de uma classe sobre a outra. No contexto da evasão universitária, seria o caso de termos a maioria dos alunos não evadindo enquanto poucos evadem a cada semestre.

Dados desbalanceados podem ser um problema tendo em vista que podem atrapalhar o aprendizado dos modelos de classificação existentes. Por exemplo, digamos que em um período da UFCG, 50 alunos evadiram e 1950 não evadiram e que queremos construir um classificador para prever a evasão. Utilizando a base desbalanceada, um classificador tenderá a classificar todos os alunos como não evasores já que isso somente ocasionará um erro de 50 em 2000 casos. Ao fazer um balanceamento, com número igual de instâncias de cada classe, reconhecer os padrões de cada classe e separá-las se torna uma tarefa menos complicada e evita casos como o do exemplo anterior.

Existem duas possibilidades para balancear as classes da base de dados, chamadas de *oversampling* e *undersampling*, respectivamente: 1) Criar artificialmente instâncias da classe minoritária (classe com menos instâncias) até atingir uma quantidade próxima a da classe majoritária (classe com mais instâncias) ou 2) Remover instâncias da classe majoritária até atingir uma quantidade próxima a da classe minoritária. Diversas são as técnicas existentes de *oversampling* e *undersampling*. Abaixo estão descritas as três técnicas utilizadas neste trabalho.

One-Sided Selection

O One-Sided Selection (OSS) [17] é uma técnica do tipo *undersampling* resultante da aplicação dos métodos Tomek Link [28] e Condensed Nearest Neighbor (CNN) [12]. A ideia é usar o OSS para remover ruídos e instâncias próximas à fronteira de decisão (Tomek Link) - Fronteira de Decisão é a função que define a qual espaço no hiperespaço pertencerá cada classe - e instâncias da classe majoritária muito distantes da fronteira de decisão (CNN) - o que as faz, teoricamente, menos relevantes para o processo de aprendizagem. As Figuras 2.3 e 2.4 ilustram o uso dos Tomek Link e do CNN, respectivamente.

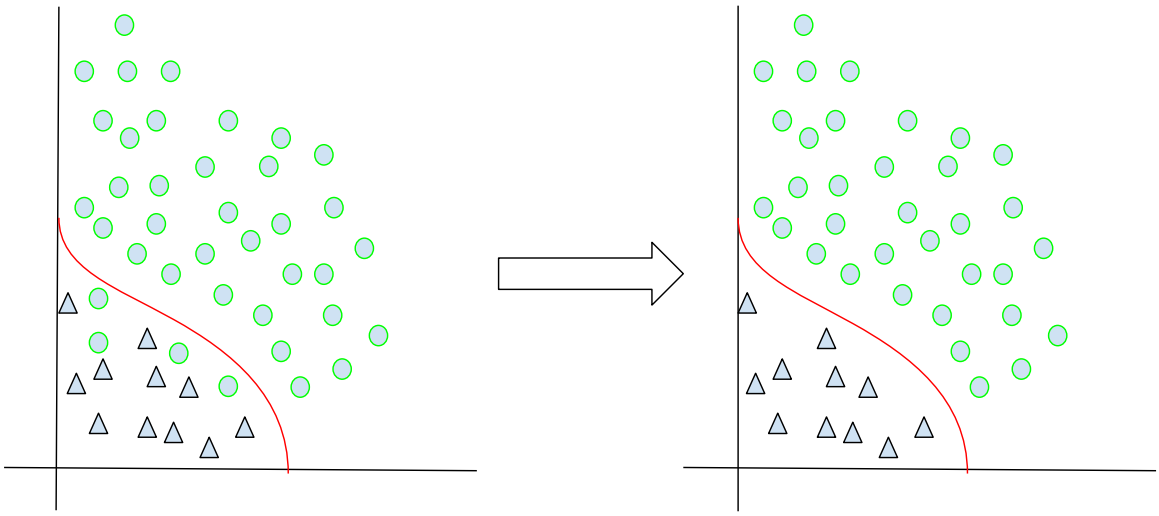

Figura 2.3: Exemplo de uso do Tomek Link.

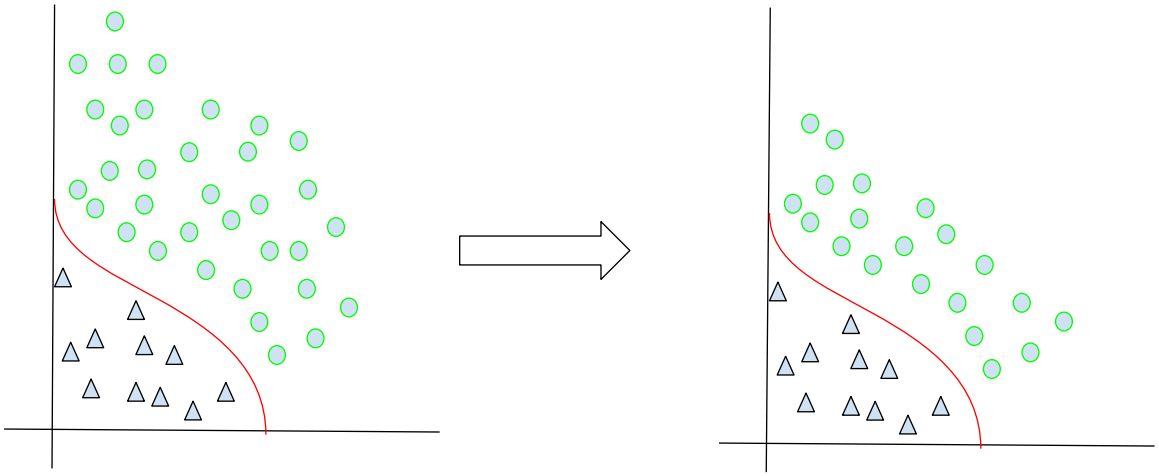

Figura 2.4: Exemplo de uso do CNN.

Random undersample e Random oversample

Ambas as técnicas são utilizadas para equiparar o número de instâncias das duas classes na base de dados. A principal diferença é que o random undersample consiste em remover instâncias da classe majoritária escolhidas aleatoriamente, enquanto o random oversample consiste em replicar aleatoriamente instâncias da classe minoritária. Os métodos de random oversample e random undersample foram escolhidos devido serem mais simples de ser

implementados e existerem estudos mostrando que métodos mais complexos não necessariamente resultaram em melhores resultados [15].

2.1.4 Árvore de Decisão

O objetivo da Árvore de Decisão é criar uma árvore composta por nós e folhas, onde os nós são utilizados para decidir qual caminho tomar e as folhas servem como definição da classe da instância que está sendo analisada. Uma Árvore de Decisão é criada seguindo o conceito de busca gulosa e com base na hipótese de que uma sequência de seleções ótimas locais levarão a uma solução ótima global. São construídas a partir da raiz em uma sequência de passos até que a árvore final seja encontrada e, em cada passo considera que a escolha do nó deve ser: 1) ótima local e 2) irrevogável.

Indução de Árvores de Decisão

A ideia é criar a árvore utilizando os atributos disponíveis mais importantes primeiro, considerando que os atributos mais importantes são aqueles que tem maior poder de classificação - a importância de cada atributo pode ser inferida utilizando algoritmos de Seleção de Atributos, como o *Information Gain*.

Para cada nó criado, o processo de inferir qual é o atributo disponível mais importante, e que deve ser utilizado no momento, é repetido até que não possa mais se expandir um nó ou que todas as instâncias pertençam a mesma classe. Como exemplo, vamos criar uma árvore de decisão usando os dados da Tabela 2.1. A Figura mostra uma tabela de atributos que queremos utilizar para criar uma árvore que classificará um animal em mamífero ou não mamífero.

Id	Nome	Temperatura do Corpo	Pernas	Dar a Luz	Mamífero
1	humano	quente	sim	sim	sim
2	baleia	quente	não	sim	sim
3	salamandra	frio	sim	não	não
4	pombo	quente	sim	não	não
5	morcego	quente	sim	sim	sim
6	sapo	frio	sim	não	não
7	tubarão-leopardo	frio	não	sim	não
8	salmão	frio	sim	não	não

Tabela 2.1: Tabela exemplo de classificação.

O primeiro passo é decidir quem deve ser escolhido como nó raiz, nesse momento estão disponíveis os atributos descritos na Figura 2.5[1]. Dentre eles, os atributos mais discriminativos se mostram sendo o "Dar a luz"e o "Temperatura do corpo- devido a capacidade de separar mais animais não mamíferos de mamíferos - e que por serem igualmente discriminativos, dão a opção do algoritmo escolher um deles ao acaso para servir como nó raiz. Ficamos com o atributo "Temperatura do corpo". É fácil de perceber que o nó filho da direita, quando a temperatura do corpo assume valor "fria", não pode mais se expandir, já que todas as classes desse nó são "não".

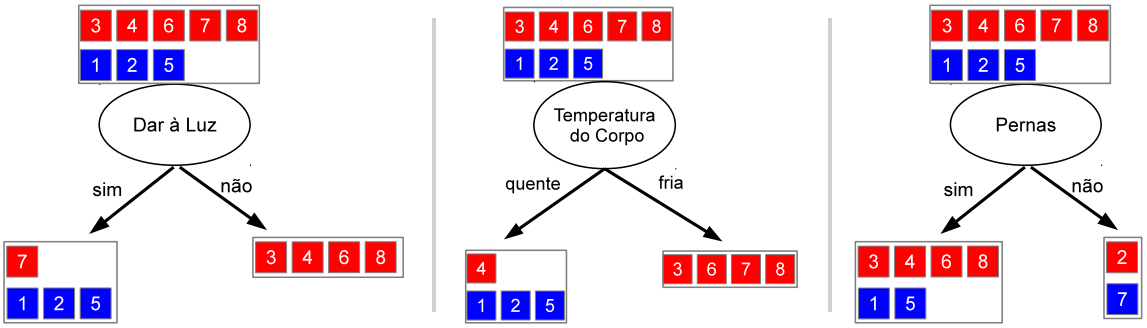

Figura 2.5: Árvore de Decisão exemplo - Escolha da Raiz.

Continuando a expansão do nó filho da raiz da esquerda - quando o "Temperatura do corpo"assume valor "quente- ainda há a disponibilidade dos atributos "Dar a luz"e "Pernas".

[1]Figuras das notas de aula cedidas pelo Prof. Dr. Leandro Balby

A Figura 2.6[1] mostra como as classes são divididas para cada atributos. Nota-se que usando o atributo "Dar a luz", as duas classes "sim"e "não"estão completamente separadas, sendo ela o atributo escolhido para continuar a expansão e finalizar a Árvore de Decisão, representada na Figura 2.7.

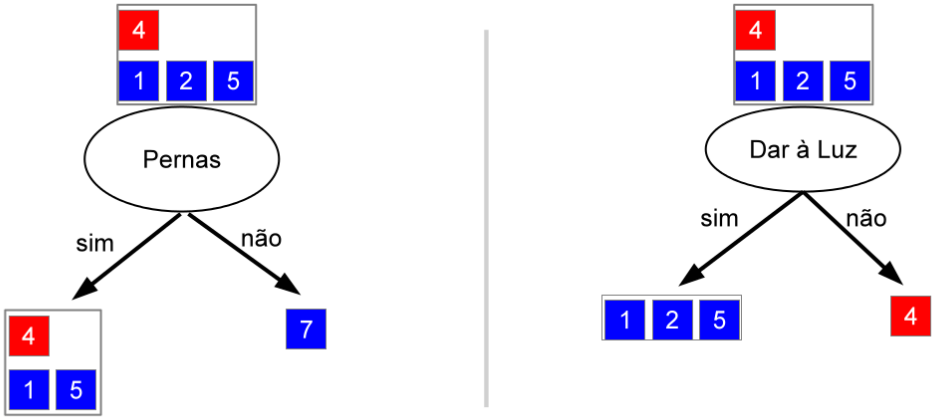

Figura 2.6: Árvore de Decisão exemplo - Expansão de nó.

Uma vantagem do uso da Árvore de Decisão é o seu poder de interpretação. Como exemplo de quão interpretável é a Árvore de Decisão, podemos tirar algumas conclusões apenas observando a árvore formada na Figura 2.7[1]: 1) Animais que tem a temperatura do corpo fria não são mamíferos, 2) Animais que dão a luz e tem a temperatura do corpo quente são mamíferos e 3) Animais que não dão a luz e tem a temperatura do corpo quente não são mamíferos.

2.1.5 Floresta Aleatória

O conceito de Floresta Aleatória surgiu como uma forma de generalização da Árvore de Decisão, tendo a finalidade de minimizar o *overfitting* - ajuste em demasia do modelo aos dados de treino que causa a perda da capacidade de generalização. A ideia é que um conjunto de treinamento seja utilizado para gerar diversos outros conjuntos de mesmo tamanho que serão utilizados para criar árvores de decisão - cada conjunto gerando uma nova árvore. Cada conjunto é criado baseado em sorteios aleatórios do conjunto original e com direito a repetição, dessa forma a variância de cada novo conjunto tende a diminuir em relação ao

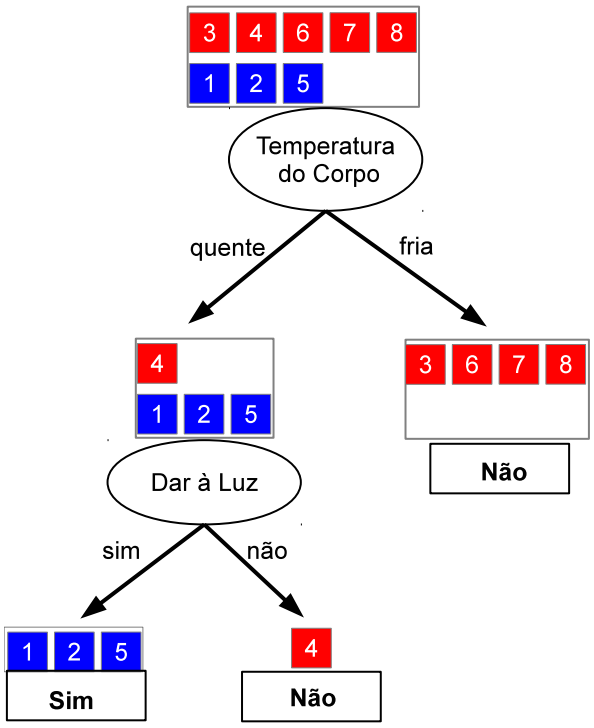

Figura 2.7: Árvore de Decisão exemplo - Completa.

original. Essa técnica é conhecida como *bagging*. A decisão da Floresta, quando necessário classificar uma nova instância, é o voto majoritário das árvores de decisão que a compõe.

2.1.6 Métricas Utilizadas

Diversas são as medidas de desempenho existentes utilizadas para avaliação de modelos de classificação. Dentre todas, devido a larga utilização na literatura, foram escolhidas a *Accuracy*, *F-measure*, *Recall*, *Precision* e o Kappa para avaliarem os experimentos realizados nesta pesquisa. Cada métrica foi escolhida por mensurar uma característica diferente dos resultados. Para introduzir essas medidas, é necessário, primeiramente, introduzir o conceito de matriz de confusão.

Uma matriz de confusão é uma tabela que permite a visualização do desempenho de um algoritmo. Cada coluna/linha representa as instâncias das classes previstas enquanto as linhas/colunas representam as verdadeiras classes das instâncias. A Tabela 2.2 mostra, com um exemplo, como pode ser estruturada uma matriz de confusão no contexto da predição de evasão de alunos. Nela é possível observar que, na classificação realizada, 16 alunos foram classificados como evasores e que 20 foram classificados como não evasores. Dos 16 classi-

ficados como evasores, 10 acertadamente eram evasores enquanto 6 na verdade eram alunos que não viriam a evadir. A mesma lógica pode ser aplicada aos alunos classificados como não evasores. Há de se notar que a diagonal da matriz representa os acertos da classificação.

Dado que identificar instâncias da classe evasor é o objetivo desta pesquisa, é possível afirmar que nesse exemplo houveram:

- 10 verdadeiro positivos (VPC). Evasores corretamente identificados como evasores;
- 12 verdadeiro negativos (VN). Não evasores corretamente identificados como não evasores;
- 8 falso negativos (FN). Não evasores identificados como evasores;
- 6 falso positivos (FP). Evasores identificados como não evasores.

		Classes previstas	
		Evasão	Continue
Classes reais	Evasão	10 (VP)	8 (FN)
	Continue	6 (FP)	12 (VN)

Tabela 2.2: Exemplo de matriz de confusão.

A partir dos valores de VP, VN, FN e FP são definidos a *Accuracy*, *Precision*, *Recall*, *F-measure* e o Kappa. Abaixo são descritos a função de cada métrica.

- *Accuracy*. Tem como função identificar a quantidade de acertos dentro do total de casos examinados e é calculado como $\frac{VP+VN}{VP+VN+FP+FN}$;
- *Precision*. Aponta quantos itens selecionados são relevantes, ou seja, quantos dos classificados como evasores são realmente evasores. Calcula-se $\frac{VP}{VP+FP}$;
- *Recall*. Aponta quantos itens relevantes, do total de itens relevantes, estão sendo identificados, ou seja, quantos evasores, do total de evasores, estão sendo identificados na classificação. Calcula-se $\frac{VP}{VP+FN}$;
- *F-measure*. Medida que sumariza a *precision* e o *recall* em um único valor. Pode ser calculado como a média harmônica da *precision* e *recall*: $2 \times \frac{precision \times recall}{precision + recall}$;

- Kappa. Avalia a qualidade do classificador indicando quanto ele difere do que seria esperado de se obter ao fazer a classificação aleatoriamente. Quanto mais próximo de 0, mais próximo está o resultado obtido com o que se obteria ao acaso e quanto mais próximo de 1, mais distante. É calculado por

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{2.1}$$

onde, $p_0 = \frac{FP+FN}{VP+VN+FP+FN}$ e $p_e = \frac{FP+FN}{VP+VN+FP+FN}$.

2.2 Matrícula e Histórico Acadêmico

A fim de obter uma graduação - primeiro título universitário dado a um indivíduo - na UFCG, cada aluno é submetido a um longo processo que se inicia no processo de seleção, baseado no desempenho do estudante na prova do Exame Nacional do Ensino Médio (ENEM), e é finalizado ao integralizar a carga horária necessária, representada por créditos, para concluir o curso pelo qual optou. Assim que ingressa na Universidade, o aluno é representado por um número de matrícula e um histórico associado, onde constará todo o seu desempenho enquanto usa essa matrícula.

Cada curso é representado por uma grade curricular própria, que idealmente (mas não obrigatoriamente) é seguida pelos alunos, composta por disciplinas, separadas por períodos, que podem ser obrigatórias, complementares ou optativas e exigem uma quantidade de créditos do aluno. Para obter o título de bacharel ou licenciado, o aluno necessariamente deve ser aprovado em todas as disciplinas obrigatórias e complementares, além de integralizar o número de créditos necessários de disciplinas optativas. Disciplinas obrigatórias e complementares são aquelas que estão explícitas no projeto político pedagógico do curso em questão, enquanto as optativas são disciplinas que o aluno está livre para escolher quando cursar e se vai cursar.

Na universidade, a nomenclatura utilizada para definir o período de tempo correspondente da matrícula até o fechamento das notas dos alunos é "Período". Essa palavra é comumente utilizada em dois contextos: 1) para definir o período letivo em vigência na universidade, por exemplo, 2014.1 e 2) contar quanto tempo uma matrícula está ativa na universidade, por exemplo, podemos dizer que um aluno está no segundo período de curso se ele tem matrícula ativa na universidade há dois períodos.

Qualquer aluno ativo da universidade, necessariamente terá disciplinas nas quais ele está matriculado no período correspondente. Ao fim do período, após o fechamento das notas do aluno, ele está sujeito a se encaixar em uma das possíveis situações para cada disciplina: 1) Aprovado, onde sua média final foi maior ou igual a 5.0, 2) Reprovado, onde a média final foi abaixo de 5.0, 3) Reprovado por falta, que significa que ele não assistiu o número de horas-aula suficiente ou 4) Trancamento, que implica que o aluno pediu para ser desvinculado da disciplina. Para obter os créditos da disciplina, o aluno deve necessariamente ser aprovado.

Um exemplo de grade curricular pode ser vista na Figura 2.8[2], onde cada coluna representa um período e apenas disciplinas obrigatórias estão sendo apresentadas. Nesse exemplo foi utilizado a grade do curso de Ciência da Computação.

Figura 2.8: Grade Curricular de Ciência da Computação.

2.2.1 Caso de uso

Podemos caracterizar o caso do modelo na universidade da seguinte forma: Inicialmente existem dados históricos de alunos da UFCG, que serão submetidos a alguns métodos a fim de criar o modelo de predição de evasão. Com o modelo pronto e com o fechamento das notas dos alunos do período corrente, essas notas podem ser submetidas ao modelo a

[2] `http://analytics.lsd.ufcg.edu.br/cursosufcg`

fim de inferir quais alunos estão em risco de evasão (isto é, não voltar para se matricular no próximo período letivo). A identificação correta dos evasores servirá para os administradores (i.e. pró-reitores, coordenadores de educação) entender as causas da evasão a fim de prevenir mais acontecimentos semelhantes. No início de cada novo período, o processo se repetirá com a finalidade de manter o modelo sempre o mais atualizado possível.

2.3 Formulação do Problema

Como já mencionado, a evasão estudantil é abordado como um problema de classificação. Classificações tipicamente considera um conjunto de vetores $X \in \mathbb{R}^m$ de atributos m-dimensionais, um conjunto de classes positivas e negativas $Y = \{+, -\}$ (no nosso caso, "evasor"e "não-evasor") e um conjunto de treinamento na forma D^{train} = $\{(\vec{x}_1,y_1),\ldots,(\vec{x}_n,y_n)\}$ onde $\vec{x} \in X$ é um vetor de atributos e $y_i \in Y$ representa a classe a que $\vec{x}_i$ pertence. A ideia é achar uma função de classificação $\hat{y} : X \to Y$ que minimiza o erro no conjunto teste D^{test} = $\{(\vec{x}_1,y_1),\ldots,(\vec{x}_p,y_p)\}$, que não está disponível durante o treinamento, ou seja, $D^{test} \cap D^{train} = \emptyset$. Mais formalmente, o objetivo é minimizar:

$$err(\hat{y}; D^{test}) = \frac{1}{|D^{test}|} \sum_{(\vec{x},y)\in D^{test}} l(y, \hat{y}(\vec{x})) \tag{2.2}$$

onde $l : Y \times Y \to \mathbb{R}$ é uma função de erro para qualquer instância do teste $(\vec{x}, y) \in D^{test}$ e representa a diferença entre o y verdadeiro e o valor predito $\hat{y}(\vec{x})$. Já que o teste está indisponível, o objetivo é minimizar o erro nos dados de treino, assumindo que tanto o conjunto de treino quanto o de teste são amostras da mesma população. Instanciou-se a classificação para o problema abordado como é descrito nas subseções seguintes.

2.3.1 Classificação por Semestre

Neste cenário, têm-se a intenção de prever se um estudante de um dado período vai evadir do seu curso. As instâncias de treino podem ser definidas por $X = \{X_t : t \in T\}$ onde $T \subseteq \mathbb{N}$ é o conjunto de períodos que o estudante pode ter cursado até o momento (por exemplo, $t = 5$ se o estudante está ativo no curso há cinco períodos). Agora, têm-se disponível um conjunto de treinamento para cada período, por exemplo, $D_t^{train} = \{(\vec{x}_1, y_1)^t, \ldots, (\vec{x}_n, y_n)^t\}$. Sendo

assim, para cada $t \in T$ há um classificador $\hat{y}^t : X^t \rightarrow Y$ que predirá se um estudante no período $t \in T$ irá evadir ou não.

2.3.2 Classificação por Curso/Semestre

Similar à classificação por período mas, nesse contexto, têm-se um conjunto de treinamento diferente (e classificadores diferentes) para cada par curso/período. Neste caso, o conjunto de instâncias é definido por $X = \{X_t^c : t \in T, c \in C\}$ onde C é o conjunto de cursos existentes na universidade. Assim, para cada $c \in C$ e $t \in T$ há um classificador $\hat{y}_c^t : X_t^c \rightarrow Y$ que predirá se um estudante de um curso $c \in C$ no período $t \in T$ irá evadir ou não.

Capítulo 3

Trabalhos Relacionados

Existem diversos trabalhos a respeito da evasão, cada um com uma perspectiva diferente de abordagem ao problema. Abaixo estão brevemente descritas aqueles trabalhos que mais se relacionam a este.

3.1 Predição de evasão

Márquez-Vera et al. [20] investigam a evasão no ensino médio em uma cidade mexicana. Eles usam diversos algoritmos de classificação populares e propõe um algoritmo genético que utiliza *cost-sensitive learning* e técnicas de balanceamento de classes. Utilizam a *accuracy*, verdadeiros positivos e verdadeiros negativos para fazer a avaliação e chegam a atingir valores de 93.4%, 94.0% e 88.3%, respectivamente. Esta dissertação apresenta um modelo baseado na evasão no ensino superior brasileiro, que pode ser considerado um trabalho relacionado mas em um contexto diferente da evasão no ensino médio.

Mustafa et al. [22] hipotetizam que dados demográficos de alunos (como idade, sexo e deficiências) nos cursos de Ciência da Computação e Engenharia da Universidade em Chittagong podem ser bons indicadores de evasão. Os autores usam árvores de decisão e concluem que os atributos mais importantes para prever a evasão são o situação financeira, idade e sexo. Também foi constatado que a *accuracy* das árvores de decisão atingiram até 38.1%. Sendo assim, mesmo com a constatação que existe o impacto da situação financeira, idade e sexo na evasão, utilizar unicamente esses atributos como entrada para o modelo é insuficiente para fazer atingir altos valores de *accuracy*. A pesquisa de Mustafata et al. se relaciona a esta

dissertação com respeito ao contexto de evasão mas se diferencia em diversos pontos como: 1) o tipo de dados utilizados, 2) a quantidade de cursos disponíveis e o tamanho da base de dados e 3) a forte consideração a respeito da divisão do curso por períodos e a utilização de um modelo preditor para cada um deles.

Pal [23] propõe fazer a predição da evasão antes mesmo dos alunos começarem os seus respectivos cursos, isso significa que o aluno já está matriculado na Universidade mas não chegou nem sequer ao período de aulas. Seguindo esse raciocínio, o autor testa quatro algoritmos de classificação usando dados sócio-econômicos e pré-universidade (por exemplo, desempenho dos alunos no ensino médio). A *accuracy* dos modelos varia de 67.7% a 85.7%. Ele conclui que o desempenho dos alunos no ensino médio é o atributo mais discriminativo para os modelos de classificação. O modelo proposto nesta dissertação difere desta abordagem no sentido que foram considerados estudantes de qualquer período da Universidade e, além disso, não houve acesso a dados sócio-econômicos ou pré-universidade do estudante, sendo assim, utilizou-se apenas a registros acadêmicos dos alunos.

Dekker et al. [8] investigam o problema da evasão em cursos de Engenharia Elétrica até o primeiro semestre de estudo. Com esse fim, ele utilizou dados de alunos durante o primeiro período de curso e dados pré-universidade como entrada para oito algoritmos de classificação, além de utilizar *cost-sensitive learning* para lidar com o desbalanceamento de classes. Foi medida a *accuracy* e foram verificadas as métricas de *precision* e *recall*, que variam de acordo com o algoritmo, e chegam a atingir valores de até 71.0% e 88.0%, respectivamente. Como conclusão, foi possível notar que dados pré-universidade não são tão efetivos para prever a evasão e que o desempenho dos alunos nas disciplinas de álgebra linear e cálculo são importantes para a predição. Neste trabalho foi seguida uma abordagem similar, mas foram considerados alunos em qualquer curso e período de um total de 76 cursos.

Balaniuk et al. [3] utilizam dados de 11.495 estudantes de três cursos (Jornalismo, Direito e Psicologia) de uma instituição de ensino superior de Brasília - Brasil. Três algoritmos de classificação foram utilizados para classificar os alunos em "evasor"e "graduado". Para treinar os modelos, foram utilizados como entrada tanto atributos com informações sócio-econômicas quanto acadêmicas dos alunos. Por fim, concluiu-se que é possível identificar estudantes com alto risco de evasão com *accuracy* de até 80.6%. As diferença entre a pesquisa de Balaniuk et al. e esta podem ser listadas: 1) Nesta pesquisa são feitas considerações

a respeito do número de períodos de curso do aluno, 2) uma base de dados com mais registros de alunos e 3) nesta pesquisa foram utilizados apenas registros acadêmicos dos alunos - apesar de não contar com dados sócio-econômicos, os resultados apresentados aqui apresentados se mostram com valores mais altos.

Manhães et al. [18] propõe uma abordagem similar a [3] com a diferença chave que, assim como na abordagem utilizada nesta dissertação, apenas atributos extraídos de registros acadêmicos dos alunos são utilizados. Nessa pesquisa foram utilizados cinco algoritmos de classificação e dados de seis cursos da Universidade Federal do Rio de Janeiro - Brasil: Engenharia Civil, Engenharia Mecânica, Engenharia de Produção, Direito, Física e Farmácia. Essa abordagem os levou a obter uma *accuracy* de pelo menos 87.0% para cada curso e uma taxa de verdadeiros positivos que varia de 66.08% em Farmácia a 88.54% em engenharia civil. Esta pesquisa é similar a de Manhães em termos de abordagem do problema, com a diferença que aqui foi considerado o número de períodos que o aluno está matriculado como um fator importante para a classificação, além de utilizar dados de mais cursos e acrescentar alguns atributos.

Pode-se destacar em alguns dos trabalhos relacionados a utilização de *cost-sensitive learning*. Nesta dissertação, o esse método não foi empregado devido ao aumento excessivo do numero de FP em relação ao pouco aumento do numero de VP, que influenciam diretamente nas métricas *precision* e *recall*. Na prática, a queda da *precision* é mais brusca que o aumento do *recall* e, a partir de um valor significante de *recall* - suficientemente alto para se ter um entendimento das causas que estão levando os alunos a evadir -, ter uma *precision* alta é importante para não aumentar o custo (tempo/recurso) de realizar a pesquisa que leva ao entendimento das causas.

A Tabela 3.1 sumariza e posiciona esta pesquisa em relação às pesquisas encontradas na revisão bibliográfica.

	Contexto	Tipos de dados	Tempo de curso	Número de cursos
Márquez-Vera	Ensino médio	pré-universidade/acadêmicos	1 ano	1
Mustafa	Universidade	demográficos	1 ano	2
Pal	Universidade	socioeconômicos/pré-universidade	Antes do 1º semestre	-
Dekker	Universidade	pré-universidade/acadêmicos	Até o 1º semestre	1
Balaniuk	Universidade	socioeconômicos/acadêmicos	2º semestre em diante	3
Manhães	Universidade	acadêmicos	Qualquer	6
Sales	Universidade	acadêmicos	Qualquer	76

Tabela 3.1: Posicionamento em relação aos trabalhos encontrados na literatura.

3.2 Ações tomadas para diminuição da evasão

Cada uma das pesquisas que serão descritas nesta seção relatam e medem o impacto de ações que foram tomadas com o objetivo de combater a evasão estudantil. É importante notar que tanto nesta dissertação quanto nas pesquisas mencionadas anteriormente, o objetivo comum é identificar quem são os prováveis evasores para, em um segundo momento, tomar alguma atitude, enquanto as pesquisas a seguir seguem o caminho inverso de tomar alguma atitude e logo após medir o impacto delas na evasão.

Zender et al. [30] assumem que a falta de adaptação dos estudantes em seu primeiro período está entre as principais razões que causadoras da evasão. Os novos alunos podem ter problemas para se adaptar a um novo ambiente, uma nova metodologia de ensino ou mesmo para criar novos laços sociais, por exemplo. A solução proposta é criar um jogo pervasivo que os estudantes jogam durante as primeiras semanas de curso. Foi descoberto que o jogo criado facilitou a adaptação dos estudantes ao novo ambiente.

Moretti et al. [21] investigam a retenção de alunos nos cursos de Ciência da Computação com respeito a clareza da transmissão do conteúdo das disciplinas e a metodologia empregada. Para isso, ele responde a três questões que envolvem saber 1) quais são as linguagens de programação, para as disciplinas introdutórias, que ajudam os alunos a entenderem as instruções de forma mais clara, 2) qual o peso que deve ser empregado nas atividades de casa, provas, questionários, projetos e atividades extras para que os alunos entendam de forma mais clara o conteúdo da disciplina, e 3) se os alunos estão mais interessados em disciplinas

que tem o currículo disponível online. Nessa pesquisa foi possível concluir que linguagens interpretadas e um peso igual para projetos e provas se correlacionam com mais clareza nas instruções.

Yadin [29] também pesquisa sobre a evasão de alunos em disciplinas introdutórias de Ciência da Computação. O autor propõe um conjunto de ações (como o uso de linguagens procedurais - Python) para fazer a disciplina mais efetiva no que se diz a transmitir as instruções mais claramente para o aluno. Foi constatado que tais medidas ajudaram a reduzir o número de reprovados na disciplina em até 77%.

Capítulo 4

Preparação dos Dados

Os dados usados nos experimentos foram fornecidos pela Pró-Reitoria de Ensino da UFCG. Os dados consistem em registros acadêmicos dos estudantes de 76 cursos da UFCG no período compreendido de 2002 a 2014, o que representa 12.5 anos de dados (ou 25 períodos). A base contempla registros de 32.342 estudantes matriculados, dentre os quais 12.560 evadem (considera-se evasão qualquer forma de evasão, com exceção de evasão por conclusão de curso), sendo assim, 61.16% chegam a concluir seus cursos e obtêm um diploma de ensino superior. A Tabela 4.1 cita e descreve os campos presentes na base.

Coluna	Descrição
Id de matrícula	Identificador único do estudante
CPF	Cadastro de Pessoa Física
Id de curso	Identificador único do curso
Id de período	Identificador do período acadêmico (por exemplo, 2014.1 significa "primeiro período de 2014")
Período de entrada	Período que o estudante começou o curso
Último período	Último período que o estudante esteve matriculado em algum curso
Id da disciplina	Identificador de uma disciplina
Créditos	Peso de cada disciplina do curso
Nota	Nota do estudante na disciplina num intervalo fechado de 0 a 10
Situação	Situação do estudante na disciplina (Aprovado, Reprovado por nota, Reprovado por falta ou Trancado)
Tipo de disciplina	Obrigatória, Complementar ou Optativa
Código de evasão	Código que identifica o tipo de evasão (por exemplo, evasão por abandono)
Créditos do curso	Total de créditos necessário para completar o curso (baseado no número de horas-aula)

Tabela 4.1: Descrição dos dados.

4.1 Pré-processamento

Antes de analisar quais atributos devem ser usadas como entradas para os modelos de classificação é necessário preprocessar os dados descritos na Tabela 4.1. Os seguintes casos foram tratados:

- Para o estudante universitário, as disciplinas complementares e obrigatórias são tratadas da mesma maneira, portanto decidiu-se transformar as disciplinas complementares em obrigatórias;

- Código de evasão. Existem diversos códigos usados na UFCG para justificar a evasão de um aluno, como evasão por abandono, por transferência ou mesmo por morte (ver Tabela 4.2). Cada código existente teve que ser mapeado para "evasor"ou "não-evasor". A fim de rotular as instâncias de treino, nós checamos para cada estudante no semestre t se ele ainda estava matriculado no semestre $t + 1$ e, caso estivesse, a instância recebia código 0 (e 1 caso contrário). Por exemplo, o estudante que evadiu de seu curso no terceiro semestre irá receber código 0 no primeiro e no segundo período e código 1 no terceiro. Estudantes que concluíram seus cursos recebem código 0 em todos os períodos;

- Cálculo de períodos. Foi calculado o período corrente de cada estudante usando o Id do período e o Período de entrada apenas contando o número de períodos que se passou desde sua entrada até o Id do período. Esse atributo foi chamado de N.PERIODOS.MATRICULADO;

- Reentrada nos cursos. Em muitas Universidades públicas brasileiras, incluindo a UFCG, é possível para os estudantes reentrarem no curso que eles estão matriculados através do Exame Nacional do Ensino Médio (ENEM). Os alunos que optam por realizar a reentrada em seus cursos recebem uma nova matrícula e um novo histórico acadêmico que conterá apenas as disciplinas que ele foi aprovado enquanto usava a matrícula antiga. Essa situação foi tratada primeiramente identificando esses estudantes, através do CPF e Id de curso, e criando um novo Id de estudante que engloba todos os registros distribuídos por todos os possíveis Ids associados a ele. Esta ação elimina os casos que chamamos de falsos calouros e falsas evasões.

Tabela 4.2: Códigos de evasão e explicação de cada código na UFCG.

Código	**Evasão**
0	Aluno regularmente matriculado
1	Graduado
2	Transferência para outra Instituição de Ensino Superior
3	Falecimento
4	Abandono de curso
5	Cancelamento de matrícula
6	Cancelamento para mudança de curso
7	Cancelamento por decisão judicial
8	Cancelamento por solicitação do aluno
9	Suspensão temporária
10	Curso concluído - não colou grau
11	Cancelamento por não cumprimento da PEC
12	Reentrada no curso (novo vestibular)
13	Cumprimento convênio
14	Novo regimento
15	Não comparecimento a cadastro
16	Remanejado de curso
17	Não compareceu ao remanejamento
18	Não compareceu à matrícula - Alunos ingressantes
19	Término de intercâmbio
20	Graduando por decisão judicial
21	Matrícula cancelada por reprovação por falta
22	Matrícula cancelada por reprovações na mesma disciplina
23	Matrícula suspensa - Débito na biblioteca
50	Aguardando cadastramento

4.2 Atributos analisados

Nesta seção é apresentado o processo de escolha dos atributos utilizados para detectar evasão. Para isso, foram considerados todos os atributos introduzidos por Manhães et al. [18] e alguns outros criados por nós.

Os atributos introduzidos por Manhães et al. [18] foram escolhidos para serem usados devido a pesquisa realizada por eles ser uma das mais recentes desenvolvidas na área e seguir uma metodologia muito próxima à empregada neste trabalho. A Tabela 4.3 mostra a lista completa de atributos utilizados no referido trabalho. Os atributos introduzidos nesta pesquisa - ou que, pelo menos, não foram encontrados em outros estudos - estão marcados na tabela por um asterisco e são explicados a seguir:

- N.PERIODOS.MATRICULADO: captura quanto tempo, em períodos, o estudante está ativo em um curso. Em algumas pesquisas esse atributo já foi considerado como um fator mas não foi utilizado da mesma forma como nós o utilizamos aqui. Em estudos anteriores ele foi utilizado para definir o escopo dos registros de alunos que seriam utilizados, como por exemplo, classificar apenas alunos do primeiro período ou primeiro ano de curso. Nesta pesquisa, o atributo foi utilizado desde o inicio até o fim do curso.

- N.CREDITOS.TOTAL: mostra quantos créditos o aluno se matriculou nesse período. É esperado que quanto mais créditos ele tenha se matriculado, maior seja o esforço que ele vai ter que empregar no período. Esse atributo é calculado somando os créditos das disciplinas que o aluno se matriculou no período. A lógica de N.CREDITOS.OBRIGATORIA, N.CREDITOS.OPTATIVA e N.CREDITOS.APROV são similares a de N.CREDITOS.TOTAL, com a diferença que esses representam os créditos para disciplinas obrigatórias, optativas e aprovadas no período corrente, respectivamente;

- N.CREDITOS.DIF: indica quanto o aluno está se desviando de sua turma. Os alunos de cada turma tendem a seguir o currículo de seus cursos de forma igual (para cada período, espera-se que os alunos de uma mesma turma tendem a se matricular nas mesmas disciplinas), sendo assim, aqueles alunos que se desviam da turma

podem estar dando um sinal que estão prestes a evadir. O atributo é calculado com base na diferença do número de créditos matriculados de um aluno e a moda de sua turma. N.CREDITOS.OBRIGATORIA.DIF, N.CREDITOS.OPTATIVA.DIF e N.CREDITOS.APROV.DIF foram criados com o mesmo propósito e apenas mudam com o fato de serem calculados usando apenas disciplinas obrigatórias, optativas e aprovadas, respectivamente;

- PORCENTAGEM.CURSO.COMPLETO: indica quanto do curso já foi integralizado pelo aluno. O atributo foi criado com a intenção de capturar quanto esforço o aluno já empregou no curso e, portanto, seria desperdiçado ao abandoná-lo. Ele é calculado dividindo o número total de créditos que o aluno já foi aprovado pelo total de créditos necessário para completar o curso;

Após escolher quais atributos seriam usados, foi aplicado com 95% de confiança, para cada período, o teste estatístico de Wilcoxon-Mann-Whitney [19] considerando duas amostras do conjunto de treinamento: as instâncias dos atributos associados a classe 0 e 1, respectivamente. Se o resultado do teste indicar que as características das diferentes amostras derivam da mesma população, então é possível concluir que esses atributos não são discriminativos e devem ser descartados. Caso contrário, pode-se entender que o atributo é um bom discriminador de evasores e não-evasores para aquele período. Para completar o conjunto de atributos, foi adicionado o STATUS.SEM[1]. A Tabela 4.4 retrata para quais períodos cada atributo foi considerado discriminativo. Por exemplo, é possível enxergar que para alunos cursando o primeiro período, todos os atributos, com exceção de N.CREDITOS.OBRIGATORIA, N.RPN e PROP.N.RPN, não são diretamente descartáveis e podem ser utilizados para identificar possíveis evasores. Também é possível notar que N.CREDITOS.OPTATIVA.DIF é um atributo discriminativo apenas nos primeiros cinco períodos de curso. Nesta pesquisa, foram utilizados dados de estudantes matriculados apenas até o décimo período de curso graças ao baixo número de evasões nos períodos posteriores.

[1]Ele não foi submetido ao teste de Wilcoxon-Mann-Whitney porque não é um atributo numérico

Tipo/Valor	Atributo	Descrição
Id (categórico)	Id de matrícula	Identificador único do estudante
Id (categórico)	Id de curso	Identificador único do curso
{1 to n} (numérico)	Id de período	Identificador do período acadêmico
{0 or 1} (categórico)	Código de evasão	Código que identifica o tipo de evasão
{0 to n} (numérico)	N.(APROV, RPN, RPF, REPR, TRAN)	Respectivamente, número de disciplinas em que o aluno foi aprovado, reprovou por nota, reprovou por falta, foi reprovado (por nota e falta) e trancadas no período
{0 to 10} (numérico)	MEDIA.APROV	Média de notas das disciplinas aprovadas no período
{0 or 1} (categórico)	STATUS.SEM	O status do período (0 se o estudante reprovou todas disciplinas, 1 caso contrário)
{0 to 10} (numérico)	MEDIA.SEM	Média de notas das disciplinas que o aluno cursou no período
{0 to 10} (numérico)	CRE	Média harmônica composta pela notas do aluno e créditos das disciplinas já cursadas
{0 to 10} (numérico)	N.PERIODOS.MATRICULADO*	número de períodos cursados até então
{0 to 1} (numérico)	PORCENTAGEM.CURSO.COMPLETO*	número de créditos aprovados dividido pelo número total de créditos necessários para concluir o curso
{0 to n} (numérico)	N.DISC.(OBRIGATORIA, OPTATIVA)	número de disciplinas obrigatórias e optativas e total de disciplinas cursadas no período
{0 to n} (numérico)	N.CREDITOS.(OBRIGATORIA, OPTATIVA, APPR)*	número de créditos de disciplinas obrigatórias e optativas e número total de créditos em disciplinas que o aluno foi aprovado no período
{0 to 1} (numérico)	PROP.N.(RPF, RPN, APROV, TRAN)	Proporção. número de N.RPF, N.RPN, N.APROV e N.TRAN dividido pelo total de disciplinas matriculadas no período
{0 to n} (numérico)	N.CREDITOS.DIF*	Diferença entre o número de créditos que o estudante está matriculado e a moda de sua turma
{0 to n} (numérico)	N.CREDITOS.(OBRIGATORIA, OPTATIVA, APPR).DIF*	Diferença entre o número de créditos de disciplinas obrigatórias, optativas e aprovadas no período e a moda de sua turma

Tabela 4.3: Descrição dos atributos usados.

Atributos/N.PERIODOS.MATRICULADO	1	2	3	4	5	6	7	8	9	10
N.CREDITOS.OPTATIVA.DIF	x	x	x	x	x					
N.RPN			x	x	x	x	x	x	x	x
N.DISC.OPTATIVA	x	x	x			x	x	x	x	x
N.CREDITOS.OPTATIVA	x	x	x			x	x	x	x	x
N.CREDITOS.OBRIGATORIA		x	x	x	x	x	x	x		x
PROP.N.RPN	x	x	x	x	x	x	x	x		
N.DISC.OBRIGATORIA	x	x	x	x	x	x	x	x		x
N.DISC	x	x	x	x	x	x	x	x	x	
N.DISC.DIF	x	x	x	x	x	x	x	x	x	
N.CREDITOS.DIF	x	x	x	x	x	x	x	x	x	
PROP.N.RPN.DIF	x	x	x	x	x	x	x	x	x	
N.CREDITOS	x	x	x	x	x	x	x	x	x	
PORCENTAGEM.CURSO.COMPLETO	x	x	x	x	x	x	x	x	x	x
N.APROV	x	x	x	x	x	x	x	x	x	x
N.RPF	x	x	x	x	x	x	x	x	x	x
N.TRAN	x	x	x	x	x	x	x	x	x	x
MEDIA.APROV	x	x	x	x	x	x	x	x	x	x
PROP.N.RPF	x	x	x	x	x	x	x	x	x	x
PROP.N.APROV	x	x	x	x	x	x	x	x	x	x
PROP.N.TRAN	x	x	x	x	x	x	x	x	x	x
MEDIA.SEM	x	x	x	x	x	x	x	x	x	x
CRE	x	x	x	x	x	x	x	x	x	x
N.CREDITOS.APROV	x	x	x	x	x	x	x	x	x	x
N.CREDITOS.APROV.DIF	x	x	x	x	x	x	x	x	x	x

Tabela 4.4: Atributos discriminativos por período.

Capítulo 5

Modelos de Classificação

Com a finalidade de criar os modelos, levou-se em consideração duas premissas básicas. A primeira assume que as razões (e atributos que capturam o sentido das razões) que levam os alunos a evadirem são as mesmas (ou variam pouco) independente de seus cursos enquanto a segunda assume que as razões variam para cada curso e, portanto, o curso que o aluno está matriculado pode influenciar positivamente no desempenho do classificador.

Para as duas premissas, foram propostos e avaliados diversos modelos e depois comparados os que tiveram melhor desempenho em cada caso para checar qual premissa descreve os dados com mais fidelidade.

5.1 Experimento

Para buscar a melhor configuração de cada um dos dois modelos, foram criados diferentes Florestas Aleatórias [14] considerando os seguintes fatores: seleção de atributos, técnicas de balanceamento de classes e número de árvores na floresta. Por exemplo, uma configuração de modelo descrita como FALSE-FALSE-10 é um modelo criado sem fazer nenhuma seleção de atributos ou balanceamento de classes e com 10 árvores; Seguindo o mesmo raciocínio, gain.ratio-Oversample-100 é um modelo criado utilizando seleção de atributos (o algoritmo *Gain Ratio*), balanceado as classes baseado no OSS e no random oversample e com 100 árvores na floresta aleatória. A Tabela 5.1 mostra os valores que cada fator pode assumir.

Foram usados todos os dados a partir de 2002.1 até 2012.2 para realizar o treinamento

Fator	Valores			
Mutual Information	*Information Gain*	*Gain Ratio*	*Symmetrical Uncertainty*	False
Balanceamento de Classes	Undersample (OSS + Undersample)	Oversample (OSS + Oversample)	FALSE	-
Número de árvores	1	10	50	100

Tabela 5.1: Fatores e possíveis valores assumidos.

de cada Floresta Aleatória[1], os dados de 2013.1 para a validação e os de 2013.2 para teste. Após a escolha dos melhores modelos, na fase de validação, os dados compreendentes ao período de 2002.1 até 2013.1 são submetidos novamente ao treinamento e testados com os dados de 2013.2. Assim, é simulado o cenário onde há estudantes matriculados no período de 2013.2 e tem-se o objetivo de prever quais são os que tem risco de evasão para que, por exemplo, a administração da Universidade que venha a usar esse serviço possa agir antes que isso aconteça. Os resultados da validação são apresentados na Figura 5.2 e Figura 5.5 e os resultados do teste na Tabela 5.5 e Figura 5.6.

Cada configuração de modelo foi executada 20 vezes a fim de calcular um intervalo de confiança. Isso foi necessário graças a aleatoriedade da Floresta Aleatória, ou seja, cada árvore da floresta usa uma amostra dos dados de treino via *bagging*, e das técnicas de desbalanceamento que podem ter resultados diferentes em cada execução. Para toda configuração, o *F-measure* foi calculado a partir da média dos *F-measures* de cada período - no caso do Modelo Global - ou de cada curso/período - no caso do Modelo Específico.

Abaixo nós descrevemos esses processos em detalhes.

5.2 Modelo Global por Semestre

Este é o modelo correspondente à primeira premissa e descrito na Subseção 2.3.1, onde a ideia é ter um classificador para cada período.

5.2.1 Seleção de Atributos

Com o objetivo de investigar se a importância dos atributos varia com o passar dos períodos, foram calculados os valores de *Information Gain*, *Gain Ratio* e *Symmetrical Uncertainty*

[1] O pacote randomForest da ferramenta R foi utilizado para criar a Floresta Aleatória.

para cada atributo da Tabela 4.4. Também foi criado um modelo para cada algoritmo de *Mutual Information* usando os 3 atributos mais importantes indicados por eles. A Tabela 5.2 mostra os atributos mais importantes por período para cada algoritmo.

Tabela 5.2: Atributos mais importantes por período.

Information Gain										
Atributos	1	2	3	4	5	6	7	8	9	10
N.APROV	0.244									
N.CREDITOS.APROV	0.324	0.212	0.197	0.239	0.258	0.266	0.285	0.276	0.317	**0.376**
PORCENTAGEM.CURSO.COMPLETO	**0.985**	**0.752**	**0.580**	**0.575**	**0.467**	**0.476**	**0.392**	**0.341**	**0.434**	0.205
N.CREDITOS.OBRIGATORIA		0.171				0.207	0.242	0.265		0.238
N.CREDITOS			0.183	0.176	0.221				0.257	

Gain Ratio										
Atributos	1	2	3	4	5	6	7	8	9	10
PORCENTAGEM.CURSO.COMPLETO	0.199	0.160	0.139	**0.134**	**0.126**	**0.123**	**0.112**	**0.110**	**0.126**	
PROP.N.RPF	0.182									
STATUS.SEM	**0.392**	**0.171**	**0.141**	0.102	0.095	0.100				
N.DISC.OPTATIVA		0.125								
N.CREDITOS.OPTATIVA			0.115							0.118
N.CREDITOS.APROV				0.077	0.089	0.094	0.099	0.106	0.116	**0.142**
N.CREDITOS.OBRIGATORIA							0.088			0.123
N.CREDITOS								0.101		
N.DISC									0.097	

Symmetrical Uncertainty										
Atributos	1	2	3	4	5	6	7	8	9	10
N.CREDITOS.APROV	0.139	0.091	0.088	0.102	0.115	0.119	0.127	0.131	0.146	**0.177**
PORCENTAGEM.CURSO.COMPLETO	**0.300**	**0.238**	**0.201**	**0.195**	**0.175**	**0.174**	**0.153**	**0.145**	**0.172**	0.120
STATUS.SEM	0.145									
N.CREDITOS.OPTATIVA		0.090	0.096							
N.CREDITOS.TOTAL				0.083	0.102			0.123	0.119	
N.CREDITOS.OBRIGATORIA						0.095	0.111			0.134

Na Tabela 5.2 é possível observar que a importância dos atributos de fato variam dependendo de qual algoritmo de seleção de atributos foi usado. Outras conclusões interessantes que podem ser vistas estão listadas abaixo:

- PORCENTAGEM.CURSO.COMPLETO e N.CREDITOS.APROV são considerados importantes para a maioria dos períodos;

- A importância de N.CREDITOS.APROV tende a crescer com o passar dos períodos. Isso é algo que faz sentido, ao pensar que com o passar dos períodos, mais esforço, que pode ser medido em créditos de disciplina, foram cursados e aprovados pelos alunos, o que implica que quanto mais perto do fim do curso os alunos estão, mais prejudicial se torna um abandono;

- STATUS.SEM aparenta ser um bom atributo para ser usado como entrada em um modelo do primeiro período. Nos períodos iniciais, e principalmente no primeiro período, o número de evasões tende a ser maior que nos períodos subsequentes. Muito disso se deve ao fato de novos alunos desistirem de seus cursos e acabarem sendo reprovados em todas, ou quase todas, as disciplinas. O STATUS.SEM é um atributo criado para identificar casos de alunos reprovados em todas as disciplinas e, portanto, potenciais evasores.

5.2.2 Balanceamento de Classes

Como retratado na Figura 5.1, a porcentagem de evasões em cada semestre é bem menor que a porcentagem de estudantes que continuam em seus respectivos cursos. Isso representa um problema conhecido na literatura como desbalanceamento de classes, um cenário onde o classificador pode ser enviesado a classificar todas as instâncias de teste com o rótulo da classe majoritária [13].

Para lidar com esse problema, foi aplicado o algoritmo One-Sided Selection [17], que é um método de *undersampling* resultante da aplicação dos métodos Tomek Link [28] e Condensed Nearest Neighbor [12]. A ideia é usar o OSS para remover ruídos, instâncias próximas à fronteira de decisão e instâncias muito distantes da fronteira de decisão do conjunto de treinamento e em um segundo momento balancear a proporção de instâncias de cada classe usando o *random undersample/oversample*.

5.2.3 Seleção de Modelos

A Figura 5.2 retrata os intervalos de confiança dos *F-measures* de cada configuração de Florestas Aleatórias. Para ter uma melhor visualização dos resultados, estão sendo exibidas apenas as configurações que obtiveram os melhores resultados.

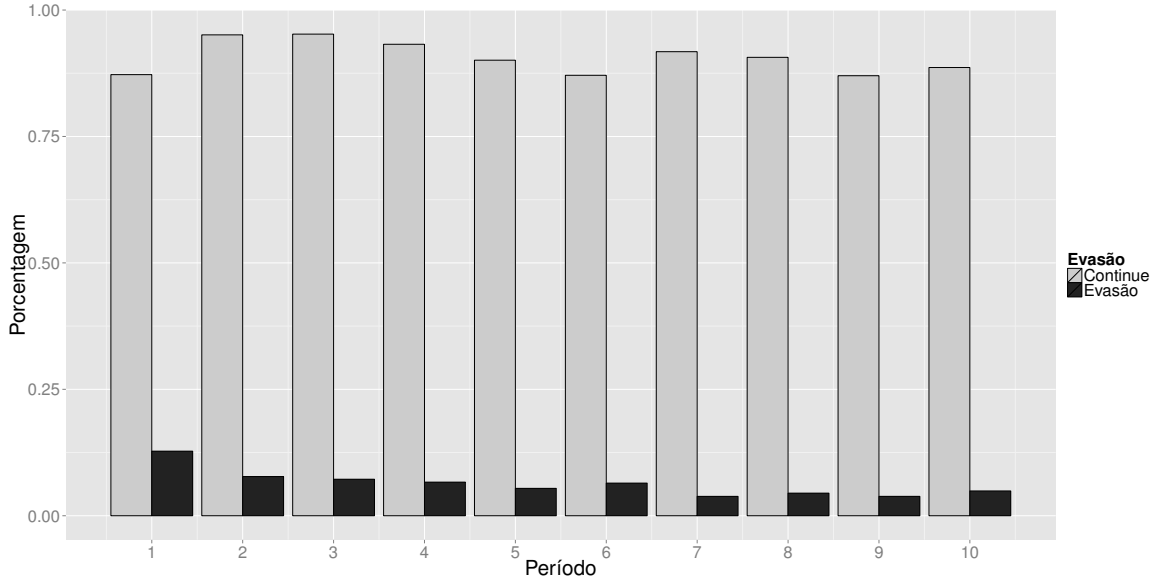

Figura 5.1: Evasão por período.

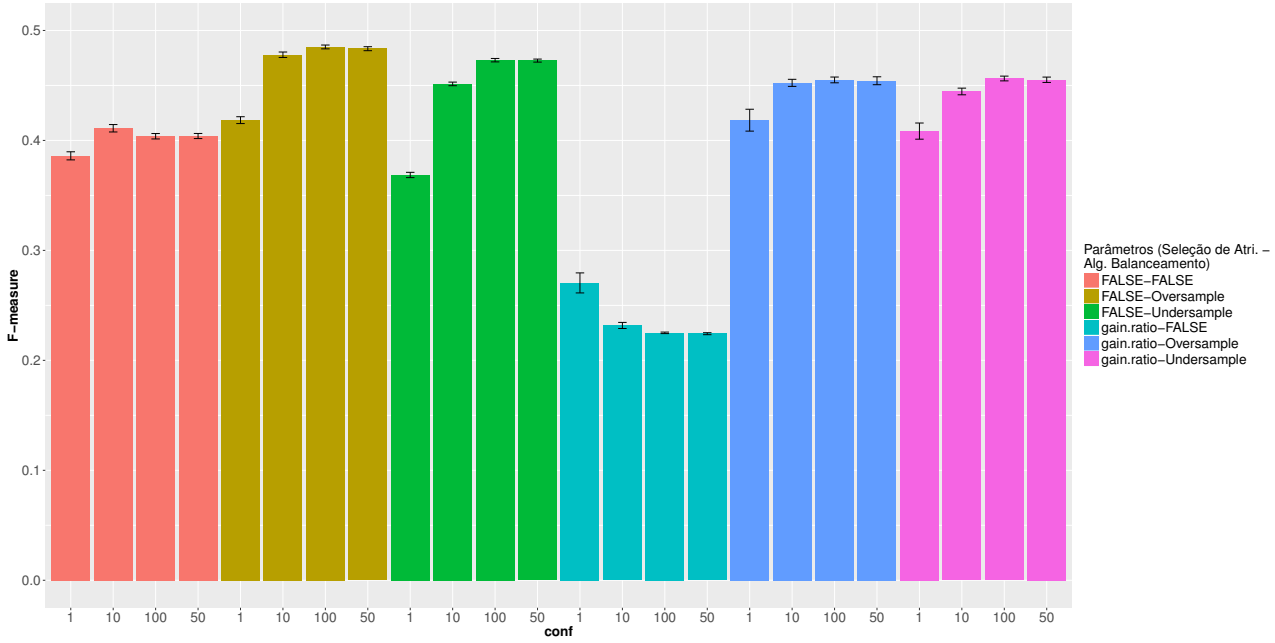

Figura 5.2: *F-measures* das configurações dos Modelos Globais

A partir da Figura 5.2 é possível notar que utilizando algum método de seleção de atributos, os melhores resultados obtidos são aqueles obtidos a partir do algoritmo *Gain Ratio*. A diferença mais notável entre a seleção de atributos usando o *Gain Ratio* e os outros algo-

ritmos é a importância dada ao STATUS.SEM. Importância que faz sentido, já que o fato de um aluno ser reprovado em todas as disciplinas no período pode ser considerado um sinal de evasão. Também é possível notar que o modelo que atingiu o maior valor de *F-measure* foi o representado pela configuração FALSE-Oversample-100, que significa que não houve seleção de atributos, houve o balanceamento das classes utilizando o OSS seguido do *Random Oversample* e a Floresta Aleatória contém 100 árvores de decisão.

Além disso, algumas outras conclusões podem ser tiradas dos resultados:

- Apesar de a seleção de atributos diminuir o valor de *F-measure* em quase 10%, um modelo mais simples ainda pode ser útil para facilitar o entendimento e a interpretação dos atributos utilizados;

- Os melhores resultados são alcançados quando não acontece a seleção de atributos. Também é notório que fazer o balanceamento surtiu o efeito esperado. Principalmente nos casos que houve seleção de atributos;

- Os resultados de *F-measure* tendem a crescer a medida que o número de árvores da Floresta Aleatória foi incrementado. Apesar disso, há de se notar que a diferença dos resultados para 50 e 100 árvores muitas vezes já não apresentam diferença estatística, insinuando que continuar a incrementar o número de árvores de decisão não vai aumentar os valores de *F-measure*;

5.3 Modelo Específico por Curso/Período

Este é o modelo correspondente à segunda premissa descrita na Subseção 2.3.2, onde a ideia é ter um classificador para cada par curso/período.

5.3.1 Seleção de Atributos

De forma semelhante ao já mencionado na descrição do Modelo Global, o objetivo da seleção de atributos é investigar se a importância dos atributos varia com o passar dos períodos de cada curso. Para isso, foram calculados o *Information Gain*, *Gain Ratio* e *Symmetrical Uncertainty* para cada par curso/período e utilizado os 3 valores mais importantes para criar

um modelo classificador daquele par. Como exemplo dos resultados, estão sendo apresentados, na Tabela 5.3, os valores de importância dos atributos para os cursos de Ciência da Computação e Farmácia. Assim, é possível observar que:

- A importância dos atributos varia ao longo dos períodos;
- A importância dos atributos varia de acordo com o algoritmo usado;
- A importância dos atributos varia de acordo com o curso;
- Alguns períodos de alguns cursos apresentam poucos, ou mesmo zero, evasões. Para esses pares curso/período, não será possível utilizar um classificador.

Tabela 5.3: Atributos mais importantes por período para Ciência da Computação e Farmácia.

Ciência da Computação - *Gain Ratio*										
Atributos	1	2	3	4	5	6	7	8	9	10
N.RPF	0.325									
PROP.N.RPF	0.304	0.282	0.182						0.219	
STATUS.SEM	0.309		0.166	0.179	0.157	0.197				
N.TRAN		0.381				0.202				
PROP.N.TRAN		0.410				0.216				
PORCENTAGEM.CURSO.COMPLETO			0.161							
N.CREDITOS.OPTATIVA.DIF				0.248						
PROP.N.REPR				0.215				0.213	0.206	
MEDIA.APROV					0.157					
N.CREDITOS.DIF					0.217					
N.CREDITOS.APROV.DIF							0.139		0.264	0.230
N.DISC.DIF							0.087			
MEDIA.SEM								0.201		
PROP.N.APROV								0.234		

Ciência da Computação - *Information Gain*										
Atributos	1	2	3	4	5	6	7	8	9	10
PORCENTAGEM.CURSO.COMPLETO	0.353	0.319	0.221			0.087				
PROP.N.APROV	0.180	0.154	0.167	0.082				0.047		
PROP.N.REPR	0.182	0.132	0.148						0.046	
N.APROV				0.090				0.043		
N.CREDITOS				0.067						
CRE					0.069					
N.CREDITOS.APROV					0.065	0.087				
N.CREDITOS.APROV.DIF					0.078		0.052		0.093	0.095
PROP.N.TRAN						0.081				
N.DISC.DIF							0.055			
MEDIA.SEM								0.049		
PROP.N.RPF									0.061	

Farmácia - *Information Gain*										
Atributos	1	2	3	4	5	6	7	8	9	10
N.CREDITOS	0.498	0.156	0.171			0.152			0.364	
N.CREDITOS.APROV	0.426	0.226	0.172	0.200	0.201	0.202	0.205	0.180		
N.DISC.OBRIGATORIA	0.499									
PORCENTAGEM.CURSO.COMPLETO		0.226	0.172	0.200	0.201	0.202	0.205	0.180		
N.CREDITOS.OBRIGATORIA							0.140	0.108		
N.APROV									0.350	
N.DISC.TOTAL									0.349	

Para cada algoritmo de *Mutual Information*, foi criado um classificador para cada par curso/semestre usando as três características mais importantes.

5.3.2 Balanceamento de Classes

Como anteriormente mencionado, devido ao número de evasões no escopo de um período de um curso ser baixo, os dados contém diversos pares curso/período com poucas, ou mesmo nenhuma, evasão. Este alto nível de desbalanceamento pode enviesar severamente o classificador mesmo usando técnicas de balanceamento. Assim, neste cenário, foi criado um classificador apenas para aqueles pares onde pelo menos 10 evasões aconteceram. É esperado, principalmente, que nos últimos períodos de curso menos evasões aconteçam. Logo, quanto maior o número de evasões, menor o número de pares curso/período disponíveis para participar do experimento, como retratado na Figura 5.3 e na Tabela 5.4. Assim, valor 10 foi empiricamente escolhido baseado no número de pares curso/período que contêm essa quantidade de evasões. Devido a esse fato, o número de pares utilizados para criar o Modelo Específico decresceu de 760 (10 períodos vezes 76 cursos) para 329 (pares com 10 ou mais evasões). Ao criar o Modelo Global não foi necessário fazer esse tipo de consideração devido não haver períodos isolados com baixo número de evasões.

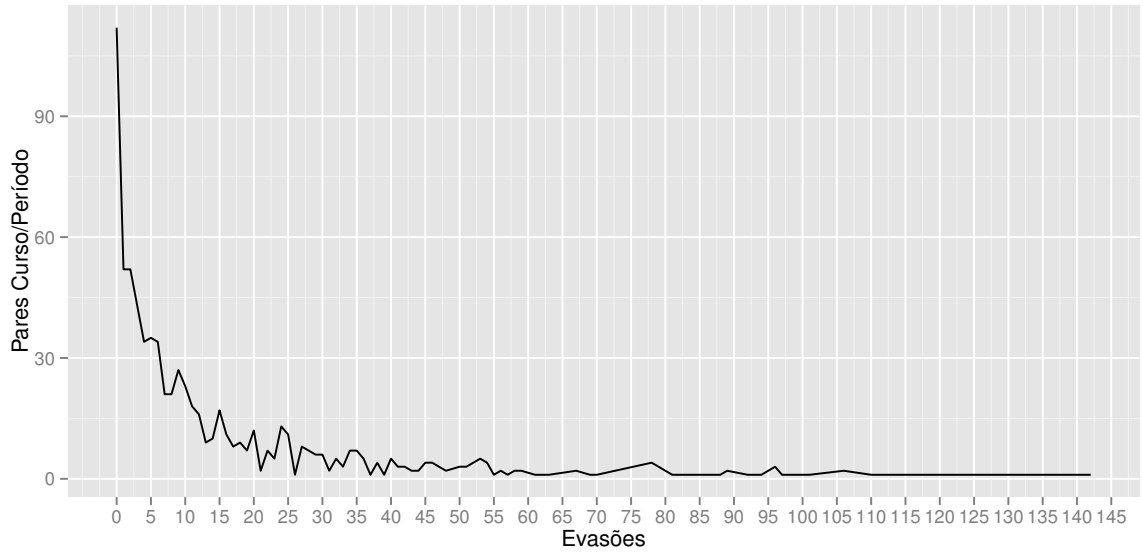

Figura 5.3: Número de pares curso/período por quantidade de evasões.

Nº evasores	0	1	2	3	4	5	6	7	8	9	10	11	12	13+
Nº Cursos/Período	112	52	52	43	34	35	34	21	21	27	23	18	16	272

Tabela 5.4: Número de pares curso/período por quantidade de evasões.

A título de exemplificação, é mostrado na Figura 5.4 o número absoluto (e porcentagens) de evasões por período nos curso de Farmácia e Ciência da Computação.

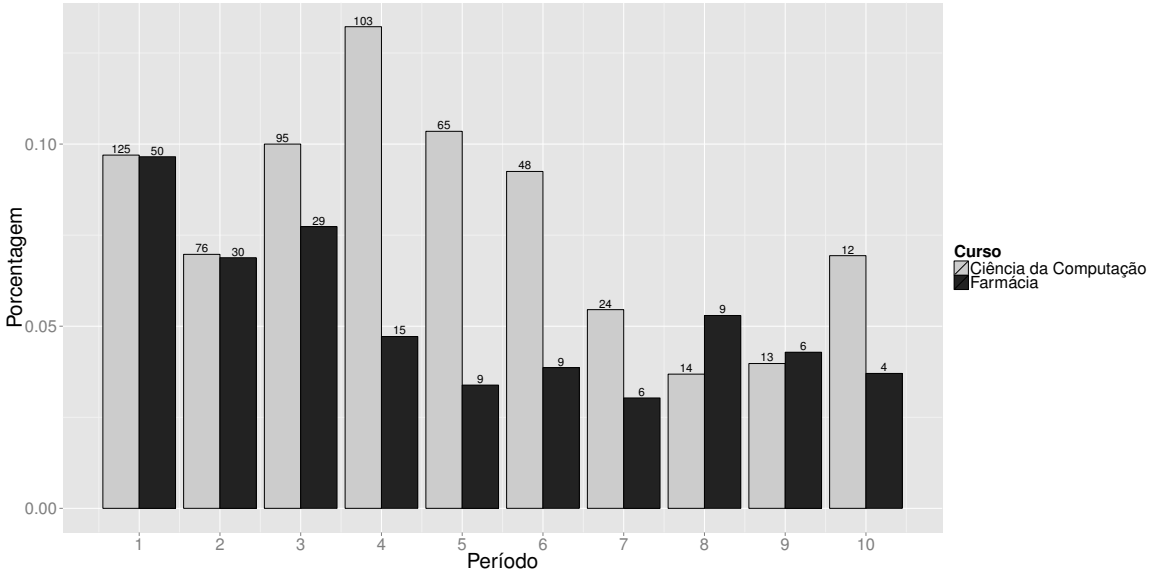

Figura 5.4: Evasão por período para cursos.

De forma similar ao Modelo Global, foi aplicado o algoritmo OSS seguido por *random undersampling*/*oversampling*.

5.3.3 Seleção de Modelos

A mesma abordagem utilizada na criação do Modelo Global foi utilizada aqui e é possível perceber que algumas configurações de Modelo Específico não tem diferença estatística e podem ser consideradas como os que obtiveram o maior valor de *F-measure*. O modelo com configuração FALSE-Oversample-100 será utilizado de agora em diante por ser o que mesmo escolhido no Modelo Global. A Figura 5.5 descreve os resultados e logo abaixo estão descritas algumas conclusões que puderam ser observadas:

- Em termos de *F-measure*, os Modelos Específicos que fizeram seleção de atributos atingiram valores bem mais baixos que os Modelos Globais. Uma possível razão pode ser atribuída ao número de instâncias extraídas de cada par curso/período ser bem menor que considerando apenas o período do Modelo Global;

- Balancear o conjunto de treinamento não surtiu o efeito esperado de aumentar o valor de *F-measure*. Esse efeito pode ser notado tanto ao comparar modelos utilizando seleção quanto comparando modelos que não utilizaram seleção de atributos e é especulado que esse não crescimento da métrica pode ser atribuído a baixa quantidade de evasores nos pares curso/período;

- Utilizar seleção de atributos causa uma piora significativa dos resultados. Assim como os casos anteriores, a baixa quantidade de alunos em cada par curso/período pode ser o causador desse efeito. Já que existe pouca informação para ser utilizada no treinamento do modelo e que o objetivo da seleção de atributos é selecionar apenas aqueles considerados mais importantes, então alguma informação é desperdiçada e o modelo tenderá a ter um *F-measure* mais baixo.

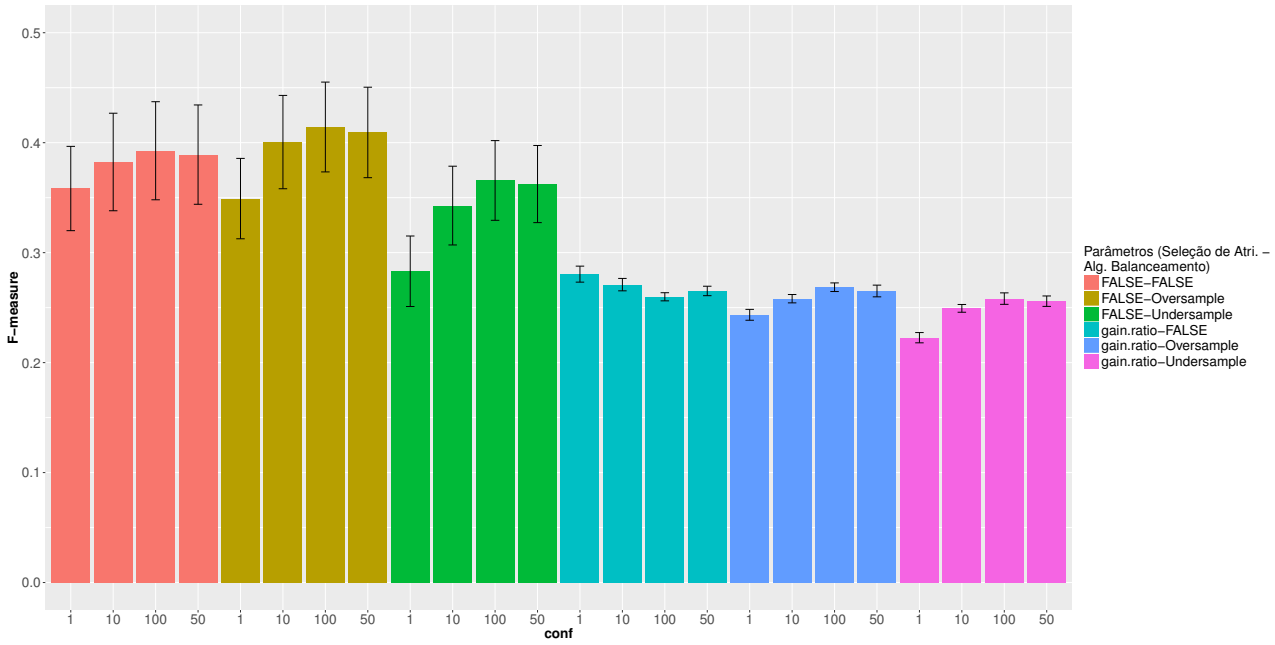

Figura 5.5: *F-measures* das configurações dos Modelos Especificos.

5.4 Modelo Global vs Modelo Específico

Para comparar os modelos Específico e Global, nós usamos as métricas *Precision*, *Recall*, *F-measure*, *Accuracy* e o Kappa. A Tabela 5.5 mostra os resultados para as configurações que atingiram os melhores resultados em cada caso e a Figura 5.6 ilustra os valores de *F-measure*, *Recall* e *Precision* de cada modelo, a fim de facilitar a visualização das diferenças entre eles. Observando os resultados do Modelo Global e do Modelo Específico é possível notar que:

- O Modelo Global atinge os maiores valores para cada métrica. O Modelo Específico, em poucos casos conseguiu ter um desempenho, para qualquer métrica, acima do Modelo Global;

- O *F-measure* tende a decrescer ao longo que os períodos vão se passando para os dois modelos. Isso é uma consequência do fato de que quanto mais se passam os períodos, menor é o número de evasões observadas, o que dificulta a aprendizagem dos modelos e que também é um sinal que os atributos criados não são suficientes para fazer uma melhor classificação. Sendo assim, o problema pode ser considerado mais complexo a medida que os períodos vão se sucedendo e mais atributos, considerando até mesmo outros contextos (e.g. social, econômico), devem ser utilizados a fim de aumentar o valor das métricas observadas.

- A medida que os períodos vão se passando, a chance dos classificadores obterem resultados melhores do que o esperado de se obter ao acaso, vai diminuindo. O Modelo Específico, principalmente, comporta-se quase como como um modelo de classificação randômico a partir do sexto período - com destaque para o período 9, onde ele acaba tendo resultados piores do que se espera obter ao acaso.

- Os modelos apresentam valores de *Recall* baixos e não muito distantes. Fica entendido que o modelo ainda pode ser melhorado a fim de identificar uma maior quantidade de evasores;

- Os dois modelos se mantém com *accuracy* considerada alta para qualquer período, mesmo com um baixo valor de *recall* (no caso do Modelo Específico, até mesmo para

períodos onde não houveram identificação de nenhum evasor), o que reforça quão desbalanceados são os dados utilizados;

- Ao verificar a métrica kappa, para o Modelo Específico, conclui-se que o modelo vai perdendo poder de predição com o passar dos períodos, de forma que, em alguns, ele chega a ter o mesmo poder preditivo que um modelo randômico (i.e. períodos 7, 9 e 10). Já para o Modelo Global, apesar de descrescer com o passar dos períodos, o kappa ainda mantém uma diferença considerável para o valor 0. Fazendo-se entender, então, que de fato o modelo aprendeu a identificar alunos evasores.

- A precisão do Modelo Global se mantem sempre com valores consideravelmente altos, o que aumenta a confiança das predições estarem corretas.

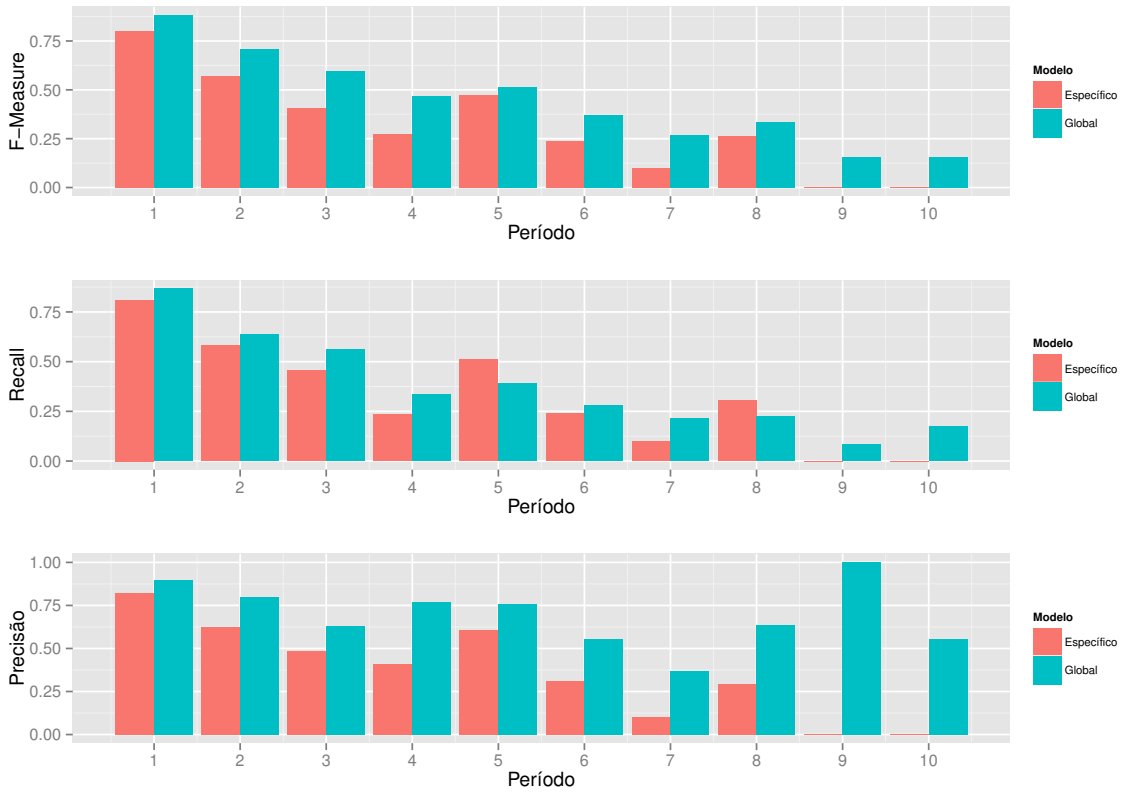

Figura 5.6: *F-measure*, *Recall* e *Precision* dos melhores modelos por período.

Modelo Global					
Nº de períodos	Accuracy	F-Measure	Recall	Precisão	Kappa
1	0.951	0.881	0.866	0.897	0.850
2	0.946	0.707	0.635	0.797	0.678
3	0.939	0.595	0.561	0.632	0.562
4	0.913	0.468	0.336	0.768	0.428
5	0.934	0.514	0.388	0.760	0.483
6	0.927	0.373	0.281	0.553	0.339
7	0.956	0.269	0.212	0.368	0.248
8	0.954	0.333	0.225	0.636	0.315
9	0.955	0.157	0.085	1.000	0.151
10	0.962	0.157	0.172	0.555	0.249
Média	**0.944**	**0.456**	**0.376**	**0.697**	**0.430**

Modelo Específico					
Nº de períodos	Accuracy	F-Measure	Recall	Precisão	Kappa
1	0.932	0.800	0.810	0.824	0.759
2	0.926	0.570	0.579	0.624	0.531
3	0.892	0.408	0.454	0.485	0.369
4	0.824	0.273	0.233	0.409	0.196
5	0.900	0.473	0.510	0.607	0.433
6	0.860	0.239	0.237	0.312	0.188
7	0.845	0.100	0.100	0.100	0.018
8	0.886	0.264	0.306	0.291	0.215
9	0.943	0.000	0.000	0.000	-0.016
10	0.968	0.000	0.000	0.000	0.000
Média	0.898	0.313	0.323	0.365	0.269

Tabela 5.5: Resultados da classificação por período.

5.4.1 Melhor Modelo

Uma das vantagens da Floresta Aleatória é dispor do alto poder de interpretação do modelo pronto. Nessa seção, será discorrido brevemente a respeito o Modelo Global, melhor modelo desenvolvido nesta pesquisa.

A Tabela 5.6 apresenta os três valores de Gini [11] mais altos, por período, dos atributos utilizados para criar a Floresta. De acordo com a tabela, percebe-se que um forte indicador da evasão é o desempenho do aluno ao longo do percurso na universidade e principalmente no semestre atual de curso, já que os MEDIA.SEM aparece em todos os 10 períodos e em quase todos é o atributo com maior importância. O Gini, seguindo a tendência das métricas já mencionadas, também decresce e evidencia que a dificuldade em identificar evasores com o passar dos períodos é maior. As Figuras 5.7 (Gini no primeiro período) e 5.8 (Gini no décimo período) destacam de maneira mais clara a importância dos atributos ao longo dos períodos e a tendência de decrescimento.

	1	2	3	4	5	6	7	8	9	10
MEDIA.SEM	**4250**	**2798**	**2698**	1528	1566	**1499**	**987**	**975**	**833**	**533**
PROP.N.APROV	3388	2441	1973	1559	**1617**	862	767		539	
CRE	3133			**1692**		1391	780	934	636	471
MEDIA.APROV		2312	1739		1346					395
PORCENTAGEM.CURSO.COMPLETO								891		

Tabela 5.6: Atributos mais importantes da Floresta Aleatória baseado no Gini.

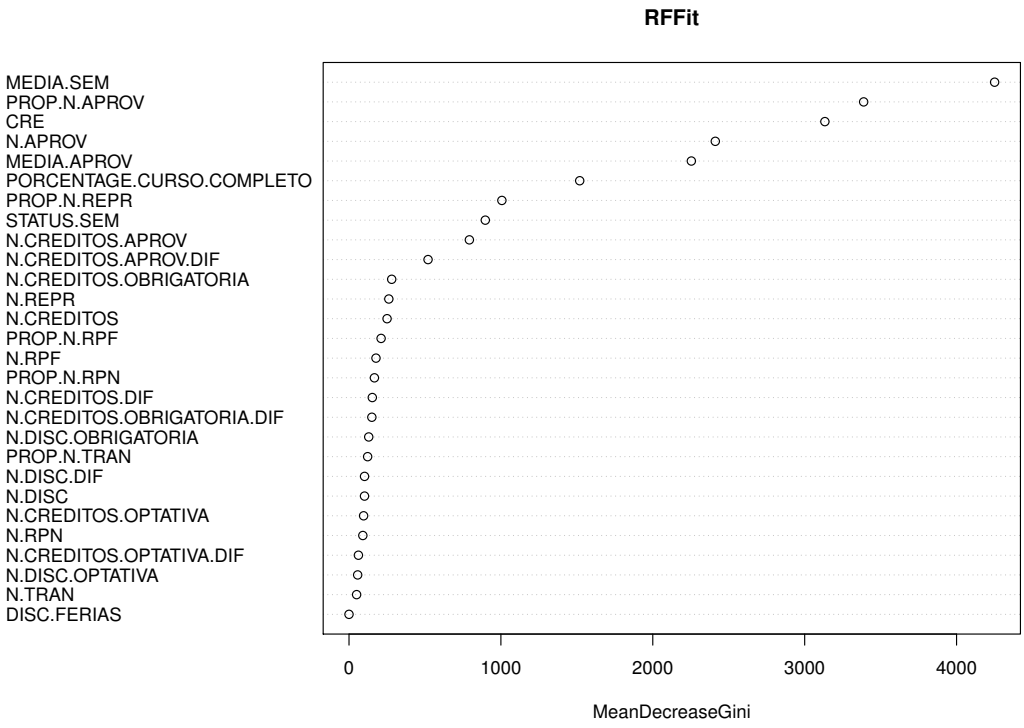

Figura 5.7: Gini dos atributos no primeiro período.

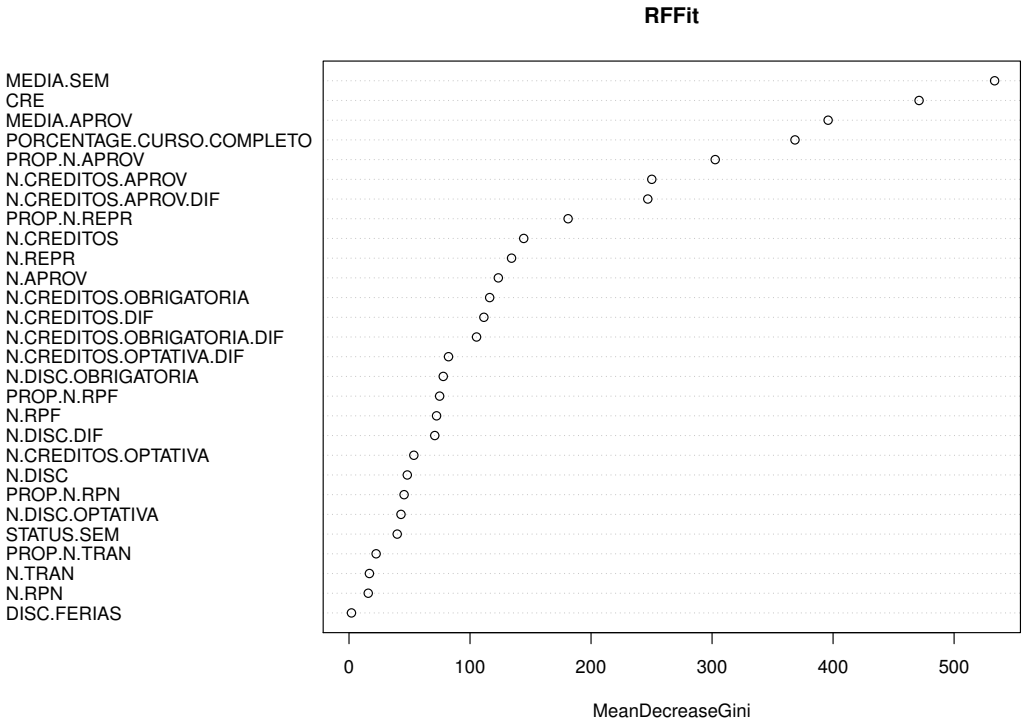

Figura 5.8: Gini dos atributos no décimo período.

Capítulo 6

Conclusão e Trabalhos Futuros

Neste trabalho o problema da evasão estudantil foi modelado como um problema de classificação. O trabalho foi aplicado no contexto da UFCG e avaliou diversas configurações de um classificador estado-da-arte, investigando o problema da evasão estudantil através de duas perspectivas: 1) as razões que levam os alunos a evadir são as mesmas independente de seus cursos, e 2) as razões variam para cada curso.

Foram utilizadas técnicas de seleção de atributos e de desbalanceamento de classes em cada perspectiva e selecionamos a melhor configuração de cada modelo para serem comparadas.

Deste trabalho, é possível tirar as seguintes importantes conclusões:

- Atributos extraídos de registros acadêmicos, por si só, já podem indicar com alta *accuracy* quais alunos tendem a evadir. Algo que se torna interessante já que registros acadêmicos estão sempre disponíveis para a universidade enquanto outros tipos de dados - como dados de condição financeira - tendem a mudar com o tempo e podem não condizer com a real situação do aluno;

- A porcentagem completa do curso pelo estudante é um fator importante para prever a evasão. O que nos faz acreditar que os estudantes levam em consideração o esforço e a facilidade que ainda terão para completar o curso e tempo passado no curso antes de evadir;

- O desempenho do estudante no período corrente (indicado por algumas métricas, como MEDIA.SEM) é um forte indício de evasão;

- A importância dos atributos variam de acordo com os cursos e períodos, e, principalmente nos períodos iniciais, um pequeno número de atributos é suficiente para atingir bons resultados. Ganhando, com menos atributos, um maior poder de interpretação do modelo.

- O Modelo Global atinge melhores resultados que o Modelo Específico - o que pode ter sido consequência da baixa quantidade de dados do Modelo Específico. Mas, ainda assim, é algo que pode ser visto como uma descoberta positiva, já que o custo computacional e a complexidade para construir 10 modelos - um para cada período - é bem menor que o custo de construir 760 modelos - 76 cursos vezes 10 períodos.

Como trabalho futuro, pretende-se estender essa abordagem a fim de considerar a utilização de dados sociais dos alunos, entender o quão impactante seria utilizar dados sociais na criação do modelo e nos resultados e vislumbrar o potencial que os resultados tem para a instituição. A hipótese levantada é que os fatores que levam os estudantes a evadir, com o passar dos períodos, são cada vez mais sociais, sendo assim, mais difíceis de serem mapeados para atributos encontrados nos dados disponibilizados para esta pesquisa. Também existe a intenção de oferecer os resultados dessa pesquisa à Pró-Reitoria de Ensino da UFCG, com o objetivo de ajudar estudantes, professores e administradores a identificar e tomar ações com a finalidade de prevenir a evasão. Além de criar uma base de dados que respeite todos os direitos de privacidade dos alunos e que possa ser utilizada abertamente.

Um exemplo de uso prático do do modelo pode ser a implantação no sistema de controle acadêmico da UFCG - sistema que faz a gestão das matrículas dos alunos da universidade -, assim, cada vez que o sistema for atualizado, automaticamente poderia se ter ideia de quem são os possíveis evasores.

Bibliografia

[1] Ana Amélia Chaves Teixeira ADACHI. *Evasão e Evadidos nos Cursos de Graduação da Universidade Federal de Minas Gerais. 2009. 214 f.* PhD thesis, Dissertação (Mestrado em Educação). Faculdade de Educação–Programa de Pós-Graduação em Educação. Universidade Federal de Minas Gerais. Belo Horizonte, 2009.

[2] Wagner Bandeira Andriola, Cristiany Gomes Andriola, and Cristiane Pascoal Moura. Opiniões de docentes e de coordenadores acerca do fenômeno da evasão discente dos cursos de graduação da universidade federal do ceará (ufc). *Ensaio: aval. pol. públ. Educ*, 2006.

[3] Remis Balaniuk, Hercules Antonio do Prado, Renato da Veiga Guadagnin, Edilson Ferneda, and Paulo Roberto Cobbe. Predicting evasion candidates in higher education institutions. In *Model and Data Engineering*, pages 143–151. Springer, 2011.

[4] Marta F Barroso and Eliane BM Falcão. Evasão universitária: O caso do instituto de física da ufrj. *ENCONTRO NACIONAL DE PESQUISA EM ENSINO DE FÍSICA*, 9:1–14, 2004.

[5] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Ensino superior mantém tendência de crescimento e diversificação, 2010.

[6] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Acesso e permanência no ensino superior, 2013.

[7] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Censo da educação superior 2013, 2013.

[8] Gerben Dekker, Mykola Pechenizkiy, and Jan Vleeshouwers. Predicting students drop out: A case study. In Tiffany Barnes, Michel C. Desmarais, Cristóbal Romero, and Sebastián Ventura, editors, *EDM*, pages 41–50. www.educationaldatamining.org, 2009.

[9] Natalícia Pacheco de L GAIOSO. *A evasão discente na Educação Superior no Brasil: na perspectiva de alunos e dirigentes*. PhD thesis, Dissertação (Mestrado)-Universidade Católica de Brasília, Brasília-DF, 2005.[Links], 2005.

[10] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[11] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[12] P. E. Hart. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:515–516, 1968.

[13] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1263–1284, September 2009.

[14] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013.

[15] Nathalie Japkowicz et al. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68, pages 10–15. Menlo Park, CA, 2000.

[16] John T Kent. Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173, 1983.

[17] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, volume 97, pages 179–186. Nashville, USA, 1997.

[18] Laci Mary Barbosa Manhães, Sérgio Manuel Serra da Cruz, and Geraldo Zimbrão. Evaluating performance and dropouts of undergraduates using educational data mining. In *Proceedings of the Twenty-Ninth Symposium On Applied Computing*, 2014.

[19] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[20] Carlos Márquez-Vera, Alberto Cano, Cristóbal Romero, and Sebastián Ventura. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence*, 38(3):315–330, 2013.

[21] Antonio Moretti, Jose Gonzalez-Brenes, Katherine McKnight, and Ansaf Salleb-Aouissi. Mining student ratings and course contents for computer science curriculum decisions.

[22] Mohammad Nurul Mustafa, Linkon Chowdhury, and MS Kamal. Students dropout prediction for intelligent system from tertiary level in developing country. In *Informatics, Electronics & Vision (ICIEV), 2012 International Conference on*, pages 113–118. IEEE, 2012.

[23] Saurabh Pal. Mining educational data to reduce dropout rates of engineering students. *International Journal of Information Engineering and Electronic Business (IJIEEB)*, 4(2):1, 2012.

[24] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986.

[25] Cristobal Romero and Sebastian Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[26] S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall series in artificial intelligence. Prentice Hall, 2010.

[27] A. Sales, L. Balby, and A. Cajueiro. Predicting student dropout: A case study in brazilian higher education. In *Proceedings of the 3rd Symposium on Knowledge Discovery, Mining and Learning*, 2015.

[28] I. Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772, Nov 1976.

[29] Aharon Yadin. Reducing the dropout rate in an introductory programming course. *Acm Inroads*, 2(4):71–76, 2011.

[30] Raphael Zender, Richard Metzler, and Ulrike Lucke. Freshup—a pervasive educational game for freshmen. *Pervasive and Mobile Computing*, 14:47–56, 2014.