Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Algoritmos de mineração de dados aplicados às turmas do Departamento de Ciência da Computação da Universidade Federal de Juiz de Fora

Lucas Carvalho Ribeiro

JUIZ DE FORA
SETEMBRO, 2021

Algoritmos de mineração de dados aplicados às turmas do Departamento de Ciência da Computação da Universidade Federal de Juiz de Fora

Lucas Carvalho Ribeiro

Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Departamento de Ciência de Computação
Bacharelado em Ciência da Computação

Orientador: Alessandreia Marta de Oliveira
Coorientador: Marcelo Caniato Renhe

JUIZ DE FORA
SETEMBRO, 2021

Algoritmos de mineração de dados aplicados às turmas do Departamento de Ciência da Computação da Universidade Federal de Juiz de Fora

Lucas Carvalho Ribeiro

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Alessandreia Marta de Oliveira
Doutora em Computação - IC/UFF

Marcelo Caniato Renhe
Doutor em Engenharia de Sistemas e Computação - COPPE/UFRJ

Luciano Jerez Chaves
Doutor em Ciência da Computação - UNICAMP

Heder Soares Bernardino
Doutor em Modelagem Computacional - LNCC

JUIZ DE FORA
03 DE SETEMBRO, 2021

Resumo

A modernização dos sistemas de gestão de instituições de ensino faz com que seja cada vez mais comum o acúmulo de dados educacionais. Este avanço tecnológico possibilita também a transformação destes dados educacionais em informações significativas a partir do uso de algoritmos de mineração de dados. Diante disso, este trabalho apresenta uma análise, através do uso dos modelos Árvore de Decisão e Floresta Aleatória, sobre dados de alocações de turmas do Departamento de Ciência da Computação da Universidade Federal de Juiz de Fora. O objetivo desta análise é criar modelos de classificação para auxiliar a identificação de fatores que influenciam os índices de aprovação de uma turma, com o intuito de auxiliar na criação de turmas com melhores rendimentos no futuro. Essa análise mostra que os principais fatores para o desempenho de uma turma são a disciplina e os cursos envolvidos, embora outras características possam se mostrar relevantes ao observar casos mais específicos.

Palavras-chave: Mineração de Dados, Mineração de Dados Educacionais, Descoberta de Conhecimento, Árvore de Decisão, Floresta Aleatória.

Abstract

The modernization of management systems in educational institutions makes the accumulation of educational data increasingly common. This technological advance also enables the transformation of educational data into meaningful information through the use of data mining algorithms. This work presents an analysis, through the use of the Decision Tree and Random Forest algorithms, on class allocation data from the Departamento de Ciência da Computação of the Universidade Federal de Juiz de Fora. This analysis's purpose is the creation of classification models to assist in identifying factors that influence the approval ratings of a class, in order to assist in creating classes with better performances in the future. This analysis shows that the main factors for a class's performance are the discipline and courses involved, although other characteristics may prove relevant when observing more specific cases.

Keywords: Data Mining, Educational Data Mining, Knowledge Discovery, Decision Tree, Random Forest.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AUC	Área sob a Curva ROC
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
DCBD	Descoberta de Conhecimento em Bases de Dados
DCC	Departamento de Ciência da Computação
EAD	Ensino à Distância
FIES	Programa de Financiamento Estudantil
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
PART	*Projective Adaptive Resonance Theory*
RIPPER	*Repeated Incremental Pruning to Produce Error Reduction*
SIGA	Sistema Integrado de Gestão Acadêmica
SVM	*Support Vector Machine*
UFJF	Universidade Federal de Juiz de Fora
UFRN	Universidade Federal do Rio Grande do Norte

1 Introdução

1.1 Apresentação do tema

Com a popularização da Internet e os baixos custos tanto de equipamentos tecnológicos quanto de armazenamento, tem se tornado comum a geração de grandes quantidades de dados por diversos sistemas. Contudo, apenas possuir dados armazenados não é suficiente, sendo necessário analisá-los para poder aproveitá-los ao máximo.

Segundo Amaral (2016), "mineração de dados são processos para explorar e analisar grandes volumes de dados em busca de padrões, previsões, erros, associações, entre outros". Os algoritmos de Mineração de Dados vêm sendo aplicadas a vários tipos de bases de dados, como na detecção de fraudes em bancos (JOHN et al., 2016), no diagnósticos de doenças (RAY, 2018) e na investigação criminal (HASSANI et al., 2016). Neste trabalho, foi explorada uma das possibilidades de uso destes algoritmos na área da educação, aplicando-os para tentar identificar quais fatores decididos durante a alocação de uma turma podem impactar o desempenho de seus alunos.

1.2 Contextualização

Com o avanço do uso de tecnologias na administração de escolas, é comum que, atualmente, as instituições tenham uma grande quantidade de dados guardados sobre os seus alunos, professores e turmas através dos anos. Ao aplicar algoritmos de Mineração de Dados sobre estas bases, é possível encontrar vários parâmetros que podem ser usados, por exemplo, para melhorar a qualidade do ensino de uma instituição.

Pelaez et al. (2019), por exemplo, analisaram parâmetros que podem levar um aluno a não conseguir concluir seu curso, chamados de *at-risk students*. Ao identificar alunos que possam estar nesta situação com antecedência, é possível fazer intervenções que os possibilitem continuar e concluir seus cursos, como é desejado.

Morsy e Karypis (2019) também tratam do uso de Mineração de Dados na

educação. Neste trabalho é feita uma análise que, com base nas notas de um determinado aluno, determina quais disciplinas têm mais chances de aumentar ou diminuir seu índice de rendimento acadêmico.

Outra aplicação de Mineração de Dados na educação pode ser vista em Liu e Tan (2020). Neste trabalho foram analisados parâmetros que podem incentivar um aluno a buscar uma carreira nas áreas de Ciência, Tecnologia, Engenharia e Matemática, para que seja possível planejar uma intervenção nos anos iniciais de ensino, com a intenção de aumentar o interesse dos alunos nessas áreas.

1.3 Problema

O Departamento de Ciência da Computação (DCC) da Universidade Federal de Juiz de Fora (UFJF) conta hoje com quatro cursos de graduação presenciais (Ciência da Computação integral e noturno, Engenharia Computacional e Sistemas de Informação), um curso de graduação à distância (Licenciatura em Computação) e dois programas de pós-graduação stricto sensu (Ciência da Computação e Modelagem Computacional).

Além destes cursos, o DCC atende toda a comunidade acadêmica do Instituto de Ciências Exatas (curso de Ciências Exatas, Física, Química, Matemática e Estatística) e da Faculdade de Engenharia (com seus diversos cursos), além dos cursos de Administração e Ciências Contábeis.

O DCC possui aproximadamente 50 professores, entre efetivos e substitutos. De acordo com o plano departamental do DCC de 2019, disponível no Sistema Integrado de Gestão Acadêmica (SIGA), foram oferecidas aproximadamente 200 disciplinas e 350 turmas. Neste cenário, apenas a tarefa de alocar recursos para todas estas turmas e cursos já é bastante complexa. Mas o objetivo não é apenas alocar recursos, e sim montar um plano da forma mais eficaz possível, de forma que as turmas alocadas consigam os melhores resultados possíveis. Para auxiliar nesta tarefa, ter informações sobre quais decisões geram os melhores resultados é extremamente interessante, podendo ajudar os responsáveis pela construção do plano departamental a tomar decisões importantes sobre alocações destas turmas.

1.4 Motivação

Durante a construção de um plano departamental existem vários aspectos a serem observados, como as disciplinas na grade de um curso, os horários das turmas de cada disciplina, os cursos com vagas alocadas em cada turma e o docente responsável por lecionar cada turma.

Com tantos fatores, é difícil dizer, a princípio, qual o impacto que cada um tem no rendimento de uma determinada turma, ou quão grande é o impacto de cada um deles. Existe também uma dificuldade ao distribuir vagas destas turmas para os cursos interessados nelas, principalmente em disciplinas presentes nas grades de vários cursos, como Algoritmos e Laboratório de Programação.

1.5 Objetivo

Neste trabalho é explorado o uso de algoritmos de Mineração de Dados em uma base de dados contendo informações sobre as várias turmas oferecidas pelo DCC/UFJF entre os anos de 2014 e 2019, assim como suas distribuições de vagas, para encontrar quais decisões tomadas durante a criação de turmas do plano departamental têm maior impacto no rendimento de alunos alocados em várias disciplinas. Com isso, espera-se encontrar formas de se fazer alocações mais eficazes, resultando em turmas com maiores índices de aprovação.

1.6 Organização do trabalho

Este trabalho está organizado em cinco capítulos, incluindo esta introdução. O Capítulo 2 apresenta a fundamentação teórica do trabalho: os conceitos básicos de mineração de dados, os modelos a serem utilizados e a metodologia utilizada para o desenvolvimento do projeto. O Capítulo 3 apresenta alguns trabalhos relacionados à aplicação de algoritmos de mineração de dados em contextos educacionais. O Capítulo 4 apresenta como foi feito o desenvolvimento do projeto, trazendo aspectos da modelagem utilizada para resolver o problema proposto neste trabalho. O Capítulo 5 mostra os resultados encontrados com os

modelos gerados, e discute o que pode ser observado. O Capítulo 6 apresenta as conclusões deste trabalho, e sugere trabalhos futuros que podem ser feitos para complementar os resultados encontrados.

2 Fundamentação teórica

Este capítulo apresenta alguns conceitos básicos da Mineração de Dados, seguido de uma revisão de trabalhos relacionados que exploram o uso de algoritmos da área no ambiente escolar. Na Seção 2.1 são apresentados os conceitos básicos de Mineração de Dados e do processo de Descoberta de Conhecimento em Bases de Dados (DCBD). Na Seção 2.2 são apresentados os modelos usados nesse trabalho. A Seção 2.3 apresenta a metodologia que foi usada para organizar e desenvolver este trabalho.

2.1 Mineração de dados e descoberta de conhecimento em bases de dados

Segundo Goldschmidt, Passos e Bezerra (2015), a Mineração de Dados é a principal etapa do processo de DCBD. Estes autores dividem o processo de DCBD em 3 partes principais:

- o pré-processamento, durante o qual o estudo e a compreensão dos dados são feitos e o conjunto de dados a ser usado é determinado, organizado, tratado e preparado para os algoritmos de Mineração de Dados;
- a Mineração de Dados em si, onde os algoritmos são aplicados à base de dados para encontrar novos conhecimentos sobre o domínio estudado;
- e o pós processamento, onde as informações obtidas do passo anterior são analisadas e interpretadas.

Além destas 3 partes principais, as tarefas envolvidas na etapa da Mineração de Dados são agrupadas em tarefas de associação, classificação, regressão, sumarização e clusterização (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Finalmente, os objetivos do modelo criado na etapa de Mineração de Dados são classificados em:

- predição, no qual o objetivo é criar um modelo que permita predizer alguns atributos de interesse com base nos dados atualmente disponíveis na base de dados;
- descrição, no qual o objetivo é esclarecer a interação e o comportamento dos atributos da base de dados.

Durante este trabalho foram criados modelos usando algoritmos de classificação, que dividem as turmas entre baixo, médio e alto desempenho. Neste caso, os modelos tentam prever a qual classe novas turmas pertencem, permitindo extrair conhecimento sobre quais decisões levam o modelo a classificar uma turma como um caso de cada classe de rendimento.

2.2 Modelos

Existem muitos algoritmos que podem ser usados para gerar modelos de classificação. Neste trabalho foi explorada a aplicação de algoritmos que geram modelos de Árvore de Decisão e de Floresta Aleatória. Estes modelos foram escolhidos por mostrarem bons resultados ao serem usados em outros trabalhos relacionados à educação com objetivos semelhantes, como pode ser visto em Carrano et al. (2019), Sales et al. (2019) e Júnior (2018), e por permitirem uma visualização de como o processo de classificação é feito, possibilitando encontrar quais informações são mais importantes neste processo. Essa análise do processo de decisão sobre o modelo de Floresta Aleatória não pode ser feita com a mesma precisão que a análise sobre a Árvore de Decisão por conta do processo de votação não ser claro, mas ainda assim é possível observar algumas tendências sobre os modelos gerados. Neste trabalho foram usadas as implementações do Scikit-learn (PEDREGOSA et al., 2011) para estes algoritmos.

2.2.1 Árvore de decisão

Árvore de Decisão é um algoritmo capaz de criar uma estrutura em forma de árvore, na qual cada nó representa um teste de uma variável e cada aresta representa um dos valores possíveis do teste. Os nós das folhas são compostos por itens de uma única classe, e

cada caminho do nó raiz até um nó folha é uma regra de classificação. Após o modelo ser construído, é possível testar novos valores ao aplicar os testes dos nós nos seus vários atributos e percorrê-la até chegar em um nó folha, que indica a qual classe este dado testado pertence (CASTRO; FERRARI, 2016).

Para a criação deste modelo é preciso selecionar um atributo inicial que é testado no nó raiz. Como o algoritmo utilizado[1] exige que todos os atributos usados sejam transformados em atributos numéricos, os testes sempre comparam se o valor é menor ou maior que uma constante. Para cada resultado possível é criado um nó, e se observa quantos objetos de cada classe estão presentes neste nó. Caso o nó tenha itens de apenas uma classe ele é um nó folha, não sendo expandido além disso. Caso o nó tenha itens de mais de uma classe, o processo anterior é repetido nele. Este processo se repete até que todos os nós folhas tenham itens de apenas uma classe. Um exemplo com os primeiros níveis da estrutura gerada por este algoritmo pode ser visto na Figura 2.1.

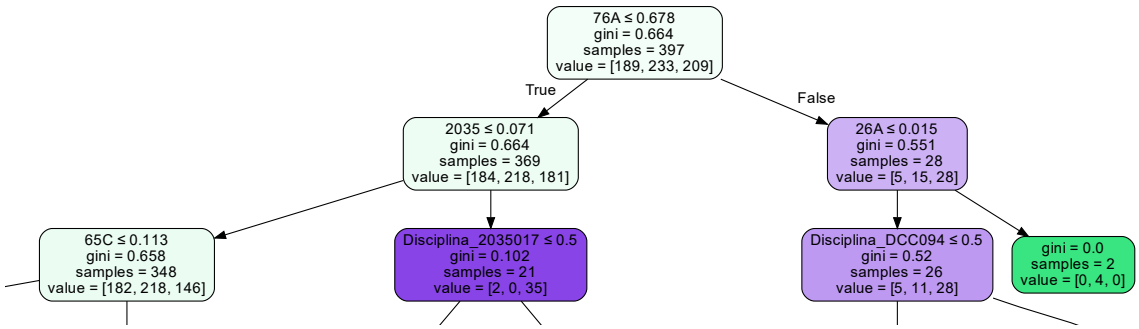

Figura 2.1: Estrutura de árvore de decisão.

2.2.2 Floresta aleatória

Floresta Aleatória é um algoritmo capaz de criar várias estruturas de árvore de decisão para um mesmo conjunto de dados, utilizando uma seleção aleatória de atributos ao criar seus nós. Para determinar a classe de um novo objeto são feitos testes com todas as árvores da floresta, observando qual classe cada árvore atribui a esta nova entrada e, em seguida, é feito um processo de votação entre todas as árvores, permititindo encontrar a qual classe este novo objeto tem mais chances de pertencer (BREIMAN, 2001).

Na implementação original, o resultado de floresta é determinado por uma votação,

[1] https://scikit-learn.org/stable/modules/tree.html

na qual cada árvore dá um voto para o resultado previsto por ela. A implementação disponibilizada no Scikit-learn[2] funciona de forma um pouco diferente. Ao invés de cada árvore simplesmente votar na classe prevista por ela, são consideradas as predições probabilísticas de cada árvore, e a média destas predições é usada para determinar a classe prevista pela floresta como um todo.

2.3 Cross-Industry Standard Process for Data Mining

Cross-Industry Standard Process for Data Mining (CRISP-DM) é uma metodologia que pode ser usada para organizar o processo de DCBD. Este é um modelo tido como um dos principais para a criação de aplicações de DCBD, e apresenta um conjunto de fases muito bem definidas (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). A CRISP-DM define 6 etapas para o processo de DCBD:

- compreensão do negócio, na qual se busca compreender o contexto dos dados a serem analisados, estabelecendo os objetivos a serem alcançados pela DCBD;
- compreensão dos dados, na qual se busca compreender quais informações estão disponíveis na sua base de dados;
- preparação dos dados, na qual os dados são organizados, tratados e preparados para que os algoritmos de Mineração de Dados possam ser aplicados;
- modelagem, na qual os algoritmos de Mineração de Dados são efetivamente aplicados, criando os modelos usados para a descoberta de conhecimento;
- avaliação, na qual é feita a análise do modelo gerado na fase anterior, permitindo estimar a qualidade do resultado obtido;
- desenvolvimento, na qual é feita a implementação dos resultados obtidos nas etapas anteriores.

[2]https://scikit-learn.org/stable/modules/ensemble.html#random-forests

2.4 Considerações finais

Neste capítulo foram apresentados conceitos básicos sobre a Mineração de Dados e o processo de DCBD. Além disso, foi apresentada a CRISP-DM, uma metodologia para desenvolver projetos de Mineração de Dados, e que é usada para organizar o desenvolvimento dos Capítulos 4 e 5 deste trabalho.

3 Trabalhos relacionados

A mineração de dados vem sendo usada na educação com vários objetivos. Este capítulo apresenta trabalhos da literatura que focam em dois objetivos principais: a detecção de alunos com risco de evasão escolar (CARRANO et al., 2019; SALES et al., 2019; COLPANI, 2019; CALIXTO; SEGUNDO; GUSMÃO, 2017) e a previsão do desempenho dos alunos (JÚNIOR, 2018; RABELO et al., 2017). Para resolver estes problemas, é comum o uso de modelos de classificação. Os trabalhos a seguir mostram exemplos de situações em que algoritmos de classificação foram usados para resolver problemas educacionais, com vários destes usando os mesmos algoritmos que são propostos por este trabalho.

3.1 Evasão escolar

Para resolver o problema da análise de alunos com risco de evasão já foram usados vários algoritmos diferentes, e todos mostraram resultados interessantes, conseguindo bons desempenhos para as métricas analisadas. Pode-se observar também que estes algoritmos podem ser utilizadas de forma satisfatória independentemente da etapa do ensino em que o aluno se encontra.

3.1.1 Evasão no ensino básico em Juiz de Fora

Em Sales et al. (2019), foram observados dados de alunos do ensino básico na rede municipal de Juiz de Fora. Para construir o modelo de predição, foi usado o algoritmo de Floresta Randômica Ponderada sobre uma base de dados disponibilizada pela prefeitura.

Para a construção deste modelo, foram usados dados sobre o nível de ensino no qual o aluno se encontrava, o turno em que ele estudava (manhã, tarde, noite ou integral), sua etnia, se ele era repetente, se ele possuía um responsável, se ele possuía necessidades especiais, se necessitava de transporte público para chegar à escola, o seu gênero e a sua situação na época (ativo na rede pública, mudou de escola ou evadiu). Embora dados como indicativos socioeconômicos, notas ou frequência dos alunos sejam interessantes

para este tipo de estudo, eles não foram disponibilizados pela prefeitura para garantir a privacidade dos alunos.

A escolha pelo algoritmo da Floresta Randômica Ponderada foi feita por permitir uma boa interpretação dos seus resultados, possibilitando recuperar quais decisões que causaram uma determinada classificação. Para realizar o treinamento, a base de dados foi dividida em duas partes: 66% dos dados foram usados para criar uma base de treinamento, enquanto os outros 34% foram usados para realizar testes no modelo produzido. Para garantir que tanto os dados de treinamento quanto os dados de teste possuíssem uma quantidade balanceada de alunos evadidos, a divisão foi feita de forma a garantir uma mesma proporção destes alunos em cada base.

Finalmente, para cada conjunto de parâmetros que foi avaliado, foram criados 30 modelos distintos para o mesmo conjunto de dados. Dentre todos os modelos criados, foi encontrada uma precisão média de 70% e uma cobertura média de 97%. O modelo de melhor precisão conseguiu alcançar 76% de precisão, e o modelo com melhor cobertura chegou a 98% de cobertura.

3.1.2 Evasão no ensino médio no Pará

Em Colpani (2019) foram usados dados do Censo Escolar de 2017, disponibilizados publicamente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) referentes ao estado do Pará. A escolha desse estado foi feita por ser o que apresentou a maior taxa de evasão escolar durante o ano analisado. Para treinar os modelos, foram usados 75% dos dados disponíveis, e os outros 25% foram usados para testar os modelos encontrados.

Para melhor organizar o trabalho, foi usada a metodologia CRISP-DM. Os indicadores utilizados para criar os modelos analisados foram a média de alunos por turma, a média de horas-aula diárias, o percentual de docentes com curso superior, a taxa de distorção idade-série, a taxa de evasão e a taxa de reprovação.

Durante a criação do seu modelo preditivo foram usados algoritmos de análise de correlação e de regressão linear. Na análise de correlação é possível encontrar quão forte é a relação entre duas variáveis. Neste estudo, foi feita a análise entre a taxa de

evasão e cada uma das outras 5 variáveis mencionadas. Esta análise permite observar que o atributo com a maior correlação foi a taxa de distorção idade-série, que mede a diferença de idade dos alunos com relação à idade esperada para a sua série atual. Já a análise de regressão linear permite estimar valores de uma variável com base nos valores de outra variável. Novamente, foi feita a análise entre a taxa de evasão e as outras 5 variáveis analisadas. A análise dos modelos de regressão linear possibilitou constatar que 33% da taxa de evasão das escolas analisadas pode ser explicada pela taxa de distorção idade-série, e cerca de 69% da variação restante não é explicada pelos fatores analisados neste estudo.

3.1.3 Evasão no ensino médio no Ceará e no Sergipe

Calixto, Segundo e Gusmão (2017) utilizaram dados disponibilizados pelo INEP, observando dados dos censos educacionais de 2014, 2015 e 2016 para escolas do Ceará e do Sergipe. Estes dois estados foram escolhidos por serem, respectivamente, os estados da região nordeste de maior e menor taxa no ranking do Índice de Desenvolvimento da Educação Básica.

Assim como em Colpani (2019), a CRISP-DM é usada para a organização do trabalho. Durante a preparação dos dados foram escolhidas 61 variáveis que são significativas para o estudo. Com os dados de 2014, 2015 e 2016 adicionados a um banco de dados, foi possível observar a evasão dos alunos entre 2014 e 2015, assim como a evasão entre 2015 e 2016. Esta análise mostra que tanto em 2015 quanto em 2016 a taxa de evasão foi de 17%.

Para criar os modelos de predição foram usados algoritmos de regressão logística e de indução de regras. O modelo criado pela regressão logística mostrou uma acurácia de 87,4% no Ceará e de 86,8% no Sergipe. Com este primeiro modelo, foi possível observar algumas variáveis que mais impactam a evasão escolar, como a idade dos alunos, a residência em área urbana, ou docentes com contrato temporário.

Já no modelo criado por indução de regras foi possível gerar várias regras que classificam um aluno como evadido ou não, possibilitando destacar que variáveis como idade, etapa do ensino e a presença do ensino médio profissionalizante apresentam uma

maior influência na evasão dos alunos.

3.1.4 Evasão na Universidade Federal de São João Del-Rei

Em Carrano et al. (2019) foi analisada a evasão na Universidade Federal de São João Del-Rei, tanto em uma visão geral quanto em uma visão específica para cada área de conhecimento. Para isso foram usados dados sobre os alunos que ingressaram entre 2010 e 2017.

Durante este trabalho foi usado o algoritmo de Árvore de Decisão para gerar os modelos preditivos. A base de dados foi dividida em vários grupos de treino e de teste para possibilitar a construção e a verificação da qualidade destes modelos. Para construir estes conjuntos, um ano foi tomado como conjunto de teste, e um conjunto de um, dois ou três anos anteriores a ele foram usados para criar o conjunto de treinamento. Com todos os modelos criados, é feita uma comparação entre as suas qualidades, e o melhor subconjunto é considerado nos próximos passos.

Para poder validar estatisticamente este resultado foram realizadas 50 repetições do experimento. Com o modelo escolhido, foi usado o algoritmo de seleção de características *infogain*, para identificar quais atributos foram os mais significativos para o modelo analisado. Nesta etapa, são criados subconjuntos mantendo 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% ou 90% dos atributos mais relevantes.

Em seguida, voltou-se a analisar todos os modelos gerados anteriormente, considerando os vários subconjuntos de atributos mais relevantes e, dentre todos os modelos e conjuntos de atributos, é encontrado aquele com maior taxa de acertos e menor taxa de erros.

Com este modelo de melhor desempenho selecionado, é possível analisar os dados de alunos ingressantes no ano observado (no caso deste trabalho, 2019) para poder prever a necessidade de ações preventivas em relação à evasão. Este modelo final também permite analisar qual foi o subconjunto de atributos mais relevantes para a evasão, permitindo promover ações preventivas para combater os principais motivos que levam alunos a evadir.

Ao analisar os vários modelos criados foi observado que os melhores modelos encontrados foram treinados apenas com os dados do ano imediatamente anterior, e con-

siderando apenas o subconjunto de 20% dos atributos mais relevantes. Com isso foi criado um modelo usando dados de 2018 para poder avaliar a evasão dos alunos em 2019 que ingressaram nos 4 anos anteriores, ou seja, em 2016, 2017, 2018 e 2019.

No total foram analisados mais de 4000 alunos, e foi encontrado que 10% deles estariam em risco de evasão. 75% dos alunos em risco de evasão estariam nas áreas de Engenharias, Ciências Agrárias, Ciências Exatas e da Terra e Ciências Sociais Aplicadas, sendo 36% deste total apenas nas Engenharias.

Quanto aos atributos mais relevantes, o *infogain* deu maiores pontuações às seguintes métricas: Média dos conceitos de autoavaliação e de avaliação sobre os docentes feitas pelos alunos; coeficiente de rendimento nos dois primeiros semestres; média de faltas do aluno no primeiro ano; média de notas do aluno no primeiro ano; volume de disciplinas cursadas; percentual de aprovações do aluno no primeiro ano; e desempenho do aluno nas 3 disciplinas de maior índice de reprovação do seu curso.

3.2 Predição de desempenho

Para o problema da predição do desempenho acadêmico também foram usadas vários algoritmos de construção de modelos de predição, e é possível observar que, de forma geral, os algoritmos testados conseguem encontrar bons resultados.

3.2.1 Predição em institutos federais de educação tecnológica e no Colégio Pedro II

Em Júnior (2018), foram analisados dados sobre os alunos dos Institutos Federais de Educação Tecnológica e do Colégio Pedro II no ENEM de 2014. Novamente foi usada a metodologia CRISP-DM para organizar a estrutura do estudo de forma geral. Ao analisar os dados disponíveis, foram encontrados 167 atributos diferentes e 25.080 candidatos que pertenciam a algum dos colégios analisados.

Neste trabalho, os alunos foram separados em um grupo de baixo-rendimento e um grupo de bom-rendimento. Para criar estes grupos foram analisadas as notas de cada aluno nas provas objetivas e na redação, e foi avaliada a média geral do aluno:

caso fosse menor que 600 pontos, o aluno era classificado como de baixo-rendimento e, caso fosse maior ou igual a 600, o aluno era classificado como de bom-rendimento. Com esta definição, uma nova variável "rendimento"foi criada para cada aluno, com o seu valor definido pela regra anterior, e este é o valor que foi categorizado pelos algoritmos utilizados.

Ao analisar a base de dados após todas as preparações, foi encontrado que 43,42% dos alunos são considerados como de bom-rendimento, enquanto 56,58% são considerados como de baixo-rendimento, o que mostra que a base de dados é pouco desbalanceada.

A separação entre os dados de teste e os dados de treinamento foi feita com o método k-*fold*, utilizando k = 10. Este método consiste em dividir o conjunto de dados em k grupos de mesmo tamanho (neste caso 10), garantindo que todos têm a mesma proporção de alunos de baixo-rendimento e de bom-rendimento. Com estes 10 grupos definidos, o processo de treinamento se repete 10 vezes, cada uma delas usando 1 conjunto diferente como conjunto de teste e os 9 conjuntos restantes como conjunto de treino.

Os modelos de classificação analisados foram criados usando os algoritmos *Projective Adaptive Resonance Theory* (PART), J48, *Random Forest*, *Naive Bayes*, *Support Vector Machine* (SVM) e JRip, que é uma implementação em Java do algoritmo *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER). Os resultados obtidos por estes algoritmos podem ser observados na Tabela 3.1.

Tabela 3.1: Comparação dos resultados encontrados pelos algoritmos em Júnior (2018)

Algoritmo	**Acurácia**	**Área sob a Curva ROC (AUC)**
JRip	70,95%	0,723
PART	68,02%	0,696
J48	69,92%	0,691
Random Forest	74,33%	0,816
Naive Bayes	73,94%	0,817
SVM	72,78%	0,716

Observando as métricas apresentadas é possível notar que os algoritmos *Random Forest*, *Naive Bayes* e SVM obtiveram os melhores resultados, com o algoritmo *Random Forest* mostrando os melhores resultados de forma geral.

Embora os algoritmos *Random Forest* e *Naive Bayes* tenham mostrado os melhores resultados, os modelos gerados por eles são "caixa-preta", ou seja, não é possível recuperar o processo de decisão que levou ao modelo gerado. Dentre os modelos "caixa-

branca" analisados, o algoritmo JRip mostrou os melhores resultados.

A análise das regras geradas pelo JRip permite observar que os atributos mais comuns entre os alunos de bom rendimento são a língua estrangeira escolhida (alunos que escolhem o inglês como língua estrangeira mostram melhores resultados), a escola estar na zona rural ou na zona urbana (escolas na zona urbana mostram melhores resultados), e o grau de relevância que o aluno dá para bolsas de estudos em instituições privadas, como o Programa de Financiamento Estudantil (FIES) (grande parte das regras que definem alunos de bom desempenho mostram a participação no FIES com relevância média ou boa).

3.2.2 Predição em turma de ensino a distância na Universidade Federal do Rio Grande do Norte

Em Rabelo et al. (2017), foram analisados dados de 13 turmas de ensino à distância (EAD) da Universidade Federal do Rio Grande do Norte (UFRN), que utiliza da plataforma Moodle para estas turmas.

Ao analisar os dados disponíveis foram encontrados 64 ações e 24 módulos associados a usuários com perfil de aluno nos logs do Moodle. Técnicas de estatística descritiva, desvio padrão, média e coeficiente de correlação foram utilizadas para encontrar, dentre estas 64 ações, quais são as mais relevantes para esta análise.

Com isso, foram encontradas oito ações principais: Ação de login do usuário; Ação de visualização do curso; Ação de visualização de recursos do curso; Ação de visualização de discussões nos fóruns; Ação de adição de postagens nos fóruns; Ação de visualização de tarefas; Ação de enviar tarefas; Ação de responder questionários.

O desempenho de um aluno pode então ser classificado com valores entre 0 e 3: valores de 0 a 1 representam um desempenho regular; valores de 1 a 2 representam um desempenho bom; e valores de 2 a 3 representam um desempenho ótimo.

Para aplicar os algoritmos de mineração de dados foi utilizada a ferramenta WEKA. Os modelos criados para analisar estes dados foram baseados nos algoritmos ID3 e J48. O modelo criado pelo algoritmo ID3 conseguiu classificar corretamente 93,97% das instâncias, enquanto o algoritmo J48 classificou corretamente 96,50% das instâncias.

A qualidade dos modelos gerados foi testada utilizando dados de 268 alunos.

Segundo o modelo preditivo, 64 teriam desempenho ótimo, 164 teriam desempenho bom e 40 teriam desempenho regular, resultando em uma previsão de 228 alunos aprovados e 40 alunos reprovados. O resultado real foi de 234 aprovações e 34 reprovações, que é bastante parecido com o resultado antecipado. Dos 234 alunos que foram de fato aprovados, 210 foram corretamente classificados, e dos 34 alunos reprovados 16 foram classificados corretamente.

3.3 Considerações finais

Neste capítulo foram apresentados trabalhos relacionados à proposta deste trabalho de conclusão de curso que exploram o uso da Mineração de Dados em contextos educacionais, de forma similar ao que é feito neste trabalho. Alguns dos principais resultados apresentados nestes trabalhos relacionados podem ser vistos na Tabela 3.2.

Tabela 3.2: Comparação dos trabalhos relacionados

Trabalho	**Nível**	**Região**	**Algoritmos**	**Dados**	**Melhores Resultados**
Carrano et al. (2019)	Universitário	São João Del-Rei	Árvores de Decisão	UFSJ	Principais fatores: Desempenho, Assiduidade e Satisfação
Sales et al. (2019)	Básico	Juiz de Fora	Floresta Randômica Ponderada	Prefeitura	Precisão: 76% Cobertura: 98%
Colpani (2019)	Médio	Pará	Análise de Correlação Regressão Linear	INEP	Principal Fator: Diferença de idade
Calixto, Segundo e Gusmão (2017)	Médio	Ceará Sergipe	Regressão Logística Indução de Regras	INEP	Ceará: 87,4% Acurária Sergipe: 86,8% Acurária
Júnior (2018)	Médio	Brasil	Random Forest, Naive Bayes etc.	ENEM	Acurácia (Random Forest): 74,32% AUC (Naive Bayes): 0,817
Rabelo et al. (2017)	Univ. (EAD)	Rio Grande do Norte	ID3 J48	UFRN	ID3: 93,97% Acurácia J48: 96,5% Acurácia

Os trabalhos permitem observar a eficiência do uso de algoritmos de Mineração de Dados na educação em vários cenários, independentemente do perfil educacional estudado, sendo válido o uso em vários níveis de ensino e em várias localizações geográficas diferentes. É possível também observar que a Mineração de Dados pode ser usada para resolver problemas similares ao proposto neste trabalho. Vale destacar que vários dos trabalhos analisados possuíam bases de dados desbalanceadas, que possuem uma quantidade consideravelmente maior de dados de uma das classes analisadas, mas mesmo assim usaram a acurácia para avaliar seus resultados. Essa métrica pode ser enviesada em bases

com essa característica, e o seu uso pode não ter sido o ideal nesses casos.

4 Desenvolvimento

Este capítulo apresenta os passos do desenvolvimento do trabalho, seguindo a estrutura da CRISP-DM. Aqui são apresentados os 4 primeiros passos: compreensão do negócio, compreensão dos dados, preparação dos dados e modelagem. A avaliação dos modelos é abordada no próximo capítulo. Para a implementação deste trabalho foram usadas a linguagem Python[3] e as bibliotecas Pandas[4] e Scikit-Learn[5].

4.1 Compreensão do negócio

Na base de dados analisada, estão presentes informações sobre as turmas do DCC da UFJF entre os anos de 2014 e 2019. Estão presentes dados de turmas pertencentes a todas as disciplinas do departamento, incluindo informações sobre os desempenhos e distribuições dos seus alunos, que pertencem a vários cursos diferentes. Ao observar estes dados, espera-se encontrar se estas características de uma turma são relacionadas a taxas de aprovação altas ou baixas, e quais destas características são as mais relevantes para esta classificação. A base de dados não inclui turmas de 2020 por conta de estas terem sido concluídas por ensino remoto, sendo com isso diferentes de turmas usuais, e não inclui turmas antes de 2014 por conta de informações sobre estas turmas não estarem disponíveis.

4.2 Compreensão dos dados

A base de dados é composta por 24 arquivos, e cada semestre tem seus dados guardados em 2 arquivos diferentes, resultando em 4 arquivos para cada ano analisado. No primeiro tipo de arquivo estão presentes as informações gerais das turmas, assim como sua divisão de vagas entre os cursos. No segundo tipo de arquivo estão presentes informações sobre a situação dos alunos de cada turma, especificando como foi o rendimento de cada aluno.

[3] https://www.python.org/
[4] https://pandas.pydata.org/
[5] https://scikit-learn.org/

Além disso, alunos de turmas aos quais conceitos não se aplicam são identificados com "sem conceito".

Cada linha do primeiro tipo de arquivo possui as seguintes informações:

- Disciplina: permite identificar a qual disciplina uma turma pertence através de um nome;
- Cod Disc: permite identificar a qual disciplina uma turma pertence através de um código único;
- T: permite identificar uma turma específica dentre as várias turmas de uma mesma disciplina em um único período;
- Cod: permite identificar a qual curso as informações de vagas disponibilizadas em uma determinada linha pertencem através de um código;
- Curso: permite identificar a qual curso as informações de vagas disponibilizadas em uma determinada linha pertencem através de um nome;
- Of.: permite identificar as vagas oferecidas para um curso em uma turma;
- Oc.: permite identificar quantas vagas foram efetivamente usadas (ocupadas) entre as vagas oferecidas para um curso em uma turma;
- Horários: permite identificar em quais dias e horas as aulas de uma turma foram lecionadas;
- Docentes: permite identificar quais docentes foram responsáveis por lecionar em uma determinada turma.

Um exemplo com algumas entradas deste primeiro tipo de arquivo pode ser visto na Tabela 4.1. Alguns dos dados foram abreviados para adequar a tabela ao tamanho da página.

Cada linha do segundo tipo de arquivo possui as seguintes informações:

- Disciplina: permite identificar a qual disciplina uma turma pertence através de um código único;

Tabela 4.1: Entradas em arquivos do primeiro tipo

Disciplina	Cod Disc	T	Cod	Curso	Of.	Oc.	Horários	Docentes
APA	DCC001	A	22A	Ciência da Computação	2	2	Ter, 21 às 23	Guilherme
APA	DCC001	A	35A	Ciência da Computação	30	10	Ter, 21 às 23	Guilherme
APA	DCC001	A	65A	Ciências Exatas	5	0	Ter, 21 às 23	Guilherme
APA	DCC001	A	65AB	Opção 2º Ciclo	10	1	Ter, 21 às 23	Guilherme
APA	DCC001	A	65AC	Opção 2º Ciclo	22	5	Ter, 21 às 23	Guilherme
APA	DCC001	A	65B	Eng. Computacional	1	0	Ter, 21 às 23	Guilherme
APA	DCC001	A	99A	Disciplinas Opcionais	52	0	Ter, 21 às 23	Guilherme
Cálc. Num.	DCC008	A	09A	Física	2	2	Seg, 10 às 12	João Carlos

- Nome: permite identificar a qual disciplina uma turma pertence através de um nome;
- Turma: permite identificar uma turma dentre as várias turmas de uma mesma disciplina em um único período;
- Oferecidas: total de vagas oferecidas para uma turma;
- Ocupadas: total de vagas efetivamente ocupadas para uma turma;
- Curso: permite identificar a qual curso as informações de situação disponibilizadas em uma determinada linha pertencem;
- Situação: identifica se os alunos representados nesta linha são classificados como aprovados, reprovados por nota, reprovados por frequência, dispensados, sem conceito ou se trancaram suas matrículas;
- Total: quantidade de alunos com a situação indicada por esta linha.

Os campos nome da disciplina, código da disciplina e código permitem encontrar quais informações do segundo arquivo correspondem a quais do primeiro. Um exemplo com algumas entradas deste segundo tipo de arquivo pode ser visto na Tabela 4.2. Novamente, alguns dos dados foram abreviados para adequar a tabela ao tamanho da página.

Tabela 4.2: Entradas em arquivos do segundo tipo

Disciplina	Nome	Turma	Oferecidas	Ocupadas	Curso	Situação	Total
DCC001	APA	A	70	18	35A	Aprovado	7
DCC001	APA	A	70	18	65AB	Aprovado	1
DCC001	APA	A	70	18	65AC	Aprovado	4
DCC001	APA	A	70	18	65C	Aprovado	1
DCC001	APA	A	70	18	22A	Rep Freq	1
DCC001	APA	A	70	18	35A	Rep Freq	1
DCC001	APA	A	70	18	35A	Rep Nota	1
DCC001	APA	A	70	18	65AC	Rep Nota	1

Para este trabalho, o campo nome da disciplina foi ignorado, uma vez que o código da disciplina é suficiente para identificá-la. Os dados sobre vagas oferecidas também foram considerados irrelevantes para esta análise, uma vez que a informação de vagas ocupadas já traz a quantidade de alunos que efetivamente se matricularam em cada turma.

4.3 Preparação dos dados

Os dados dos arquivos mencionados na seção anterior foram organizados em um dataset no qual cada linha corresponde aos dados de uma turma. Para cada combinação única de código de disciplina e código de turma nos arquivos de semestres diferentes foi criada uma linha, identificando uma turma da base de dados.

Também foram criadas novas colunas para armazenar informações adicionais sobre cada turma:

- período: permite identificar a qual semestre e a qual ano uma turma pertenceu;
- vagas ocupadas: possui um somatório das vagas ocupadas por cada curso em uma turma, permitindo saber o número total de alunos;
- aprovados: guarda valores entre 0 e 1, que representam qual porcentagem dos alunos de uma turma foram aprovados ao fim do semestre.

Além disso foram criadas algumas colunas para tornar dados nominais em dados numéricos, já que a implementação utilizada não aceita dados nominais. Os dados sobre a distribuição de alunos de cada curso foram representados com uma nova coluna para cada curso presente na base de dados. Estas colunas foram preenchidas com números entre 0 e 1, representando qual porcentagem da turma é composta por alunos de cada curso. Os dados sobre os docentes foram representados com uma nova coluna para cada docente presente na base de dados. Estas colunas foram preenchidas com 1 caso aquele docente lecionasse naquela turma, e 0 caso contrário. Nos casos em que 2 docentes eram responsáveis pela mesma turma, ambos receberam o valor 1. As informações de horário foram separadas em dias e horas, e ambos foram tratados de forma similar aos dados de docentes.

Ao analisar os dados presentes neste *dataset* foi possível observar que existe um total de 1959 turmas na base de dados. Deste total, 689 turmas possuíam 4 ou menos vagas ocupadas. De forma geral turmas com menos de 5 pessoas são canceladas pelo departamento, a não ser que o docente insista em lecioná-las, ou elas sejam turmas de monografia, que sempre têm apenas 1 aluno. Além disso, grande parte destas turmas tinha aprovação de 0% ou de 100%, o que pode afetar os modelos gerados. Por isso, este grupo de turmas foi removido da base de dados.

4.4 Modelagem

Para este trabalho foram criados 2 tipos de modelos. Foram criados dois modelos de Árvore de Decisão para disciplinas específicas, um para a disciplina DCC119 (Algoritmos) e outro para a disciplina DCC 120 (Laboratório de Programação). Além disso, foi criado um modelo de Floresta Aleatória contemplando todas as turmas com 5 ou mais alunos. Nestes modelos, as turmas são divididas em 3 classes. Na primeira classe estão turmas com menos de 50% dos alunos aprovados. Na segunda classe estão turmas entre 50% e 70% de aprovação, enquanto na terceira classe estão turmas com mais de 70% de aprovação. Estes valores foram escolhidos por dividirem as turmas em 3 classes de tamanhos aproximadamente iguais, com 400 turmas na primeira classe, 444 na segunda e 426 na terceira.

Para separar a base de dados em base de treino e de teste foram usados os anos das turmas. A base de teste foi composta pelas turmas de 2019. Foram construídas 5 bases de treinamento para o modelo de Floresta Aleatória com o restante dos dados, com as bases consistindo de turmas nos intervalos fechados [2014, 2018], [2015, 2018], [2016, 2018], [2017,2018] e [2018, 2018], nos quais um intervalo do tipo [A, 2018] consiste de todos os anos entre A e 2018, incluindo os extremos, e o intervalo [2018, 2018] é composto apenas por turmas do ano 2018. Por terem quantidades menores de dados de turmas disponíveis os modelos de Árvore de Decisão foram treinados unicamente com o intervalo [2014-2018].

Para o modelo com o algoritmo de Floresta Aleatória, que engloba todas as disciplinas, foram criadas florestas com tamanhos entre 2 e 30 árvores com cada base

de treinamento. Para cada tamanho foram criados 500 modelos. Para os modelos com algoritmo de Árvore de Decisão, que incluem dados de apenas 1 disciplina, foram criadas 100 árvores com a base de treinamento.

4.5 Considerações finais

Neste capítulo foram apresentadas as primeiras etapas do desenvolvimento, seguindo os passos sugeridos pela metodologia CRISP-DM. Foram apresentadas a compreensão do negócio, explicando o contexto dos dados usados e o que se espera atingir como objetivo. Em seguida é apresentada a etapa da compreensão dos dados, explicando como a base de dados está organizada. Depois segue a etapa da preparação dos dados, explicando o que foi feito sobre a base original para chegar nos conjuntos de treinamento e de teste usados para as análises. Finalmente é explorada a etapa da modelagem, explicando quais modelos foram criados.

5 Análise de resultados

Este capítulo apresenta os resultados encontrados ao analisar os modelos de árvore de decisão e de floresta aleatória mencionados no capítulo anterior, destacando algumas das relações encontradas nestes modelos.

5.1 Conjunto de todas as disciplinas

O modelo a seguir foi construído analisando turmas de todas as disciplinas presentes na base de dados. Ele foi criado com o algoritmo de floresta aleatória e permite uma visualização das várias estruturas de árvores geradas para que se tenha uma ideia do processo de decisão que foi tomado.

5.1.1 Determinando o grupo de treinamento

Como mencionado no capítulo anterior, foram criados grupos de treinamento com as turmas a partir de 2014 a 2018. Já o grupo de teste em todos os casos foi formado pelas turmas do ano de 2019. Para escolher o melhor grupo de treinamento foram criados 500 modelos de árvore de decisão com cada conjunto. Os resultados gerados, assim como a quantidade de turmas com cada faixa de aprovação presentes em cada grupo de treinamento, são apresentados na Tabela 5.1.

Tabela 5.1: Comparação dos resultados encontrados com cada grupo de treino

Intervalo	**Ac. Média**	**Ac. Máxima**	**<50% de Apr.**	**>50% e <70% de Apr.**	**>70% de Apr.**
2014, 2018	0,4770	0,5215	601	395	610
2015, 2018	0,4545	0,4928	489	326	482
2016, 2018	0,5096	0,5502	374	255	353
2017, 2018	0,4775	0,5072	275	181	238
2018, 2018	0,4632	0,5215	152	89	132

Como é possível notar na Tabela 5.1, tanto no caso médio quanto no melhor caso, os melhores modelos encontrados foram criados com os dados do intervalo [2016, 2018]. Por isso, este é o conjunto utilizado para treinar os modelos de floresta aleatória.

5.1.2 Modelos de floresta aleatória

Com o grupo de treinamento selecionado, é possível criar os modelos de floresta aleatória. Como mencionado na Seção 4.4, foram criados modelos contendo entre 2 e 30 árvores, sendo 500 modelos para cada tamanho de floresta, gerando os resultados apresentados na Tabela 5.2. Para os testes de acurácia foram usadas as turmas do ano de 2019.

Tabela 5.2: Comparação dos resultados encontrados com cada tamanho de floresta

Quantidade de Árvores	Acurácia Média	Acurácia Máxima
2	0,4614	0,5837
3	0,4869	0,5789
4	0,5054	0,5837
5	0,5110	0,6029
6	0,5152	0,6124
7	0,5241	0,6077
8	0,5314	0,6124
9	0,5311	0,6220
10	0,5368	0,6172
11	0,5365	0,6077
12	0,5420	0,6172
13	0,5476	0,6459
14	0,5482	0,6172
15	0,5478	0,6220
16	0,5484	0,6220
17	0,5509	0,6363
18	0,5520	0,6268
19	0,5520	0,6268
20	0,5536	0,6268
21	0,5556	0,6172
22	0,5571	0,6268
23	0,5594	0,6220
24	0,5588	0,6411
25	0,5590	0,6316
26	0,5605	0,6268
27	0,5624	0,6363
28	0,5619	0,6268
29	0,5629	0,6316
30	0,5644	0,6316

Ao observar a Tabela 5.2, é possível notar que florestas com tamanhos maiores tendem a apresentar uma acurácia média melhor, por mais que em alguns casos tenha ocorrido uma pequena queda de acurácia em relação ao tamanho anterior. Observa-se ainda que a acurácia máxima se mantém relativamente estável em florestas com mais de 5 árvores, com valores máximos variando entre 0,6029 e 0,6459. Por mais que o resultado médio esteja melhorando com o tamanho das florestas, parece que os melhores modelos mantém um desempenho semelhante. Dentre os modelos gerados, o melhor deles foi um modelo com 13 árvores, com valor de semente para o gerador de 7216, e conseguindo uma acurácia de 0,6459.

Ao observar as árvores geradas na Figura 5.1, é possível notar que grande parte das decisões feitas durante a classificação são baseadas nos campos curso, docente ou

disciplina. Vale mencionar que, na Figura 5.1, caminhos à esquerda apresentam resultado verdadeiro para a verificação, enquanto caminhos à direita apresentam resultado falso. Além disso, a disciplina DCC008 é a disciplina de Cálculo Numérico, 75C, 75E e 75O são códigos de cursos de Licenciatura em Computação pela UAB nos polos de Bicas, Ilicínea e Ubá, respectivamente, 69B é o curso de Engenharia Elétrica com habilitação em Robótica Industrial, e 24A é o curso de Engenharia Civil.

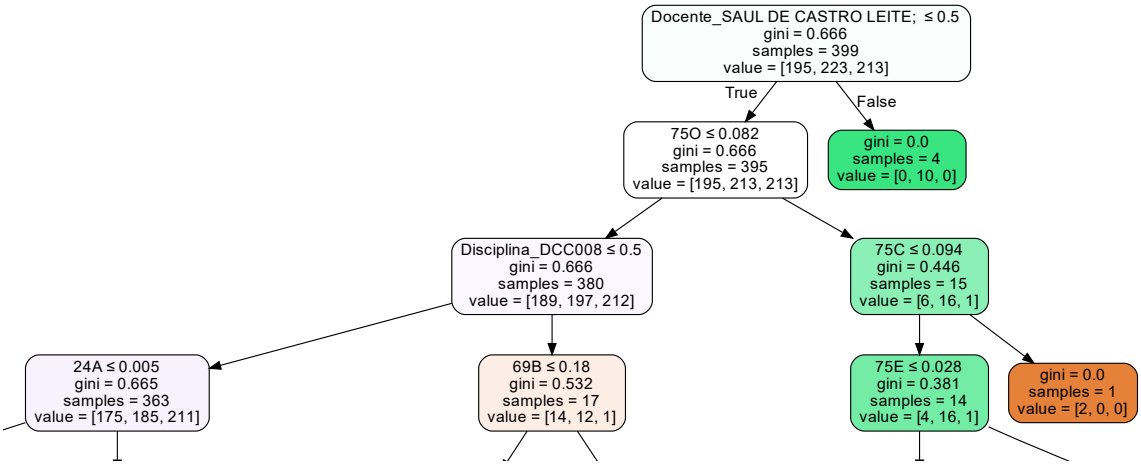

Figura 5.1: Primeiros níveis de uma das árvores geradas.

Os dados referentes ao dia e à hora da turma também aparecem nas árvores geradas, mas em geral são aplicados em grupos menores de dados. A Figura 5.2 mostra outra parte da árvore apresentada na Figura 5.1, com um trecho contendo uma decisão baseada no horário de aula e outra no dia de aula associados a uma turma. A comparação por horário foi feita em um grupo de 18 turmas, enquanto a comparação por dia foi feita em um grupo de 10 turmas. Na Figura 5.1 65A é o curso de Ciências Exatas, 65B é o curso de Engenharia Computacional, 69 é o curso de Engenharia Elétrica com habilitação em Sistemas de Potência, 76A é o curso de Sistemas de Informação e 35A é o curso noturno de Ciência da Computação.

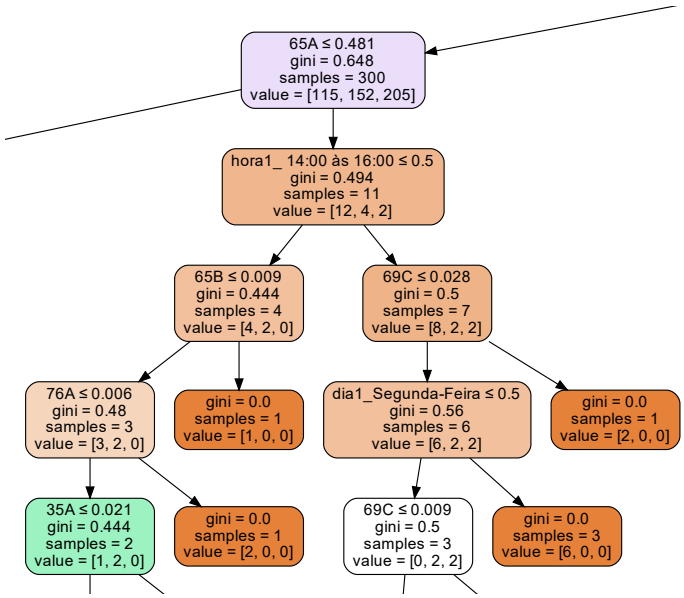

Figura 5.2: Classificações baseadas em campos de dia e horário.

Em alguns raros casos, os dados de dia e horário são usados para dividir conjuntos maiores de turmas, como pode ser visto na Figura 5.3, que mostra os primeiros níveis de outra árvore gerada pelo modelo. Neste caso, a comparação foi usada para separar um grupo de 54 turmas. Na Figura 5.3 a disciplina 2035002 é a disciplina de Análise e Projeto de Algoritmos do mestrado em Ciência da Computação, 65E é o curso de Física, 65G é o curso de Química, 76A é o curso de Sistemas de Informação e 78A é o curso de Ciências Contábeis.

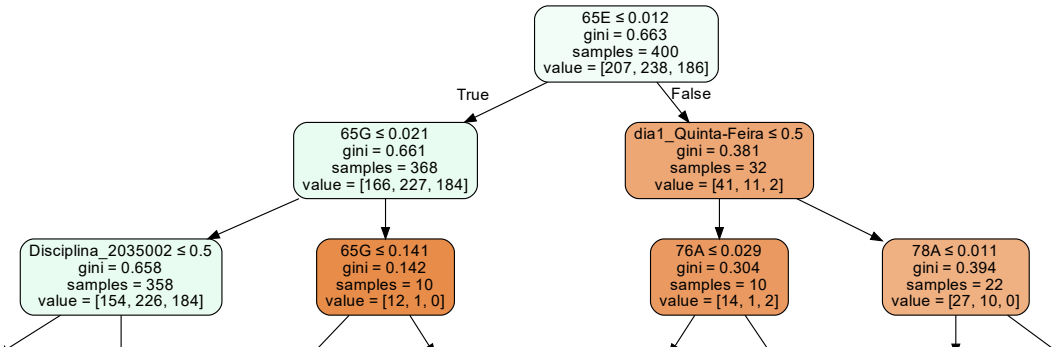

Figura 5.3: Classificação por dia envolvendo um conjunto maior de turmas.

Os dados de vagas ocupadas também aparecem em algumas comparações, mas em geral são menos utilizados que os demais valores. A Figura 5.4 mostra um trecho de uma árvore contendo uma comparação baseada no total de vagas ocupadas, que foi feita para dividir um grupo de 4 turmas. Na Figura 5.4 a disciplina EADDCC013 é a disciplina

de Seminário Integrador 1 e 75I é o curso de Licenciatura em Computação UAB no polo de Sete Lagoas.

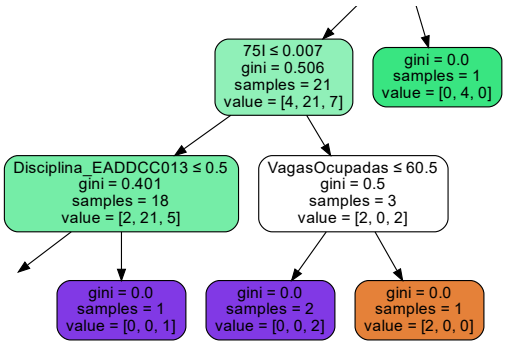

Figura 5.4: Classificações baseadas em vagas ocupadas.

Os modelos criados com o conjunto total dos dados não apresenta resultados com acurácia muito alta, mas ainda assim podem ter algumas aplicações. Eles permitem encontrar algumas tendências gerais, como quais as disciplinas ou quais os cursos com melhores ou piores rendimentos, permitindo que esforços para a melhoria do ensino sejam direcionados para estes casos de maior necessidade.

5.2 Turmas de disciplinas específicas

Os modelos a seguir foram gerados usando dados das turmas de apenas uma disciplina de cada vez, sendo elas DCC119 (Algoritmos) e DCC120 (Laboratório de Programação). Os modelos foram criados usando o algoritmo de árvore de decisão e permitem a visualização da estrutura utilizada para o processo de classificação. Como os dados para estes modelos são mais escassos, o treinamento foi feito apenas com o conjunto de turmas no intervalo [2014, 2018], e os testes foram feitos com as turmas de 2019. Para cada disciplina foram criados 100 modelos diferentes, e o melhor modelo criado para cada disciplina é apresentado a seguir.

5.2.1 DCC119 - Algoritmos

O melhor modelo encontrado para Algoritmos apresentou uma acurácia de 0,8461. Grande parte das decisões feitas foram baseadas nos cursos dos alunos matriculados e algumas

poucas no docente responsável pela turma e no número de vagas ocupadas. A Figura 5.5 mostra os primeiros níveis da árvore de decisão gerada para esta disciplina. Na Figura 5.5 65G é o curso de Química, 82A é o curso de Matemática, 65A é o curso de Ciências Exatas e 76A é o curso de Sistemas de Informação.

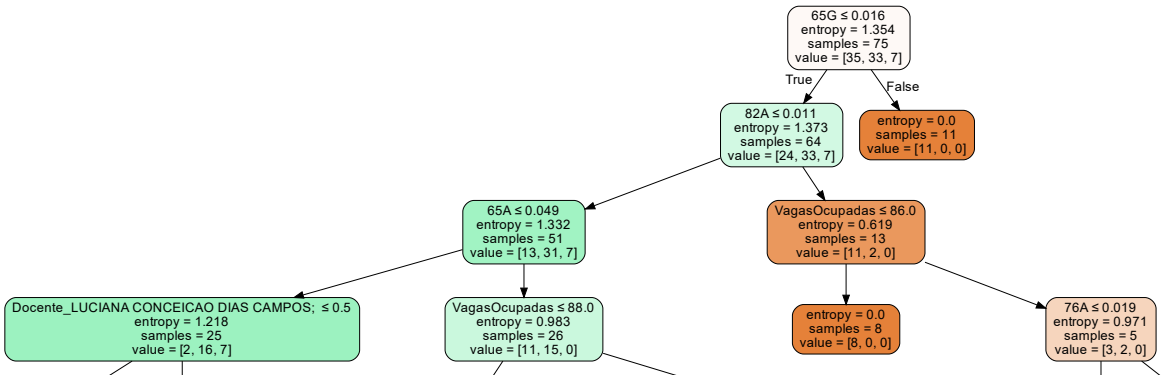

Figura 5.5: Primeiros níveis da árvore de decisão para a disciplina DCC119

O modelo gerado para Algoritmos já mostra uma acurácia maior que o gerado com o conjunto de todas as disciplinas, e permite uma visão mais detalhada dos resultados. Novamente é mostrado que o principal fator para o rendimento é o curso que está sendo analisado, e as demais variáveis do sistema não possuem uma grande importância.

5.2.2 DCC120 - Laboratório de Programação

O melhor modelo encontrado para Laboratório de Programação apresentou uma Acurácia de 0,77272. Novamente, grande parte das decisões são feitas pelos cursos associados aos alunos. Por outro lado, desta vez existe um destaque para decisões por dia, hora ou total de vagas ocupadas. A Figura 5.6 mostra os primeiros níveis da árvore de decisão gerada para esta disciplina. Na Figura 5.6 65E é o curso de Física, 67A é o curso de Engenharia Ambiental e Sanitária, 69A é o curso de Engenharia Elétrica com habilitação em Sistemas Eletrônicos e 49A é o curso de Engenharia de Produção.

O modelo gerado para Laboratório de Programação também mostra resultados com uma acurácia maior que o gerado com o conjunto de todas as turmas, e permite uma análise mais detalhada dos resultados. Neste modelo muitas decisões ainda são feitas sobre os cursos, mas existem decisões significativas feitas com outros dados do sistema também. As turmas com horário entre 10:00 e 12:00 têm alunos dos cursos de Engenharia Ambiental

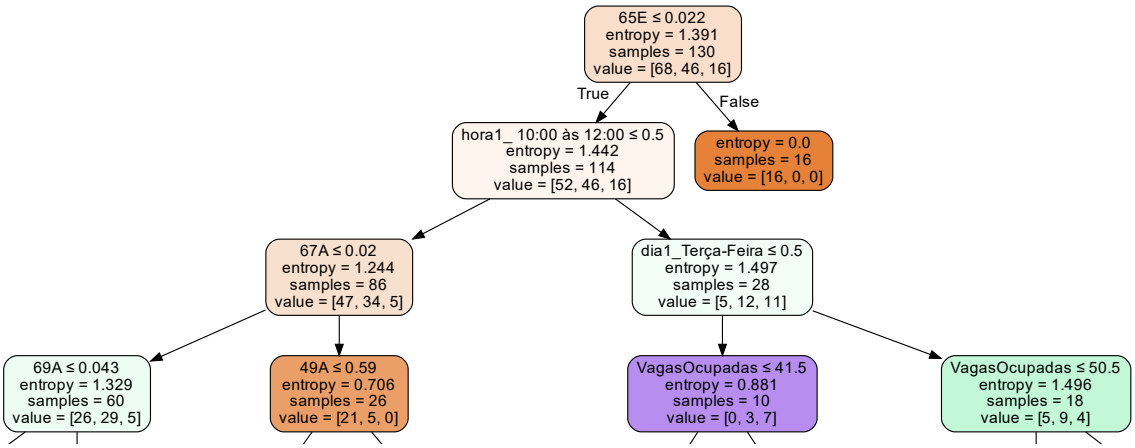

Figura 5.6: Primeiros níveis da árvore de decisão para a disciplina DCC120

e Sanitária, Engenharia Civil e algumas das habilitações de engenharia Elétrica, e pode ser que a presença de alunos desses cursos seja a verdadeira motivação para o desempenho dessas turmas, e não o horário em si.

5.3 Considerações finais

Neste capítulo foram apresentados os modelos gerados com a base de dados das turmas do DCC, apresentando algumas de suas principais características e a acurácia de cada modelo, assim como os testes e as decisões que impactaram na escolha dos grupos de teste.

É possível notar que o modelo gerado com o conjunto de todas as disciplinas não possui uma acurácia tão alta quanto os modelos gerados com dados de apenas uma disciplina específica. Isso provavelmente ocorre por conta de várias disciplinas terem poucas turmas, com muitos casos de uma ou duas turmas por ano. Ao observar o modelo de forma geral é possível notar que a grande maioria das decisões tomadas são baseadas ou nos cursos ou na disciplina de uma dada turma. Com a disciplina como um dos principais fatores determinantes para o desempenho de uma turma, a falta de dados com um determinado valor desta variável pode causar classificações de baixa acurácia no modelo. Ainda assim é possível uma visão geral da situação dos vários cursos e disciplinas, permitindo identificar quais estão mais propícios a apresentar cada classe de desempenho.

Nos modelos de disciplinas específicas, que continham quantidades mais significativas de turmas para cada disciplina, foram observadas acurácias mais altas, e algumas

decisões são feitas com variáveis que eram raras no modelo feito com o conjunto completo de dados. Isso provavelmente ocorre por este modelo não ter decisões pela disciplina, uma vez que todas as turmas possuem o mesmo valor para esta coluna, possibilitando que outros fatores tenham maior relevância para o modelo gerado.

6 Conclusão

Neste trabalho foram analisados estudos que exploram o uso de algoritmos de mineração de dados em aplicações relacionadas ao cenário educacional brasileiro, observando principalmente os tipos de estudantes avaliados e os algoritmos utilizados para a análise sobre os dados disponíveis. Os problemas revisados têm como foco dois cenários relacionados à educação, sendo eles a evasão escolar, no qual se tenta prever se o aluno faz parte de um grupo com tendências a evadir ou não, e a predição de desempenho dos alunos, no qual se tenta prever se um aluno faz parte de um grupo de bom ou mau desempenho. Os cenários observados nestes estudos englobam alunos de vários níveis, tendo casos de nível básico, médio e universitário, assim como de várias regiões do país, com estudos a níveis municipais, estaduais e nacionais.

As análises de dados nestes trabalhos foram feitas com vários algoritmos de mineração de dados, como regressão linear, regressão logística, J48, e floresta aleatória. Todos os trabalhos observados buscavam classificar estudantes em grupos, embora alguns tivessem como objetivo final encontrar o principal fator divisor destes grupos, enquanto outros tentavam apenas encontrar boas previsões para novos alunos, baseando a qualidade da sua análise em alguma métrica, com a acurácia do modelo sendo a mais comumente utilizada. Ao observar os resultados encontrados é possível notar que os estudos feitos com este tipo de algoritmo apresentaram resultados interessantes, mostrando que a mineração de dados pode ser aplicada com sucesso em vários cenários educacionais, particularmente em aplicações que envolvem a classificação de alunos em alguns grupos.

Além destas análises foi apresentado um estudo com um conjunto de dados do DCC da UFJF, buscando classificar turmas com base no seu desempenho, utilizando-se de modelos de árvore de decisão e de floresta aleatória. Diferentemente dos demais estudos observados nesta área, este projeto criou seus modelos de classificação a partir de características das turmas em geral, e não de cada aluno específico. Esta análise foi feita tanto em um escopo com turmas de todas as disciplinas do departamento de forma simultânea quanto com as disciplinas Algoritmos e Laboratório de Programação de forma

isolada, já que estas são as disciplinas com mais turmas no departamento.

Ao observar as análises feitas neste estudo foi possível notar que os fatores com maior influência no desempenho de uma turma foram os cursos que as compunham e a disciplina lecionada. Nos modelos gerados para disciplinas específicas é mais comum encontrar decisões tomadas sobre outras variáveis, mas ainda assim os cursos que fazem parte de cada turma são responsáveis pela maior parte das decisões no processo de classificação. O modelo criado com o conjunto completo de disciplinas não teve uma acurácia muito boa, chegando a 0,6459 no melhor caso. Embora este modelo não seja muito eficiente em prever o resultado de uma turma, ele ainda pode ser utilizado para visualizar disciplinas e cursos que possuem uma tendência a um melhor ou pior desempenho. Já os modelos criados baseando-se em apenas uma disciplina mostraram acurácias consideravelmente melhores, chegando a 0,8461 para Algoritmos e 0,7727 para Laboratório de Programação. Novamente, a principal informação a ser observada é acerca de quais cursos tiveram melhores ou piores desempenhos em cada uma destas disciplinas, embora seja um pouco mais comum encontrar decisões tomadas sobre outras variáveis. O modelo gerado para Laboratório de Programação, por exemplo, apresenta decisões importantes sendo tomadas com base na hora e no dia em qua a turma foi alocada.

Ao observar os resultados encontrados neste trabalho é possível visualizar quais cursos e quais disciplinas possuem melhores e piores resultados dentre as turmas do DCC, permitindo identificar em quais cenários são necessárias mudanças para atingir melhores resultados, e quais cenários podem servir como exemplo de excelência. Como trabalho futuro seria interessante buscar por mais informações sobre turmas com diversos desempenhos, analisando características como sua metodologia de ensino, por exemplo, para que sejam identificados modelos de sucesso que possam ser aproveitados em casos de menor rendimento. Finalmente, após estas mudanças serem implementadas, seria interessante um novo estudo semelhante a este, para que seja possível observar se houve uma melhoria significativa dos resultados do DCC.

Bibliografia

AMARAL, F. *Aprenda Mineração de Dados*. [S.l.]: Alta Books, 2016.

BREIMAN, L. *Machine Learning*, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001.

CALIXTO, K.; SEGUNDO, C.; GUSMÃO, R. P. de. Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. In: *Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017)*. [S.l.]: Brazilian Computer Society (Sociedade Brasileira de Computação - SBC), 2017.

CARRANO, D. et al. Combinando técnicas de mineração de dados para melhorar a detecção de indicadores de evasão universitária. In: *Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE 2019)*. [S.l.]: Brazilian Computer Society (Sociedade Brasileira de Computação - SBC), 2019.

CASTRO, L. N. de; FERRARI, D. G. *Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações*. [S.l.]: Editora Saraiva, 2016.

COLPANI, R. Mineração de dados educacionais: um estudo da evasão no ensino médio com base nos indicadores do censo escolar. *Informática na educação: teoria & prática*, Universidade Federal do Rio Grande do Sul, v. 21, n. 3, mar 2019.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. *Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações*. 2. ed. [S.l.]: Elsevier, 2015. ISBN 978-85-352-7822-4.

HASSANI, H. et al. A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Wiley, v. 9, n. 3, p. 139–154, apr 2016.

JOHN, S. et al. Realtime fraud detection in the banking sector using data mining techniques/algorithm. In: *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*. [S.l.]: IEEE, 2016.

JÚNIOR, W. M. d. S. *Mineração em dados do ENEM para a predição do desempenho acadêmico no âmbito da Rede Federal de Educação Tecnológica*. Dissertação (mathesis) — Universidade Federal de Pernambuco, jan. 2018. Disponível em: ⟨https://repositorio.ufpe.br/handle/123456789/29994⟩.

LIU, R.; TAN, A. Towards interpretable automated machine learning for stem career prediction. Zenodo, 2020.

MORSY, S.; KARYPIS, G. Will this course increase or decrease your gpa? towards grade-aware course recommendation. Zenodo, 2019.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PELAEZ, K. et al. Using a latent class forest to identify at-risk students in higher education. Zenodo, 2019.

RABELO, H. et al. Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos de EaD em ambientes virtuais de aprendizagem. In: *Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017)*. [S.l.]: Brazilian Computer Society (Sociedade Brasileira de Computação - SBC), 2017.

RAY, R. Advances in data mining: Healthcare applications. *International Research Journal of Engineering and Technology(IRJET)*, v. 5, issue 3, mar. 2018.

SALES, F. et al. Evasão no ensino básico da rede pública municipal de juiz de fora: uma análise com mineração de dados. In: *Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE 2019)*. [S.l.]: Brazilian Computer Society (Sociedade Brasileira de Computação - SBC), 2019.