INSTITUTO POLITÉCNICO DE BRAGANÇA
Escola Superior de Tecnologia e Gestão

Modelo para Classificação do Risco de Abandono Escolar em Cursos de Engenharia com Base em Métodos de *Academic Analytics*

Jhonny de Lima

Dissertação apresentada à Escola Superior de Tecnologia e Gestão para obtenção do Grau de Mestre em Sistemas de Informação

Trabalho realizado sob a orientação de:

Professor Doutor Paulo Alexandre Vara Alves

Professora Doutora Maria João Varanda Pereira

Professora Doutora Simone de Almeida

Bragança

Junho de 2018

Dedicatória

À meus pais,

Tudo lhes devo.

Agradecimentos

Primeiramente gostaria de agradecer a Deus por me guiar, iluminar e dar forças para persistir com os meus sonhos e objetivos e não desanimar diante das dificuldades.

Aos meus pais e irmão, por todo apoio que me deram em todas as fases da minha vida, fator que se revelou fundamental para a elaboração desta dissertação. Agradeço-lhes por toda a compreensão e motivação oferecidas a mim ao longo deste trabalho.

Aos meus orientadores, Professor Doutor Paulo Alexandre Vara Alves, Professora Doutora Maria João Varanda Pereira e Professora Doutora Simone de Almeida, por todo apoio e paciência que tiveram comigo. Agradeço-lhes pelas palavras de incentivo e otimismo e, acima de tudo, pelas sugestões, comentários e conselhos que contribuíram diretamente para que fosse possível atingir os objetivos desta dissertação.

Aos meus verdadeiros amigos, mesmo não havendo a necessidade de citar seus nomes, pois quem me conhece, também os conhece.

À Universidade Tecnológica Federal do Paraná e ao Instituto Politécnico de Bragança, que viabilizaram e tornaram possível a realização deste mestrado.

Resumo

O enorme aumento na quantidade de dados educacionais gerados e armazenados possibilita obter informações de extrema importância sobre o processo de ensino e aprendizagem, nomeadamente informações dos estudantes que possam estar em risco de abandono ou que necessitam de atividades específicas para aumentar seu sucesso.

Neste sentido, o presente estudo busca identificar o perfil dos estudantes de licenciatura em engenharia da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Bragança (IPB) que abandonaram seus cursos e, a partir disso, propor um modelo de classificação do risco de abandono escolar baseado na árvore de decisão C5.0.

Para a execução deste estudo, foram utilizados dois conjuntos de dados obtidos a partir da base de dados Oracle do IPB, um conjunto para os estudantes diplomados e outro para os estudantes que abandonaram seus cursos (*dropouts*), ambos para o período de 2007 a 2015.

Manipular e analisar grandes volumes de dados costuma ser um processo lento e complexo e, por esse motivo, foram adotados métodos e técnicas de *Academic Analytics*, em busca de compreender os dados utilizados neste estudo.

Os principais resultados permitem concluir que a maior parte dos *dropouts* ocorre antes mesmo de os estudantes estarem registrados por um ano em seus cursos. Em muitos casos, os estudantes que abandonam possuem um rendimento semelhante ao dos diplomados, o que pode indicar a existência de fatores externos ao IPB que favorecem ao abandono. O número total de *dropouts* em cada curso é no mínimo 30% dentro do intervalo de 9 anos. 25% dos estudantes abandonaram seus cursos sem ter cursado ao menos uma disciplina.

A criação do modelo baseado em árvore de decisão tem como principais objetivos a geração de instruções padronizadas, facilidade de interpretação e permitir a adição de vários cenários possíveis, contribuindo com diversas vantagens ao processo de tomada de decisão.

Palavras Chave: Ensino Superior; Abandono Escolar; Engenharia; Mineração de Dados; *Academic Analytics*.

Abstract

The enormous increase in the amount of educational data generated and stored enables us to obtain extremely important information about the teaching and learning process, namely information from students who may be at risk of dropout or who need specific activities to increase their success.

In this sense, the present study aims to identify the profile of the engineering students of the Polytechnic Institute of Bragança (IPB) who dropped out and, from this, propose a risk classification model of school drop-out based on the decision tree C5.0.

For the execution of this study, we used two data sets obtained from the IPB Oracle database, one for the graduate students and another for the dropout students, both for the period from 2007 to 2015.

Manipulating and analyzing large volumes of data is often a slow and complex process and, for this reason, Academic Analytics methods and techniques have been adopted in order to understand the data used in this study.

The main results allow us to conclude that most dropouts occur before students are enrolled for one year in their courses. In many cases, students who drop out have a similar performance to the graduates, which may indicate the existence of factors external to the IPB that favor dropout. The total number of dropouts in each course is at least 30% over the 9 years. 25% of students dropped out without attend at least one subject.

The creation of the model based on decision tree has as main objectives the generation of standardized instructions, easy interpretation and allow the addition of several possible outcomes, contributing with several advantages to the decision-making process.

Keywords: Higher Education; Dropout; Engineering; Data Mining; Academic Analytics.

Índice Geral

Lista de Siglas/Abreviaturas

AA – *Academic Analytics*

AG – Algoritmo Genético

BD – Base de Dados

BDR – Base de Dados Relacional

BI – *Business Intelligence*

CRAN – *Comprehensive R Archive Network*

DCBD – Descoberta de Conhecimento em Bases de Dados

DM – *Data Mining* (Mineração de Dados)

DW – *Data Warehouse*

ECTS – *European Credit Transfer System* (Sistema Europeu de Transferência de Créditos Acadêmicos)

EDM – *Educational Data Mining*

ESTiG – Escola Superior de Tecnologia e de Gestão

ETL – *Extract, Transform and Load* (Extração, Transformação e Carregamento)

IES – Instituições de Ensino Superior

IPB – Instituto Politécnico de Bragança

K-NN – *K-Nearest Neighbor*

NoSQL – *Not only SQL*

OLAP – *Online Analytical Processing* (Processamento Analítico Online)

OLTP – *Online Transaction Processing* (Processamento de Transações em Tempo Real)

PB – Processo de Bolonha

RB – Redes Bayesianas

RNA – Rede Neural Artificial

SGBD – Sistema Gerenciador de Base de Dados

SGBDR – Sistema Gerenciador de Base de Dados Relacional

SI – Sistemas de Informação

SQL – *Standard Query Language*

TIC – Tecnologia da Informação e Comunicação

Índice de Figuras

Índice de Tabelas

Índice de Listagens

Capítulo 1 Introdução

1.1. Motivação

Abandono escolar e problemas de desempenho são questões de grande preocupação para as Instituições de Ensino Superior, uma vez que o aumento do número de fatores internos e externos às instituições vêm contribuindo significantemente para tais problemas.

O enorme aumento na quantidade de dados educacionais gerados e armazenados nas bases de dados das instituições de ensino superior possibilita obter informações valiosas sobre o processo de ensino e aprendizagem, dentre elas, informações dos estudantes que podem estar em risco de abandono ou que necessitam de atividades específicas para aumentar seu sucesso. Dessa forma, nota-se que a análise de dados é um fator primordial para se compreender com precisão a situação de cada estudante e, poder escolher adequadamente a melhor abordagem para se prosseguir.

Lidar com grandes volumes de dados costuma ser um processo demorado e complexo. A área de *Data Mining* auxilia no processo de descoberta de conhecimento por meio de seus diversos algoritmos e ferramentas que processam esses dados de forma a buscar correlações importantes entre eles.

Neste sentido, a busca em como compreender melhor os dados gerados pelos estudantes, como predizer o comportamento dos mesmos, e como melhorar a forma de ensino e aprendizagem, tornam a área de *Educational Data Mining* essencial nos dias de hoje.

1.2. Objetivos

Este trabalho busca identificar padrões relativos ao desempenho escolar dos estudantes dos cursos de engenharia da Escola Superior de Tecnologia e Gestão (ESTiG) do Instituto Politécnico de Bragança (IPB). O objetivo é descobrir os motivos que levaram e podem levar os estudantes ao abandono escolar e, dessa forma, definir o perfil dos estudantes que possam estar em risco de abandono. Para atingir tal objetivo, haverá a necessidade de se recorrer a técnicas de *Educational Data Mining* e *Academic Analytics*.

Para dar resposta ao principal objetivo do presente trabalho, teve-se como objeto de estudo duas bases de dados, uma para os estudantes que abandonaram os cursos (745 estudantes), e outra para os diplomados (1099 estudantes), ambas para o período de 2007 a 2015, totalizando uma amostra de 1844 estudantes. Para tanto, este trabalho busca cumprir aos seguintes objetivos específicos:

- Caracterizar os cursos quanto à sua taxa de abandono;
- Analisar a evolução do abandono em cada curso ao longo dos 9 anos;
- Identificar os fatores que mais influenciam o abandono escolar;
- Identificar as disciplinas com maior taxa de reprovação entre os estudantes que abandonaram;
- Caracterizar o perfil dos estudantes que abandonaram quanto ao número de disciplinas feitas, quais as disciplinas com maior índice de reprovação, número de disciplinas até o abandono.

Diante disso, pretende-se alertar os professores e diretores de curso para a necessidade de definir estratégias que visem a redução do abandono escolar, nomeadamente prestar um maior apoio aos estudantes que se encontram em risco, ajudando-os a ultrapassar dificuldades e a gerir de uma forma eficiente o seu percurso acadêmico.

Deste modo, este trabalho encontra-se estruturado em 4 principais pontos. O primeiro é uma introdução ao ensino superior em Portugal e à problemática do abandono escolar; o segundo apresenta os métodos, ferramentas, algoritmos e metodologia utilizada por meio da revisão da literatura; no terceiro ponto serão apresentados e analisados os resultados do estudo; e por fim, apresentam-se as principais conclusões e perspectivas para trabalhos futuros.

Capítulo 2 Abandono Escolar no Ensino Superior

2.1. Ensino Superior em Portugal

O ensino superior português é organizado em universitário e politécnico e é ministrado em instituições públicas e privadas. O ensino universitário é guiado por uma ótica de promoção da investigação e de criação do saber, enquanto o politécnico, é norteado por uma perspectiva de investigação aplicada e de desenvolvimento de projetos [1].

De modo a implementar o Processo de Bolonha, em 2005 foi iniciado um processo de reforma da Lei de Bases do Sistema Educativo, em que foi introduzido o *European Credit Transfer System* (ECTS) nos ciclos de estudo, mecanismos de mobilidade, entre outros [1].

O Processo de Bolonha (PB) pode ser considerado um modelo, um paradigma de como as relações entre o mercado passaram e continuam a passar por grandes mudanças. De forma geral, adota-se o modelo 3+2+3, isto é, para a maioria dos cursos predominam três anos de licenciatura, dois de mestrado e três de doutoramento [2]. Esta estrutura foi introduzida em 2006 e totalmente implementada, em Portugal, a partir do ano letivo de 2009/2010 [1].

Os compromissos resultantes do PB acarretaram alterações na organização e no desenvolvimento do currículo das Instituições de Ensino Superior (IES), consequências que correspondem, para muitos, a uma mudança de paradigma, como é apresentado no Decreto-Lei nº 74 [3]:

> *Questão central no Processo de Bolonha é o da mudança de paradigma de ensino de um modelo passivo, baseado na aquisição de conhecimentos, para um modelo baseado no desenvolvimento de competências, onde se incluem quer as de natureza genérica – instrumentais, interpessoais e sistémicas – quer as de natureza específica associadas à área de formação, e onde a componente experimental e de projeto desempenham um papel importante. Identificar as competências, desenvolver as metodologias adequadas à sua concretização, colocar o novo modelo de ensino em prática são os desafios com que se confrontam as Instituições de Ensino Superior.*

Não somente em Portugal, mas o ensino em toda a Europa foi e continua sendo organizado de forma a facilitar a mobilidade e a equalização de todos os estudantes, diplomados e servidores de ensino superior. Algumas das principais exigências para que essa mobilização possa ocorrer é o reconhecimento mútuo de graus e de outras qualificações de ensino superior, organização numa estrutura de três ciclos e a cooperação europeia na garantia da qualidade [4].

2.2. Abandono Escolar

Sabe-se que a educação é responsável por uma grande parte do desenvolvimento social, ou seja, oferece sustentabilidade para uma sociedade que deseja evoluir de maneira intelectual, econômica, humana e estrutural [5].

Em Portugal, o acesso de diferentes públicos ao ensino superior colocou às IES novos desafios e novas responsabilidades, em especial a de todos assegurar condições de igualdade de oportunidades de permanência e de sucesso acadêmico. Neste novo cenário, as IES tiveram de saber responder às exigências derivadas do PB, de modo especial, ao nível da reorganização curricular dos cursos, aos resultados de aprendizagem e das metodologias de ensino e de avaliação [6].

Em 2008, a União Europeia adotou um valor referência para o ensino superior, afirmando que, até 2020, pelo menos 40% da população adulta na faixa etária dos 30 a 34 anos deverão possuir um diploma de nível superior ou equivalente, encorajando grande parte da sociedade a ingressar no ensino superior e reduzindo as taxas de abandono [7].

O termo abandono escolar além de permitir diversas interpretações, pode ser aplicado em diversos tipos de contextos com significados levemente distintos. Em certos casos considera-se como abandono o desligamento do curso pelo estudante, independentemente da frequência do mesmo, ou seja, compreende-se como abandono a simples suspensão do vínculo entre o estudante e a instituição antes do término do processo para a conclusão do curso [8].

Em outros casos, pode-se diferenciar o abandono escolar segundo períodos médios para a conclusão do curso. Contudo, também se considera como abandono as ocorrências de interrupção do ciclo de estudos, independentemente do nível em que se localize os estudos, bem como a duração de tal interrupção. Alguns autores definem abandono como a desistência definitiva após certo contato com o curso, onde termos como perda ou fuga são, muitas vezes, considerados como sinônimos do termo abandono [8].

Sabe-se que a transição para o ensino superior é uma etapa crucial na vida acadêmica dos estudantes, não se limitando, assim, a uma simples transição de ano letivo, já que pode significar um período crítico para o ajuste e desenvolvimento acadêmico dos estudantes [6]. De acordo com Cunha e Carrilho [9], é preciso olhar o estudante de forma diferenciada e acolhedora, em especial no momento do seu ingresso, por ser o primeiro ano de graduação um período crítico para o seu ajuste e desenvolvimento acadêmico.

As IES de vários países identificam os estudantes em risco e tentam implementar processos que os possam ajudar a concluir com sucesso os seus estudos e, assim, evitar o abandono [10].

Tanto o abandono como os problemas de desempenho, apesar do interesse que têm vindo a adquirir, não são problemáticas novas na sociedade portuguesa, e são aspectos que necessitam de grande atenção na educação, especialmente em modalidades de cursos que possuem como características inerentes a distância física entre professores e estudantes e a comunicação por meio de recursos tecnológicos [11].

Segundo Gibson [12], existem três categorias de fatores que podem identificar os motivos de abandono do curso:

- Fatores do estudante: motivação e vocação, autoconfiança, além da preparação educacional anterior;

- Fatores situacionais: apoio e suporte familiar e do empregador, além de situações da vida pessoal;
- Fatores do sistema educacional: metodologia e qualidade do ensino, condições das infraestruturas, como com o auxílio oferecido pela instituição.

De acordo com Duran e Gaioso, citado por Fernandes et al. [13], as causas mais frequentes para o abandono no ensino superior estão relacionadas à fatores descritos na Tabela 1.

Tabela 1: Categorias de causas de abandono (adaptado de Fernandes [13]).

Categoria	**Causa**	**Exemplos**
Psicológica	Comportamento do indivíduo	Reprovações sucessivas, falta de referencial familiar e imaturidade
Sociológica	Influência do meio social	Falta de orientação vocacional, deficiência da educação básica e imposição familiar
Organizacional	Influência da instituição	Desconhecimento da metodologia do curso, concorrência de outras IES e corpo docente
Interacional	Interação com outras pessoas	Ausência de laços afetivos com as IES, mudança de endereço e exclusão social
Econômica	Relação econômico-financeira	Busca de herança profissional, falta de perspectiva profissional e problemas financeiros

Já para Quinn [14], são seis os fatores-chave que levam os estudantes a abandonarem os cursos, são eles: sócio-culturais, estruturais, políticos, institucionais, pessoais e de aprendizagem. O mesmo autor afirma que todos esses fatores estão inter-relacionados, por exemplo, fatores pessoais, como trabalhar durante os estudos, são determinados por fatores estruturais, como a pobreza.

Apesar das IES tratarem o abandono de estudantes a partir de perspectivas diferentes, todas partilham do mesmo objetivo que mira reduzir as taxas de abandono. As diferentes abordagens buscam tratar questões inerentes às instituições ou elaborar estratégias orientadas aos estudantes. As que se focam nas instituições procuram descobrir uma forma de motivá-las a reduzir a taxa de abandono [10].

De acordo com Benavente et al. [15], o perfil do estudante em risco de abandono demonstra, em geral, um atraso escolar significativo, ausência de ambições escolares, ausência de interesse pela escola, pelas disciplinas e pelas aulas. O estudante em risco é,

em geral, mais velho que os outros colegas do mesmo grau de ensino, não se sente apoiado pela família, vive em um meio familiar intelectualmente desfavorecido e tem, claro, um rendimento escolar insuficiente. O mesmo autor também afirma que: "Os estudantes que abandonam têm problemas com a escola e foram já por ela abandonados, em muitos casos. Só ocasionalmente se encontra um bom estudante, entusiasmado, com projetos escolares, que renuncia à escola".

Para muitos jovens, a entrada no ensino superior está longe de ser um ponto de chegada; para alguns ela corresponde a um ponto de partida para um percurso acadêmico diferente [16]. Portanto, entende-se que o abandono escolar é um fenômeno complexo sendo influenciado por diversas variáveis, as quais compreendem fatores internos e externos às IES.

Capítulo 3 Evolução dos Sistemas de Apoio à Decisão

3.1. Introdução aos Sistemas de Apoio à Decisão

Tal como ocorre com atividades de comércio, entretenimento e comunicação, cada vez mais o processo educacional tem utilizado novas tecnologias como por exemplo as plataformas de apoio ao processo de aprendizagem. Atualmente as IES enfrentam uma avalanche de informações, seja sobre os estudantes ou sobre os seus demais assuntos acadêmicos, tornando visível a necessidade de tratar e administrar este volume de informações por meio de sistemas de informações (SI) eficazes.

Segundo Benson e Standing [17], um sistema é basicamente "uma coleção de partes que trabalham juntas para alcançar algum propósito". Outros consideram um sistema como um conjunto de ideias relacionadas e, que possuem três principais características: propósito ou função, contexto ou um ambiente em que ele possui aplicabilidade e normalmente entradas e saídas.

Para entender o que são sistemas de informação, é necessário antes introduzir alguns termos e estabelecer alguns conceitos, como [17]:

- Dados: são as entradas brutas dos SI;
- Informações: são os dados processados. Da perspectiva de entrada e saída, dados são inseridos, processados de alguma maneira e depois demonstrados;
- Conhecimento: o conhecimento é uma coisa humana, de natureza subjetiva e baseado na experiência. É basicamente entender o que a informação significa ou implica.

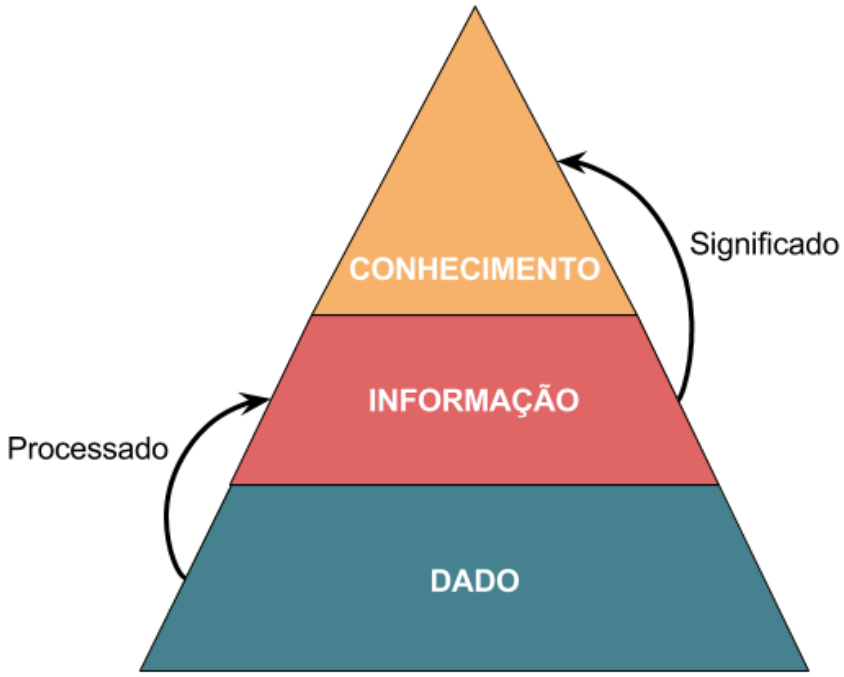

Figura 1: O relacionamento entre dado, informação e conhecimento (reproduzido de Benson e Standing [17]).

Basicamente as principais características dos SI são a geração e gerenciamento de informações. Dito isso, pode-se pensar na informação como um recurso ativo ou estratégico, onde quanto mais atualizada e precisa a mesma, maior será a vantagem.

3.2. *Business Intelligence*

O termo *Business Intelligence* (BI) já não é um conceito precisamente recente, sendo o Gartner Group responsável pela atribuição desta nomenclatura na década de 80. A árdua e intensa programação, indicavam altos custos no desenvolvimento da BI na época de seu surgimento. Além disso, com a chegada das interfaces gráficas e da fixação da arquitetura cliente/servidor, foram disponibilizados no mercado produtos voltados às análises de negócios, de maneira acessível e amigável aos gestores [18].

Para Santos e Ramos, citado por Alves [19], os sistemas de BI combinam um conjunto de ferramentas que são capazes de gerar relatórios para disponibilizar informações relevantes a partir dos dados obtidos nas organizações, informações estas, que serão utilizadas pela gestão de topo para a tomada de decisão.

Por sua vez, o Gartner Group [20], descreve BI como um conceito abrangente que contém aplicações, infraestruturas, ferramentas e boas práticas que possibilitam o acesso e análise de informação para melhorar e otimizar o desempenho das organizações e instituições, bem como as decisões.

Atualmente, é grande a quantidade de produtos de BI disponíveis no mercado, estes vão desde pacotes configuráveis até ferramentas isoladas. De qualquer modo, essas ferramentas possuem como principal característica a transformação dos dados em informações que possibilitarão à organização a melhor tomada de decisão.

A BI oferece suporte à tomada de decisão pois permite observar de forma clara e objetiva, o contexto de determinada organização diante do mercado ou sociedade. BI incorpora elementos de *data mining*, previsão, otimização, processamentos analíticos online (OLAP), *data warehouse* e gerenciamento do conhecimento [18].

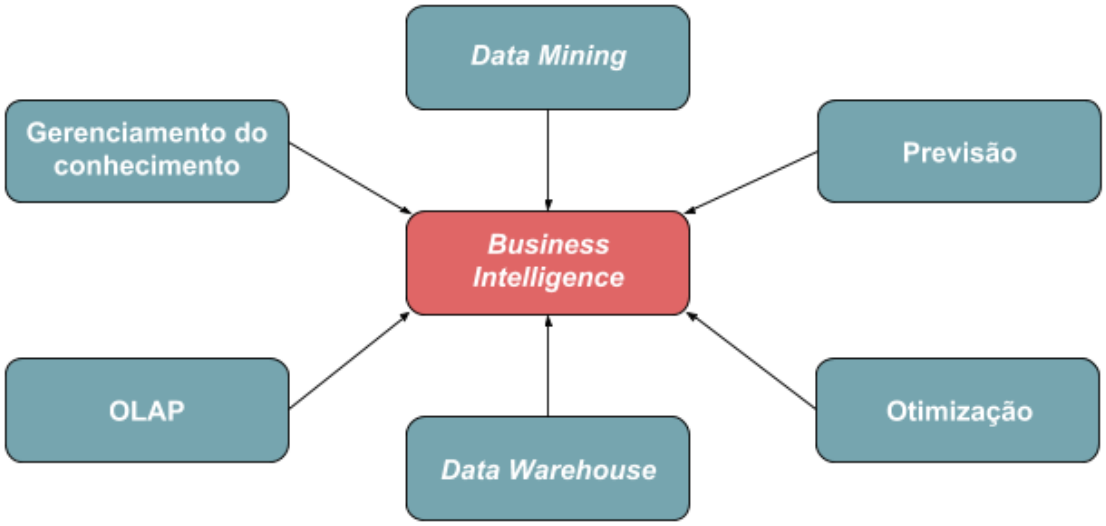

Figura 2: Relação da BI com outros Sistemas de Informação.

Como pode-se observar na Figura 2, a BI é um conceito amplo e versátil que suporta uma grande quantidade de sistemas e tecnologias, cumprindo assim, com os requisitos e necessidades do meio em que está a ser implementado tornando-o mais eficiente e útil.

3.3. *Data Mining*

Com a enorme difusão do uso de sistemas informatizados em inúmeras áreas, aumenta a cada dia o volume de dados gerados e armazenados em bases digitais, necessitando assim, da utilização de maneiras de se explorar e analisar tais bases. Segundo Witten e Frank [21], bases nas áreas de comércio, finanças, saúde, educação e ciência, por exemplo, têm sido muito valorizadas e, com isso, esforços vêm sendo aplicados para analisá-las, visto que diversos novos padrões relevantes, a partir dos dados disponíveis, podem ser identificados.

O termo "*data mining*" foi usado de maneira igualmente crítica pelo economista Michael Lovell em um artigo publicado na *Review of Economic Studies*, de 1983. *Data Mining* (DM) ou Mineração de Dados, é uma área disciplinar que surgiu da convergência de três áreas: estatística clássica, inteligência artificial e bases de dados. Esta pode ser entendida como parte de um processo maior conhecido como Descoberta de Conhecimento em Bases de Dados (DCBD), do inglês *Knowledge Discovery in Databases* (KDD), cujo principal objetivo, é a obtenção não trivial de conhecimento implícito em bases de dados, previamente desconhecido, por meio da utilização de algoritmos específicos para a extração de padrões destas bases [8]. Tal conceito pode ser ressaltado por Fayyad et al. [22], ao afirmar que KDD é "o processo não trivial de identificar padrões válidos, desconhecidos, potencialmente úteis e, ao final, compreensíveis em dados".

Segundo Sferra e Corrêa [23], DM pode ser compreendido como o processo automatizado de extração e análise de informações implícitas de uma grande base de dados, tendo como objetivo representar características do passado e predizer tendências para o futuro e, com isso, aplicar tais decisões em diversas áreas que utilizam o conhecimento como base, como instituições de pesquisa, empresas e indústrias.

Dito isso, as técnicas de DM são utilizadas em grandes volumes de dados, tendo como objetivo descobrir padrões escondidos e relacionamentos que possam ser úteis na tomada de decisões. A Figura 3 exibe a sequência de passos utilizados no processo de extração do conhecimento de dados.

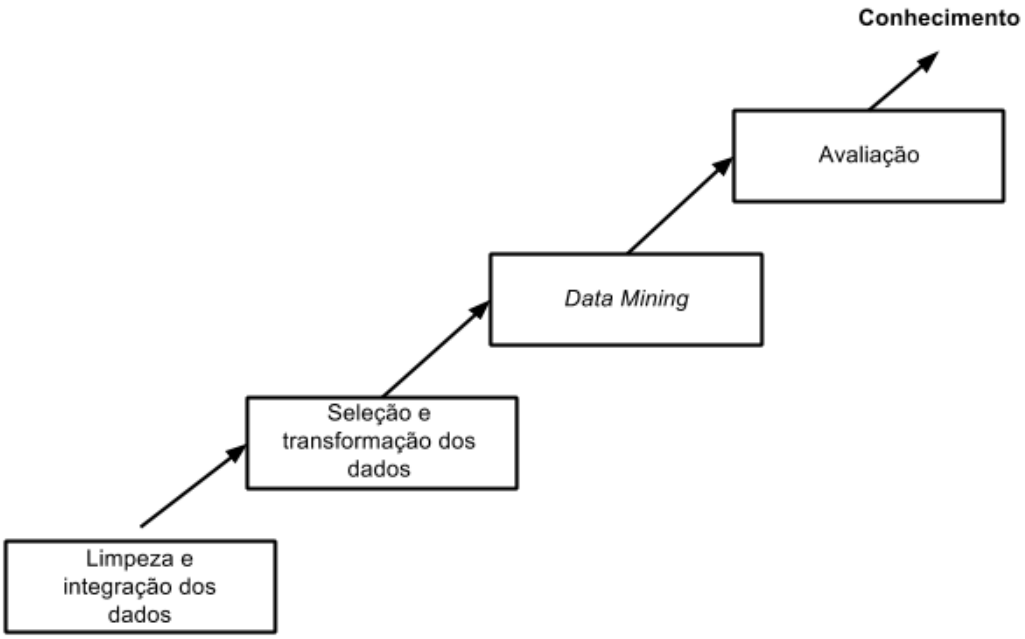

Figura 3: Etapas da extração do conhecimento de dados (adaptado de Ahmed e Elaraby [24]).

As funções ou técnicas de DM geralmente se enquadram em duas categorias de aprendizagem: supervisionada e não supervisionada. As noções de aprendizagem supervisionada e não supervisionada são derivadas do *Machine Learning* (Aprendizagem de Máquina), que é conhecida como uma subárea da inteligência artificial. A *Machine Learning* lida com técnicas que permitem que os dispositivos aprendam com seu próprio desempenho e modifiquem seu próprio funcionamento.

A construção de um modelo supervisionado envolve treinamento, processo pelo qual o algoritmo analisa muitos casos em que o valor alvo já é conhecido. No processo de treinamento, o modelo "aprende" a lógica para fazer a previsão. A aprendizagem supervisionada geralmente resulta em modelos preditivos, enquanto que, a não supervisionada tem como objetivo a detecção de padrões [25].

Em contraste à aprendizagem supervisionada, na aprendizagem não supervisionada não existe distinção entre atributos dependentes e independentes, dessa forma, não há nenhum resultado conhecido anteriormente para guiar o algoritmo na construção do modelo [25]. A aprendizagem não supervisionada nos permite abordar problemas com pouca ou nenhuma ideia do que nossos resultados devem se parecer.

Diversos algoritmos e técnicas como classificação, agrupamento, associação, regressão, árvores de decisão, algoritmo genético, etc., são utilizados no processo de descoberta de

conhecimento em bases de dados [24]. Contudo, sistemas educativos apresentam características que requerem um tratamento diferente do problema de mineração, exigindo diferentes métodos. Consequentemente, é necessário a utilização de técnicas específicas de DM que possam ser aplicadas na aprendizagem e a outros tipos de dados sobre estudantes, surgindo assim a *Educational Data Mining* (capítulo 4).

3.4. *Data Warehouse*

A história do *Data Warehouse* (DW) surge com a evolução da informação e dos sistemas de suporte à decisão. Bill Inmon definiu na década de 1990, um modelo conhecido como *Data Warehouse* que, segundo ele, é uma coleção de dados baseada em assuntos, integrada, variável em relação ao tempo e não volátil. Baseada em assuntos, significa que o DW é identificado ou desenvolvido fundamentado no principal tema de um determinado ambiente.

De acordo com Lima [26], diversos autores definiram DW segundo Inmon. DW é uma coleção de dados que:

- é baseada em assuntos: os dados e a informação devem ser apresentados organizados por assuntos, seguindo as necessidades dos usuários finais;
- é integrada: o DW deve ser uma extensa e única fonte de informação para e sobre o negócio;
- varia em relação ao tempo: o DW possui o histórico das informações, o que possibilita a análise da evolução histórica com diversas linhas temporais;
- não é volátil: o DW contém informações estáveis. Após inseridos no DW, os dados não são apagados.

Conforme Das e Mohapatro [27], um DW é um sistema de bases de dados relacional utilizado para armazenar, analisar e reportar funções. Um DW concentra seu foco no armazenamento de dados. Os dados de fontes diferentes são limpos, transformados e carregados em um *warehouse*, de modo que estejam disponíveis para *data mining* e funções analíticas *on-line*. O DW é um sistema de bases de dados baseado em SQL (*Standard Query Language*) e, as duas principais abordagens para se armazenar dados em um DW são as seguintes:

- Dimensional: os dados da transação são divididos em tabelas “fato”, que contém as informações de referência que dão contexto aos fatos;
- Normalizado: as tabelas são organizadas de acordo com as categorias de dados, tais como dados sobre produtos, clientes e assim por diante. Uma estrutura normalizada separa os dados em entidades, que criam várias tabelas em uma base de dados relacional.

Além de uma base de dados relacional, um ambiente de DW inclui uma solução de extração, transporte, transformação e carregamento (ETL), um mecanismo de processamento analítico *online* (OLAP), ferramentas de análise de clientes e outras aplicações que gerenciam o processo de coleta de dados [28].

A Figura 4 exibe uma arquitetura básica para um DW. Os usuários finais podem acessar diretamente dados derivados de vários sistemas de origem por meio do DW.

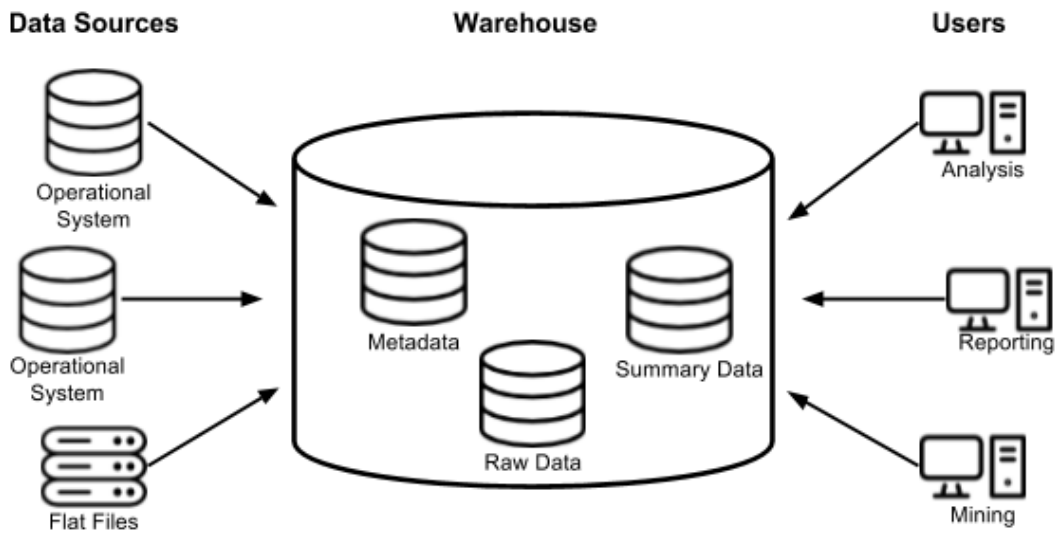

Figura 4: Arquitetura básica de um DW (adaptado de Oracle [28]).

Portanto, o DW proporciona um ambiente confiável para o processo de tomada de decisão, facilitando a aplicação de técnicas estatísticas e análises para identificar relações, que a primeira vista podem parecer ocultas.

3.5. *Big Data*

Big Data surgiu em 2011 com o objetivo de representar novas e incomuns fontes de dados (redes sociais, por exemplo), suportadas por tecnologias avançadas e combinações raras de habilidades do usuário. Alguns autores preveem que a popularidade de *big data* cairá tão rápido quanto ele emergiu, devido à sua complexidade e na escassez de trabalhadores qualificados [29]. Por outro lado, nota-se que a utilização de técnicas de *big data* tem se tornado uma tendência cada vez maior em diversas áreas de investigação.

A priori, *big data* surgiu como consequência de um conjunto de problemas gerados por excessivas quantidades de dados produzidas por empresas e organizações. Mais tarde, o termo passou a descrever as soluções que buscam tratar tais dados, desde o armazenamento até a transformação em informações relevantes nas tomadas de decisões.

Dito isso, o conceito de *big data* está relacionado à capacidade de processar e analisar grandes volumes de informações que permitam a extração de conhecimento para melhorar o processo de tomada de decisão [30]. No ambiente acadêmico, as técnicas relacionadas à *big data* podem melhorar os processos de avaliação, *feedback* e entrega do conteúdo. Tal prática possibilita obter diversas informações a respeito dos estudantes e sobre a sua interação com os conteúdos disponibilizados, ambientes de aprendizagem e sobre o processo de avaliação [31].

> *Big data é mais do que uma simples questão de tamanho, é uma oportunidade de encontrar insights em novos e emergentes tipos de dados e conteúdos, para tornar seu negócio mais ágil e para responder a perguntas que foram anteriormente consideradas fora de seu alcance* [32].

Na educação este conceito pode auxiliar a responder complexas e importantes questões, como: O que um estudante sabe? O que ele não sabe? Até que ponto um conteúdo ou uma estratégia pedagógica pode funcionar para um grupo de estudantes? A falha do estudante ocorreu porque ele não aprendeu ou ela se deveu à distração, esquecimento ou a uma pergunta mal formulada? Como aumentar as chances de um estudante obter sucesso em uma avaliação? [31]

Para Ferreira [33], os principais tipos de informação que têm obtido o foco na coleta e análise das aplicações que processam grandes volumes de dados educacionais dizem respeito aos:

- Dados de identidade dos usuários: quem são eles, de onde fazem o acesso e quais são suas permissões;
- Dados de interação do usuário com o ambiente de aprendizagem: basicamente se refere à experiência que os estudantes tiveram ao utilizar o ambiente;
- Dados inferidos sobre conteúdo: refere-se às evidências que sejam capazes de apontar o nível existente na relação entre um conteúdo escolar e os ganhos de aprendizagem de estudantes;
- Dados relacionados ao sistema: em sua maior parte se refere aos dados que já são coletados pelos mecanismos da instituição, como notas, registros disciplinares e de assiduidade;
- Dados inferidos sobre os estudantes: são os dados mais difíceis de se obter, porque dizem respeito ao comportamento do estudante perante o processo de aprendizagem, que pode ser afetado em diferentes momentos e por diferentes fatores.

Dispor de mais conhecimentos sobre o processo de aprendizagem é essencial, uma vez que com a análise de grandes volumes de dados as instituições de ensino podem melhorar seus modelos de gestão de aprendizagem, construção de novas práticas de ensino, auxílio na atuação de educadores e, consequentemente, favorecer melhorias nos ganhos de aprendizagem [31].

Capítulo 4 Ciência de Dados Educacionais

4.1. Introdução à Ciência de Dados Educacionais

De acordo com Souza [34], o grande crescimento das bases de dados em decorrência do armazenamento de informações referentes a registros de acessos (*logs*), conversas entre estudantes e professores, atividades, criação de fóruns, etc., proporcionou aos pesquisadores a oportunidade de obter informações que pudessem ser pertinentes ao processo de ensino, tendo como intuito compreender adequadamente os fatores que podem estar relacionados ao sucesso ou fracasso de cursos, alunos e práticas pedagógicas.

Apesar do elevado número de dados a respeito dos estudantes em suas bases de dados, as IES têm sido ineficientes no uso dos mesmos. Tal problema ocorre devido ao constante crescimento na quantidade de dados, o que torna necessário a criação de novas técnicas e abordagens para entender os padrões existentes [8].

De maneira geral, a análise de dados educacionais representa uma área de pesquisa emergente para o desenvolvimento de métodos que exploram dados provenientes de ambientes educacionais e também administrativos com o objetivo de compreender melhor os estudantes e os cenários em que eles aprendem [35].

4.2. *Educational Data Mining*

A evolução da Tecnologia da Informação e Comunicação (TIC) contribuiu substancialmente para o aumento da quantidade de dados gerados e disponibilizados. Considerando-se as limitações de processamento ou demais restrições associadas, atualmente a capacidade de geração de dados é significativamente maior do que a capacidade de se analisar os mesmos.

É possível notar o mesmo cenário de ampla geração de conjuntos de dados diversos na área da educação. Devido a vasta difusão do uso de sistemas informatizados nas escolas e universidades, torna-se cada vez maior o volume de dados gerados e armazenados em bases de dados. Como principais exemplos disso, tem-se a consolidação da modalidade de ensino a distância e do *blended learning*, além da incorporação de sistemas integrados de gestão em instituições educacionais [8].

A *Educational Data Mining* (EDM) é uma área de investigação ainda em ascensão que busca desenvolver ou adaptar métodos e algoritmos de mineração já existentes, permitindo compreender de forma mais clara os dados em contextos educacionais, produzidos especialmente por estudantes e professores, tendo em conta os ambientes nos quais eles interagem. Tais estratégias têm como objetivo, por exemplo, compreender melhor o estudante no seu processo de aprendizagem, analisando a sua interação com o ambiente [36].

A *International Educational Data Mining Society* define EDM como:

> *"uma disciplina emergente preocupada em desenvolver métodos para explorar os dados únicos e cada vez mais em grande escala que vêm de ambientes educacionais, e usar esses métodos para entender melhor os estudantes e as configurações em que eles aprendem"* [37].

Segundo Romero e Ventura [38], EDM pode ser definida como a união de três principais áreas (Figura 5): ciência da computação, educação e estatística. A junção dessas áreas gera outras subáreas intimamente relacionadas à EDM, como: a educação baseada em computador, *data mining* e *machine learning* e *learning analytics*.

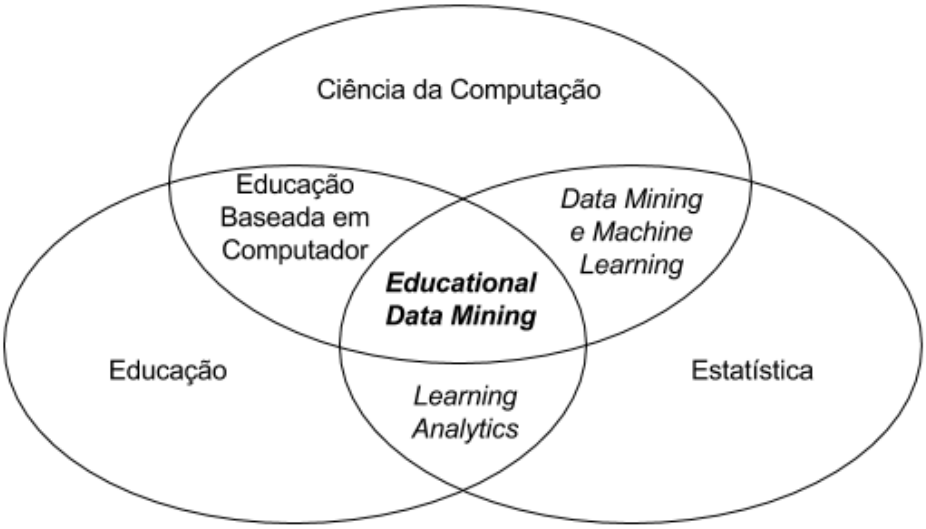

Figura 5: Principais áreas relacionadas à EDM (adaptado de Romero e Ventura [38]).

Dito isso, a EDM possibilita compreender de forma mais clara e eficaz os estudantes e o papel do contexto onde a aprendizagem ocorre, além dos demais fatores que influenciam o processo de aprendizagem. Por exemplo, pode-se identificar qual abordagem instrucional (individual ou colaborativa) oferece melhores benefícios ao estudante, além de, verificar se o estudante encontra-se desmotivado ou confuso, para assim, personalizar e adaptar os ambientes e métodos de ensino para oferecer melhores condições de aprendizagem [39].

4.3. Métodos de *Educational Data Mining*

Atualmente, existe uma diversidade de métodos de EDM, cabendo ao responsável pelo processo identificar qual é o mais adequado para o propósito que se precisa. Alguns desses métodos são vastamente conhecidos como universais, independentemente do tipo de mineração, já outros, possuem destaque especial somente em EDM [40].

Diversos métodos utilizados em EDM são derivados da área de mineração de dados. Contudo, devido às particularidades e exigências em ambientes educacionais e seus dados, na maioria das vezes tais métodos necessitam ser adaptados para seu melhor funcionamento [39].

Romero e Ventura [38] apresentam os seguintes métodos como os mais utilizados em EDM:

Tabela 2: Métodos de EDM, suas descrições e aplicações (adaptado de Romero e Ventura [38]).

Método	**Descrição**	**Aplicação**
Prediction	Existem três tipos de métodos de *prediction:* - *Classification*: quando a variável preditiva é categórica; - *Regression*: quando a variável preditiva é contínua; - *Density Estimation*: quando a variável preditiva é uma função de densidade de probabilidade.	Prever o desempenho do estudante e identificar os seus comportamentos.
Clustering	Forma grupos de dados, de maneira que os objetos contidos nos dados fiquem agrupados de acordo com a semelhança entre eles.	Agrupar matérias semelhantes dos cursos ou agrupar os estudantes baseado na sua aprendizagem e padrões de interação.
Relationship Mining	Existem quatro tipos: - *Association Rule Mining:* relação entre variáveis; - *Correlation Mining:* correlação linear entre as variáveis; - *Sequential Pattern Mining:* associações temporais entre as variáveis; - *Causal Mining:* relação de causalidade entre as variáveis.	Identificar relações nos padrões de comportamento dos estudantes e determinar as dificuldades de aprendizagem ou erros que ocorrem frequentemente.
Outlier Detection	Descobre pontos de dados que são significativamente diferentes dos demais. Um *outlier* é uma observação diferente (ou medida) que normalmente assume um valor maior ou menor que os demais dados.	Detectar os estudantes com dificuldades de aprendizagem, desvios nas ações ou comportamentos dos estudantes ou professores, e identificar irregularidades nos processos de aprendizagem.
Social Network Analysis (SNA)	Compreende e mede as relações entre entidades em informação em rede. SNA vê relações sociais como nós (atores individuais dentro da rede) e conexões ou ligações (relações entre indivíduos, tais como amizade, parentesco, etc.)	Interpretar e analisar a estrutura e relações em tarefas colaborativas e interações com as ferramentas de comunicação.
Process Mining	Consiste em três subáreas: - Verificação de conformidade; - Descoberta de modelos; - Extensão de modelos.	Refletir o comportamento do estudante quanto à sua evolução e desempenho ao longo da sua jornada acadêmica.
Text Mining	Algumas tarefas típicas desse método são: - Categorização de textos; - Agrupamento de textos; - Análise de sentimentos; - Síntese de documentos.	Analisar o conteúdo dos fóruns de discussão, *chats*, páginas *web*, documentos, e assim por diante.
Distillation of Data for Human Judgment	Representa os dados de forma mais legível e visual para facilitar a compreensão humana e assim apoiar decisões importantes baseadas em dados.	Ajudar os educadores a visualizar e analisar as atividades de curso dos estudantes e o uso de informação.

Método	Descrição	Aplicação
Discovery with Models	Utiliza um modelo gerado por um método de *prediction*, ou por um método de *clustering* e, em seguida, utiliza tal modelo como componente, ou ponto de partida, em outra técnica de *prediction* ou *relationship mining*.	Apoiar a identificação de relações entre os comportamentos dos estudantes e as características dos estudantes ou variáveis contextuais.
Knowledge Tracing	Estima a competência do estudante em determinadas áreas do conhecimento.	Acompanhar o conhecimento do aluno ao longo do tempo.
Nonnegative Matrix Factorization (NMF)	Interpreta, de maneira simples e direta, em termos de matrizes quadradas, também conhecido como modelo de transferência. NMF consiste numa matriz de números positivos, resultante do produto de duas matrizes menores.	A matriz M (M = Q * S) representa um determinado resultado observado, que pode ser decomposta em duas matrizes: - Q: matriz que representa a matriz dos itens; - S: domínio de competências de cada estudante

De acordo com Baker, Isotani e Carvalho [39], algumas das principais utilizações dos métodos de EDM são: fornecer *feedback* aos professores, orientar os estudantes com recomendações, identificar grupos de estudantes com características em comum e prever o risco de abandono. Tais informações podem propor mudanças significativas tanto no curso quanto na metodologia de ensino, ou mesmo um contato individual com estudantes desmotivados ou com baixo rendimento acadêmico [41].

4.4. Técnicas e Algoritmos de *Educational Data Mining*

Em sua grande parte, as técnicas utilizadas em EDM são originais da área de DM, sendo necessário, na maioria das vezes, adaptá-las devido às particularidades existentes em ambientes e dados educacionais.

Segundo Faria [40], algumas das técnicas e algoritmos de EDM mais utilizadas são: árvores de decisão, regressão linear, redes bayesianas, redes neurais artificiais, algoritmo genético, algoritmo *k-nearest neighbor* e algoritmo *k-means*.

4.4.1. Árvore de Decisão

Árvores de decisão são modelos estatísticos com estrutura semelhante à de uma árvore, onde cada nó interno (não-folha), pode ser entendido como um atributo teste, cada nó-

folha (nó-terminal) possui um rótulo de classe e cada sub-árvore representa o resultado de um teste [42]. Sua utilização recomenda o treinamento supervisionado para classificação e predição dos dados. O nó de mais alto nível numa árvore de decisão é chamado de nó-raiz. A Figura 6, apresenta uma classificação utilizando um algoritmo de árvore de decisão, para verificar se as condições climáticas são favoráveis para a prática de tênis.

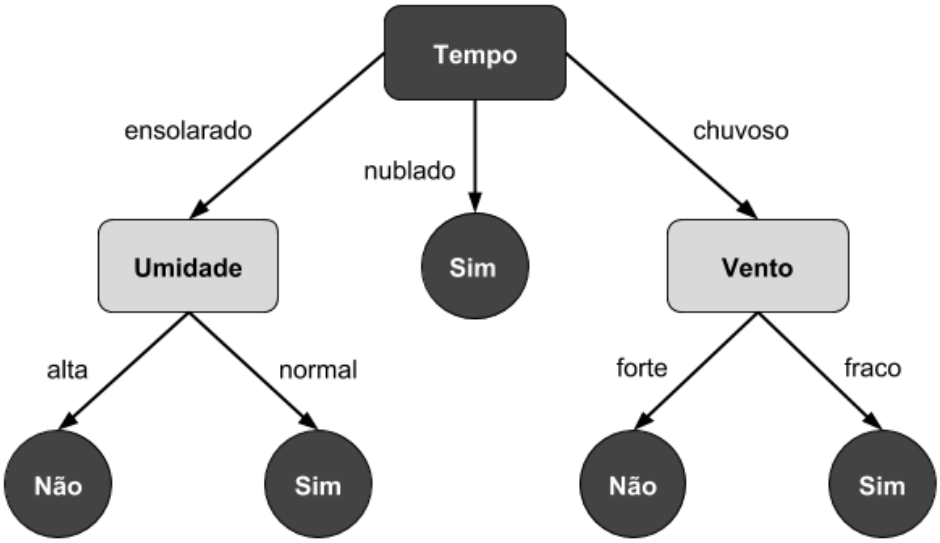

Figura 6: Exemplo de visualização de uma árvore de decisão.

Analisando a Figura 6, observa-se que existem três condições climáticas favoráveis para se praticar tênis: ensolarado com índice de umidade normal, nublado e chuvoso com pouco vento; enquanto que, com tempo ensolarado com alto índice de umidade, e com tempo chuvoso e com ventos fortes, a prática de tênis não é aconselhável.

Para Costa et al. [36], após o treinamento de um algoritmo de árvore de decisão, as instâncias são classificadas de acordo com o caminho que satisfaz às condições desde o nó-raiz até o nó-folha, sendo rotuladas de acordo com o nó-folha de tal caminho. Alguns dos algoritmos mais populares de árvore de decisão são o ID3, C4.5 e o CART.

O ID3 (Quinlan, 1986) foi o primeiro algoritmo de árvores de decisão. Baseado em busca heurística, é um algoritmo recursivo que busca pelos atributos que melhor dividem os exemplos, gerando assim, sub-árvores. A principal limitação do ID3, é que este só manipula atributos categóricos não-ordinais, tornando impossível utilização de conjuntos de dados com atributos contínuos, por exemplo.

O algoritmo C4.5 é um dos algoritmos mais utilizados, uma vez que apresenta ótimos resultados com problemas de classificação. Também proposto por Quinlan (1993), o C4.5 representa uma significativa evolução do ID3. As principais diferenças em relação ao ID3 são: manipula tanto atributos categóricos como atributos contínuos, trata valores desconhecidos, utiliza a medida de razão de ganho para selecionar o atributo que melhor divide os exemplos e apresenta um método de pós-poda das árvores geradas.

Apresentado pelos autores Breiman et al. [43], o algoritmo CART (*Classification and Regression Trees*) consiste numa técnica que resulta tanto em árvores de classificação quanto em árvores de regressão. As árvores geradas pelo algoritmo CART são sempre binárias, as quais podem ser percorridas por completo respondendo à simples questões do tipo “sim” ou “não”.

4.4.2. Regressão Linear

Segundo Han e Kamber [42], a regressão linear é uma técnica de predição que descreve o relacionamento entre duas variáveis, uma preditiva y, e uma única preditora x, modelando y como uma função linear de x. Isto é,

Listagem 1: Relacionamento entre variável preditiva e variável preditora.

$$y = b + wx$$

onde a variância de y é constante e b e w são os coeficientes de regressão. Tais coeficientes podem ser resolvidos pelo método dos mínimos quadrados, o qual estima qual é a reta mais apropriada para apresentar os dados, ou seja, aquela que minimiza o erro entre os dados atuais e a estimativa da reta.

Alguns exemplos comuns de aplicação da regressão em EDM são: tentar explicar certas variações do desempenho de alunos em função do aumento da carga horária de estudos ou do número de disciplinas e, na predição de valores futuros aplicando testes a estudantes para avaliar o potencial sucesso na escola/universidade [44].

4.4.3. Redes Bayesianas

Redes Bayesianas (RB) também conhecidas como redes casuais, são modelos probabilísticos utilizados para descrever a relação entre variáveis aleatórias. Para Russel e Norvig [45], RBs representam, de forma compacta, a distribuição conjunta de probabilidades de um conjunto de variáveis. Estas estruturas são amplamente utilizadas na área de modelagem de usuário (*User Modeling*), visto que são capazes de lidar com incerteza, característica comum dos dados referentes ao conhecimento humano [46].

De acordo com Neapolitan [47], a técnica de RB nasceu no contexto onde há um grande número de variáveis e o objetivo é de verificar qual a influência probabilística não direta de uma para as demais. Dessa forma, a teoria de Redes Bayesianas combina princípios de teoria dos grafos, teoria da probabilidade, ciência da computação e estatística [48].

As RBs são representadas por meio de grafos acíclicos direcionados, onde os vértices representam as variáveis e as arestas, as relações de dependência entre as variáveis. Cada variável pode assumir um conjunto finito de valores, chamados estados, que possuem determinadas probabilidades de ocorrência. Estas probabilidades são definidas a partir dos cálculos estatísticos, e cada variável possuirá uma tabela de valores de probabilidades para que suas possíveis ações sejam realizadas [49]. A Figura 7 exibe um exemplo de estrutura de RB de uma pessoa que deseja verificar se possui câncer de pulmão e, para isso, verificam-se dois aspectos: se o paciente é fumante, e se o mesmo costuma ficar exposto a um alto nível de poluição; além disso, é necessário um raio-x para confirmar a causa de tal desconforto.

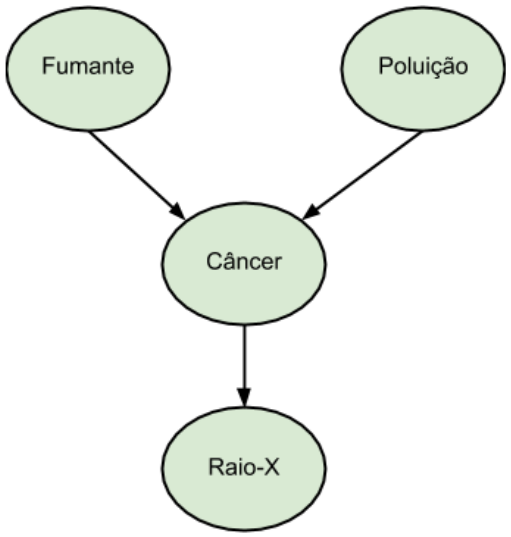

Figura 7: Exemplo de estrutura de RB.

4.4.4. Redes Neurais Artificiais

O cérebro é considerado o centro de inteligência e aprendizagem do corpo humano, sendo ele responsável por aprender, memorizar e colocar em prática toda a informação absorvida. Capacidade de tratar informações inconsistentes, alta flexibilidade para se adaptar a situações supostamente pouco definidas e tolerância a falhas, são características humanas desejáveis a qualquer sistema artificial.

Uma rede neural artificial (RNA) é um sistema computacional baseado nos conceitos de como o cérebro humano está organizado e como ele aprende. De acordo com Côrtes, Porcaro e Lifschitz [50], existem duas estruturas principais: o nó, que corresponde ao neurônio; o link, que corresponde às conexões entre neurônios.

As principais topologias de RNAs são: redes de propagação para frente (*feedforward*) e redes realimentadas (*recurrent*), ver Figura 8; sendo distinguidas de acordo com o método de propagação da informação recebida. Todas as camadas intermediárias representam os diferentes níveis de conhecimento obtidos no seu processamento, numa tentativa de imitar o cérebro humano [50].

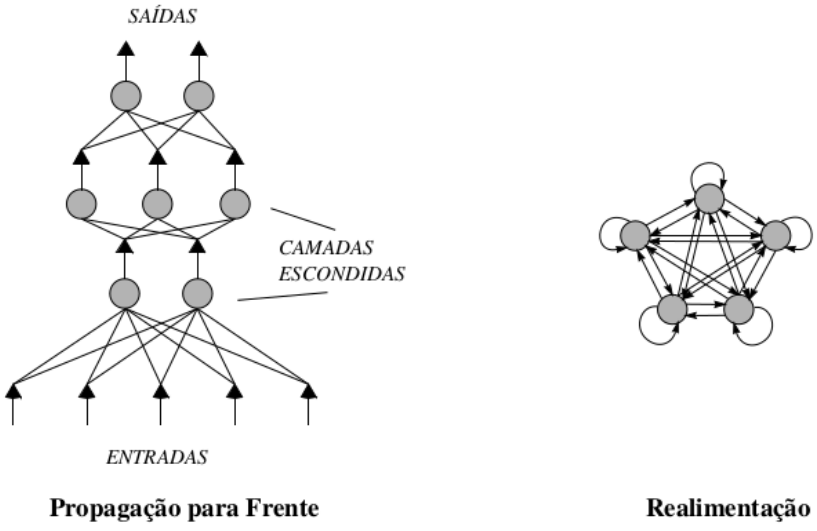

Figura 8: Principais topologias de redes neurais artificiais (reproduzido de Rauber [51]).

As redes de propagação para frente possuem fluxo de informação unidirecional, onde os neurônios que recebem informação ao mesmo tempo são agrupados em camadas, e as camadas que não possuem relação com as entradas ou saídas da rede chamam-se "camadas escondidas". De forma distinta, as redes realimentadas possuem um

comportamento dinâmico, uma vez que as ligações entre os neurônios não possuem restrições [51].

4.4.5. Algoritmo Genético

Propostos inicialmente por Holland, algoritmos genéticos (AGs) são modelos computacionais de busca e otimização baseados na teoria da evolução das espécies de Charles Darwin. Além de ser uma estratégia de gerar-e-testar muito elegante, AGs são capazes de identificar e explorar fatores ambientais e convergir para soluções ótimas, ou aproximadamente ótimas em níveis globais.

AGs buscam reproduzir um ambiente natural, onde somente os indivíduos mais adaptados prosperam e reproduzem, transmitindo seu código genético para as próximas gerações [36].

O princípio básico do funcionamento dos AGs é que um critério de seleção vai fazer com que, depois de muitas gerações, o conjunto inicial de indivíduos gere indivíduos mais adaptados. Um método muito utilizado é o Método da Roleta, onde indivíduos de uma geração são escolhidos para fazer parte da próxima geração, por meio de um sorteio de roleta. O fluxograma da Figura 9 apresenta a estrutura de um algoritmo genético.

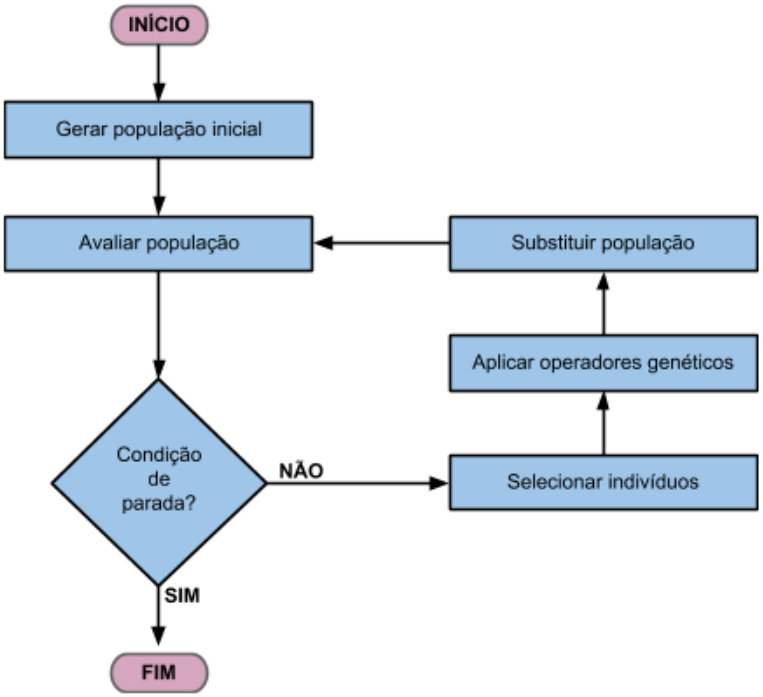

Figura 9: Fluxograma de um algoritmo genético (adaptado de Costa et al. [36]).

De acordo com Elmasri e Nevathe [52], AGs são utilizados na solução e agrupamento de problemas e sua capacidade de resolver problemas em paralelo fornece uma ferramenta poderosa para a DM.

4.4.6. Algoritmo *k-Nearest Neighbor*

Proposto por Fukunaga e Narendra em 1975, o algoritmo *k-Nearest Neighbor* (k-NN) é um algoritmo de aprendizagem supervisionada, cujo objetivo é determinar o rótulo de classificação de uma amostra baseado nas amostras vizinhas vindas de um conjunto de treinamento. É comum considerar que proximidade entre as características das amostras pertencentes a uma classe específica de um problema variem pouco, ou seja, amostras mais próximas possuem maior probabilidade de serem do mesmo tipo da amostra que se pretende classificar [53].

No k-NN, para se classificar uma nova amostra são recuperados os k vizinhos mais próximos, buscando sempre o que tem a menor distância. Feito isso, é atribuída à nova amostra, a classe mais frequente entre esses k vizinhos. Na Figura 10 tem-se um pseudocódigo do algoritmo k-NN.

```
1  início
2      Informar o valor de K
3      para cada nova amostra faça
4              Calcular distância para todas as amostras
5              Determinar o conjunto das K distâncias mais próximas
6              O rótulo com mais representantes no conjunto dos K
7              vizinhos será o escolhido
8      fim para
9      retornar: conjunto de rótulos de classificação
10 fim algoritmo
```

Figura 10: Pseudocódigo do algoritmo k-NN (adaptado de Moreira, Costa e Aguiar [54]).

De acordo com Cunha [55], existem diversas formas para se calcular a distância entre a amostra que se deseja classificar e as demais amostras, contudo, a mais simples e

utilizada é a distância euclidiana. Considerando os pontos $p = (p_1, p_2, p_3, \ldots, p_{n-1}, p_n)$ e $q = (q_1, q_2, q_3, \ldots, q_{n-1}, q_n)$ a distância euclidiana entre eles pode ser definida por:

Listagem 2: Fórmula para cálculo da distância euclidiana.

$$d(p_i, q_i) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

A Figura 11 ilustra o funcionamento do algoritmo k-NN, onde tem-se o k = 5 e amostras de treinamento representadas pelos quadrados vermelhos e pelos quadrados azuis. A variável k representa a quantidade de vizinhos mais próximos que serão utilizados para classificar a nova amostra, aqui representada pela estrela. Com isso, como 3 são do rótulo vermelho e 2 do rótulo azul, a nova amostra receberá o rótulo vermelho.

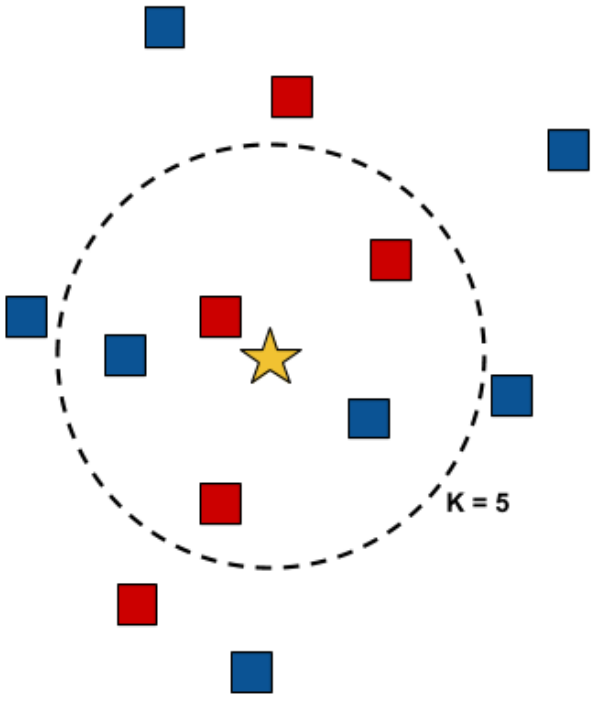

Figura 11: Exemplo do funcionamento do k-NN.

4.4.7. Algoritmos *k-Means* e *k-Means++*

O algoritmo *k-Means* é um algoritmo simples e eficiente cujo objetivo é dividir um determinado número de objetos em k *clusters*, onde a distância de cada objeto para o centróide de cada *cluster* é o que determinará onde o objeto deverá ser alocado [56]. Segundo Araar & Haddad [57], *clusters* são áreas compostas por uma coleção de objetos similares entre si e diferentes dos objetos pertencentes aos outros *clusters*. O valor de k é um parâmetro definido pelo usuário que determina o número de *clusters* em que os objetos serão divididos.

A escolha do parâmetro k pode ser um problema, uma vez que ao início, pode não se saber quantos *clusters* existem. De acordo com Souza et al. [58], a escolha inicial da posição dos *clusters* pode obedecer aos seguintes critérios: *(i)* seleção das k primeiras instâncias; *(ii)* seleção de k instâncias de forma aleatória; e *(iii)* seleção de k instâncias que possuam alto grau de dissimilaridade. A Figura 12 apresenta um pseudocódigo do algoritmo *k-Means*.

```
1  início
2      Informar o valor de K
3      Definir os K pontos como clusters
4      Definir os centróides
5      enquanto todos os centróides continuarem a mudar faça
6          Atribuir cada objeto ao cluster do centróide mais próximo
7          Recalcular as posições dos centróides
8      fim enquanto
9      retornar: conjunto de clusters
10 fim algoritmo
```

Figura 12: Pseudocódigo do algoritmo *k-Means* (adaptado de Arthur e Vassilvitskii [59]).

Para Linden [60], uma das características que torna este algoritmo viável é sua velocidade, geralmente convergindo em poucas iterações para uma configuração estável. Contudo, embora possa ser provado que este algoritmo sempre termina, ele não necessariamente encontra a configuração ótima de *clusters*, uma vez que ele é bastante sensível ao conjunto de centroides inicialmente escolhidos.

Para evitar um possível efeito negativo devido à má escolha dos centroides iniciais, foi proposto o algoritmo *k-Means++* [59], onde centroides iniciais não são escolhidos aleatoriamente. A ideia da modificação é selecionar um bom conjunto de centroides iniciais, sendo essa a única diferença entre eles.

Segundo Faria [61], algumas das vantagens encontradas com as modificações atribuídas ao *k-Means++* são:

- Melhoria no tempo de execução;
- Melhoria na qualidade do resultado;
- Melhoria nos resultados à medida que o número de *clusters* aumenta.

Com isso, nota-se que EDM é uma área de pesquisa de extrema importância para questões e problemas educacionais, uma vez que tem como objetivos fazer descobertas sobre o comportamento dos estudantes e seu percurso académico, e seus métodos e técnicas são essencias para a Análise de Dados Acadêmica, ou ainda *Academic Analytics* (seção 4.3).

4.5. *Academic Analytics* no Ensino Superior

Há vários anos, as empresas implementam sistemas tecnológicos para análise avançada dos dados disponíveis. Esses sistemas têm se tornado fundamentais para o processo de tomada de decisão, requisito essencial no gerenciamento eficaz de informação. Pode-se dizer que, o interesse das IES nesta nova onda de como utilizar os dados está apenas no início e elas estão aos poucos tentando implementar seus sistemas de análise [62].

Ao passo que as IES coletam mais e mais dados a respeito de seus estudantes e suas bases de dados tornam-se mais complexas e acessíveis, entramos numa nova era de uso de dados para melhorar o sucesso dos estudantes, simplificar processos e utilizar recursos de maneira mais eficiente. Uma vez que os dados são analisados, é possível obter melhores processos de colocação de estudantes, previsões de matrículas mais precisas e sistemas de aviso prévio que identificam e ajudam os estudantes em risco de falhar ou abandonar o curso [63].

O termo *Analytics* é um neologismo na educação e é importado de outras áreas, em particular do campo da gestão:

> *"Analytics" é um termo usado nos negócios e na ciência para se referir ao suporte computacional para a captura de dados digitais para ajudar a tomada de decisão. Com o crescimento de grandes conjuntos de dados e do poder computacional, isso se estende ao projeto de infraestruturas que necessitam de feedback rápido para informar intervenções cujo impacto pode, por sua vez, ser monitorado. As organizações têm um "sistema nervoso digital" cada vez mais sensível, fornecendo feedback em tempo real sobre o ambiente externo e os efeitos das ações* [64].

Analytics – tomada de decisão baseada em dados em que a informação é utilizada para fundamentar as decisões em todos os níveis da empresa – no Ensino Superior pode ocorrer em dois planos (Tabela 3) [65]:

- *Learning Analytics:* centra-se no processo de ensino e aprendizagem a uma escala institucional ou infraestrutural (estudante, unidade curricular, curso, universidade, etc.);
- *Academic Analytics:* tem o foco na escala institucional ou suprainstitucional, ou seja, se baseia no uso de dados para apoiar o gerenciamento das atividades das IES em suas diversas formas (financeira, educacional, marketing, etc.) e, em monitorar questões relacionadas ao sucesso acadêmico de estudantes (estudantes em risco de retenção, por exemplo).

Tabela 3: *Learning* e *Academic Analytics* (adaptado de Ferreira e Andrade [65]).

Tipo de *Analytics*	**Objeto de análise**	**Beneficiado**
Learning Analytics	Redes sociais, desenvolvimento conceitual, predição e padrões de sucesso/insucesso	Estudantes e universidade
Academic Analytics	Perfis dos estudantes, desempenho escolar, fluxo de conhecimento	Administradores, financiadores, marketing, etc.

Para Long e Siemens [66] as principais diferenças entre *Learning Analytics* e *Academic Analytics* são:

> *Learning Analytics é mais específico do que Academic Analytics: o foco do primeiro é exclusivamente no processo de aprendizagem. O Academic Analytics reflete o papel da análise de dados em nível institucional, enquanto o Learning Analytics se concentra no processo de aprendizagem (que inclui a análise da relação entre estudante, conteúdo, instituição e educador).*

Dito isso, Análise de Dados Acadêmicos, tradução para *Academic Analytics*, é um ramo da análise de dados que surgiu no ensino superior após a prática de técnicas de mineração de dados e ferramentas de BI no campo da gestão. Pode referir-se vastamente a práticas de tomada de decisão para fins operacionais em nível de instituição, mas também pode ser aplicado a questões de ensino e aprendizagem de estudantes [67].

Academic Analytics (AA) faz uso de métodos de análise estatística, mineração de dados e modelagem preditiva para revelar e reconhecer padrões ocultos em amplas bases de dados educacionais. Esses padrões nos permitem compreender melhor diversos aspectos educacionais, como o comportamento do estudante e os resultados de aprendizado [68].

AA tem como foco o uso de dados provenientes dos sistemas de informação da IES para tentar compreender os dados cadastrais dos estudantes e outros que se relacionam com a experiência acadêmica do estudante na instituição [67], [69].

Em outras palavras, AA é a aplicação de ferramentas e estratégias de BI para orientar as práticas de tomada de decisão em instituições educacionais. O objetivo de um programa de AA é auxiliar os encarregados do planejamento estratégico em um ambiente de aprendizado a medir, coletar, decifrar, relatar e compartilhar dados de maneira eficaz, de modo que os pontos fortes e fracos operacionais e estudantis possam ser identificados.

As primeiras iniciativas de AA buscam prever quais estudantes estão em dificuldade acadêmica, permitindo que os professores e conselheiros personalizem os caminhos de aprendizagem ou forneçam instruções adaptadas às necessidades específicas de aprendizado [69].

Como resultado, as práticas analíticas entraram em ação para viabilizar os meios de medir e melhorar o desempenho, enquanto hospedam novos softwares e serviços profissionais para produzir um serviço inteligente acessível para toda a instituição [70]. Mas o que fazer para colocar mais ação no *analytics*? De acordo com Norris et al. [71], seis ações principais são necessárias para evoluir da geração atual de *academic analytics* (ferramentas, soluções e serviços) para a *action analytics*:

- Focar em processos, soluções e comportamentos;
- Incorporar fatores da força de trabalho nos currículos e ofertas educacionais;
- Utilizar a nova geração de análise de arquitetura aberta para melhorar o acesso, acessibilidade e o sucesso para os estudantes;

- Incorporar comparações interinstitucionais e intersetoriais em soluções;
- Desenvolver novas práticas/soluções que se adequem com os objetivos e estratégias institucionais;
- Desenvolver a capacidade organizacional e mudar a cultura para incentivar o comportamento baseado em evidências e a inovação focada na ação para melhorar o desempenho.

Simples, mas transformadora, *analytics* fornece um novo modelo para os líderes das IES melhorarem o ensino, a aprendizagem, a eficiência organizacional e a tomada de decisões e, como consequência, servir de base para a mudança sistêmica. O crescimento contínuo da quantidade de dados cria um ambiente no qual abordagens novas e inovadoras são necessárias para compreender os padrões de valor existentes nos dados [66].

O seguinte ciclo para reflexão de análise de aprendizado é proposto por Long e Siemens [66]:

1. Nível do curso: percurso de aprendizagem, análise de redes sociais, análise de discurso;
2. *Educational Data Mining*: modelagem primitiva, *clustering*, mineração de padrões;
3. Currículo inteligente: o desenvolvimento de recursos curriculares semanticamente definidos;
4. Conteúdo adaptativo: sequência adaptativa de conteúdo baseada no comportamento do estudante, sistemas de recomendação;
5. Aprendizagem adaptativa: processo de aprendizagem adaptativa (interações sociais, atividade de aprendizagem, apoio ao estudante, não apenas conteúdo).

Sabemos que a avaliação frequente e antecipada possui um grande impacto, mas quando expressamos tal impacto em termos de retenção de estudantes, valor muito significativo, torna-se mais fácil incentivar a reestruturação de cursos e integrar avaliações mais formativas. Estudantes precisam de *feedback* rápido, e os estudantes com maior dificuldade, em especial, precisam ser orientados para conteúdos importantes para ajudá-los a corrigir seus equívocos e concluir cursos com mais sucesso [62]. Portanto, AA pode fornecer aos profissionais uma ligação essencial entre instrução, avaliação e esforço do estudante.

Capítulo 5 Ferramentas de Suporte ao *Academic Analytics*

5.1. Bases de Dados

O crescente armazenamento de dados em bases de dados informatizadas tornou necessária para instituições, empresas e usuários a busca por segurança no armazenamento, além de cópias de segurança para corrigir eventuais erros, ataques cibernáticos ou desastres. Com isso, foi preciso que as instituições e empresas buscassem soluções para otimizar sua performance, já que, como resultado dessa inovação, as bases de dados se tornaram significativamente volumosas e pouco eficientes [72].

Uma Base de Dados (BD) pode ser definida como uma coleção de dados logicamente coerentes entre si. Tais dados são gerenciados, interpretados e manipulados de acordo com uma necessidade específica [72]. Segundo Elmasri e Navathe [52], BD é um dos componentes essenciais da vida em sociedade e a maioria das pessoas encontra-se diariamente em diversas situações nas quais elas interagem com uma BD. Para os autores, a definição de uma BD passa pelo processo de definir os tipos, estruturas e restrições lógicas dos dados a serem armazenados.

De acordo com Date [73], “uma base de dados é uma coleção de dados persistentes, utilizada pelos sistemas de aplicação de uma determinada empresa”. Em outras palavras, uma BD é formada por um conjunto organizado de informações relacionadas, criadas com o objetivo específico de atender usuários, onde ao agrupá-las para uma mesma finalidade, é possível determiná-la como BD [74]. Dito isso, é possível classificar uma BD em duas principais categorias: relacional e não relacional.

Uma base de dados relacional (BDR) modela seus dados de forma que eles sejam compreendidos pelo usuário como relações. Um Sistema Gerenciador de Base de Dados

Relacional (SGBDR) é um software responsável por controlar o armazenamento, recuperação, exclusão, segurança e integridade dos dados em uma BDR. A arquitetura de uma BDR pode ser descrita como: relações (tabelas), tuplas (resgistros), atributos (colunas) e chaves (primária e estrangeira) [52], [73], [75].

Para Rezende [75], os objetivos de um SGBDR são: isolar o usuário dos detalhes internos da BD e proporcionar a independência dos dados em relação às aplicações, ou seja, tornar a estratégia de acesso e a forma de armazenamento independente da aplicação.

A imensa quantidade de dados gerados diariamente em vários domínios de aplicação traz enormes desafios na forma de manipulação, armazenamento e processamento de consultas em diversas áreas da computação, em especial nas áreas de bases de dados, mineração de dados e recuperação de informação. Nesse contexto, os SGBDRs não são os mais adequados, ou "completos", às necessidades do domínio do problema de *Big Data* [76].

Uma das tendências para solucionar os diversos problemas e desafios gerados pelo contexto *Big Data* é o movimento denominado NoSQL (*Not only* SQL). O NoSQL oferece diversas soluções inovadoras de armazenamento e processamento de grandes volumes de dados [76]. Algumas das características que tornam as BD NoSQL tão diferentes em relação às relacionais são: escalabilidade horizontal, ausência de esquema ou esquema flexível, suporte nativo de replicação, API simples para acessar a BD e consistência eventual (teorema CAP – *Consistency*, *Availability e Partition tolerance*) [77].

As organizações públicas e privadas percebem cada vez mais o valor dos dados que possuem à sua disposição, e compreendem o quanto estes são um bem de extrema importância para o aumento da produtividade, eficiência e competitividade. Como consequência, a exploração de enormes volumes de dados assume um papel cada vez mais importante na sociedade atual. Por exemplo, a utilização de um DW central permite às organizações executarem operações de análise e, assim, obterem informações que são de importância estratégica e tática para as suas atividades [78], [79].

No entanto, de acordo com Oliveira, Rodrigues e Henriques [80], uma boa parte dos dados apresenta erros ou anomalias. As anomalias dos dados criam problemas à sua

utilização adequada, influenciando negativamente a veracidade dos resultados e conclusões obtidas. Assim sendo, antes da aplicação de qualquer ferramenta de análise, os dados devem ser "limpos" a fim de remover e reparar quaisquer anomalias que possam existir.

Para Rahm e Hai Do [81], a limpeza de dados (*data cleaning*) lida com a detecção e remoção de erros e inconsistências dos dados com o objetivo de aumentar/melhorar a qualidade dos mesmos. Tipicamente, o processo de limpeza de dados não pode ser executado sem o envolvimento de um especialista, contudo, tal processo deve ser o mais automatizado possível em virtude dos grandes volumes de dados geralmente processados e do tempo necessário para que um especialista conduza a sua limpeza manual [80].

Outro processo de extrema importância para o sucesso da transição dos dados dos sistemas de origem para o DW é o ETL (*Extract, Transform and Load*). Tal processo é responsável pelo tratamento e limpeza dos dados provenientes dos sistemas OLTP (*Online Transaction Processing*) para a inserção, geralmente, em um DW ou *Data Mart* [82]. Assim sendo, pode-se representar o ETL de acordo com a figura abaixo:

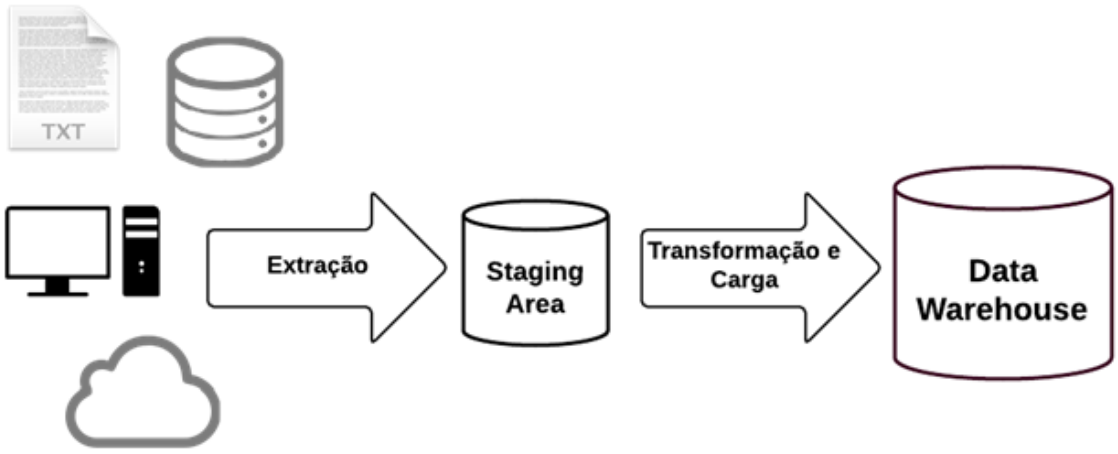

Figura 13: Representação do processo de ETL (reproduzido de Elias [82]).

De acordo com Elias [82], pode-se entender a etapa de extração como a fase onde os dados são extraídos dos OLTPs e conduzidos para uma *staging area*, onde a conversão desses dados para um único formato se faz necessária devido a heterogeneidade existente nas informações. Após a extração, é possível iniciar a etapa de transformação e limpeza dos dados e, por fim, a carga dos dados no DW é iniciada.

Tendo em vista tamanha complexidade, a base de dados Oracle é a mais indicada para grandes empresas ou grandes aplicações, que exigem requisitos de negócios mais complexos, e que possuem capital para investir nos recursos de segurança e performance adicionais que ela oferece [74].

5.1.1. Oracle Database

A Oracle Corporation, comumente conhecida como Oracle, tem como principal produto a sua base de dados, a Oracle Database. A base de dados Oracle que já se encontra na versão 18.1 da *release* 18c, segue o modelo relacional evoluindo continuamente a cada versão, além de oferecer ferramentas de gestão dos dados armazenados e prover escalabilidade, segurança e alto desempenho para o armazenamento de dados [83].

De acordo com Alhadi e Ahmad [84], a base de dados Oracle é a BD de grande escala mais utilizada no mundo dos negócios e seu desempenho influencia diretamente a eficiência das aplicações. Para os mesmos autores, essa BD é apropriada para gerenciar e trocar dados corporativos, especialmente em organizações com BDs de grande porte.

A DB-Engines, iniciativa que busca coletar e apresentar informações sobre SGBDs, apresenta mensalmente o DB-Engines Ranking (Figura 14), uma lista de SGBDs classificados por sua popularidade atual. O cálculo desse *ranking* considera o número de menções do sistema em *websites*, buscas no Google e Bing, discussões técnicas sobre o sistema em alguns fóruns, o número de vagas de trabalho oferecidas e o número de profissionais que mencionam o uso do SGBD em seus perfis [85].

Rank			DBMS	Database Model	Score		
Jun 2018	May 2018	Jun 2017			Jun 2018	May 2018	Jun 2017
1.	1.	1.	Oracle	Relational DBMS	1311.25	+20.84	-40.51
2.	2.	2.	MySQL	Relational DBMS	1233.69	+10.35	-111.62
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1087.73	+1.89	-111.23
4.	4.	4.	PostgreSQL	Relational DBMS	410.67	+9.77	+42.13
5.	5.	5.	MongoDB	Document store	343.79	+1.67	+8.79
6.	6.	6.	DB2	Relational DBMS	185.64	+0.03	-1.86
7.	7.	↑ 9.	Redis	Key-value store	136.30	+0.95	+17.42
8.	↑ 9.	↑ 11.	Elasticsearch	Search engine	131.04	+0.60	+19.48
9.	↓ 8.	↓ 7.	Microsoft Access	Relational DBMS	130.99	-2.12	+4.44
10.	10.	↓ 8.	Cassandra	Wide column store	119.21	+1.38	-4.91

Figura 14: 10 primeiras posições do DB-Engines Ranking em junho de 2018 (reproduzido de DB-Engines [85]).

Para além, a base de dados Oracle também oferece suporte básico e avançado à BI, um dos diferenciais que levam as organizações optarem pela escolha da mesma. Como dito anteriormente, BI busca alcançar a inteligência por trás dos dados, de forma a suportar a tomada de decisão. Dito isso, três das principais opções oferecidas pela base de dados Oracle são: a *Oracle Predictive Analytics Operations*, a *Oracle Database Native Analytics* e a *Oracle Advanced Analytics Option*.

Oracle Predictive Analytics Operations, como o próprio nome diz, oferece operações de análise preditiva, tecnologia que captura processos de mineração de dados em rotinas simples. Por vezes chamada de "mineração de dados em um clique", a análise preditiva simplifica e automatiza o processo de mineração de dados e, apesar de utilizar mineração de dados, não é necessário conhecimento prévio de tal tecnologia para utilizá-la [86]. As três operações de análise preditiva disponibilizadas são:

- *EXPLAIN*: explica como preditores individuais (colunas) afetam a variação de valores em uma coluna de destino;
- *PREDICT*: para cada caso (linha), prevê os valores em uma coluna de destino;
- *PROFILE*: cria um conjunto de regras para casos (linhas) que implicam o mesmo valor de destino.

A base de dados Oracle suporta uma variedade de recursos analíticos nativos (*Oracle Database Native Analytics*), e como todos esses recursos fazem parte de um servidor em comum, é possível combiná-los com eficiência. Algumas das análises nativas suportadas são descritas na Tabela 4:

Tabela 4: *Oracle Database Native Analytics* (adaptado de Oracle [86]).

Característica Analítica	**Descrição**
Transformações de dados complexos	A transformação de dados é aspecto importante das aplicações analíticas e do ETL. Pode-se usar expressões SQL para implementar transformações de dados ou pode-se usar o pacote DBMS_DATA_MINING_TRANSFORM.
Funções estatísticas	Fornece uma longa lista de funções estatísticas com suporte para: teste de hipóteses, cálculo de correlação, estatística de tabela cruzada e estatística descritiva.
Funções SQL analíticas e de janela	Suporta funções analíticas e de janela para calcular conjuntos cumulativos, móveis e centralizados.
Álgebra linear	O pacote UTL_NLA expõe um subconjunto das bibliotecas populares BLAS e LAPACK (versão 3.0) para operações com vetores e matrizes representadas como VARRAYS.
OLAP	Oracle OLAP suporta análise multidimensional e pode ser utilizado para melhorar o desempenho de consultas multidimensionais.

Característica Analítica	Descrição
Análise espacial	O Oracle *Spatial* oferece recursos espaciais avançados para suportar soluções GIS e LBS.
Mineração de texto	O Oracle *Text* usa o SQL padrão para indexar, pesquisar e analisar textos e documentos armazenados na base de dados Oracle, em arquivos e na *web*.

O *Oracle Advanced Analytics Option* possui dois componentes, contudo, daremos ênfase em apenas um deles, o *Oracle Data Mining*. Este componente oferece um conjunto abrangente de algoritmos para executar diversas tarefas de mineração, como classificação, regressão, detecção de anomalias, extração de recursos, agrupamento e análise de mercado. A Tabela 5 apresenta alguns desses algoritmos e suas respectivas características:

Tabela 5: Algoritmos disponibilizados pelo *Oracle Data Mining* (adaptado de Oracle [25]).

Algoritmo	Função/Técnica	Tipo de Aprendizagem
Apriori	Associação	Não supervisionada
Árvore de Decisão	Classificação	Supervisionada
Análise Semântica Explícita	Classificação	Supervisionada
k-Means	Agrupamento	Não supervisionada
Máquina de Vetores	Detecção de Anomalias	Não supervisionada
Maximização de Expectativas	Agrupamento	Não supervisionada
Modelos Lineares Generalizados	Classificação e Regressão	Supervisionada
Naive Bayes	Classificação	Supervisionada
Random Forest	Classificação	Supervisionada
Rede Neuronal	Classificação e Regressão	Supervisionada

A ampla gama de ferramentas oferecidas, em especial aquelas relacionadas à BI, são algumas das razões pelas quais inúmeras organizações têm adotado a base de dados Oracle como sua base de dados. Tendo em vista o crescimento exponencial na quantidade de dados gerados, mais do que ser capaz de armazenar esses dados, é essencial para as organizações ser capaz de compreendê-los corretamente para utilizá-los de maneira eficaz.

5.2. Análise de Dados com R

Como dito anteriormente, o enorme volume de dados gerados, coletados e armazenados tornou necessário o desenvolvimento de novos métodos, tecnologias e ferramentas para a análise eficiente dos mesmos, e uma dessas ferramentas é o R. Por muito tempo, essa ferramenta foi utilizada por especialistas na área da estatística, uma vez que o objetivo da mesma era atender tal área. Contudo, com o passar dos anos, diversos desenvolvedores passaram a utilizar essa ferramenta, tornando-a uma das ferramentas mais utilizadas para a análise de grandes volumes de dados [87].

Desenvolvido originalmente por Ross Ihaka e Robert Gentleman (Universidade de Auckland, Nova Zelândia) na década de 90, R é uma linguagem e ambiente de desenvolvimento voltado, em especial, para computação estatística e produção de gráficos. Inspirado nas linguagens S (sintaxe) e Scheme (implementação e semântica), R encontra-se disponível como software livre e multiplataforma. R oferece uma grande variedade de técnicas estatísticas (modelagem linear e não-linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento, etc.) e técnicas gráficas extensíveis [88]–[90].

Podendo ser facilmente estendida por meio de seus inúmeros pacotes disponíveis no CRAN[1] (*Comprehensive R Archive Network*), a linguagem R vem sendo utilizada em diversos projetos e estudos, sendo um deles o *Data Mining*. Um pacote do R é um conjunto de funções que têm um tema em comum e, ao serem carregadas em memória, tornam-se disponíveis para uso [88], [89].

Atualmente, o repositório de pacotes CRAN possui 12620 pacotes disponíveis. Além disso, existem muitos pacotes fornecidos e em desenvolvimento em outros sites, como o Bioconductor[2], o R-Forge[3] e o GitHub[4] [92], [93].

Algumas das capacidades adquiridas com a utilização dos pacotes são a implementação de técnicas estatísticas especializadas, dispositivos gráficos, capacidades de importação e exportação, ferramentas de relatórios, etc. Alguns exemplos de pacotes do R são: o `dplyr` (oferece uma série de funções que facilitam a manipulação de dados), o `readxl`

[1] https://cran.r-project.org/
[2] https://www.bioconductor.org/
[3] https://r-forge.r-project.org/
[4] https://github.com/

(funções para leitura de arquivos excel) e o `ggplot2` (pacote para criação de gráficos) [91].

Para auxiliar os usuários a descobrir quais pacotes utilizar, o CRAN oferece um guia de tarefas para diferentes aplicações. Alguns tópicos relacionados à mineração de dados são [94]:

- Aprendizagem de Máquina e Aprendizagem Estatística;
- Análise de Agrupamento e Modelos de Mistura Finita;
- Análise de Séries Temporais;
- Processamento de Linguagem Natural;
- Estatística Multivariada;
- Análise de Dados Espaciais.

Com isso, conclui-se que o R é uma linguagem de programação de extrema importância para aqueles que possuem bases de dados e precisam transformá-las em conhecimento, visto que ela permite realizar análises de forma rápida, flexível e eficiente.

Capítulo 6 Modelo para Classificação do Risco de Abandono Escolar em Cursos de Engenharia

6.1. Metodologia

O presente estudo é de natureza quantitativa descritiva, onde busca-se investigar os dados relativos aos estudantes dos cursos de licenciatura em engenharia, pós Bolonha com a duração de 3 anos, do Instituto Politécnico de Bragança (IPB). A partir disso, criou-se um modelo que permite aplicar metodologias de *Academic Analytics* para a identificação do perfil de estudantes em risco de abandono.

Em busca de cumprir com os objetivos estabelecidos, tomou-se como referência os cursos de Engenharia Civil, Engenharia de Energias Renováveis, Engenharia Eletrotécnica e de Computadores, Engenharia Informática, Engenharia Mecânica e Engenharia Química, da Escola Superior de Tecnologia e Gestão (ESTiG) do IPB. A análise proposta foi reportada ao período de 2007 a 2015, levando em consideração os estudantes já diplomados e os estudantes que abandonaram seus respectivos cursos (identificados neste trabalho como "*dropouts*").

Uma situação de *dropout* é caracterizada por um estudante que não renovou sua matrícula no ano letivo atual e, consequentemente, não concluiu o curso. De modo contrário, um estudante diplomado é aquele que cursou e foi aprovado em todas as disciplinas de seu curso.

Os dados foram obtidos da base de dados Oracle do IPB por meio de processos de consulta em SQL e exportação para CSV. Os dados foram tratados de forma anônima e de acordo com o Regulamento Geral de Proteção de Dados. Para a análise dos dados foram usadas as ferramentas Excel e R. O Excel foi usado para a análise estatística e o R para *data mining* recorrendo aos algoritmos *k-Means* e C5.0 (árvore de decisão).

A partir dos dados fornecidos, constatou-se que entre os anos de 2007 e 2015, cerca de 745 estudantes abandonaram seus cursos, dentre os quais 610 são do sexo masculino e 135 do sexo feminino. Por outro lado, 1099 estudandes concluíram os estudos, destes 822 são do sexo masculino e 277 são do sexo feminino. Esta informação pode ser visualizada na Tabela 6.

Tabela 6: Número total de estudantes por sexo.

Curso	**Diplomados Sexo Masculino**	**Diplomados Sexo Feminino**	***Dropouts* Sexo Masculino**	***Dropouts* Sexo Feminino**	**Total**
Engenharia Civil	183	91	154	52	**480**
Engenharia de Energias Renováveis	106	35	68	20	**229**
Engenharia Eletrotécnica e de Computadores	146	9	110	4	**269**
Engenharia Informática	160	34	143	21	**358**
Engenharia Mecânica	203	22	117	8	**350**
Engenharia Química	24	86	18	30	**158**
Total Geral	**822**	**277**	**610**	**135**	**1844**

Tomando como base a idade de acesso aos estudos, por curso e ao longo dos nove anos, pode-se construir uma tabela com os principais dados estatísticos relativos a esta informação (Tabela 7) e, dessa forma, traçar um paralelo entre a idade de acesso aos estudos dos estudantes diplomados e dos *dropouts*.

Tabela 7: Idade de acesso aos estudos.

	Diplomados						
Curso	**Mínimo**	**Máximo**	**1º Quartil**	**Mediana**	**3º Quartil**	**Média**	**Desvio Padrão**
Engenharia Civil	18	45	20	22	25	23,14	5,06
Engenharia de Energias Renováveis	18	31	18	19	20	19,42	2,19

Diplomados							
Curso	**Mínimo**	**Máximo**	**1º Quartil**	**Mediana**	**3º Quartil**	**Média**	**Desvio Padrão**
Engenharia Eletrotécnica e de Computadores	18	37	20	23	25	23,41	4,15
Engenharia Informática	18	34	19	20	22	20,89	3,11
Engenharia Mecânica	18	49	19	20	23	21,91	4,86
Engenharia Química	18	31	20	23	26	23,12	3,43
Dropouts							
Engenharia Civil	18	68	19	21	28	24,5	7,97
Engenharia de Energias Renováveis	17	60	18	20	24	22,7	7,24
Engenharia Eletrotécnica e de Computadores	18	67	19	22	29	25,1	8,1
Engenharia Informática	18	41	19	20	22	21	3,58
Engenharia Mecânica	18	50	19	20	24	22,6	5,85
Engenharia Química	18	30	19	20	22	20,9	2,82

Verifica-se na Tabela 7 que a idade de acesso aos estudos dos estudantes que abandonaram os estudos tende a ser superior à idade dos diplomados, indicando que este é um fator que está estritamente relacionado ao abandono. Dessa forma, torna-se possível obter respostas ao seguinte objetivo específico: identificar os fatores que mais influenciam o abandono escolar.

6.2. Evolução do Abandono Escolar

Realizada a análise global e a caracterização da amostra no ponto anterior, busca-se agora responder aos demais objetivos específicos deste estudo para, dessa forma, identificar o perfil dos estudantes em risco de abandono. Assim, com base na amostra gerada, são produzidos gráficos (Figuras 15 - 20) para expressar a evolução do número de estudantes de cada curso ao longo dos nove anos e, consequentemente, atingir o segundo objetivo específico: analisar a evolução do abandono em cada curso ao longo dos 9 anos.

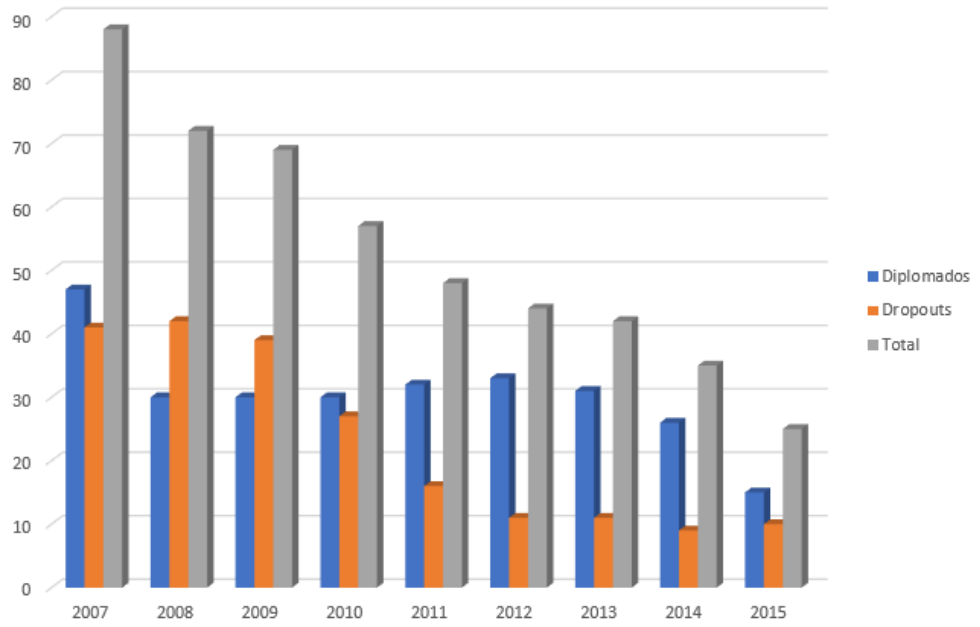

Figura 15: Evolução do número de estudantes de Engenharia Civil.

Com base na Figura 15, pode-se notar a diminuição do número total de estudantes de Engenharia Civil que deixaram o IPB ao longo dos anos, assim como a diminuição do número de *dropouts* a partir do ano de 2009. Nota-se também que, o número de diplomados volta a superar o número de *dropouts* a partir de 2010, contudo, este número volta a cair a partir de 2013.

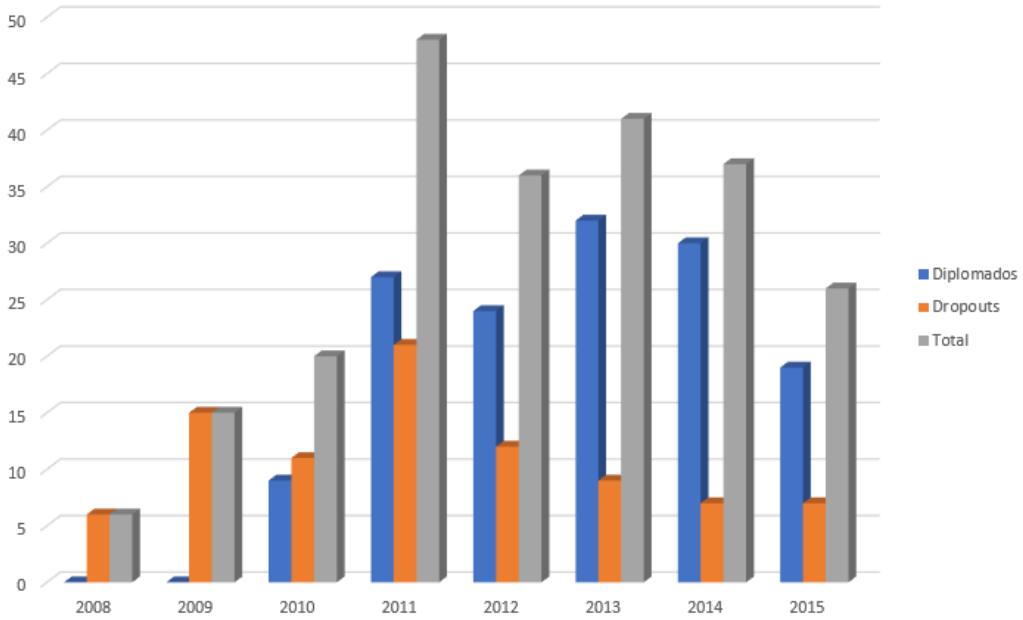

Figura 16: Evolução do número de estudantes de Engenharia de Energias Renováveis.

Conforme ilustrado na Figura 16, nota-se que entre os anos 2008 e 2010, o número de *dropouts* foi maior que o número de diplomados no curso de Engenharia de Energias Renováveis, contudo, deve-se ressaltar que este curso teve início no IPB no ano de 2008 e, portanto, não haveria a possibilidade de existir diplomados nos seus dois primeiros anos de existência. Observa-se que o ano de 2011 foi o ano em que o número de diplomados ultrapassou o número de *dropouts*, o qual, a partir dos anos seguintes, passou a diminuir significativamente.

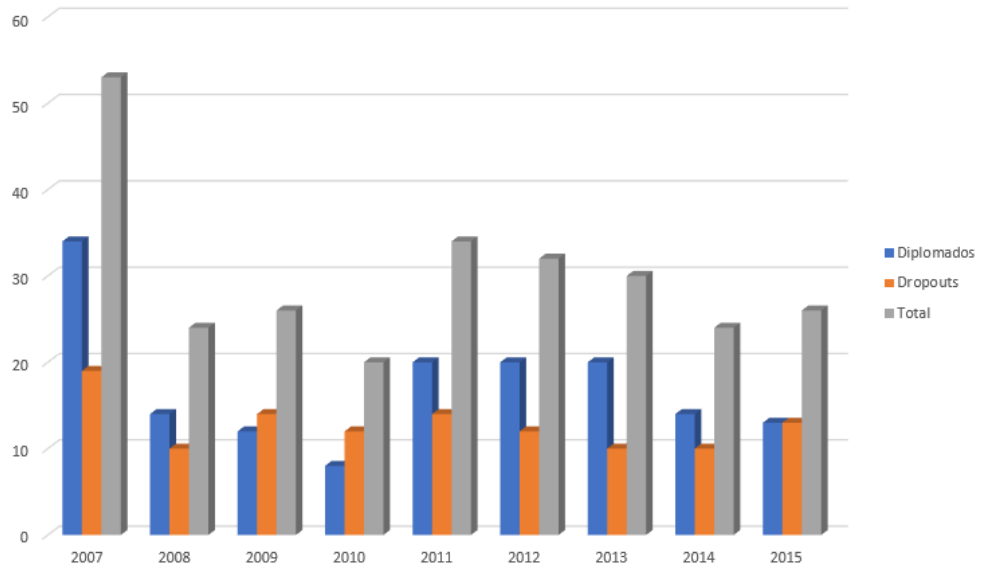

Figura 17: Evolução do número de estudantes de Engenharia Eletrotécnica e de Computadores.

Relativamente ao número total de estudantes que deixam o curso de Engenharia Eletrotécnica e de Computadores, o gráfico apresentado anteriormente nos indica que tal valor sofre muitas variações de ano para ano, bem como o número de estudantes que abandonam o curso, que ultrapassa o número de diplomados nos anos de 2009 e 2010. Com relação aos diplomados, o número desses alunos sofre uma significante queda até o ano de 2010, após isso, tal valor começa a crescer e permanece constante até o ano de 2013, onde volta a sofrer uma queda.

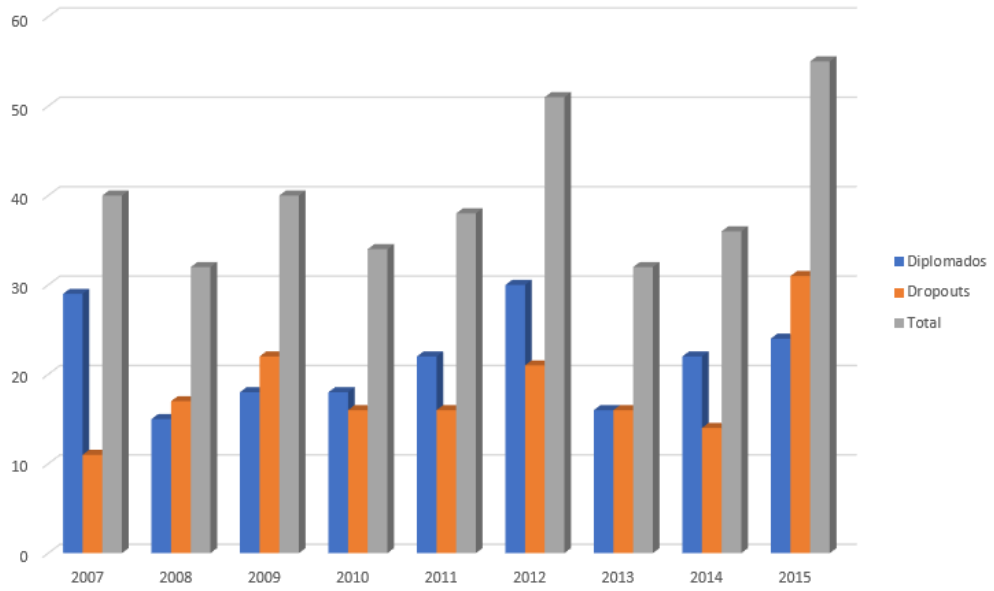

Figura 18: Evolução do número de estudantes de Engenharia Informática.

O gráfico apresentado pela Figura 18 revela que o comportamento demonstrado pelo curso de Engenharia Informática é quase idêntico ao comportamento apresentado pelo curso de Engenharia Eletrotécnica e de Computadores. Por outro lado, nota-se que de 2014 para 2015 ocorreu um ligeiro aumento de diplomados, mas um aumento mais acentuado de *dropouts*. Além disso, pode-se observar que houve uma queda no número de *dropouts* nos dois anos seguintes à mudança do plano de estudos do curso em 2012.

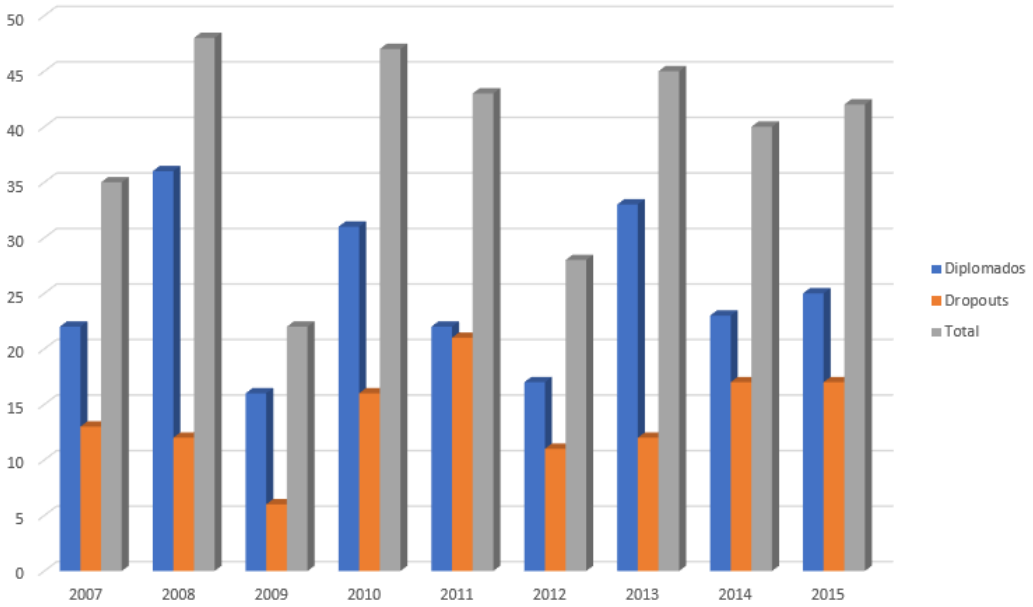

Figura 19: Evolução do número de estudantes de Engenharia Mecânica.

Assim como no curso de Engenharia Informática, o curso de Engenharia Mecânica (Figura 19) apresenta variações significativas no número de estudantes, em especial nos diplomados. Contudo, vale ressaltar que este é o único curso onde o número de diplomados em todos os anos é superior ao número de *dropouts*.

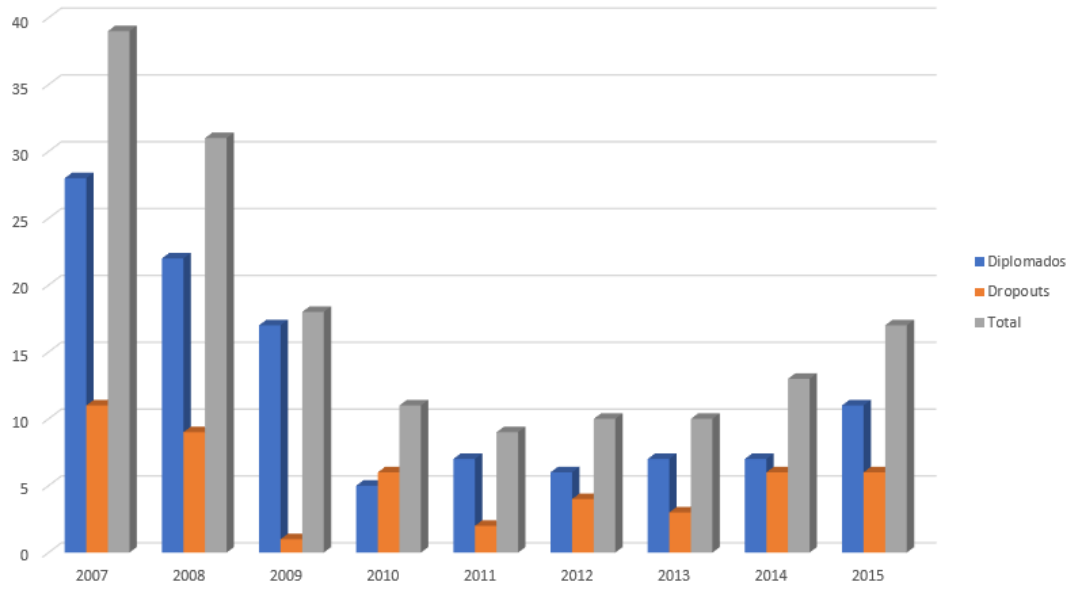

Figura 20: Evolução do número de estudantes de Engenharia Química.

Analisando a Figura 20, é possível observar que, no geral, o número de estudantes que deixam o curso de Engenharia Química tende a ser menor que os demais cursos. Para mais, nota-se que até o ano de 2011 houve uma queda significativamente elevada no número total de estudantes. Além disso, constata-se que 2010 foi o único ano em que o número de *dropouts* ultrapassou o número de diplomados.

Os gráficos apresentados anteriormente nos mostram que, em geral, na maioria dos anos os cursos possuem um número de diplomados superior ao número de *dropouts*, contudo, deve-se destacar o intervalo entre os anos de 2008 e 2010, em que na maioria dos cursos o número de *dropouts* superou o de diplomados em certos anos.

Tendo como base as informações apresentadas anteriormente, e em busca de responder ao nosso primeiro objetivo específico (caracterizar os cursos quanto à sua taxa de abandono), o seguinte gráfico foi gerado:

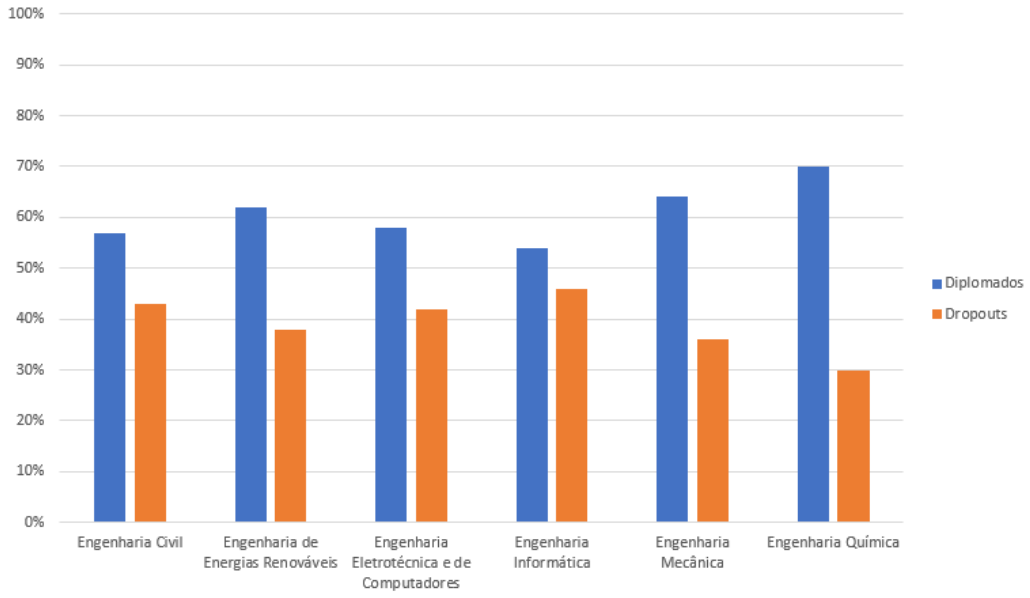

Figura 21: Percentagem de diplomados e *dropouts* por curso.

A Figura 21 nos mostra que, em geral, todos os cursos possuem mais de 50% de estudantes diplomados, contudo, todos também possuem um elevado índice de *dropouts*, em especial, os cursos de Engenharia Civil, Engenharia Eletrotécnica e de Computadores e Engenharia Informática, que possuem mais de 40% de *dropouts* ao longo dos nove anos. Portanto, pode-se concluir que os cursos com o maior e menor número de *dropouts* são Engenharia Informática e Engenharia Química, respectivamente.

Uma análise mais detalhada e tendo por base a quantidade de anos registrados dos *dropouts*, ao longo dos nove anos e por curso (Figura 22), torna visível que a elevada percentagem dos abandonos ocorre antes mesmo dos estudantes estarem registrados por um ano em seus respectivos cursos. Por outro lado, pode-se notar que grande parte dos estudantes abandonam seus cursos mesmo após estarem registrados a um tempo muito superior ao necessário para a conclusão do curso.

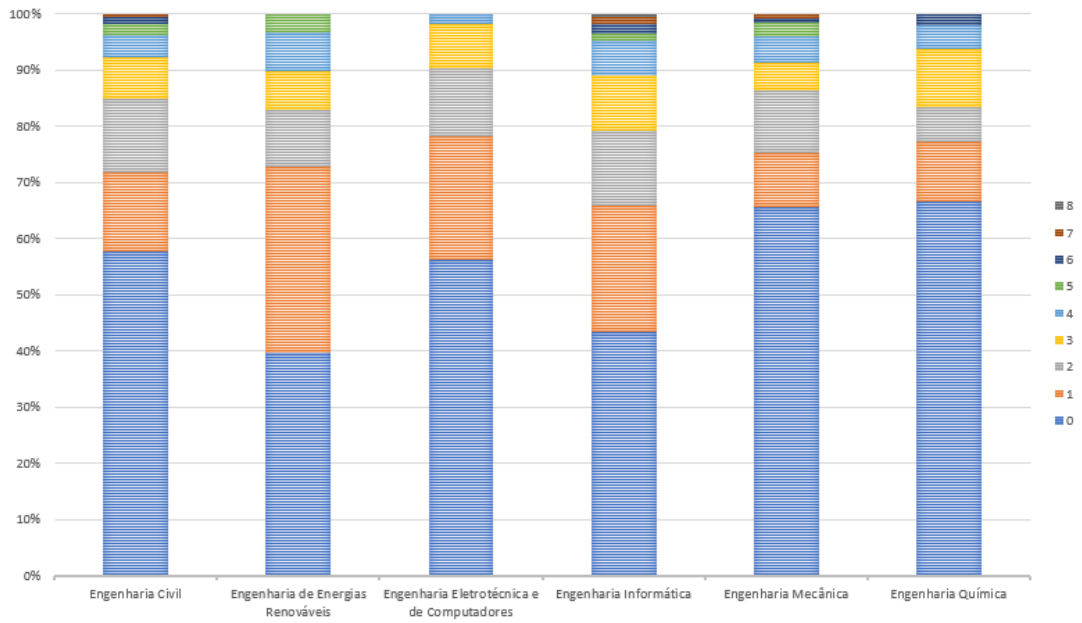

Figura 22: Percentagem de estudantes por anos registrados até o momento de abandono.

Outra análise de suma importância é possível ser feita a partir das disciplinas cursadas e do número de tentativas para aprovação nas mesmas. Por meio dos resultados obtidos com esta análise, pode-se verificar quais são as disciplinas em que os estudantes tendem a ter maior dificuldade para aprovação, independentemente se o estudante encontra-se em risco de abandono ou não. A Tabela 8 apresenta 5 disciplinas em comum entre as 10 disciplinas com maior número de tentativas para aprovação para cada curso entre diplomados e *dropouts*.

Tabela 8: Disciplinas com maior número de tentativas para aprovação.

Engenharia Civil		
Disciplina	**Diplomados Média de Tentativas**	***Dropouts* Média de Tentativas**
Álgebra Linear e Geometria Analítica	2,54	2,04
Cálculo I	2,4	2
Cálculo II	2,3	1,88
Estatística	2,59	1,94
Estruturas I	2,82	2,03
Engenharia de Energias Renováveis		
Eletromagnetismo e Máquinas Elétricas	2,26	2,2
Estatística	2,13	2,7
Física	2,53	2,52
Mecânica dos Fluidos	2,79	3,21
Química-Física e Termoquímica	2,36	1,88

Engenharia Eletrotécnica e de Computadores		
Disciplina	**Diplomados Média de Tentativas**	***Dropouts* Média de Tentativas**
Álgebra Linear e Geometria Analítica	2,87	2,92
Cálculo I	3,25	3,37
Cálculo II	3,67	3,09
Controlo de Sistemas	2,64	2,11
Estatística	3,12	2,69
Engenharia Informática		
Álgebra Linear e Geometria Analítica	2,8	1,85
Algoritmos e Estruturas de Dados	2,48	2,86
Cálculo I	3,51	1,79
Estatística	2,56	2,05
Física	2,17	2,22
Engenharia Mecânica		
Álgebra Linear e Geometria Analítica	2,72	2,07
Cálculo I	2,67	2,25
Cálculo II	2,95	2,29
Mecânica dos Fluídos	1,8	1,89
Métodos Numéricos	2,2	2
Engenharia Química		
Cálculo I	2,72	2,21
Estatística	3,11	2,18
Física	3,16	2,09
Mecânica dos Fluídos	3,25	2,38
Termodinâmica Química I	3,09	2,2

A partir das informações apresentadas na Tabela 8, pode-se notar que muitas das disciplinas em que os estudantes encontram maiores dificuldades para aprovação são disciplinas de formação base, ou seja, disciplinas comuns para todos os cursos de engenharia. Além disso, tais informações nos possibilitam responder a um dos objetivos específicos: identificar as disciplinas com maior taxa de reprovação entre os estudantes que abandonaram os estudos. Dito isso, a Tabela 9 apresenta um comparativo das médias das notas de aprovação entre diplomados e *dropouts* em 5 das principais disciplinas de formação base para os cursos de engenharia abordados neste estudo.

Tabela 9: Média das notas de aprovação em disciplinas de formação base.

Diplomados						
Disciplina	**Civil**	**Energias Renováveis**	**Eletrotécnica e de Computadores**	**Informática**	**Mecânica**	**Química**
Álgebra Linear e Geometria Analítica	11,25	11,37	11,12	11,6	11,73	11,38
Cálculo I	11,05	11,65	11,09	11,8	11,24	11,26
Estatística	11,4	11,92	11,92	11,8	12,22	11,32
Física	11,88	11,86	11,33	11,67	11,94	11,8
Informática	11,44	12,17	11,96	12,95	12,08	11,58
Dropouts						
Álgebra Linear e Geometria Analítica	11,45	11,04	10,71	11,73	11,85	11,33
Cálculo I	11,17	11,45	10,89	11,48	11,18	11,14
Estatística	11,48	12	11,44	12	12,15	11,45
Física	11,53	12	11,08	11,26	11,27	11,36
Informática	11,25	11,26	12	13,13	11,87	11,17

Baseado nas informações apresentadas nas Tabelas 8 e 9, nota-se que, em certos momentos os estudantes que abandonaram possuíram um rendimento escolar superior ao dos diplomados, contudo, algum outro fator acabou levando-os ao abandono. Dessa forma, pode-se notar a influência que fatores externos aos estudos, como àqueles apresentados na seção 2.2 podem ter em situações como esta.

Dito isso, tomando-se como referência os *dropouts*, a Tabela 10 busca analisar e comparar a média de disciplinas aprovadas e a média de tentativas para aprovação nas mesmas. Devido a alta concentração de estudantes no primeiro quartil, decidiu-se efetuar a distribuição por decis, o que proporciona a melhor visualização e comparação dessas variáveis.

Tabela 10: Média de disciplinas aprovadas e tentativas para aprovação dos *dropouts*.

Decil	**Média de Disciplinas**	**Média de Tentativas**
1	0	0
2	0	0
3	0,51	0,51
4	1,49	1,61
5	2,84	3,56
6	4,51	5,08
7	5,55	6,89
8	7,72	10,01

Decil	Média de Disciplinas	Média de Tentativas
9	11,22	16,55
10	22,74	34,46

Observa-se por meio da Tabela 10 que, nos três primeiros decis as médias de disciplinas e tentativas possuem os mesmos valores, estes sendo iguais ou muito próximos de zero. A partir disso e com base no número total de *dropouts*, pode-se constatar que dos 745 alunos, 187 (25%) nunca fizeram nenhuma tentativa, ou seja, estes abandonaram seus cursos sem ter cursado ao menos uma disciplina.

Com o objetivo de se fazer uma análise comparativa entre os *dropouts* de todos os cursos, e levando-se em consideração a média das notas de todas as disciplinas dos cursos, bem como a média do número de tentativas para aprovação nas disciplinas, foram criados o gráfico e a tabela a seguir (Figura 23 e Tabela 11).

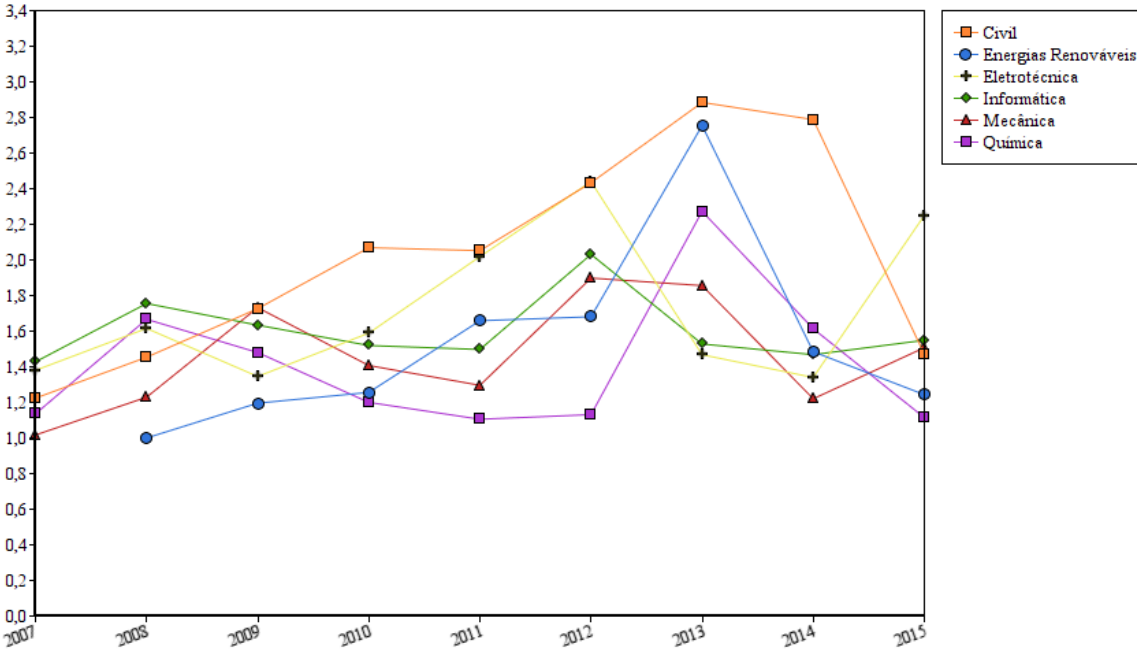

Figura 23: Média de tentativas para aprovação nas disciplinas dos *dropouts*.

A Figura 23 nos permite observar muitas variações em todos os cursos, contudo, vale destacar o curso de Engenharia Civil onde na maioria dos anos, este apresentou a média de tentativas superior à média de todos os outros cursos. Deve-se também destacar os anos de 2011 e 2012, em que a média de tentativas sofre um crescimento significativo.

Da mesma forma, a média de tentativas de muitos cursos passou a sofrer uma grande queda entre os anos de 2012 e 2013.

Tabela 11: Média das notas em todas as disciplinas dos cursos.

Ano	Civil	Energias Renováveis	Eletrotécnica e de Computadores	Informática	Mecânica	Química
2007	11,29		11,94	12,09	11,65	11,64
2008	11,59	11,76	11,58	12,52	11,75	12,1
2009	11,6	11,26	11,45	11,77	10,95	12,2
2010	11,66	11,39	11,62	12,43	11,74	12,43
2011	11,67	10,9	11,35	11,98	11,98	13,16
2012	11,8	11,32	12	12,6	11,89	13,36
2013	11,85	11,86	11,88	12,67	12,99	12,68
2014	11,67	12,55	11,8	12,84	11,83	12,77
2015	12,71	11,97	11,81	11,78	12,71	12,97

Baseado na Tabela 11, verifica-se que não existem diferenças significativas entre as médias dos cursos e, até mesmo, entre as médias do próprio curso. Apesar disso, é possível observar que o curso de Engenharia Química é o curso que, no geral, costuma possuir a maior média dentre os cursos.

A diminuição do número de tentativas para aprovação às disciplinas depende de vários fatores, nomeadamente de uma melhor preparação dos estudantes, adoção de novas estratégias de ensino, entre outros. Verificou-se nos últimos anos uma adoção mais generalizada de metodologias de avaliação contínua, o que tem impacto direto no número médio de tentativas para fazer as disciplinas.

6.3. Modelo para Previsão de Estudantes em Risco de Abandono Escolar

Como consequência das análises realizadas anteriormente, pode-se concluir que, relativamente aos cursos de engenharia e aos fatores internos ao IPB, as principais características que podem ajudar a distinguir entre um estudante com tendências a concluir o curso e um estudante que pode estar em risco de abandono são: a idade de

acesso aos estudos, o número de disciplinas aprovadas, o número de tentativas para aprovação e a média das notas.

Dito isso, baseado nas notas dos *dropouts* e no sistema de classificação nacional, o gráfico abaixo foi gerado com o objetivo de identificar a percentagem de estudantes que se enquadra em cada situação. Segundo o sistema de classificação nacional, estudantes com notas entre 10 e 13 são classificados como "suficiente", entre 14 e 17 como "bom" e entre 18 e 20 como "muito bom". A partir disso, observa-se que 81,7% dos estudantes são classificados como "suficiente", enquanto que, menos de 1% são classificados como "muito bom".

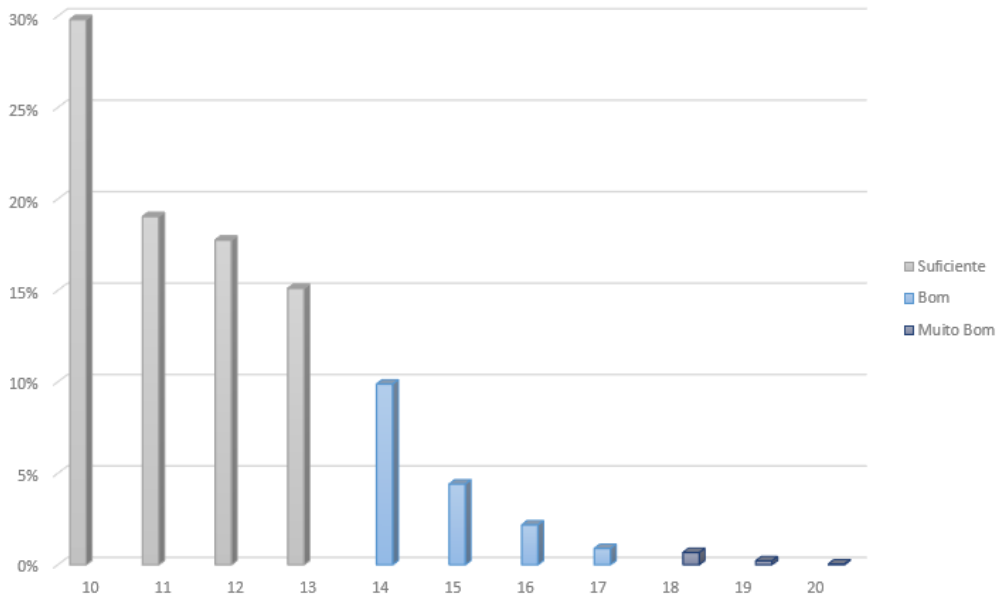

Figura 24: Classificação dos *dropouts* nas disciplinas em que foram aprovados.

Em seguida, com a finalidade de agrupar os estudantes baseado nas demais características citadas acima, utilizou-se o algoritmo *k-Means* (apresentado na seção 4.4.7) para identificar esses grupos e as características inerentes a cada um deles. Em busca de determinar o número ideal de *clusters* (k), foi aplicado o *Elbow Method* (Figura 25), método projetado para ajudar a encontrar o número apropriado de *clusters* em um conjunto de dados.

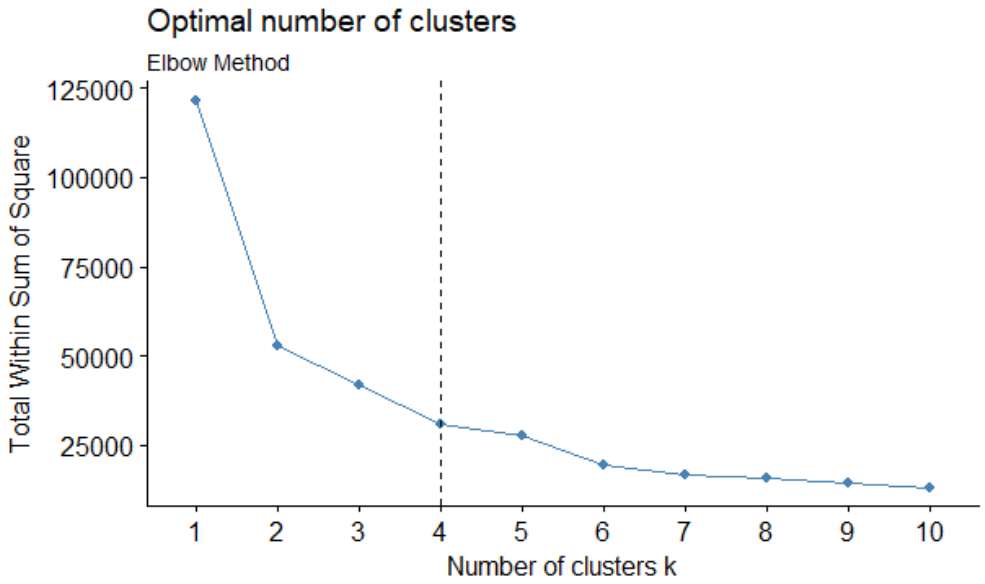

Figura 25: Definição do número ideal de *clusters*.

Por meio da aplicação do *Elbow Method*, nota-se que o número ideal de *clusters* para agrupar os estudantes é 4. Dessa forma, como resultado do agrupamento em função da idade de acesso aos estudos, a Tabela 12 apresenta os valores dos centroides obtidos e que podem ser melhor visualizados nos gráficos contidos na Figura 26.

Tabela 12: Centroides resultantes do algoritmo *k-Means*.

Cluster	**Acesso aos Estudos**	**Número Disciplinas**	**Número Tentativas**
1 (vermelho)	36,61	7,25	7,68
2 (azul)	20,17	3,61	4,21
3 (verde)	25,62	23,46	35,97
4 (ciano)	21,34	10,15	15,55

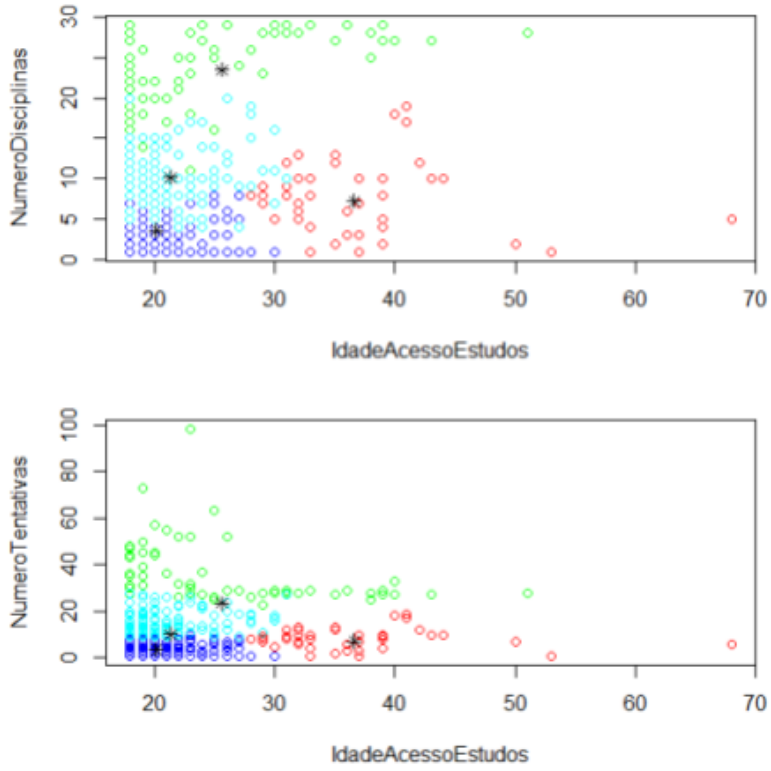

Figura 26: Agrupamento resultante do algoritmo *k-Means*.

Ao analisar os gráficos da Figura 26, é possível identificar uma grande diferença na situação do *cluster* azul em relação à do verde. Apesar da idade de ingresso aos estudos de ambos não serem tão distintas, nota-se que os estudantes pertencentes ao *cluster* azul costumam tentar menos, contudo, também possuem menos disciplinas. Por outro lado, os estudantes localizados no *cluster* verde, possuem um elevado número de tentativas e de disciplinas, chegando, em alguns casos, a estar muito próximo do número total de disciplinas que um curso de engenharia do IPB possui em média (30 disciplinas).

Com base nisso, para a implementação do modelo de classificação proposto, recorreu-se à utilização da árvore de decisão C5.0, uma evolução da árvore de decisão C4.5 (apresentada na seção 4.4.1). A ideia é que, baseado na idade de acesso aos estudos, na média das notas e na média de tentativas para aprovação nas disciplinas, seja possível identificar se um estudante apresenta características de um estudante diplomado ou um estudante que abandonou os estudos e, a partir disso, verificar se este encontra-se em risco de abandono. O quadro apresentado a seguir contém o código da implementação do algoritmo utilizado.

Listagem 3: Implementação da Árvore de Decisão C5.0.

```
1 install.packages("C50")
2 require(C50)
3
4 model <- C5.0(students[1:floor(nrow(students)*0.7), -4], +
5               students[1:floor(nrow(students)*0.7), 4])
6 model
7 summary(model)
8
9 predict_result <- predict(model, students[(floor(nrow(students)*0.7) + 1):nrow(students),])
10 predict_result
11
12 table(students[(floor(nrow(students)*0.7) + 1):nrow(students), 4], +
13                 Predicted = predict_result)
```

Como pode-se observar no código acima, para a fase de treinamento, foi utilizado 70% do conjunto de dados total. Além disso, com o objetivo de facilitar a visualização completa da árvore (Figura 29), foram utilizadas letras para representar os estudantes diplomados (D) e os estudantes que abandonaram os cursos (A). A Figura 27 apresenta os resultados obtidos na fase de treinamento do algoritmo.

```
Evaluation on training data (1159 cases):

        Decision Tree
      ----------------
      Size      Errors

        12  150(12.9%)   <<


       (a)   (b)    <-classified as
      ----  ----
       203   132    (a): class A
        18   806    (b): class D


Attribute usage:

100.00% MediaNotas
100.00% MediaTentativas
 31.06% IdadeAcesso
```

Figura 27: Resultado da fase de treinamento.

Observa-se que o tamanho da árvore é 12 e, dos 1159 casos, o algoritmo classificou corretamente 1009, possuindo uma margem de erro de 12,9%. É possível notar que dos 335 casos de abandono, o algoritmo classificou corretamente 203, enquanto que, dos 824 casos de diplomados, foram classificados de maneira precisa 806. Ao concluir o treinamento, aplicou-se os outros 30% do conjunto de dados na fase de testes do

algoritmo. A Figura 28 exibe o resultado da comparação entre os resultados preditos e os valores reais do conjunto de dados.

	Predicted	
	A	D
A	150	73
D	19	256

Figura 28: Comparação entre os resultados preditos e os valores reais.

Ao analisar a Figura 28, conclui-se que o modelo proposto por este estudo tem uma precisão de 67,3% para identificar características relativas aos *dropouts*, enquanto que, 93,1% para os diplomados.

A Figura 29 demonstra a relação entre os elementos do modelo proposto (nós e folhas) e seus atributos durante a fase de treinamento. Nesse caso, em cada nó folha pode-se observar a existência de duas cores, o cinza claro (*dropouts*) e o cinza escuro (diplomados) bem como seus respectivos percentuais de acerto. Além disso, a quantidade de acertos em cada nó será representada pela variável n.

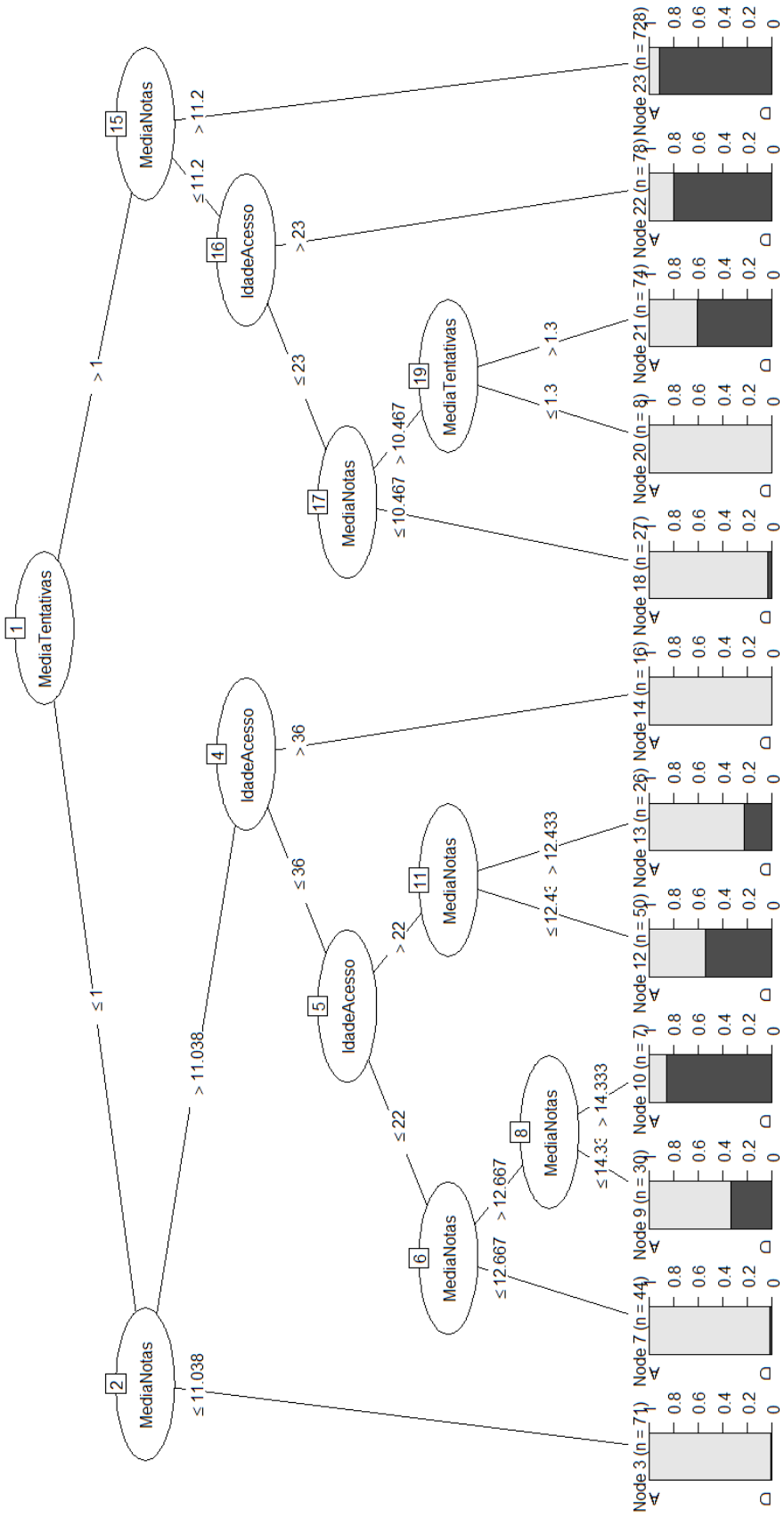

Figura 29: Árvore de decisão para classificação de estudantes em risco de abandono escolar.

Capítulo 7 Conclusões

7.1. Considerações Finais e Perspectivas Futuras

A análise apresentada anteriormente teve como principal objetivo identificar o perfil dos estudantes dos cursos de licenciatura em engenharia da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Bragança e, com isso, propor um modelo de previsão com base naqueles que possam estar em risco de abandono.

O conjunto de dados utilizado nesse estudo foi extraído da própria base de dados do IPB, e levou em consideração os cursos de Engenharia Civil, Engenharia de Energias Renováveis, Engenharia Eletrotécnica e de Computadores, Engenharia Informática, Engenharia Mecânica e Engenharia Química.

Como principais conclusões deste trabalho e para o período em análise (2007 a 2015) pode-se concluir que:

- A maior parte dos *dropouts* ocorre antes mesmo de os estudantes estarem registados por 1 ano em seus respectivos cursos;
- Em geral, as disciplinas com maior taxa de reprovação tendem a ser as disciplinas de formação base;
- Nota-se em muitos casos que, os estudantes que abandonaram os cursos possuem um rendimento escolar muito semelhante ao dos estudantes diplomados;
- Em média os estudantes diplomados costumam possuir mais tentativas que os *dropouts*;
- Na maioria dos anos o número de diplomados foi superior ao número de *dropouts*, contudo, o número total de *dropouts* é elevado;

- Menos de 1% dos estudantes que abandonaram são classificados como “muito bom” segundo o sistema de classificação nacional;
- 25% dos estudantes nunca fizeram nenhuma tentativa, ou seja, abandonaram seus cursos sem ter cursado ao menos uma disciplina;
- 6% dos estudantes abandonaram faltando 5 disciplinas ou menos para a conclusão do curso;
- O modelo proposto tem uma precisão de 67,3% para identificar características relativas ao abandono.

O presente estudo busca contribuir com as diretrizes definidas pelo Instituto Politécnico de Bragança, para que assim, seja possível identificar ações preventivas de combate ao abandono, tais como: monitorar os anos posteriores a 2015 para verificar o comportamento dos estudantes que abandonaram o IPB; e identificar o estudante em risco de abandonar os estudos e acompanhá-lo a fim de tentar reverter a situação.

Como perspectivas para trabalhos futuros, destacam-se alguns desafios:

- Acrescentar dados qualitativos à mineração;
- Aplicar outras técnicas de mineração de dados como por exemplo o *Deep Learning* para encontrar outras variáveis que caracterizam o perfil dos estudantes em risco de abandono;
- Utilizar técnicas de *Learning Analytics* no sentido de analisar o percurso de aprendizagem dos estudantes, usando outras fontes de informação como por exemplo as presenças nas aulas e a utilização do ambiente virtual de aprendizagem.

O desenvolvimento de um modelo baseado em árvore de decisão proporciona a criação de instruções padronizadas, facilidade de interpretação e permite a adição de vários cenários possíveis para a identificação do perfil de estudantes e para a classificação do risco de abandono, contribuindo significativamente com as IES no processo de tomada de decisão.

Bibliografia

[1] Direção-Geral do Ensino Superior, "Sistema de Ensino Superior Português." [Online]. Available: https://www.dges.gov.pt/pt/pagina/sistema-de-ensino-superior-portugues. [Accessed: 17-May-2018].

[2] L. Bianchetti, "The Bologna Process and the Subjugation of Higher Education to the Training for the Market," vol. 1, pp. 426–435, 2014.

[3] Ministério da Ciência Tecnologia e Ensino Superior, "Decreto-Lei nº 74/2006," pp. 2242–2257, 2006.

[4] I. A. N. Cruz and F. G. R. P. Junior, "O Processo de Bolonha e o Espaço Europeu de Ensino Superior: Notas Introdutórias Sobre o Quadro Europeu de Qualificações (QEQ) e os Quadros Nacionais de Qualificação (QNQ)," vol. 7, pp. 59–68.

[5] A. L. Prim and J. D. Fávero, "Motivos da evasão escolar nos cursos de ensino superior de uma faculdade na cidade de Blumenau," *E-Tech Tecnol. para Compet. Ind.*, vol. especial, pp. 53–72, 2013.

[6] F. Ferreira and P. Fernandes, "Fatores que Influenciam o Abandono no Ensino Superior e Inicitaivas para a Sua Prevenção: O Olhar de Estudantes," *Educ. Soc. Cult.*, vol. 45, pp. 177–197, 2015.

[7] J. M. Riiheläinen, "Structural Indicators on Higher Education in Europe – 2016," 2016.

[8] S. J. Rigo, W. Cambruzzi, J. L. V. Barbosa, and S. C. Cazella, "Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios," *Rev. Bras. Informática na Educ.*, vol. 22, no. 01, p. 132, 2014.

[9] S. M. Cunha and D. M. Carrilho, "O Processo de Adaptação ao Ensino Superior e o Rendimento Acadêmico: Adaptação e Rendimento Acadêmico," *Psicol. Esc. e Educ.*, vol. 9, no. 2, pp. 215–224, 2005.

[10] Comissão Europeia, "A Modernização do Ensino Superior na Europa: Acesso,

Retenção e Empregabilidade," 2014.

[11] A. J. C. Kampff, "Mineração de Dados Educacionais para Geração de Alertas em AVAs como Apoio à Prática Docente," 2009.

[12] C. C. Gibson, "The distance learner's academic self-concept," in *Distance Learners in Higher Education: Institutional Responses for Quality Outcomes*, 0th ed., 1998.

[13] J. Fernandes, A. da S. Ferreira, D. C. de O. Nascimento, E. Shimoda, and G. F. Teixeira, "Identificação de fatores que influenciam na evasão em um curso superior de ensino à distância," *Perspect. Online*, pp. 80–91, 2010.

[14] J. Quinn, "Drop-out and Completion in Higher Education in Europe: among students from under-represented groups," no. October, p. 104, 2013.

[15] A. Benavente, J. Campiche, T. Seabra, and J. Sebastião, *Renunciar à Escola: O Abandono Escolar no Ensino Básico*, no. May 2015. Fim de Século, 1994.

[16] N. Alves, A. N. Almeida, and M. M. Vieira, "Da Normatividade da Definição de Sucesso Escolar no Ensino Superior à Pluralidade das Vivências Estudantis: Trajetórias e Perfis de Mobilidade na Universidade de Lisboa," *CLABES*, 2013.

[17] S. Benson and C. Standing, *Information Systems: A Business Approach.*, 3rd Editio. Australia: John Wiley & Sons Australia, Ltd, 2008.

[18] C. E. Cayres, J. R. D. Oliveira, and A. Marini, "Business Intelligence Na Era Da Informação E As Vantagens Do Oracle Na Efetivação Dessa Tecnologia," vol. IV, pp. 59–73, 2009.

[19] J. L. G. Alves, "Sistema de Business Intelligence no Projeto Educativo de Guimarães," 2015.

[20] Gartner IT Group, "Business Intelligence - BI," 2014. [Online]. Available: http://www.gartner.com/it-glossary/business-intelligence-bi/. [Accessed: 09-Oct-2017].

[21] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Second. San Francisco: Elsevier, 2005.

[22] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[23] H. H. Sferra and Â. M. C. J. Corrêa, "Conceitos e Aplicações de Data Mining," *Rev. Ciência Tecnol.*, pp. 19–34, 2003.

[24] A. B. E. D. Ahmed and I. S. Elaraby, "Data Mining : A prediction for Student ' s Performance Using Classification Method," *World J. Comput. Appl. Technol.*, vol. 2, no. 2, pp. 43–47, 2014.

[25] Oracle, "Oracle Data Mining Basics," 2018. [Online]. Available: https://docs.oracle.com/en/database/oracle/oracle-database/18/dmcon/data-mining-basics.html#GUID-2116E665-721E-4EBA-AFE1-A30D6E8078C6. [Accessed: 25-May-2018].

[26] M. C. C. Lima, "Business intelligence no ensino superior – um caso de estudo," 2012.

[27] T. K. Das and A. Mohapatro, "A Study on Big Data Integration with Data Warehouse," *Int. J. Comput. Trends Technol.*, vol. 9, no. 4, pp. 188–192, 2014.

[28] Oracle, "Data Warehousing Concepts." [Online]. Available: https://docs.oracle.com/cd/B10500_01/server.920/a96520/concept.htm. [Accessed: 12-Oct-2017].

[29] B. Wixom *et al.*, "The Current State of Business Intelligence in Academia : The Arrival of Big Data," vol. 34, no. January, pp. 1–13, 2014.

[30] M. Bienkowski, M. Feng, and B. Means, "Enhancing teaching and learning through educational data mining and learning analytics: An issue brief," *Washington, DC SRI Int.*, pp. 1–57, 2012.

[31] P. D. Scaico, R. J. G. B. de Queiroz, and A. Scaico, "O conceito big data na educação," *An. do Work. Informática na Esc.*, vol. 20, no. 1, pp. 328–336, 2014.

[32] IBM, "What is big data?," 2011. [Online]. Available: http://www-01.ibm.com/software/data/bigdata/. [Accessed: 13-Oct-2017].

[33] J. Ferreira, "Big Data in Education: The 5 Types That Matter," 2013. [Online].

Available: http://www.knewton.com/blog/knewton/from-jose/2013/07/18/big-data-in-education. [Accessed: 13-Oct-2017].

[34] R. C. de Souza, "Aplicação de Learning Analytics para Avaliação do Desempenho de Tutores a Distância," Universidade do Estado do Rio Grande do Norte e Universidade Federal Rural do Semi-Árido, 2016.

[35] B. K. Daniel, *Big Data and Learning Analytics in Higher Education: Current Theory and Practice*, 1st ed. Springer, 2017.

[36] E. Costa, R. S. J. Baker, L. Amorim, J. Magalhães, and T. Marinho, "Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações," vol. d, pp. 1–29, 2012.

[37] International Educational Data Mining Society, "Educational Data Mining." [Online]. Available: http://educationaldatamining.org/%0A. [Accessed: 17-Oct-2017].

[38] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 12–27, 2013.

[39] R. Baker, S. Isotani, and A. Carvalho, "Mineração de Dados Educacionais: Oportunidades para o Brasil," *Rev. Bras. Informática na Educ.*, vol. 19, no. 02, pp. 3–13, 2011.

[40] S. M. S. M. L. de Faria, "Educational Data Mining e Learning Analytics na melhoria do ensino online," 2014.

[41] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, "Data mining algorithms to classify students," *Educ. Data Min. 2008*, no. November, pp. 8–17, 2008.

[42] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Second. Elsevier Inc., 2006.

[43] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont CA: Wadsworth Inc, 1984.

[44] R. L. Rodrigues, F. P. A. De Medeiros, and A. S. Gomes, "Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem," no. Cbie, pp. 607–616, 2013.

[45] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Third. Pearson, 2009.

[46] H. M. Seffrin and P. Jaques, "Avaliando o Conhecimento Algébrico do Estudante através de Redes Bayesianas Dinâmicas," no. Cbie, pp. 782–791, 2015.

[47] R. E. Neapolitan, *Learning Bayesian Networks*. Pearson, 2003.

[48] I. Ben-Gal, "Bayesian Networks," p. 7, 2008.

[49] H. Ferreira, R. D. Araújo, F. Dorça, and R. Cattelan, "Uma Abordagem Híbrida para Acompanhamento da Aprendizagem do Estudante Baseada em Ontologias e Redes Bayesianas em Sistemas Adaptativos para Educação," no. Cbie, p. 447, 2016.

[50] S. D. C. Côrtes, R. M. Porcaro, and S. Lifschitz, "Mineração de Dados – Funcionalidades, Técnicas e Abordagens," *PUC-Rio Informática*, p. 35, 2002.

[51] T. W. Rauber, "Redes Neurais Artificiais," no. May, p. 228, 2014.

[52] R. Elmasri and S. B. Navathe, *Fundamentals of Database Systems*, Sixth. Pearson, 2010.

[53] L. P. Reis, J. Vieira, P. Lemos, R. Novais, and B. M. Faria, "Higher Education Access Prediction using Data-Mining," 2017.

[54] M. I. G. Moreira, A. C. da R. Costa, and M. S. Aguiar, "Legislative Virtual Learning Environment: A Case Study in the Management of Distance Professional Education," vol. 14, pp. 264–283, 2017.

[55] A. C. Cunha, "Localização de Dispositivos Móveis Usando Roteadores com Antenas Direcionais e Classificação de Dados," Universidade Federal do Amazonas, 2014.

[56] M. F. Pinheiro, L. C. F. Neto, H. N. de S. Júnior, E. C. da Mata, A. F. L. J. Jr., and Á. de L. Santana, "Identificação de Grupos de Alunos em Ambiente Virtual de Aprendizagem : Uma Estratégia de Análise de Log Baseada em Clusterização," *3º Congr. Bras. Informática na Educ. - Work. (WCBIE*, no. Cbie, pp. 582–591, 2014.

[57] A. Araar and A. Haddad, “Identification of New Connections for IP Intrusion Detections using WEKA Platform and KDD Cup 99 1 1,” *J. Emerg. Trends Comput. Inf. Sci.*, vol. 6, no. 1, pp. 7–14, 2015.

[58] R. Souza, F. M. Neto, A. Santos, L. Fontes, E. Naassom, and R. Valentim, “Um Ambiente Inteligente de Avaliação de Comportamentos de Tutores e Turmas no Ambiente Virtual de Aprendizagem Moodle,” no. Cbie, p. 417, 2016.

[59] D. Arthur and S. Vassilvitskii, “K-Means++: The Advantages of Careful Seeding,” *Proc. eighteenth Annu. ACM-SIAM Symp. Discret. algorithms*, pp. 1027–1025, 2007.

[60] R. Linden, “Técnicas de Agrupamento,” *Rev. Sist. Informação da FSMA*, vol. 4, pp. 18–36, 2009.

[61] M. M. Faria, “Detecção de Intrusões em Redes de Computadores com Base nos Algoritmos KNN, K-Means++ e J48,” Faculdade Campo Limpo Paulista, 2016.

[62] S. A. Ferreira and A. Andrade, “Academic analytics: Anatomy of an exploratory essay,” *Educ. Inf. Technol.*, vol. 21, no. 1, pp. 229–243, 2014.

[63] F. Matsebula and E. Mnkandla, “A Big Data Architecture for Learning Analytics in Higher Education,” *IEEE Africon 2017 Proc.*, 2017.

[64] S. B. Shum, “Policy Brief: Learning Analytics,” *UNESCO*, 2012.

[65] S. A. Ferreira and A. Andrade, “Academic Analytics: Mapeando o Genoma da Universidade,” *Rev. Iberoam. Tecnol. del / da Aprendiz. / Aprendiz.*, vol. 1, no. 3, pp. 167–174, 2013.

[66] P. Long and G. Siemens, “Penetrating the Fog: Analytics in Learning and Education,” *Educ. Rev.*, vol. 46, pp. 30–32, 2011.

[67] P. Baepler and C. J. Murdoch, “Academic Analytics and Data Mining in Higher Education,” *Int. J. Scholarsh. Teach. Learn.*, vol. 4, no. 2, pp. 1–9, 2010.

[68] M. Joshi, P. Bhalchandra, A. Muley, and P. Wasnik, “Analyzing Students Performance Using Academic Analytics,” *Proc. 2016 Int. Conf. ICT Business, Ind. Gov. ICTBIG 2016*, pp. 0–3, 2016.

[69] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, “Academic Analytics: A New Tool for a New Era,” no. August 2007, 2007.

[70] U. Bin Mat, N. Buniyamin, P. M. Arsad, and R. A. Kassim, “An Overview of Using Academic Analytics to Predict and Improve Students’ Achievement: A Proposed Proactive Intelligent Intervention,” *2013 IEEE 5th Int. Conf. Eng. Educ. Aligning Eng. Educ. with Ind. Needs Nation Dev. ICEED 2013*, pp. 126–130, 2013.

[71] D. Norris, L. Baer, J. Leonard, L. Pugliese, and P. Lefrere, “Action Analytics: Measuring and Improving Performance That Matters in Higher Education,” 2008.

[72] A. M. de Oliveira, F. S. dos S. Canedo, R. Y. C. Himeno, and G. Bruschi, “Análise de Performance do Banco de Dados Oracle 11G R2 com Diferentes Tecnologias de Armazenamento,” *Cad. Estud. Tecnológicos*, vol. 3, no. 1, pp. 1–26, 2015.

[73] C. J. Date, *Introdução a Sistemas de Bancos de Dados*, 8th ed. Campus, 2003.

[74] F. G. Rodrigues and E. D. O. Silva, “Aplicação da automação de testes de software em SGBD’s : Um estudo de caso utilizando o banco de dados Oracle,” no. 1999, 2015.

[75] R. Rezende, “Conceitos Fundamentais de Banco de Dados,” 2014. .

[76] M. R. Vieira, J. M. De Figueiredo, G. Liberatti, and A. F. M. Viebrantz, “Bancos de Dados NoSQL: Conceitos, Ferramentas, Linguagens e Estudos de Casos no Contexto de Big Data,” no. 1, pp. 1–30, 2012.

[77] M. Cavalcante, “Banco de dados NoSQL: Um novo paradigma - Revista SQL Magazine 102.” [Online]. Available: https://www.devmedia.com.br/banco-de-dados-nosql-um-novo-paradigma-revista-sql-magazine-102/25918. [Accessed: 24-May-2018].

[78] D. P. Ballou and G. K. Tayi, “Enhancing Data Quality in Data Warehouse Environments,” *Commun. Acm*, vol. 42, no. 1, pp. 73–78, 1999.

[79] W. H. Inmon, K. Rudin, C. K. Buss, and R. Sousa, *Data Warehouse Performance*. Wiley, 1998.

[80] P. J. Oliveira, F. Rodrigues, and P. R. Henriques, "Limpeza de Dados - Uma Visão Geral," *SDDI - II Simpósio Doutoral do DI*, 2004.

[81] E. Rahm and H. Hai Do, "Data Cleaning: Problems and Current Approaches," *IEEE*, pp. 1–11, 2000.

[82] D. Elias, "Entendendo o processo de ETL," 2014. [Online]. Available: https://canaltech.com.br/business-intelligence/entendendo-o-processo-de-etl-22850/. [Accessed: 25-May-2018].

[83] Oracle, "Database Concepts," 2018. [Online]. Available: https://docs.oracle.com/en/database/oracle/oracle-database/18/cncpt/introduction-to-oracle-database.html#GUID-A42A6EF0-20F8-4F4B-AFF7-09C100AE581E. [Accessed: 24-May-2018].

[84] N. Alhadi and K. Ahmad, "Query Tuning in Oracle Database," *J. Comput. Sci.*, vol. 8, no. 11, pp. 1889–1896, 2012.

[85] DB-Engines, "DB-Engines Ranking," 2018. [Online]. Available: https://db-engines.com/en/ranking. [Accessed: 24-May-2018].

[86] Oracle, "Introduction to Oracle Data Mining," 2018. [Online]. Available: https://docs.oracle.com/en/database/oracle/oracle-database/18/dmapi/introduction-to-oracle-data-mining.html#GUID-429CF74D-C4B7-4302-9C33-5292A664E2AD. [Accessed: 25-May-2018].

[87] G. Lima, "Análise de dados na prática com R Studio." [Online]. Available: https://www.devmedia.com.br/analise-de-dados-na-pratica-com-r-studio/39279. [Accessed: 26-May-2018].

[88] M. de S. Lauretto, "Introdução à Análise de Dados Utilizando o Ambiente R," 2015. [Online]. Available: http://each.uspnet.usp.br/lauretto/cursoR2015/cursoR2015.pdf. [Accessed: 26-May-2018].

[89] I. Cegatta, "Introdução à linguagem R para análise de dados," 2018. [Online]. Available: https://italocegatta.github.io/cursoR_2017_10/. [Accessed: 26-May-2018].

[90] IBPAD, “O que é Programação ou Linguagem em R?,” 2017. [Online]. Available: https://www.ibpad.com.br/blog/o-que-e-programacao-ou-linguagem-em-r/. [Accessed: 26-May-2018].

[91] Full Join, “Kit de Sobrevivência em R - Parte 3: Pacotes,” 2016. [Online]. Available: http://fulljoin.com.br/blog/2016/04/03/kit-de-sobrevivencia-em-r-parte-3/. [Accessed: 26-May-2018].

[92] CRAN, “Contributed Packages,” 2018. [Online]. Available: https://cran.r-project.org/web/packages/index.html. [Accessed: 30-May-2018].

[93] Y. Zhao, “R and Data Mining: Examples and Case Studies,” *Acad. Press*, no. December 2012, pp. 1–160, 2015.

[94] CRAN, “CRAN Task Views,” 2018. [Online]. Available: https://cran.r-project.org/. [Accessed: 30-May-2018].