CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA

FACULDADE DE TECNOLOGIA DE INDAIATUBA

DR. ARCHIMEDES LAMOGLIA

CURSO DE TECNOLOGIA EM

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

GABRIEL VIEIRA SANTELLO

Aplicação de técnica de *Big Data* para análise comparativa do desempenho acadêmico dos alunos formados pela FATEC Indaiatuba nos cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial

INDAIATUBA
2019

CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA

FACULDADE DE TECNOLOGIA DE INDAIATUBA

DR. ARCHIMEDES LAMOGLIA

CURSO DE TECNOLOGIA EM

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

GABRIEL VIEIRA SANTELLO

Aplicação de técnica de *Big Data* para análise comparativa do desempenho acadêmico dos alunos formados pela FATEC Indaiatuba nos cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial

Trabalho de Graduação apresentado por Gabriel Vieira Santello como pré-requisito para a conclusão do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, da Faculdade de Tecnologia de Indaiatuba, elaborado sob a orientação do Profa. Dra. Maria das Graças J. M. Tomazela.

INDAIATUBA
2019

CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA

FACULDADE DE TECNOLOGIA DE INDAIATUBA

DR. ARCHIMEDES LAMOGLIA

CURSO DE TECNOLOGIA EM

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

GABRIEL VIEIRA SANTELLO

Banca Avaliadora:

Profª. Dra. Maria das Graças J. M. Tomazela	Orientadora
Prof. Dr. Aldo Nascimento Pontes	Professor Convidado
Ramon Pansonato	Profissional da Área

Data da defesa: 09/12/2019

Dedico este trabalho a todos que me acompanharam nessa jornada,
sendo estes amigos, colegas, professores e minha família.

AGRADECIMENTOS

À Profª. Dra. Maria das Graças J. M. Tomazela, pela orientação, confiança, oportunidade, sabedoria e por incentivar o desenvolvimento deste trabalho.

Ao meu amigo e colega de curso Lucca Antonio Moreira Ecclissi, pela amizade, apoio, discussões e disponibilidade em diversas etapas do trabalho.

Ao membro da banca de PTG, Prof. Carlos Cesar Farias de Souza, pelas valiosas sugestões e questionamentos.

A todos amigos deste curso que participaram dos Projetos Interdisciplinares junto a mim durante a graduação, pelo apoio, pela amizade e pela ajuda no dia-a-dia, em especial, Marina Estarópolis dos Reis, Paulo Roberto de Lima, Guilherme Pedrozo Abacherli, Jean Carlos de Souza e Jefferson Lopes dos Santos Ribeiro.

A todos da minha família, especialmente aos meus pais e irmãs, pelo incentivo, amor e carinho e por me apoiarem a cada dia ao longo da caminhada em busca deste objetivo.

“Na vida, não existe nada a temer, mas a entender.”
Marie Curie

RESUMO

Os modelos analíticos para interpretação de dados extraídos de plataformas educacionais podem auxiliar na análise e visualização de informações, ajudando a prever o desempenho dos alunos, gerando recomendações, fornecendo *feedback*, dentre outras aplicações. Atualmente, muitas instituições de ensino usam a análise de dados para melhorar os serviços que fornecem, tanto direta quanto indiretamente. Um fator importante em relação ao sucesso de instituições de ensino, como faculdades, é o índice de retenção de alunos no curso com bom desempenho destes na área de formação, que também é afetado diretamente pela compreensão do currículo de disciplinas fornecido pelo curso. Nesse contexto, a implementação de técnicas de *Big Data* oferece uma grande oportunidade, pois possibilita trabalhar com um grande volume de dados e métricas, como os dados que relacionam o desempenho do aluno durante o curso e os pontos a serem melhorados. Sendo assim, este trabalho tem como objetivo aplicar ferramentas de *Big Data* para analisar o desempenho acadêmico dos alunos da FATEC Indaiatuba dos cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial, obtidos no cadastro das disciplinas no sistema SIGA das FATECs, comparando parâmetros relevantes entre os cursos, visando a apoiar a tomadas de decisão na faculdade em ações educacionais e para redução de evasão de alunos. Na fundamentação teórica são apresentados alguns trabalhos relacionados à análise de dados e aplicação das técnicas de *Big Data* em instituições de ensino, que sustentam a pesquisa. Para alcançar o objetivo proposto, foi realizada uma pesquisa experimental, composta pela revisão bibliográfica de artigos relacionados, aplicação da técnica de técnicas de *Big Data* por meio das bibliotecas Pandas e Matplotlib para Python para processamento dos dados e visualização dos resultados. Assim foi possível determinar a relação entre o resultado acadêmico dos alunos (aprovações, reprovações e desistências), informações sobre quais semestres são mais propensos para o aluno desistir do curso e a influência da origem do aluno (escola pública ou particular) com os níveis de evasão, possibilitando maior engajamento do corpo discente e trazendo novos direcionamentos para as Faculdades de Tecnologia do Estado de São Paulo (FATECs).

Palavras chave: Análise de dados; FATEC; comparação.

SUMÁRIO

INTRODUÇÃO

A cada ano está havendo um aumento acentuado na quantidade de dados gerados nos mais diversos sistemas criados em todo o mundo. A disseminação de diferentes tipos de dispositivos eletrônicos e seus usos geram continuamente enormes quantidades de dados, como por exemplo dados de radiofreqüência obtidos por RFID, dados de sensores, dados de interação de redes sociais e dados de Internet móvel. Dentre esses, encontram-se dados estruturados, semiestruturados (ou fracamente estruturados) e desestruturados, essenciais para significância dos estudos, mas que torna padronizações e análises difíceis com ferramentas até então tradicionais (HU, 2016). Em essência, foram gerados aproximadamente 2,3 trilhões de gigabytes por dia em dados no mundo no ano de 2017 (LYNCH, 2017), sendo que entre 2015 e 2017, o volume de dados na Internet aumentou em 5 zettabytes em comparação com o ano anterior que atingiu 14,5 zettabytes (KOCHETKOV et al, 2017). Desse modo, surgiram novas técnicas para manipulação de quantidades massivas de dados, caracterizando-se assim o *Big Data*. Este termo, por sua vez refere-se a quantidades de dados, que são muito grandes e/ou complexas para serem tratadas com eficácia e eficiência por teorias, tecnologias e ferramentas tradicionais.

Dentre as inúmeras possíveis aplicações, a tecnologia de análise de *Big Data* pode fornecer análise de dados para o gerenciamento de assuntos educacionais. Um dos objetivos da aplicação de *Big Data* no sistema de educação inteligente é a visualização de dados, que permite aos usuários observar diretamente os resultados do processamento de dados. Outro objetivo é fornecer parâmetros para tomada de decisão, dando um julgamento prospectivo em um certo grau de acordo com os dados visualizados, que é onde o presente trabalho foca seus esforços. A tecnologia de *Big Data* pode melhorar a eficiência operacional das instituições de ensino, minerando a informação e o conhecimento escondidos em uma vasta quantidade de dados e fornecendo a base para atividades sociais e econômicas de seres humanos. A análise de grandes volumes de dados também pode ser aplicada para melhorar a qualidade do ensino nas universidades, analisar comportamentos de usuários na rede universitária e predizer o comportamento de determinado grupo de alunos em vários aspectos (HU, 2016)**.**

Os dados são considerados como o novo petróleo e ativo estratégico, e impulsionam ou até determinam o futuro da ciência, da tecnologia, da economia e possivelmente de tudo em nosso mundo hoje e no futuro (CAO, 2017)**.** *Big Data* têm sido usado por professores e setores administrativos para obter uma visão geral de como o processo educacional está sendo

conduzido na instituição. Pode ser usado como um meio de aumentar o desempenho acadêmico dos alunos, bem como aumentar a eficácia de professores. (KOCHETKOV e PROKHOROV, 2017). Entretanto, há problemas a serem resolvidos incluindo a disponibilidade de armazenamento, expressão, gerenciamento, confiabilidade e transmissão efetiva das grandes quantidades de dados (HU, 2016).

Portanto, a partir do que foi apresentado institui-se a questão norteadora que constitui o problema desta pesquisa: “Como a comparação de desempenho acadêmico entre cursos de Exatas e Humanas, por meio das técnicas de *Big Data*, pode ajudar as instituições de ensino superior na tomada de decisão para melhorar o ensino?”. Desse modo, visando a responder tal questão, o objetivo deste trabalho foi aplicar ferramentas de *Big Data* para analisar as informações dos alunos da FATEC Indaiatuba dos cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial, obtidos no cadastro das disciplinas no sistema SIGA das FATECs, comparando parâmetros relevantes entre os cursos. Sendo assim, esse trabalho apresenta a hipótese que se as informações dos alunos da FATEC Indaiatuba dos cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial forem analisadas por meio de técnicas de *Big Data*, será possível auxiliar a gestão da faculdade nos processos educacionais e diminuir a evasão escolar durante a graduação.

A metodologia a ser utilizada para desenvolver este estudo foi a pesquisa experimental que, de acordo com Gil (2002), consiste na determinação de um objeto de estudo, fazer a seleção de variáveis que sejam capazes de influenciá-lo, definir meios para controlar e observar os efeitos que esta variável manipulada possa produzir nesse objeto.

Com relação à estrutura da pesquisa desenvolvida, esta será apresentada da seguinte forma: no Capítulo I, será exposta a fundamentação teórica, que traz os conceitos chave e os trabalhos relacionados. No Capítulo II, será é apresentada a metodologia adotada e os passos percorridos para a realização desta pesquisa. No Capítulo III são apresentados os desafios da pesquisa e os resultados obtidos e discutidos. Por fim, são apresentadas as considerações finais e as referências.

CAPÍTULO I

Fundamentação Teórica

1.1 – Conceitos-chave

1.1.1 *Big Data*

Big Data é uma área de estudos que utiliza quantidades de dados massivas, que também podem ser complexas, e as trata por meio de tecnologias e ferramentas para apoiar decisões de negócios ou resolver problemas. Os grandes volumes de dados podem ser obtidos por meio de repositórios de dados públicos ou privados, redes sociais, sensores, ferramentas de IoT, RFID, dentre outros, nos mais diversos formatos e extensões de arquivo, como texto, vídeo, áudio e imagens (CAO, 2017).

Em meio aos dados estruturados, semiestruturados e desestruturados, a dificuldade em encontra padrões para normatização torna as análises difíceis com ferramentas tradicionais (HU, 2016)

Desse modo, *Big Data* é definido pelos chamados 5 V's (KOCHETKOV e PROKHOROV, 2017) (LYNCH, 2017):

- **Volume** - relacionado a uma grande quantidade de informação, tendo no mínimo tamanho de alguns terabytes;
- **Velocidade** - taxa de obtenção de novos dados armazenados e a velocidade de seu processamento;
- **Variedad**e - conteúdo de informações estruturadas e não estruturadas em um grande cluster de dados;
- **Veracidade** - informações relevantes para a análise, tratando de questões de confiança e descarte de um conjunto de dados;
- **Valor** – a utilidade e disponibilidade de informações, que se analisadas, terão impacto positivo na atividade para o usuário final.

Além desses, há um sexto "V" posterior, a "Visualização", que expõe os dados analisados para auxiliar na sua compreensão. Para isso, são utilizados tabelas, gráficos e outras ferramentas visuais, muito comumente relacionadas a *Business Intelligence*.

1.1.2 Análise de *Big Data*

Muitas organizações usam dados para tomar melhores decisões estratégicas e operacionais. A utilização de dados para tomar decisões não é algo novo: as organizações armazenam e analisam grandes volumes de dados desde o advento dos sistemas de *data warehouse* no início dos anos 90. Entretanto, a natureza dos dados disponíveis está mudando e as mudanças trazem consigo a complexidade na gestão dos volumes na análise desses dados (DANIEL, 2015).

É importante notar que há duas entidades técnicas. Primeiro, há *Big Data* para quantidades maciças de informações detalhadas. Em segundo lugar, há análises avançadas, realizadas por uma coleção de diferentes tipos de ferramentas, incluindo aquelas baseadas em análises preditivas, mineração de dados, estatísticas, inteligência artificial, processamento de linguagem natural, e assim por diante. De acordo com Russom (2011), o *Big Data* em conjunto com as ferramentas de análise (*Big Data Analytics*) representam as novas práticas de BI (*Business Intelligence*) da atualidade.

De acordo com Daniel (2015), há três etapas necessárias para que o *Big Data* agregue valor às atividades de gestão:

- **Coleção**

A coleta de dados é o primeiro passo para obter valor a partir do *Big Data*. Isto exige a identificação de dados que podem revelar informações úteis e valiosas. Os dados devem ser filtrados e só depois armazenados de forma que seja útil para a tomada de decisão;

- **Análise**

Uma vez que os dados foram organizados em uma forma utilizável, eles devem ser analisados. No entanto, com a crescente diversidade na natureza dos dados, o gerenciamento e a análise desses conjuntos de dados diversificados está se tornando um processo muito complexo. A análise precisa incluir vinculação, correlacionando diferentes conjuntos de dados para que seja possível entender a informação que deve ser transmitida por esses dados.

- **Visualização e aplicação**

Nessa etapa, os dados analisados são disponibilizados aos usuários em uma forma que é interpretável e integrado nos processos existentes e, em última instância, usado para orientar nas tomadas de decisão.

Daniel (2015) afirma que, atualmente a análise de *Big Data* está sendo explorada principalmente em negócios, governo e cuidados de saúde devido à grande quantidade de dados coletados e armazenados nesses ambientes. Já em relação ao ensino superior há poucas pesquisas sobre o tema, apesar do interesse crescente na exploração dos dados disponíveis nessa área.

1.1.3 *Big Data* na Educação

É importante ressaltar que a pesquisa sobre *Big Data* destina-se principalmente a examinar como agregar eficientemente e correlacionar volumes maciços de dados para identificar padrões comportamentais recorrentes e tendências significativas ao invés de catalogar o estado atual. Dessa maneira a utilização de tecnologias de análise de *Big Data* (*Data analytics* e *Data mining*) pode subsidiar a gestão das instituições de ensino superior de forma que essas possam responder efetivamente às mudanças que acontecem dentro e fora da instituição, bem como permanecer adequadas às necessidades da sociedade que elas servem.

Big Data já é usado na educação há algum tempo, visando a obter uma visão global de como o processo educacional está sendo levado em algumas instituições. As ferramentas podem servir como meio para aumentar o desempenho acadêmico dos alunos e professores no momento de ensinar e aprender. (KOCHETKOV e PROKHOROV, 2017).

Instituições educacionais pelo mundo utilizam plataformas online que servem como locais de coleta e armazenamento de dados. Uma vez que se trabalha e analisa tais dados, pode-se notar problemas de aprendizado, desempenho e comportamento dos alunos. Por meio de análises preditivas pode-se até mesmo evitar que situações indesejadas aconteçam. (WASSAN, 2014; WEST, 2012).

A utilização de algoritmos computacionais para prever quais alunos precisam de apoio diferenciado para que não abandonem seu curso já é uma realidade. Por isso, auxiliando na tomada de decisão dos gestores e professores de uma escola ou universidade, evita-se perda

de recursos, tanto da instituição quanto do aluno, em uma situação de desmotivação inevitável (GULWANI, 2017).

Sin e Muthu (2015) listam as maneiras que as técnicas de *Big Data* podem ser utilizadas para dar suporte às atividades de planejamento nas instituições de ensino superior:

- **Previsão de desempenho** - o desempenho do aluno pode ser previsto por meio da análise da interação do aluno em um ambiente de aprendizagem com outros alunos e professores;
- **Detecção de risco de evasão** - ao analisar o comportamento do aluno, o risco de estudantes evadirem do curso pode ser detectado e ações podem ser implementadas para reter esses alunos;
- **Visualização de dados** - o relatórios sobre dados educacionais tornam-se cada vez mais complexo à medida que os dados educacionais crescem em tamanho. Com a utilização de ferramentas de *Big Data*, os dados podem ser visualizados a partir do uso de técnicas de visualização de dados, possibilitando a identificação das tendências e as relações entre esses dados;
- **Feedback inteligente** - os sistemas de aprendizagem podem fornecer comentários precisos aos alunos, em resposta aos seus questionamentos, melhorando assim a interação e o desempenho dos alunos;
- **Recomendação do curso** - novos cursos podem ser recomendados aos estudantes com base nos interesses identificados pela análise de suas atividades. Isso possibilitará que as opções dos alunos não sejam equivocadas no momento da escolha de campos de interesse;
- **Estimativa de habilidades estudantis** - estimativa das habilidades adquiridas pelo aluno no decorrer do curso;
- **Detecção de comportamento** - Detecção de comportamentos de estudantes em comunidade, atividades ou jogos que ajudam no desenvolvimento de um aluno.

1.2 – Trabalhos relacionados

Nesta seção são apresentados estudos e pesquisas relacionados à temática do *Big Data*, bem como suas aplicações. O levantamento realizado foi orientado pela busca de pesquisas científicas e/ou tecnológicas que têm em seus objetivos a implementação e análise de ferramentas e tecnologias para a padronização dos conceitos acerca do tema, servindo de base para o estudo a ser desenvolvido.

A ferramenta que serviu de referência para isso foram as bases de dados internacionais, tais como ScienceDirect, Google Scholar e IEEE, por meio do qual se buscou mapear as pesquisas dessa natureza circunscritas nos últimos anos.

Segundo Cao (2017), o termo *Big Data* refere-se às quantidades de dados que são muito grandes e/ou complexas para serem tratadas com eficácia e/ou eficiência por teorias, tecnologias e ferramentas tradicionais relacionadas a dados. Entretanto, essa área de pesquisa traz consigo outras definições relevantes para podermos destinar esforços e tipos de profissionais diferentes em busca da resolução de um problema. Por sua vez, *Data analysis* refere-se ao processamento de dados por teorias, tecnologias e ferramentas tradicionais (por exemplo, estatística, matemática ou lógica clássica) para obter informações úteis e para fins práticos. Já *Data analytics* refere-se às teorias, tecnologias, ferramentas e processos que permitem uma compreensão profunda e descoberta de *insights* em dados. A análise de dados consiste em análise descritiva, análise preditiva e análise prescritiva. O desenvolvimento de *data mining* e *machine learning*, juntamente com a análise de dados original e analítica descritiva da perspectiva estatística, formam o conceito geral de "análise de dados". Ou seja, a análise de dados é a ciência multidisciplinar de examinar quantitativamente e qualitativamente dados com o propósito de tirar novas conclusões ou insights (exploratórios ou preditivos), ou extrair e provar hipóteses (confirmatórias ou baseadas em fatos) sobre informações para as tomadas de decisão. Desse modo, foi concluído que *Data science* é um termo que engloba as definições anteriores e se mostra por ser um novo campo interdisciplinar que se forma a partir de estatística, informática, computação, comunicação, gerenciamento e sociologia para estudar dados e seus ambientes, a fim de transformar dados em percepções e decisões.

Os trabalhos de Kochetkov e Prokhorov (2017) e Lynch (2017) trazem definições relevantes e determinantes para o *Big Data*. São assim estruturados os 5 V's, sendo estes Volume, Velocidade, Variedade, Veracidade e Valor. Volume está relacionado a uma grande quantidade de informação, tendo no mínimo tamanho de alguns terabytes. Velocidade é a taxa de obtenção de novos dados armazenados e a velocidade de seu processamento. Já variedade se diz do conteúdo de informações estruturadas e não estruturadas em um grande *cluster* de dados. Em geral, em grandes clusters de dados são dados semiestruturados, a porcentagem de informações úteis é de cerca de 5 a 7% por *cluster*. Por sua vez, a veracidade dos dados obtidos é importante para podermos saber se estamos lidando com informações relevantes para a análise. O problema da confiabilidade da informação dificulta a análise de *Big Data* e, como consequência, leva a resultados errados da análise. Por último, o valor é a utilidade de

informações para o usuário final. Caracteriza-se pela disponibilidade de informações cuja análise e resultados terão impacto positivo na atividade de um usuário.

O trabalho de Larson e Chang (2016) diz que a visualização, embora não seja um novo conceito, é um componente-chave da análise rápida. A visualização como parte de um serviço permite que os usuários compreendam rapidamente conjuntos de dados complexos criados a partir de análises estatísticas ou modelos analíticos. O *Big Data* é geralmente definido não apenas pela velocidade, variedade e volume, mas também pela validade, veracidade, valor e visibilidade - o que implica a importância da visualização.

A pesquisa de Schultz e O'Neil (2013) sustenta as etapas e técnicas a serem utilizadas neste trabalho. Eles dizem que na fase de modelagem, design e desenvolvimento, a modelagem é utilizada de duas maneiras: modelagem analítica em ciência dos dados e modelagem de dados para descrever os dados usados em análises rápidas. A modelagem analítica inclui análise descritiva, preditiva e prescritiva usando algoritmos de aprendizado de máquina, como regressão, agrupamento ou classificação.

Dentro deste contexto, os trabalhos de Wassan (2014) e West (2012) mostram que várias instituições educacionais também desenvolveram softwares de *dashboard* e armazenamento de dados que permitem acompanhar os problemas de aprendizado, desempenho e comportamento dos alunos. Sendo assim, a partir de tais dados coletados é possível começar o trabalho de análises preditivas.

Gulwani (2017) escreve sobre a utilização de algoritmos computacionais para prever quais alunos em uma universidade precisam de apoio diferenciado para que não abandonem seu curso. Para isso, foi utilizada a técnica CART, que implementa árvores de decisão binárias, uma estrutura estatística capaz de detectar alunos com problemas ou deficiências em matérias-chave. Com isso, evita-se gasto de tempo e dinheiro da instituição e dos alunos com algo que não será o campo de destaque do próprio aluno. Por isso, é possível assim também detectar pessoas que não tem interesse e/ou aptidão para o curso e podem ser ajudados para obter esta constatação o quanto antes.

Hu et al (2017) apresentaram para analisar sistematicamente os diversos dados relacionados à educação, provenientes de duas universidades públicas com o objetivo de auxiliar os alunos na tomada de decisão, usando de informações, sobre suas escolhas futuras ao decorrer do curso. A pesquisa experimental foi realizada nas faculdades em questão, sendo que se conclui que o modelo proposto pode ser expandido a outras instituições de ensino (generalista) e, por causa disso, é de grande vantagem atribuir esse tipo de análise à educação.

Birjali et al (2017) realizaram análise no âmbito emocional de usuários de uma rede social com base em dados coletados na própria rede, por meio de três etapas: processamento, análise e visualização de dados. Para isso, os dados (textos das postagens) foram analisados a partir do serviço Flume (utilizado para coletar, agregar e mover de modo eficiente grandes quantidades de dados), processados usando o script Jaql e armazenados e analisados no *Hadoop Distributed File System* (HDFS). Para a análise, as palavras negativas e positivas obtidas por meio do banco de dados formado foram submetidas ao método de *MapReduce.* Por último, os resultados foram formatados e visualizados como gráficos por meio da ferramenta *BigSheets BigInsights* (IBM). Os autores mostraram o tempo de processamento para análise de dados maciços de uma rede social e revelaram que é possível realizar uma análise de dados consistente, tendo como base o tipo de filtro a ser utilizado de acordo com cada solicitação.

A seguir, no Quadro 1, são compilados os trabalhos apresentados até então, destacando suas principais características.

Quadro 1: Trabalhos Relacionados e Destaques

Autor (Ano)	**Objetivo da Pesquisa**	**Destaques**	**Tipo**
Cao, L. (2017)	Definir e apresentar conceitos referentes ao *Big Data* e suas aplicações, descrevendo a história e o possível futuro das aplicações da tecnologia.	São definidos e diferenciados termos como: *Big data, Data analytics, Data analysis e Data Science.*	Conceito
Kochetkov, O. T. e Prokhorov, I. V. (2017)	Discutir as abordagens para o uso de *Big Data* no processo educacional de instituições de ensino superior	O artigo possui conceitos importantes utilizados na área de *Big Data*, tais como os 5 V's.	Conceito

(Cont.) Quadro 2: Trabalhos Relacionados e Destaques

Autor (Ano)	Objetivo da Pesquisa	Destaques	Tipo
Lynch, C. F. (2017)	Discutir como *Big Data* e *Data Mining* podem ser utilizados na educação, bem como seus conceitos e aplicações.	Este trabalho traz conceitos relevantes sobre os tipos de dados existentes na área de *Big Data*, incluindo *Rich Data* e *Data as a Service*.	Conceito
Larson, D., Chang, D. (2016)	Propor uma estrutura ágil para entrega de BI, análise rápida e *Data Science*, considerando o impacto do *Big Data*.	Revelar a importância da visualização também frente aos outros 5 V's.	Conceito
O'Neil, C., Schutt, R. (2013)	Sintetizar e apresentar conceitos e técnicas sobre *Data Science* apresentado modelos estatísticos e algoritmos em linguagens de programação como o R.	Diferenciam e caracterizam os tipos de modelagem, agrupando-as por semelhança e aplicabilidade.	Aplicação
Wassan, J. T. (2014)	Focar a experimentação de modelagem educacional baseada em técnicas de *Big Data*	Explica a utilização de bancos de dados NoSQL para a aplicação da modelagem (MongoDB).	Aplicação
Vyas, M.S. e Gulwani, R. (2017)	Utilizar a técnica CART, que implementa árvores de decisão binárias, uma estrutura estatística capaz de detectar alunos com problemas ou deficiências em matérias-chave.	Evita que seja gasto de tempo e dinheiro da instituição e dos alunos com algo que não será o campo de destaque do próprio aluno.	Aplicação
Ramirez-Gallego, S. et al (2017)	Explicar os conceitos envolvidos na metodologia de *MapReduce* e alternativas.	Mostra como a técnica de *MapReduce* funciona e auxilia em trabalhos que lidam com grandes quantidades de dados.	Aplicação
Hu, Q. et al (2017)	Analisar sistematicamente o grande volume de dados são passíveis de coleta nas instituições de ensino, visando auxiliar os alunos na tomada de decisão sobre escolhas futuras no curso.	Analisa o aspecto emocional dos alunos com base em previsões, para auxiliar na tomada de decisões.	Aplicação
Birjali et al (2017)	Analisar os usuários em âmbito emocional, com base em dados coletados por meio de redes sociais.	Explica procedimentos e métodos utilizados para análises robustas no campo de estudos de *Big Data*.	Aplicação

Fonte: Elaborado pelo autor

Destes, os que mais se aproximam do presente projeto são os trabalhos de Hu, Q. et al (2017) e Vyas, M.S. e Gulwani, R. (2017). O primeiro analisa uma grande quantidade de dados proveniente das plataformas educacionais da universidade, traçando um paralelo relevante com o presente trabalho, pois a partir da coleta de dados, as análises auxiliam na tomada de decisão por parte da instituição de ensino e também dos alunos. Já o segundo trabalho foca sua análise para resolução de problemas que levem a uma melhor alocação de recursos da instituição frente ao desempenho e necessidade dos alunos.

O presente trabalho está inserido no contexto dos trabalhos apresentados nessa sessão por meio da relevância que as respostas para a hipótese inicial trarão às instituições de ensino. Os recursos das instituições de ensino, para serem bem aplicados, devem ser utilizados de acordo com as respostas que a análise de dados feita para cada uma das escolas traz, onde os responsáveis pela gestão decidem pontos focais de atuação para evitar a evasão escolar e o melhor aproveitamento dos alunos. O índice de evasão e a diversidade de alunos são fatores que estão inseridos nos objetivos centrais das escolas, faculdades e universidades, de acordo com os trabalhos citados anteriormente. Os dados coletados diariamente por essas instituições podem ser úteis para sua própria gestão. O impacto da pesquisa se dá no auxílio da tomada de decisão da faculdade, pela destinação de recursos financeiros e educacionais.

CAPÍTULO II

Metodologia

2.1 – Natureza da Pesquisa

O desenvolvimento deste estudo teve como foco a abordagem experimental que, segundo Gil (2017), consiste em determinar um objeto de estudo, selecionar as variáveis capazes de influenciá-lo e definir meios de controle e observação dos efeitos que essa variável produz nesse objeto. Assim, deve-se definir a hipótese, entender as regras ambientais e comportamentais para a execução do projeto, executar o experimento e analisar os resultados obtidos.

2.2 – Variáveis de Análise

Durante o desenvolvimento da análise, foram se definidas as variáveis para tornar a coleta e a análise de dados relevante:

1. Dados pessoais dos alunos da FATEC Indaiatuba quanto a identificação e sua caracterização;
2. Categorias de dados a serem analisados para caracterizar o desempenho de cada aluno;
3. Dados para a comparação dos cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial da FATEC Indaiatuba, durante sua coexistência.

2.3 – Ferramentas Utilizadas

As ferramentas utilizadas na execução deste trabalho foram selecionadas de acordo com a necessidade de cada etapa da análise. Optou-se então pelos serviços e bibliotecas Google Colab Notebook, Matplotlib e Pandas, que trazem a disponibilidade do trabalho em nuvem e manipulação ágil dos dados.

Cada uma destas ferramentas possui suas funções descritas de maneira mais detalhada nos tópicos a seguir:

Google Colab Notebook[1]: é um serviço em nuvem que funciona como o Jupyter Notebook. Ele é utilizado na execução de clusters de Apache Spark e Pandas, para seleção, transformação, estatística e visualização de dados. Com execução em tempo real de cada fragmento de código, o trabalho contorna erros com muito mais facilidade e velocidade. Para este trabalho, o processamento foi realizado em uma máquina na nuvem com processador Intel(R) Xeon(R) CPU @ 2.20GHz, cache de 56320 KB e memória RAM de 13.341992 GB.

Matplotlib[2]: é uma biblioteca Python dedicada à plotagem de gráficos 2D a partir de arrays. Ela é bastante utilizada para visualização de dados, apresentando uma série de opções, como gráficos de barra, linha, pizza, histogramas, dentre outros.

Pandas[3]: é uma biblioteca Python *open source* de ótimo desempenho capaz de simplificar tarefas de manipulação de dados. Ela fornece ferramentas de análise de dados e estruturas de dados de alta performance, compreendendo manipulação, leitura e visualização de dados.

2.4 – Experimento da Pesquisa

Para a realizar este trabalho foi feita a análise de dados coletados por meio dos sistemas de gerenciamento dos alunos da FATEC Indaiatuba (SIGA), com foco nos cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial. Os dados para a realização desta pesquisa foram obtidos e cedidos pela FATEC Indaiatuba no formato padrão de planillhas do Microsoft Office Excel (.xlsx). A coleta se deu pelos desenvolvedores da plataforma SIGA, com a autorização do diretor geral desta instituição e repassado pela professora orientadora deste projeto.

Inicialmente foi feita uma varredura geral para definição de quais seriam as colunas presentes na tabela relevantes para as respostas buscadas pelo projeto. As colunas foram então

[1] Conforme disponível em:< https://colab.research.google.com>
[2] Conforme disponível em: <https://matplotlib.org/>
[3] Conforme disponível em: <https://pandas.pydata.org/>

mapeadas e descritas para que fosse possível serem utilizadas durante a codificação nas etapas posteriores.

Como se trata de uma análise comparativa entre os cursos de Gestão Empresarial e Análise e Desenvolvimento de Sistemas, foi necessário padronizar o período temporal utilizado nas análises. Desse modo, optou-se pelo período de coexistência dos cursos na FATEC Indaiatuba, sendo que todos os dados considerados são de alunos do período noturno. Foram selecionados então dados partindo do primeiro semestre de 2012 até o segundo semestre de 2018.

Para a coluna CURSO foi feita a filtragem visando a selecionar apenas os cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial. Quanto ao desempenho acadêmico, foram selecionados os campos DISCIPLINA, NOTA, FREQUENCIA, STATUS_ALUNO, SEMESTRE_ANO e CONCEITO (de aprovação ou reprovação na disciplina). Ademais, a tabela apresenta, dentre outros fatores, as colunas TURNO (turno em que o aluno estuda), RA (número de registro acadêmico do aluno – utilizado como chave primária em certas análises), NOME (nome do aluno), ESCOLA_PUBLICA (expõe se o aluno é proveniente de uma escola pública ou particular) e NOTA_VESTIBULAR (desempenho do aluno no vestibular).

A partir da seleção e limpeza inicial dos dados, eles foram importados na nuvem por meio do Google Colab Notebook, não sendo necessário criar virtualizações do sistema, uma vez que esta própria ferramenta já funciona como um sistema isolado.

A limpeza e processamento inicial dos dados foi feita utilizando também a biblioteca *Pandas* para Python, sendo estes então armazenados em um arquivo .csv com todos os dados padronizados e corretamente distribuídos entre as colunas, eliminando campos com dados faltantes. Uma vez que os dados estavam prontos para uso, organizados e processados, esses foram armazenados na nuvem em servidores Google e trabalhados por meio da ferramenta *Google Colab Notebook*, na qual cada fração de código é executada individualmente. Para a visualização dos resultados, a biblioteca *Matplotlib* para Python foi essencial, uma vez que com ela é possível a geração de gráficos e adaptação destes às necessidades do trabalho, tais como modificações de estilo e cor.

Abaixo é apresentado um exemplo da arquitetura utilizada neste projeto:

Figura 1: Arquitetura de ferramentas utilizadas nas análises

Fonte: Elaborado pelo autor

CAPÍTULO III

Análise de Dados

3.1 – Coleta e Modelagem

A importação dos dados e bibliotecas a serem utilizadas, bem como a chamada da tabela, foi realizada por meio de comandos no Google Colab Notebook, como mostrado na Figura 2:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import colorsys
plt.style.use('seaborn-talk')
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
```

```
df = pd.read_csv('/content/drive/My Drive/Colab Notebooks/Indaiatuba_BigData.csv', low_memory=False)
```

Figura 2: Importação das bibliotecas necessárias e carregamento dos dados

Fonte: Elaborado pelo autor

3.2 – Manipulação dos Dados

A tabela concedida pela FATEC Indaiatuba continha inicialmente dados de todos os cursos oferecidos pela instituição, totalizando 19380 linhas. Como filtragem, realizou-se seleções buscando somente pelos dados dos cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial que estavam dentro do período proposto e que estivessem consistentes. Como a planilha foi cedida em formato estruturado (.xlsx), foi realizada a conversão para o formato .csv, suportado pela biblioteca Pandas para a manipulação dos dados. É importante notar que, no geral, as ferramentas focadas em *Big Data* foram desenvolvidas para trabalhar com dados desestruturados.

3.3 – Parâmetros Utilizados

Buscando sanar as perguntas decorrentes da hipótese deste projeto, definiu-se quais indicadores de desempenho seriam utilizados para comparar os resultados dos cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial da FATEC Indaiatuba. Assim, foram definidos os seguintes índices:

- Medianas das notas do vestibular;
- Quantidade de alunos que concluíram o curso (3 anos ou mais de curso);
- Número de semestres para conclusão do curso e índice de evasão;
- Porcentagem inicial de alunos vindos de escola pública e particular;
- Porcentagem de alunos egressos vindos de escola pública e particular;
- Desistências por semestre;
- Aprovação em disciplinas por categoria (Humanas, Exatas e Tecnológicas).

3.4 – Notas do Vestibular e Dados Gerais

A comparação entre os cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial da FATEC Indaiatuba, apresentados a partir deste ponto pelas siglas “ADS” e “GE” respectivamente, iniciou-se pela obtenção de dados relevantes entre os dois cursos. Primeiramente, optou-se por descobrir qual era o total de alunos analisados para cada um dos cursos. Utilizando o método *describe* da biblioteca Pandas, concluiu-se que seriam analisados os dados de 1047 alunos do curso de Gestão Empresarial e 700 alunos do curso de Análise e Desenvolvimento de Sistemas.

Sabendo disso, o próximo passo dado buscou comparar o desempenho desses alunos no vestibular, por meio do agrupamento de dados utilizando a coluna NOTA_VESTIBULAR. Obteve-se então a seguinte distribuição, apresentada nos gráfico histograma da Figura 3:

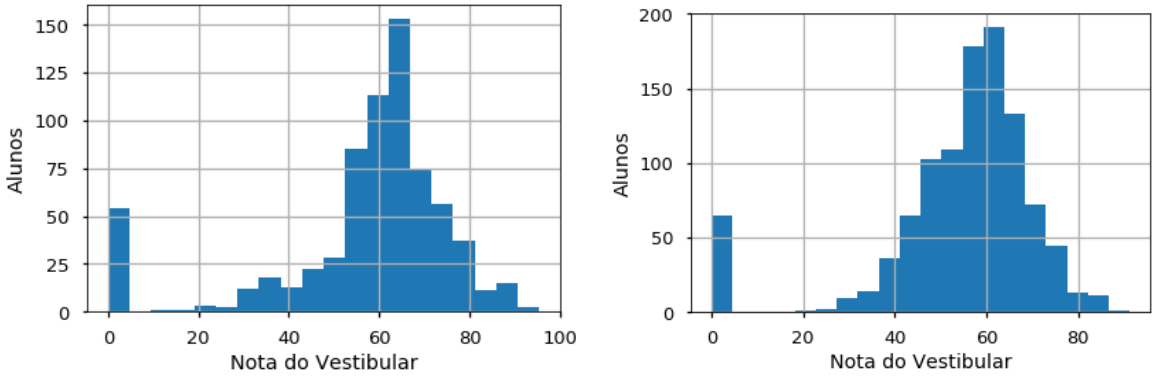

Figura 3: Histogramas das notas de vestibular dos alunos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial, respectivamente. O eixo X mostra a nota dos alunos de 0 a 100, cada barra vertical apresenta agrupamentos de notas de 5 em 5 pontos e o eixo Y a quantidade de alunos.

Fonte: Elaborado pelo autor

Nota-se que há uma concentração de alunos com nota 0 (zero) no vestibular, explicada pelo número de alunos decorrentes de transferência entre instituições de ensino superior, que passaram a integrar o quadro de alunos da FATEC Indaiatuba. No banco de dados, esses alunos estão representados pelo número do RA seguido pela letra “T” (exemplo: 123456789T). Como esses alunos não prestaram o vestibular comum para ingressarem em cada um destes cursos, a nota atribuída a eles foi zero. Mesmo assim, é importante manter seus dados na análise para fins de comparação de desempenho.

Buscando obter respostas mais relevantes, decidiu-se então por obter as medianas das notas de vestibular de cada um dos cursos. A nota mediana de GE foi de 57,75 pontos, enquanto a mediana de ADS foi de 61,875 pontos. Esse dado não corrobora intuitivamente com a concorrência histórica entre os dois cursos, onde GE possui mais candidatos/vagas do que ADS.

A partir deste ponto de corte, algumas análises subsequentes foram realizadas com dados totais e dados apenas de alunos que estavam abaixo da mediana na nota do vestibular. Isso porque, intuitivamente, espera-se que alunos com menores notas tenham mais dificuldades no decorrer do curso. Tivemos então dados de 517 alunos abaixo da mediana de notas no vestibular para GE e de 349 alunos de ADS. Quando a análise exige que sejam feitas comparações apenas entre alunos que concluíram os cursos, tem-se dados de 273 alunos para GE (183 abaixo da mediana) e 77 para ADS (44 abaixo da mediana).

A figura 4 a seguir mostra a situação dos alunos de cada curso no momento desta análise. Os gráficos foram obtidos por meio do agrupamento de linhas do banco de dados considerando o STATUS_ALUNO e RA, para que cada um dos RA fossem contados apenas

uma vez. É importante notar que, apesar de um alto número de matrículas canceladas e baixo número de conclusões de curso, ainda há um grande número de alunos em curso, que podem ter sua evasão evitada se medidas forem tomadas.

Os status a seguir, presentes na Figura 4, se dividem entre Cancelado, Em Curso, Concluído, Transferido (para outras instituições de ensino), Trancado 1 (alunos que estão no primeiro trancamento de matrícula) e Trancado 2 (alunos que estão no segundo trancamento de matrícula). Uma vez que um aluno retorna de um trancamento de matrícula, ele passa a figurar no banco de dados como “Em Curso”, assim não é mais possível saber se ele já foi um aluno em trancamento ou não para aquele momento da análise.

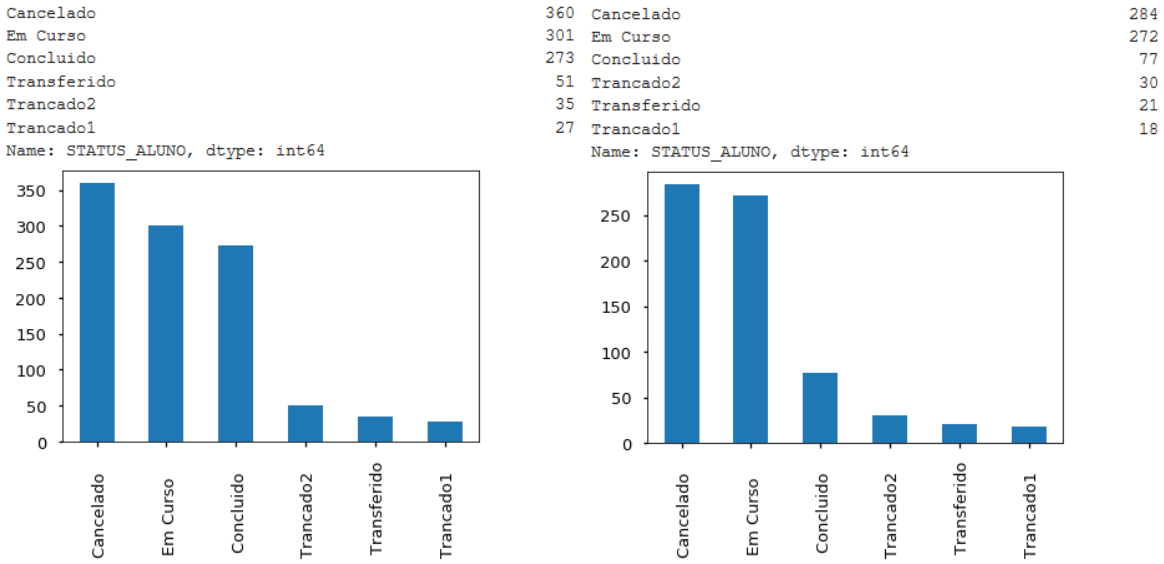

Figura 4: Status da matrícula dos alunos dos cursos de Gestão Empresarial e Análise e Desenvolvimento de Sistemas, respectivamente. O eixo X mostra o status dos alunos e o eixo Y a quantidade de alunos para aquele status.

Fonte: Elaborado pelo autor

3.5 – Índice de Evasão e Tempo de Formação

Com os dados acima obtidos, viu-se que a evasão do curso de Gestão Empresarial ficou em 34,38% e de Análise e Desenvolvimento de Sistemas em 40,57%, desconsiderando os possíveis cancelamentos que ainda podem ocorrer dos alunos classificados como “Em Curso”.

Buscou-se saber então o comportamento dos alunos que concluíram os cursos, considerando o semestre de formação. Na Figura 5 é possível notar que a maior parte dos

alunos de GE se formam entre o 6º e o 7º semestres. Como os cursos da FATEC Indaiatuba possuem 6 semestres, esse resultado mostra que grande parte desses alunos se forma permanecendo geralmente apenas mais um semestre para conclusão do curso. Como contraponto, há o curso de ADS, em que a maior parte de alunos formados se encontra entre o 6º e 8º semestres. Vê-se então que esses últimos alunos demandam mais recursos da faculdade, uma vez que demoram mais tempo para concluírem o curso.

Nota-se que, para ambos cursos analisados, há alunos que concluíram o curso em até 11 semestres e em menos de 6 semestres. Os primeiros podem ser explicados por trancamentos realizados, em que posteriormente o aluno voltou ao curso e o concluiu. Já os últimos são explicados por alunos provenientes de transferência, onde provavelmente houve aproveitamento de matérias para conclusão do curso.

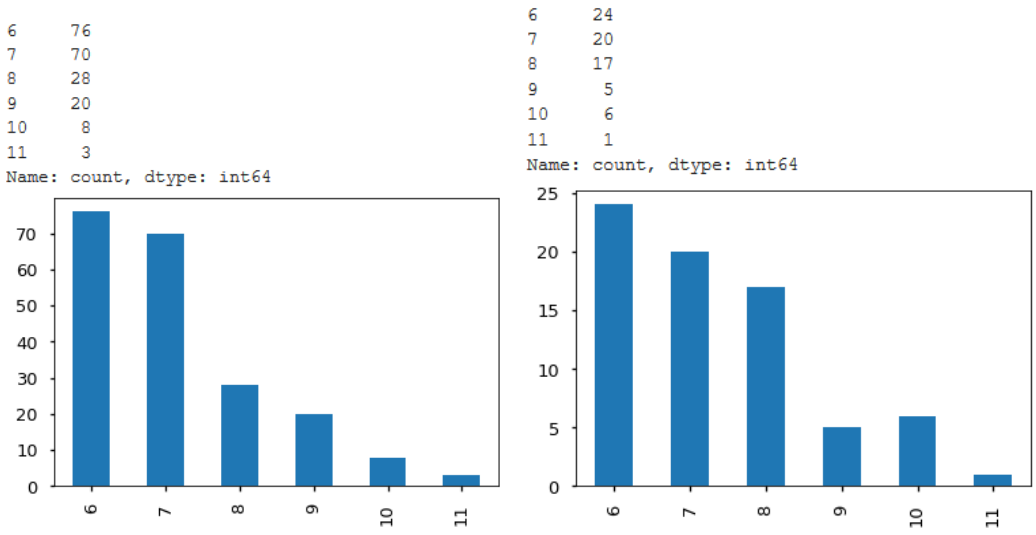

Figura 5: Tempo para conclusão de curso dos alunos dos cursos de Gestão Empresarial e Análise e Desenvolvimento de Sistemas, respectivamente. Para essa análise foi considerado o total de alunos formados em cada curso no período analisado. O eixo X mostra o semestre de conclusão do curso e o eixo Y a quantidade de alunos formados naquele semestre.

Fonte: Elaborado pelo autor

3.6 – Escola Pública e Particular

Uma vez que se obteve os índices de evasão dos alunos e seu comportamento para conclusão do curso, decidiu-se então por comparar a esfera da escola de ensino médio da qual tais alunos eram provenientes. As comparações foram feitas buscando correlações entre a proporção de alunos que ingressavam na FATEC Indaiatuba e destes que concluíam o seu

respectivo curso, sendo que parte vinha de escolas públicas e outra parte de escolas particulares.

As análises foram feitas por meio do agrupamento de alunos pela RA, ESCOLA_PUBLICA e STATUS_ALUNO e a contagem feita pelo número de linhas obtidas e agrupadas para a coluna ESCOLA_PUBLICA.

A Figura 6 mostra para o período analisado que, no total, o curso de Análise e Desenvolvimento de Sistemas teve cerca de 74% de seus alunos ingressantes vindos de escola pública, sendo que os 26% restantes provém de escolas particulares. Quando se olha para os alunos que concluíram o curso, nota-se que a maior desistência vem dos alunos de escolas particulares, que passam à proporção de cerca de 22% dos concluintes.

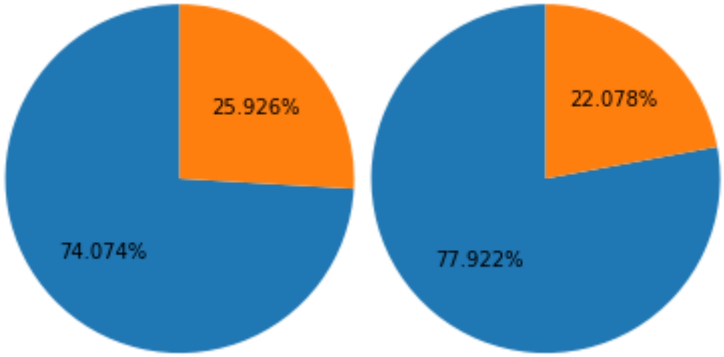

Figura 6: Porcentagem de alunos provenientes de escola pública (azul) e escola particular (laranja) ingressantes e formados no curso de Análise e Desenvolvimento de Sistemas, respectivamente.

Fonte: Elaborado pelo autor

Em uma análise mais profunda, fez-se o mesmo experimento apenas com os alunos que concluíram o curso de ADS que estavam abaixo da mediana nas notas do vestibular. Tal resultado é apresentado pela Figura 7. Dessa vez, surpreendentemente a proporção entre alunos de escola pública e particular praticamente não se modifica, permanecendo com cerca de 72% e 28%, respectivamente. Esse resultado sugere que, uma vez que um aluno com maiores dificuldades didáticas passa no vestibular e estuda para permanecer no curso até sua conclusão, o fator Escola Pública x Escola Particular não influencia na sua capacidade de concluir o curso, ou seja, seu *background* não se faz mais relevante do que os estudos no decorrer de sua formação universitária.

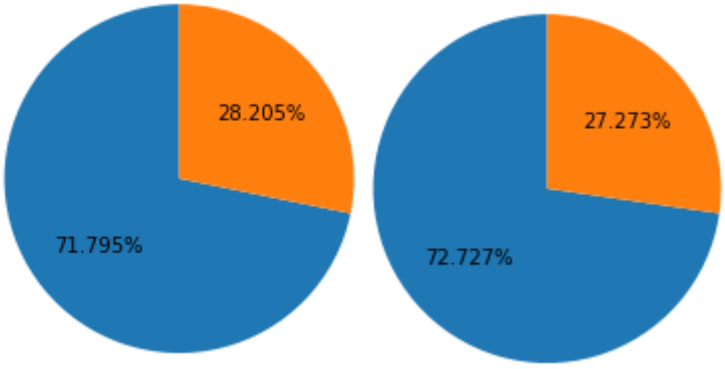

Figura 7: Porcentagem de alunos abaixo da mediana nas notas do vestibular, provenientes de escola pública (azul) e escola particular (laranja) ingressantes e formados no curso de Análise e Desenvolvimento de Sistemas, respectivamente.

Fonte: Elaborado pelo autor

Em seguida, trabalhou-se no mesmo tipo de comparação para os alunos do curso de Gestão Empresarial. Primeiramente, a comparação foi feita considerando o total de alunos ingressantes neste curso e do total de alunos que o concluíram. Dessa vez, nota-se por meio da Figura 8 que, em relação ao curso de ADS, GE possui mais alunos provenientes de escola particular, sendo estes cerca de 37% do total de alunos ingressantes. Quando se olha para os alunos que concluíram esse curso, vê-se que há uma desistência maior dos alunos vindos de escola pública, já que os alunos de escola particular passam a ser quase 41% do total.

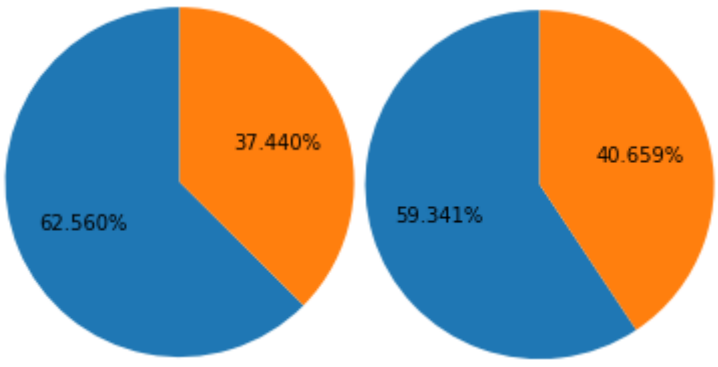

Figura 8: Porcentagem de alunos provenientes de escola pública (azul) e escola particular (laranja) ingressantes e formados no curso de Gestão Empresarial, respectivamente.

Fonte: Elaborado pelo autor

Observa-se que o mesmo ocorre quando consideramos alunos que foram aprovados no vestibular com notas abaixo da mediana para este curso. A Figura 9 mostra que a proporção de alunos de escola particular com desempenho inferior no vestibular aumenta para cerca de 44% entre os ingressantes e 47% entre os que concluíram o curso. Com este resultado, conclui-se que, ao contrário do esperado, o desempenho de alunos de escola particular no vestibular não é maior do que o desempenho dos alunos de escola pública, porém esses são os que mais permanecem no curso até sua conclusão. Desse modo, vê-se que é importante focar recursos da instituição de ensino para permanência de alunos provenientes de escolas públicas até o final do curso.

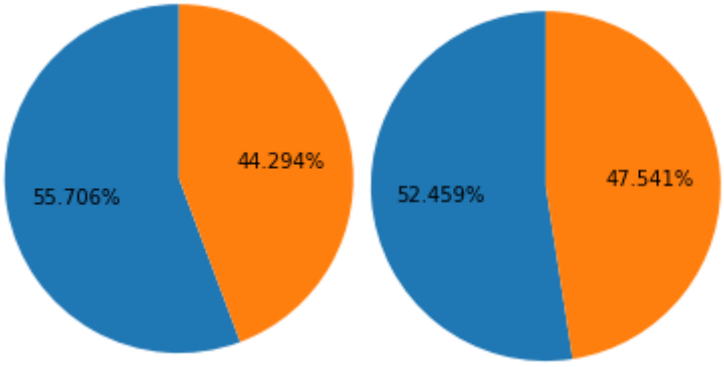

Figura 9: Porcentagem de alunos abaixo da mediana nas notas do vestibular, provenientes de escola pública (azul) e escola particular (laranja) ingressantes e formados no curso de Gestão Empresarial, respectivamente.

Fonte: Elaborado pelo autor

3.7 – Desistências

Quando se volta o olhar para as desistências dos cursos de Gestão Empresarial e Análise e Desenvolvimento de Sistemas, vê-se que é importante saber quando essa grande quantidade de alunos decide por abandonar o curso e seguir novos caminhos. Buscando por tais respostas, decidiu-se filtrar os dados a fim de obter a quantidade de alunos desistentes por semestre, contando “1” como o primeiro semestre de aulas do aluno, “2” o segundo semestre e assim sequencialmente. Nota-se por meio da Figura 10 que GE apresentou alunos desistentes até o 12º semestre, ou seja, houve casos em que o aluno ficou vinculado à faculdade por 6 anos e acabou por abandonar o curso. Já para ADS obteve-se comportamento semelhante nesse quesito, com alunos desistindo até o 11º semestre.

Em análise mais minuciosa, observa-se que a maior parte dos alunos que optam por abandonar o curso tomam essa decisão já no 1º semestre de aulas. Isso pode ocorrer por diversos fatores, tais como oportunidades em outras universidades, decepção com a escolha feita quanto ao curso, decepção quanto ao conteúdo do curso/professores, dificuldade no acompanhamento das aulas desde o primeiro momento, falta de recursos ou apoio para seguir com os estudos, dentre outros. Uma vez que isso foi observado, faz-se necessário descobrir os motivos principais de desistência e atuar como instituição para coibir a desistência desde o início do aluno na faculdade.

Uma vez que analisou-se o primeiro semestre de ambos os cursos separadamente, vemos que o restante das desistências se distribui pelos demais semestres e que o comportamento se dá pela diminuição de abandonos semestre a semestre.

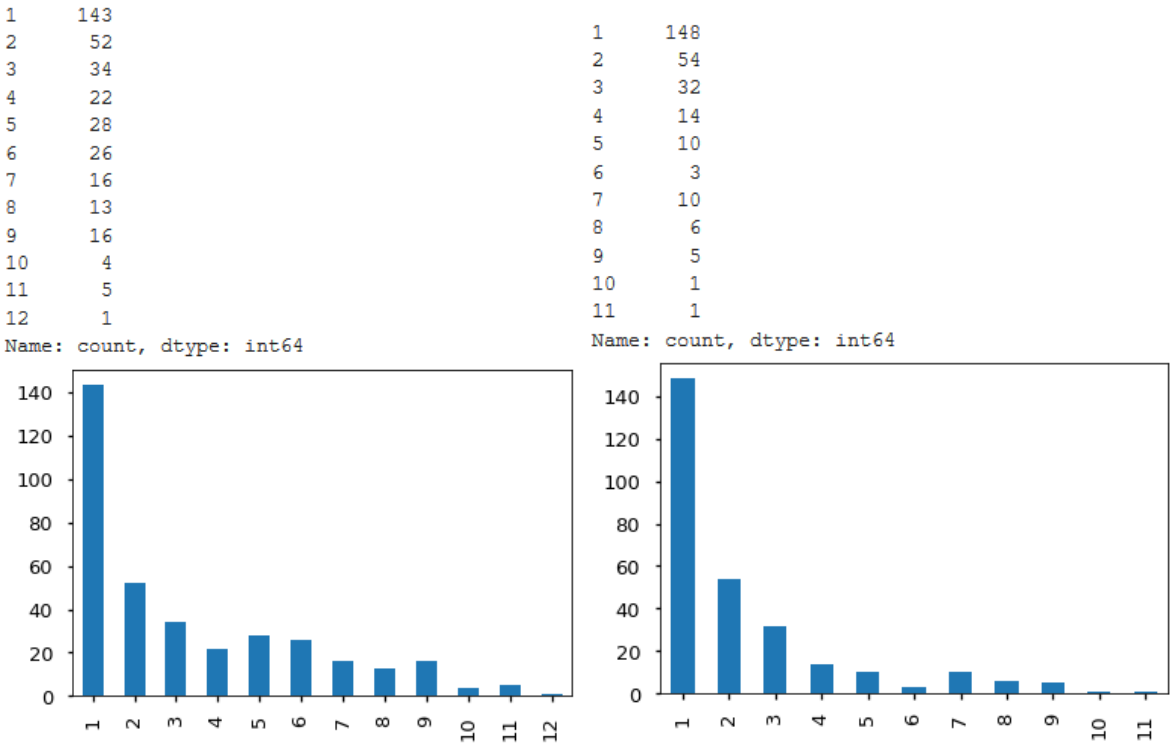

Figura 10: Quantidade de alunos desistentes por semestre dos cursos de Gestão Empresarial e Análise e Desenvolvimento de Sistemas, respectivamente. Para essa análise foi considerado o total de alunos com matrícula cancelada em cada curso no período analisado. O eixo X mostra o semestre de desistência do aluno no curso e o eixo Y a quantidade de alunos que cancelaram a matrícula naquele semestre.

Fonte: Elaborado pelo autor

3.8 – Índices de Aprovação (Humanas, Exatas e Tecnológicas)

Reunindo todos os dados de desistências e evasão escolar obtidos até o momento, definiu-se que o próximo e último passo seria descobrir se as disciplinas de alguma área específica estariam contribuindo para esses casos. Para isso, distribuiu-se as matérias de cada curso em três grandes áreas: Humanas, Exatas e Tecnológicas.

Os gráficos das figuras 11, 12 e 13 a seguir mostram o comportamento das aprovações e reprovações nas disciplinas quando comparamos os cursos de Gestão Empresarial e Análise e Desenvolvimento de Sistemas. Intuitivamente, espera-se que o curso da área de Humanas, como Gestão Empresarial, tenha maior afinidade com a área de Humanas, bem como Análise e Desenvolvimento de Sistemas tenha maior propensão a ter um bom desempenho nas matérias Tecnológicas e Exatas.

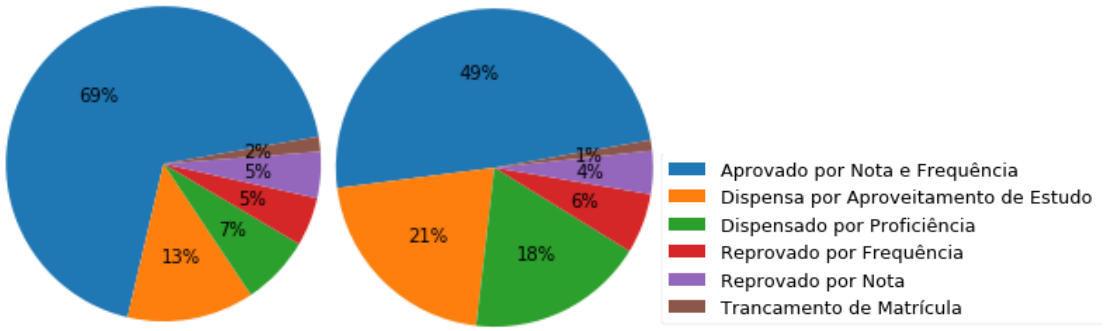

Figura 11: Índices de aprovação, dispensa e reprovação nas disciplinas de Humanas dos alunos dos cursos de Gestão Empresarial e Análise e Desenvolvimento de Sistemas, respectivamente.

Fonte: Elaborado pelo autor

Primeiramente, nota-se que GE possui melhor desempenho nas matérias denominadas como Humanas quando comparados com os alunos de ADS. Tem-se um índice de aprovação de 69% para GE frente a apenas 49% para alunos de ADS. Desse modo, sugere-se que alunos de ADS necessitam de maior empenho e cuidado durante o desenvolvimento de tais disciplinas. Ao mesmo tempo, a instituição de ensino tem a oportunidade de criar alternativas que tenham como objetivo trazer o aluno de ADS de maneira mais efetiva para os conteúdos abordados em disciplinas de Humanas.

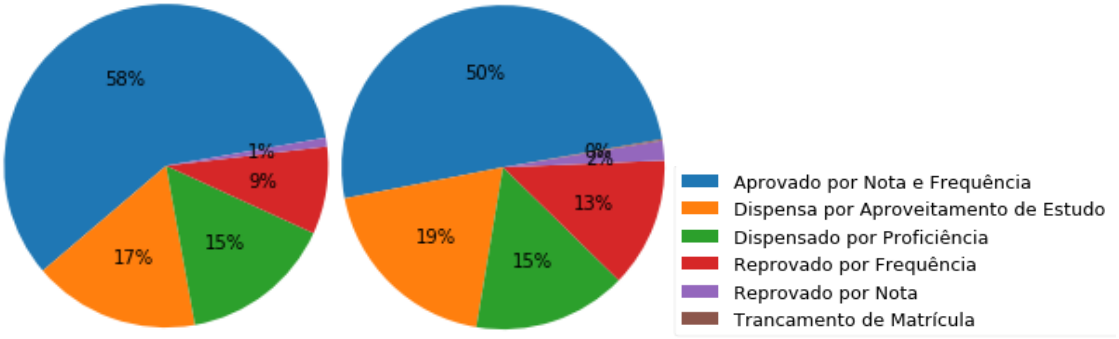

Figura 12: Índices de aprovação, dispensa e reprovação nas disciplinas de Exatas dos alunos dos cursos de Gestão Empresarial e Análise e Desenvolvimento de Sistemas, respectivamente.

Fonte: Elaborado pelo autor

Já para matérias denominadas como Exatas, os alunos de GE possuem desempenho também melhor quando comparados com os alunos de ADS, porém com comportamento mais próximo. Dessa vez, o índice de aprovação é de 58% para GE contra 50% para alunos de ADS. Mais uma vez vê-se que alunos de ADS precisam de um acompanhamento maior durante os estudos de disciplinas desta categoria, provalvemente onde a instituição trabalhe mais a base de conhecimento dos alunos e forneça melhores condições para monitoria e acompanhamento discente.

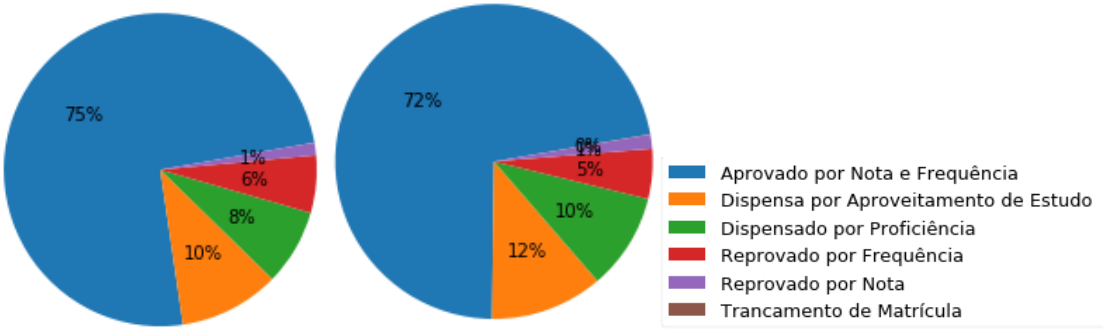

Figura 13: Índices de aprovação, dispensa e reprovação nas disciplinas de Tecnológicas dos alunos dos cursos de Gestão Empresarial e Análise e Desenvolvimento de Sistemas, respectivamente.

Fonte: Elaborado pelo autor

Por fim, tem-se as matérias denominadas como “Tecnológicas”, que mostram índice de aprovação elevado para ambos cursos. O índice de aprovação se apresenta em 75% para GE e 72% para alunos de ADS, desconsiderando alunos já aprovados por dispensa por proficiência ou aproveitamento. Nota-se então que as matérias dessa categoria provavelmente

não impactam significantemente na desistência de alunos, quando comparadas com “Humanas” ou “Exatas.

Assim, as ações a serem tomadas pela instituição de ensino devem buscar para ambos os cursos um enfoque maior em causas dos índices de reprovas e soluções que amenizem isso frente às disciplinas de “Humanas” e “Exatas”, sendo que ADS demonstra maior dificuldade em ambas as áreas. Ao mesmo tempo, retomando os resultados obtidos com as notas dos vestibulares, vê-se que os alunos com melhor desempenho no vestibular (ADS) não são aqueles que apresentam melhor desempenho nas disciplinas do curso. Os motivos a serem analisados, dentre outros, passam pelo vestibular não estar selecionando adequadamente alunos para este curso até disciplinas dos cursos de ADS serem relativamente mais difíceis do que as do curso de GE.

CONSIDERAÇÕES FINAIS

Inicialmente neste trabalho foi definido o problema de pesquisa dado pela seguinte questão: “Como a comparação de desempenho acadêmico entre cursos de Exatas e Humanas, por meio das técnicas de *Big Data*, pode ajudar as instituições de ensino superior na tomada de decisão para melhorar o ensino?”. Com isso em mãos, buscou-se trabalhos relacionados publicados pela comunidade acadêmica que fossem recentes e utilizassem técnicas de *Big Data* para obtenção de respostas para problemas de cunho educacional em instituições de ensino.

Uma vez que as perguntas e estudos prévios foram concluídos, selecionou-se então os materiais que seriam utilizados na presente análise. Assim, foram definidas ferramentas que fossem de livre e fácil acesso e que suprissem a necessidade das análises. Optou-se então por utilizar a linguagem Python com as bibliotecas Pandas e Matplotlib, executadas e armazenadas na nuvem por meio do Google Colab Notebook. Os dados para a análise foram fornecidos pela FATEC Indaiatuba sendo então preparados e filtrados até que se extraísse informação suficiente para responder as questões de interesse deste trabalho. Dentre diversas informações presentes na base de dados, encontram-se as notas de vestibular dos alunos, sua origem acadêmica, seu status de matrícula e desempenho individual em disciplinas cursadas.

O comparativo entre os cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial da FATEC Indaiatuba trouxe resultados importantes para ajudar a instituição no planejamento futuro visando a possíveis ações de melhoria educacional que possam evitar a evasão do aluno, principalmente nos primeiros períodos dos cursos.

Ao trabalhar-se as notas de vestibular dos alunos de ambos cursos, nota-se que o desempenho dos melhores alunos no vestibular não se mantém após a entrada na instituição. Adicionalmente, alunos que provêm de situações mais precárias, como os de escola pública, concluem em maior proporção o curso ADS, aquele com maior nota de corte no vestibular, quando comparados aos alunos de escola particular, sendo que o inverso ocorre para o curso de GE.

Explorando o tempo para formação dos alunos e os semestres com maior desistência, notou-se que predominantemente se desiste dos cursos logo nos primeiros semestres, o que não dá tempo hábil à instituição comprovar a importância da formação acadêmica para a vida pessoal e profissional. Os motivos para evasão podem ser estudados em trabalhos posteriores,

porém afirma-se que, independemente de quais sejam, devem ser trabalhados e evitados desde o princípio.

Por fim, completando o objetivo deste trabalho e afirmando a hipótese inicial do projeto, viu-se que os alunos de ADS possuem maiores dificuldades em todas as grandes classes de disciplinas (Humanas, Exatas e Tecnológicas), sendo que as matérias de Humanas são as que mais contribuem com a reprovação do aluno. Dessa maneira, foram encontrados pontos em que a instituição de ensino pode focar seus esforços visando a uma maior quantidade de alunos formados e que um menor número de desistências ocorra.

Trabalhos futuros devem ser desenvolvidos para encontrar quais são as dificuldades que a instituição de ensino superior tem que mais afetam a permanência dos alunos nos cursos, sendo aconselhado o *benchmark* com outras FATECs e faculdades públicas que ofereçam cursos de tecnologia. Assim, espera-se que os recursos possam ser melhor alocados e que a satisfação da comunidade seja maior com os serviços prestados e disponibilizados à população.

REFERÊNCIAS

BIRJALI, M., BENI-HSSANE, A., ERRITALI, M. Analyzing Social Media through Big Data using InfoSphere BigInsights and Apache Flume. In: PROCEDIA COMPUTER SCIENCE, 113., 2017. **The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks**, Moroco, 2017. p. 280-285.

CAO, L. 2017. **Data science: A comprehensive overview**. ACM Computing Surveys. v. 50, n. 3, p. 42, 2017.

DANIEL, B. Big Data and analytics in higher education: opportunities and challenges. **British journal of educational technology**, v. 46, n. 5, p. 904-920, 2015.

GIL, A. C. **Como elaborar projetos de pesquisa**. 6 ed. São Paulo. Atlas, 2017.

HU, Q., POLYZOU A., RANGWALA H. Enriching Course-Specific Regression Models with Content Features for Grade Prediction. In: ICDSA, **International Conference on Data Science and Advanced Analytics**, Fairfax/Minneapolis, 2017. p. 504-513.

KOCHETKOV, O. T., PROKHOROV, I. V. **The reseach of approaches of applying the results of big data analysis in higher education**. Conference Proceedings. v. 1797, n. 020008, 2017.

LARSON, D., CHANG, V. **A review and future direction of agile, business intelligence, analytics and data science**. Jornal International Journal of Information Management. v. 36, p. 700-710, 2016.

LYNCH, C. F. **Who prophets from big data in education? New insights and new challenges**. Theory and Research in Education. v. 15, n. 3, p. 249–271, 2017.

O'NEIL, C., SCHUTT, R. **Doing Data Science: Straight Talk from the Frontline**. 1 ed. California. O'Reilly, 2013.

RAMIREZ-GALLEGO, S., FERNÁNDEZ, A., GARCIA, S. Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce. **Journal Information Fusion**, Granada, Out. 2017, vol. 42, p. 51-61.

RUSSOM, P. Big Data Analytics. **TDWI best practices report, fourth quarter**, v. 19, p. 40, 2011.

SIN, K. MUTHU, L. Application of Big Data in education data mining and learning analytics: a literature review. **ICTACT journal on soft computing**, v. 5, n. 4, 2015.

VYAS, M. S., GULWANI, R. Predicting Student's Performance using CART approach in Data Science. In: ICECA 2017, **International Conference on Electronics, Communication and Aerospace Technology**, Mumbai, 2017. p. 58-61.

WASSAN J. T. **Discovering Big Data Modelling for Educational World**. Procedia - Social and Behavioral Sciences. v. 176, p. 642 – 64, 2015.

WEST, D. M. **Big Data for education: Data mining, data analytics, and web dashboards**. Governance Studies at Brookings. v. 9, p. 1 – 10, 2012.