Refas Revista Fatec Zona Sul
1ª. Edição Outubro 2014
Análise e Desenvolvimento de Sistemas
ISSN 2359-182X

ANÁLISE DE BIG DATA NO CENÁRIO EDUCACIONAL: UTILIZAÇÃO DE MODELOS PREDITIVOS NAS FATECS DO CENTRO PAULA SOUZA

BIG DATA ANALYTICS IN THE EDUCATIONAL SCENARIO: PREDICTIVE MODELS UTILIZATION IN CENTRO PAULA SOUZA'S FATECS

Francisco Ariel Campos Florencio[1]
Bruno Amadio de Araulu[2]
Maria das Graças Junqueira Machado Tomazela[3]
Michel Moron Munhoz[4]

RESUMO

Existe expressiva quantidade de dados educacionais gerados, mas que, em geral, não estão disponíveis aos gestores, no tempo e formato adequados. Neste contexto, a utilização de Big Data e suas ferramentas de análise potencializa a transformação das atividades de planejamento e tomada de decisão. Assim, o objetivo do trabalho foi aplicar ferramentas de análise de Big Data em dados da área educacional visando a dar profundo e amplo conhecimento a respeito do cenário educacional das Fatecs. Para atingir os objetivos propostos, optou-se pelo campo das pesquisas de natureza explicativa de abordagem experimental. Após revisão bibliográfica na área de Big Data, foi realizada pesquisa e avaliação de ferramentas para análise de Big Data. Na sequência foram realizadas atividades de pré-processamento dos dados e finalmente o uso de regressão linear e correlação, a fim de identificar a potencialidade e as possibilidades das ferramentas de análise de Big Data a partir de dados educacionais da Fatec de Indaiatuba. Como resultado dos procedimentos realizados obteve-se um modelo de regressão com aproximadamente 90% de precisão e a partir das análises dos seus coeficientes foi possível identificar as disciplinas que mais influenciam na retenção do aluno.

Palavras-chave: Análise de big data. Tecnologia. Dados educacionais. Planejamento estratégico.

ABSTRACT

There is a vast amount of educational data that is generated, but is generally not available to managers in the appropriate time and format. In this context, the use of Big Data and its analysis tools has the potential to transform planning and decision-making activities. Thus, the objective of this work was to apply Big Data analysis tools in educational data in order to give a deep and broad knowledge about the educational scenario of Fatecs. To achieve the objectives of this research, it was opted for the research field of explanatory nature of experimental approach. After a bibliographic review in the area of Big Data, it was carried out research and evaluation of tools for analysis of Big Data. Subsequently, data pre-processing activities were performed and finally the use of linear regression and correlation to identify the potential and possibilities of Big Data analysis tools based on educational data from Fatec Indaiatuba. As a result of the procedures performed it was obtained a regression model with approximately 90% accuracy and from the analysis of its coefficients it was possible to identify the disciplines that most influence student retention.

Keywords: big data analytics, technology, educational data, strategic planning.

[1] Fatec Indaiatuba. E-mail: frankcmps76@gmail.com.
[2] Fatec Indaiatuba. E-mail: brunooyeah@gmail.com.
[3] Fatec Indaiatuba. E-mail: gtomazela@fatecindaiatuba.edu.br.
[4] Fatec Indaiatuba. E-mail: michel@fatecindaiatuba.edu.br.

`

1 INTRODUÇÃO

Instituições de ensino superior trabalham em um ambiente cada vez mais complexo e competitivo. Existe uma necessidade crescente para responder a mudanças econômicas, políticas e sociais. Os desafios são, entre outros, manter as demandas para o vestibular, reduzir a evasão e os níveis de reprovação, bem como garantir a qualidade da aprendizagem (DANIEL, 2015).

Existe uma vasta quantidade de dados educacionais gerados, mas que, em geral, não estão disponíveis aos gestores, no tempo e formato adequados. Neste contexto, a utilização de Big Data e suas ferramentas de análise tem o potencial de transformar as atividades de planejamento e tomada de decisão (SIN; MUTHU, 2015).

Segundo Gandomi e Haider (2015), tamanho é a primeira característica que vem à mente considerando a questão: O que é big data? Trata-se de grandes volumes de dados provenientes de várias fontes, tais como redes sociais, sensores, dispositivos, terceiros, aplicativos da Web e mídias sociais, e em uma variedade de formatos, como texto, vídeo, áudio, diagramas, imagens e combinações de dois ou mais formatos.

Geralmente, Big Data é identificado por um conjunto de características fundamentais (DANIEL, 2015; GANDOMI e HAIDER, 2015; RUSSOM,2011): 1) Volume - grande quantidade de informações; 2) Velocidade - relativa à taxa crescente em que a informação flui dentro de uma organização; Veracidade - refere-se aos preconceitos, ao ruído e à anormalidade nos dados; 3) Variedade - referindo-se a dados em diversos formatos estruturados e desestruturados; 4) Verificação - refere-se à verificação e segurança de dados; 5)Valor –diz respeito aos resultados do processo de Big Data, ou seja, se os dados foram utilizados para gerar valor nos processos.

Segundo Daniel (2015), 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.

As Faculdades de tecnologia do Centro Paula Souza, como escola pública, precisam prestar conta à sociedade a respeito de seus resultados. Dessa maneira, as ações e as atividades de planejamento, devem ser subsidiadas por informações precisas para atingir suas metas acadêmicas. Assim, o objetivo deste trabalho foi aplicar ferramentas de análise de Big Data em dados da área educacional visando a dar um profundo e amplo conhecimento a respeito do cenário educacional das Fatecs e, desta forma, possibilitar a proposição ações assertivas na gestão das diversas unidades.

2 REFERENCIAL TEÓRICO

Para a realização deste projeto, alguns trabalhos relacionados foram analisados por meio de pesquisas disponibilizadas nas bases de dados *Science Direct*, IEEE e Web of Science filtradas dos últimos 4 anos. Os resultados foram obtidos em uma visita técnica realizada na UNICAMP a fim de selecionar artigos que fossem relacionados ao tema abordado neste

trabalho, isto é, análise dos conceitos relacionados a *Big Data* e suas ferramentas na área educacional. Esses trabalhos são apresentados a seguir:

Jin et al (2015) desenvolveram um trabalho para definir conceitos relacionados a *Big Data*, bem como suas características e valores. Os autores apresentaram uma breve revisão sobre as oportunidades e a importância do *Big Data*, assim como alguns desafios que o *Big Data* traz. Para a realização desse trabalho uma pesquisa experimental foi realizada. O trabalho teve como foco a aplicação na indústria, e como resultado constatou-se que uma abordagem de engenharia integrada deve ser empregada no gerenciamento de um projeto de *Big Data* para assim, tornar os processos melhores.

O trabalho de Gandomi e Haider (2015) teve como objetivo exibir, de uma forma descritiva e conceitual, como as ferramentas atualmente disponíveis tratam os dados não estruturados, para que assim gerem informações que possam ser utilizadas e analisadas. Uma característica deste artigo é o foco em análises relacionadas a dados não estruturados, que constituem 95% dos *Big Data*s. Os autores destacam a necessidade de desenvolver métodos analíticos eficientes e adequados para alavancar enormes volumes de dados heterogêneos em formatos de texto, áudio e vídeo não estruturados.

Bussaban e Waraporn (2015), realizaram um estudo com o objetivo de mostrar a importância de implementar o estudo de Data Science nos cursos de Ciência da computação e Matemática. Os autores afirmam que o cursos de Matemática podem ser pensados como tendo cinco componentes: visualização de dados (por exemplo, gráficos de dados, elementos de percepção visual), manipulação de dados (por exemplo, SQL, fusão, agregação e iteração), estatísticas computacionais (por exemplo, intervalos de confiança através do bootstrap , simulação, regressão, seleção de variáveis), mineração de dados / aprendizagem em máquina (por exemplo, classificação, validação cruzada) e tópicos adicionais (por exemplo, mineração de texto, mapeamento, expressões regulares, ciência da rede). Os autores afirmam ainda que em cursos de Ciência da computação as tecnologias de dados têm como objetivo conscientizar os alunos sobre a gama de tarefas que um computador é capaz de realizar (além de fornecer ferramentas concretas para executar tarefas específicas). Temas específicos incluem: como escrever código de computador; publicação de dados na World Wide Web (HTML); descrição de dados e marcação semântica (XML); armazenamento de dados (formatos de arquivo, planilhas, bancos de dados); gerenciamento de dados e resumo (consultas de banco de dados, SQL); processamento de dados (R).

Siddiqa et al. (2016), desenvolveram uma revisão que se centrou principalmente em aspectos de *Big Data* no contexto de gerenciamento de dados. Os autores descreveram as técnicas de gerenciamento de *Big Data*, explicitando as técnicas existentes para armazenamento, pré-processamento, processamento e também segurança de dados. Aspectos críticos dessas técnicas foram analisados por meio da elaboração de uma taxonomia para identificar os problemas e as propostas feitas para amenizar esses problemas. Esta pesquisa também pretendeu ser um guia para desafios e soluções no gerenciamento de Big Data e também um ponto de referência para trabalhos futuros. Os autores concluíram que, é crucial o desenvolvimento e utilização de técnicas e tecnologias de gerenciamento eficazes para lidar com os desafios de *Big Data*.

Os objetivos do trabalho de Chang e Larso (2016) foram: 1) revisar o alinhamento entre princípios ágeis e entrega de BI, analisar a ciência de dados. 2) analisar metodologias ágeis e como elas foram aplicadas com BI e estão surgindo em Big Data. 3) revisar os componentes e as melhores práticas da distribuição do Agile em BI, considerando o impacto do Big Data. 4) propor uma estrutura ágil para entrega de BI, análise rápida e ciência de dados. Os autores concluem que os princípios ágeis para entrega de BI, como o Agile, mudaram a forma que os

dados são tratados, análises rápidas e ciência de dados foram incluídas e as ideias ágeis foram encaixadas no mundo de BI.

O trabalho de Sivarajah et al. (2017), objetivou apresentar e sintetizar uma análise estruturada de ponta da literatura normativa sobre *Big Data* e *Big Data Analytics* para apoiar a sinalização de futuras direções de pesquisa. Para a realização deste trabalho, foi realizada uma revisão sistemática da literatura. De acordo com os autores, os resultados dessa revisão ajudarão os acadêmicos e profissionais de *Big Data* e de *Big Data Analytics* a desenvolver novas soluções com base nos desafios associados a Big Data, que foram agrupados em três categorias principais, com base no ciclo de vida dos dados: desafios de dados, processos e gerenciamento.

Birjali, Beni-Hssane e Erritali (2017), realizaram uma análise emocional com base em dados coletados em redes sociais utilizando processamento, análise e visualização de dados. Para isso, os dados dos *tweets* foram analisados a partir do *Flume* (serviço distribuído, confiável e disponível para coletar, agregar e mover de modo eficiente grandes quantidades de dados de eventos de fluxo), processados usando o script Jaql, armazenados e analisados no HDFS. Para a realização da análise, as palavras negativas e positivas passaram pelos métodos do *MapReduce* em seguida, os resultados foram exibidos como gráficos usando a ferramenta *BigSheets BigInsights* da IBM. Os autores concluíram que essa arquitetura não é apenas aplicável para streaming, processamento, análise e visualização dos dados do twitter, mas também para melhorar a aplicação de outros tipos de Big Data de várias fontes.

O trabalho de Gulwani (2017) relatou sobre a utilização de algoritmos computacionais para prever quais alunos de uma universidade, analisando um curso específico, precisavam de apoio diferenciado para que não abandonassem o curso. A técnica utilizada foi o CART, que implementa árvores de decisão binárias. Nesse trabalho, os autores montam uma estrutura estatística capaz de detectar alunos com problemas ou deficiências em matérias chave. Como resultado, os autores destacam que uma vez que a universidade for equipada com uma boa base de programação e/ou qualquer base de TI, naturalmente os alunos desenvolverão mais interesse em seu campo de estudo.

O trabalho de Kumar et al (2017) objetivam explanar sobre o uso geral da Mineração e Ciência de Dados no campo da Educação, citando todos os possíveis usos e localizando o desenvolvimento científico alcançado até o momento. Como resultado os autores esperam que esse projeto sirva como um primeiro passo na construção das bases do TGDS (*Theory-guided Data Science*) e encoraje outros trabalhos de acompanhamento para desenvolver em profundidade as formalizações teóricas desse paradigma.

Flath e Stein 2018, desenvolveram um toolbox de Data Scince para melhorar a utilização de tarefas de previsão na área de fabricação de produtos, para preencher a lacuna entre a pesquisa de aprendizagem de máquina e as necessidades práticas. Apresentam também diretrizes e práticas recomendadas para modelagem, engenharia de recursos e interpretação, alavancando ferramentas de sistemas de informações empresariais, bem como aprendizado de máquina. Os autores concluem que simplesmente mergulhar uma enorme quantidade de dados em algoritmos inteligentes não é a solução que muitos pesquisadores e profissionais esperam que seja. Em vez disso, eles mostram que a melhoria constante, a engenharia de recursos e a consolidação complementam o poder preditivo de um sistema analítico de negócios.

Uma pesquisa que apresentou uma introdução das características de metodologias de Map Reduce foi o trabalho de Fernández, Fernández, Garcia et al (2018). O artigo traz uma projeção de novos algoritmos neste campo de pesquisa e um estudo experimental que permitirá contrastar os problemas de escalabilidade para cada tipo de fusão de processo no MapReduce

para Big Data Analytics. Como resultado, os autores trazem conceitos de aplicação utilizando o processo Map Reduce, decorrentes de um estudo prático, além disso, listam diversas formas de utilização do Map Reduce e suas ferramentas.

3MÉTODO

Para atingir os objetivos desta pesquisa, optou-se pelo campo das pesquisas de natureza explicativa, que têm como preocupação central identificar os fatores que determinam ou que contribuam para a ocorrência dos fenômenos (GIL, 2002). Neste campo, a referência foi a abordagem experimental.

As etapas de execução deste trabalho foram: 1) Revisão Bibliográfica: verificação do "estado da arte" na área de *Big Data*; 2) Pesquisa e avaliação de ferramentas para análise de *Big Data*; 3) Modelagem, coleta, limpeza e estruturação dos dados; 4) Realização de atividades de regressão linear e correlação, para identificar as potencialidades e as possibilidades das ferramentas de análise de Big Data a partir de dados educacionais das Fatecs

Para realizar a análise de grandes massas de dados é comumente utilizado o sistema Spark, pois este oferece um ambiente clusterizado em que o processamento pode ser dividido entre várias máquinas e realizado de forma mais eficiente, neste caso foi utilizado o Pyspark, módulo e API do Spark para Python. Os códigos utilizados para essa análise foram executados numa máquina local, mas possuem suporte para serem *clusterizados* se necessário.

Foram utilizados para a realização dessa análise os módulos de python Pyspark, Pandas, MatPlotLib e SciPy.

4 RESULTADOS E DISCUSSÃO

Nas subseções a seguir são apresentadas as atividades desenvolvidas em cada etapa prevista no cronograma do projeto, a saber: 1) revisão sistemática da literatura sobre *Big Data*; 2) seleção e análise das ferramentas de *Big Data* e; 4) início do processo de Modelagem, coleta, limpeza e estruturação dos dados

4.1 Descrição dos dados

Os dados para a realização deste projeto foram concedidos pela Fatec Indaiatuba em forma de planilha eletrônica (formato xlsx), as planilhas foram enviadas pelos desenvolvedores do SIGA (Sistema Integrado de Gestão Acadêmica) com a autorização do diretor da Fatec Indaiatuba.

Para esta análise foram utilizados os dados referentes ao desempenho dos alunos da Fatec Indaiatuba. A base de dados possui uma linha com dados do aluno para cada disciplina cursada e o status de matrícula do aluno: se ele foi aprovado, transferido ou teve a matrícula cancelada.

Os atributos escolhidos, visando a dar suporte às análises: DISCIPLINA, SIGLA (da disciplina) NOTA, FREQUENCIA, STATUS_ALUNO, SEMESTRE_ANO e CONCEITO, para verificar o desempenho do aluno em cada matéria. Além disso, a planilha também conta com atributos como o campo TURNO, para definir em qual turno o aluno estuda, RA, que exibe o número de registro do aluno, NOME, que identifica o nome do aluno, ESCOLA_PUBLICA, que define se o aluno ingressante na FATEC é proveniente de uma escola pública ou particular, RAÇA, que pode ser negra, parda, branca amarela ou não declarada conforme preenchida pelo aluno, NOTA_VESTIBULAR, que exibe o desempenho do aluno no vestibular e o campo DATA_NASCIMENTO do aluno.

4.2 Preparação de Dados

Primeiramente foi definida a realização de um modelo de regressão que pudesse prever qual seria a situação de um aluno com base nos dados disponibilizados e quais desses dados seriam mais relevantes nessa previsão, a partir dos coeficientes da regressão.

Para realização da regressão os dados foram agrupados por nome de aluno usando o Power Query para que cada aluno apareça apenas uma vez na tabela, para as colunas “Frequência” e “Nota” foi calculada a média dos dados agrupados que representa respectivamente a frequência média do aluno no curso e a média de notas do aluno, ou, o seu rendimento.

As colunas necessárias para a regressão foram transformadas em colunas numéricas atribuindo um valor para cada item da coluna.

Foram realizadas as seguintes atribuições: “ESCOLA_PUBLICA” com valor 1 caso o aluno seja de escola pública e 0 um caso não seja, “RAÇA” com valor entre -1 e 1 para as raças Amarela, Branca, Indígena, Parda e Negra; “Frequência” de 0 a 100 que representa a porcentagem da frequência do aluno nas aulas, “Média de Notas” de 0 a 100 que representa a média de todas as notas do aluno e “Nota Vestibular” de 0 a 100 que representa a nota que o aluno obteve no vestibular da Fatec.

A tabela foi exportada para um arquivo do tipo CSV que pode ser lido pelos métodos do módulo pyspark e armazenada em um *data frame*.

4.3 Realização da Regressão Linear

Na aplicação da regressão foi utilizado o objeto “*LinearRegression*” da subpackage Mllib do pyspark. Mllib, é uma biblioteca de machine learning integrada ao sistema Spark que permite a criação de rotinas, modelos e pipelines de aprendizado de máquina com a possibilidade de clusterização utilizando os recursos do Spark. Esse objeto recebe como argumento dois arrays, uma com os dados a serem previstos e a outra com uma lista de todos os dados de entrada para essa previsão, para deixar os dados nesse formato é usado o objeto “*VectorAssembler*”.

Criou-se então um objeto “*VectorAssembler*” que recebe um array com o nome das colunas numéricas que serão usadas na regressão e o nome de uma coluna de saída que concatena todas as entradas.

Os dados preparados pelo “*VectorAssembler*” foram então separados em 75% para treinamento do modelo e 25% para teste.

Após o treinamento do modelo de regressão e seus testes foi possível extrair informações como, o erro médio quadrado, o erro médio absoluto, os coeficientes da regressão, as previsões dadas pelo modelo para cada uma das linhas de teste e os seus resíduos (diferença entre o valor real e a previsão).

Na Figura 1 se encontram o erro médio absoluto e o erro médio quadrado.

Figura 1 - Cálculo de erro médio absoluto e erro médio quadrado

```
print(pred_results.meanAbsoluteError,',',pred_results.meanSquaredError)
```

```
0.26922217439633794 , 0.109731568939899
```

Fonte: autores

Os coeficientes da regressão podem ser observados na Figura 2, estes estão respectivamente relacionados com as colunas “ESCOLA_PUBLICA”, “RAÇA”, “NOTA_VESTIBULAR”, “MediaNotas”, “Frequência”, ou seja, as colunas usadas para criar o objeto “*LinearRegression*” na ordem que foram dadas.

Figura 2 - Coeficientes de regressão

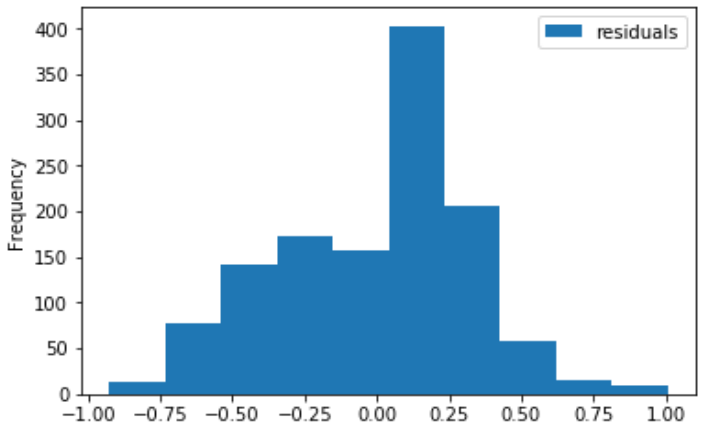
```
regressor.coefficients
```

```
DenseVector([0.0041, 0.0086, 0.0009, 0.1173, -0.0009])
```

Fonte: autores

A fim de compreender a extensão dos erros da regressão foi elaborado um histograma com os resíduos do modelo como mostra a Figura 3.

Figura 3 - Histograma de resíduos



Fonte: autores

Ao analisar este histograma foi possível ver que a grande maioria dos testes tiveram uma divergência com o valor real entre -0,75 e 0,5, ou seja, levando em consideração que os valores reais são sempre um ou zero, essa diferença apresenta grande impacto. O ideal seria que a maioria dos dados se encontrasse no meio, próximo de zero, representando o menor resíduo possível.

Com esse modelo de regressão linear atingiu-se uma previsão da situação do aluno de aproximadamente 49% de precisão, observa-se então que as entradas não descrevem com exatidão o status do aluno, ou elas por si só não são suficientes para prevê-lo. Então para entender melhor a influência das variáveis de entrada sem a interferência da precisão de uma regressão linear foi realizado como segundo passo uma análise de correlações.

O Quadro 1 apresenta as principais linhas de código utilizadas para a realização dessa análise.

Quadro 1 - Fragmentos de código da regressão linear

Importação dos dados	data = spark.read.csv ('Indaiatuba_BigData_Prep.csv', inferSchema=True, header=True)
Criação do Vector Assembler	featureAssembler=VectorAssembler(inputCols=['ESCOLA_PUBLICA','RAÇA','NOTA_VESTIBULAR','MediaNotas','Frequencia'],outputCol='Idependent_Features')
Criação do "output" e "finalized_data"	output=featureAssembler.transform(data) finalized_data = output.select('Idependent_Features','STATUS_ALUNO')
Criação de dados de treino e teste	train_data,test_data = finalized_data.randomSplit([.75,.25])
Criação do Linear Regression	regressor = LinearRegression(featuresCol='Idependent_Features', labelCol='STATUS_ALUNO')
Treinamento do módulo	regressor=regressor.fit(train_data)
Teste do módulo	pred_results=regressor.evaluate(test_data)

Fonte: autores

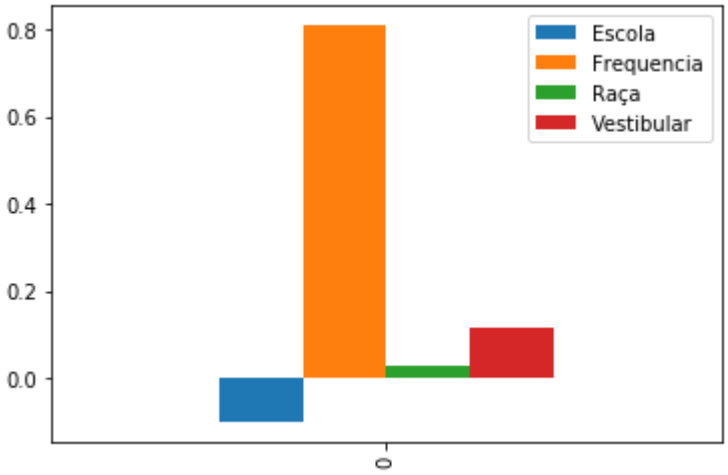
4.3.1 Análise de Correlação

Para essa análise foi feito tanto o cálculo de correlação com o alvo final, a situação do aluno, quanto com o seu rendimento, ou seja, sua média de notas no curso. Essa análise foi realizada com a utilização do módulo Scipy, que possui métodos para o cálculo de diversos coeficientes de correlação.

Fez-se um estudo das características dos dados e a partir dele se escolheu o cálculo de coeficiente de Spearman. O coeficiente de Spearman avalia com que intensidade a relação entre duas variáveis pode ser descrita pelo uso de uma função monótona (função entre dois conjuntos que preserva a relação de ordem entre estes). Logo o quão mais próximo esse coeficiente for de 1 ou -1 mais forte é essa relação.

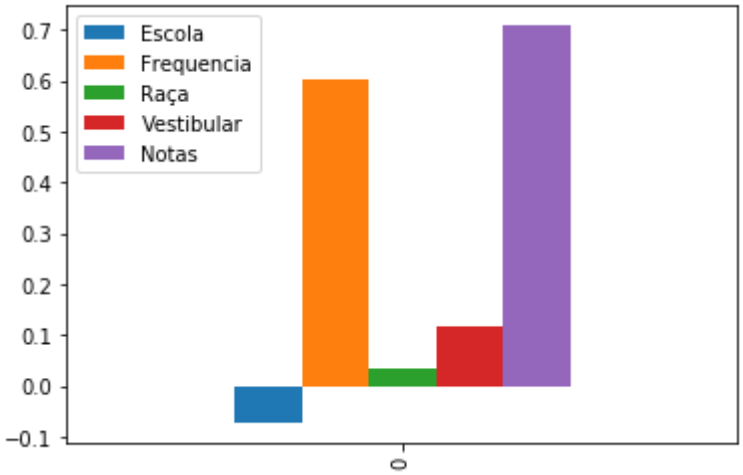
Com o método Spearman foi calculado o coeficiente de Spearman de cada uma das colunas previamente usadas na regressão linear primeiramente com o rendimento do aluno, Figura 4, e posteriormente com sua situação final, Figura 5.

Figura 4 - Gráfico de correlações com o rendimento do aluno



Fonte: do autores

Figura 5 - Gráfico de correlações com o status do aluno



Fonte: autores

Com a análise dos gráficos ficou claro que há uma grande correlação entre a frequência de um aluno e o seu rendimento escolar, bem como entre a frequência e suas notas para com sua situação final. Mas essa conclusão é um tanto óbvia e não necessita de uma análise de dados complexa para ser evidenciada, porém o que os gráficos também demonstram é que não há correlação entre os demais dados e o rendimento ou situação final do aluno o que tem grande valor, ao contrário do que se esperava pelo senso comum é possível dizer que a procedência de escola pública ou particular não tem impacto no aproveitamento escolar do aluno. O mesmo pode ser dito para raça e nota no vestibular.

4.3.2 Realização da Segunda Regressão Linear

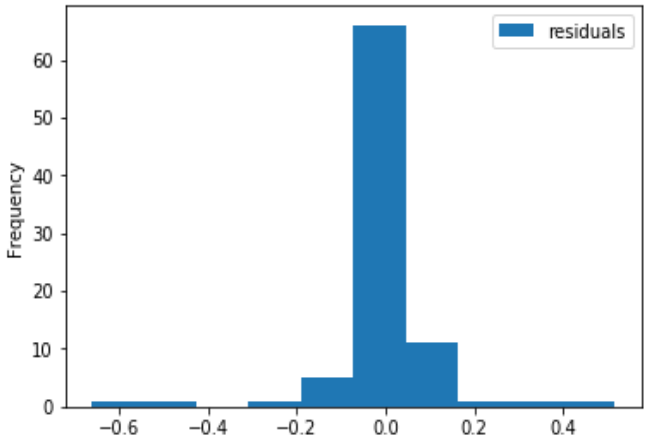
Posteriormente foi realizada uma segunda regressão linear, desta vez levando em consideração os dados de cada disciplina separadamente, para essa análise foi utilizado apenas o curso de Análise e Desenvolvimento de Sistemas para reduzir os valores nulos em matérias de alunos em cursos diferentes. Foi utilizado para cada disciplina a nota máxima atingida pelo aluno, a nota mínima e o número de vezes que o aluno a cursou.

Desta vez o agrupamento de dados por nome do aluno foi realizado com Pyspark assim como o pivot que transpõe cada matéria da coluna “DISCIPLINA” para três novas colunas com os cálculos max, min e count citados anteriormente.

Antes de realizar o procedimento já descrito para realização da regressão linear foi utilizado o método “fillna” para substituir os valores nulos do *data frame* por zeros, já que o processo da regressão linear não aceita valores vazios.

Após o treinamento e os testes do modelo foi calculado a precisão dele que atingiu um percentual de aproximadamente 90%. Assim como na regressão anterior foi plotado um gráfico com os resíduos do modelo, Figura 6, e nele pode-se observar que os dados estão muito mais concentrados no centro do eixo X próximo a zero e são muito poucos os testes que se distanciam disso.

Figura 6 - Histograma de resíduos 2



Fonte: autores

Com este modelo foi analisado os coeficientes para tentar encontrar as principais causas de reprova dos estudantes.

Para isso, foi necessário relacionar cada um dos coeficientes com seu respectivo nome de coluna, então foi criado um *data frame* com os valores dos coeficientes e para cada um deles em uma nova coluna o seu nome. Após isso o *data frame* foi colocado em ordem crescente baseado no valor dos coeficientes assim pode-se observar quais dados são mais relevantes para o resultado final, Figura 7.

Figura 7 - Coeficientes em Ordem Crescente

```
-0.06552986921393407   Sociedade e Tecnologia_count(NOTA)
-0.024086323032674054   Redes de Computadores_count(NOTA)
-0.023124930508496237   Programação WEB_count(NOTA)
-0.022546008752036256   Programação em Microinformática_count(NOTA)
-0.019391953881494967   Organização de Computadores_count(NOTA)
-0.018012189202490158   Laboratório de Hardware_count(NOTA)
-0.013162240785949127   Tópicos Especiais em Informática (Escolha 2)_count(NOTA)
-0.012789961171152218   Espanhol I_count(NOTA)
-0.010946848938387243   Trabalho de Graduação I_count(NOTA)
-0.009649722199548974   Engenharia de Software II_count(NOTA)
-0.008668055033597225   Linguagem de Programação_count(NOTA)
```

Fonte: autores

Porém encontrou-se um problema, para cada vez que o modelo era treinado e se ordenava os coeficientes, a ordem era diferente. A precisão do modelo se mantinha consistente, porém não a forma que ele encontrava seu resultado, ou seja, o modelo era bom para fazer previsões, mas não para se analisar os coeficientes.

Para tentar encontrar algum padrão, o treinamento do modelo foi executado várias vezes, e seus coeficientes gravados em um arquivo CSV para análise. A partir deste arquivo foi feito uma tabela com a média dos coeficientes, como mostra a Tabela 1.

Tabela 1 - Média de coeficientes

index	media	nome_coluna
130	9.19048	[140, Sistemas de Informação_min(NOTA)]
85	19.5238	[95, Laboratório de Engenharia de Software_min...
127	20.3333	[137, Sistemas Operacionais II_min(NOTA)]
97	21.3333	[107, Metodologia da Pesquisa Científico-Tecno...
145	22.2857	[155, Ética e Responsabilidade Profissional_mi...
94	23.1905	[104, Matemática Discreta_min(NOTA)]
100	23.8095	[110, Organização de Computadores_min(NOTA)]
141	30.9524	[151, Trabalho de Graduação II_count(NOTA)]
134	36.0476	[144, Sociedade e Tecnologia_max(NOTA)]
31	38.0476	[41, Engenharia de Software II_min(NOTA)]
93	38.2381	[103, Linguagem de Programação_count(NOTA)]
136	39.6667	[146, Trabalho de Graduação I_min(NOTA)]

Fonte: autores

Na Tabela 1 pode-se observar o índice utilizado para encontrar o nome da coluna. Os coeficientes estão ordenados do menor para o maior, ou seja, as colunas com menores médias são aquelas com os menores coeficientes ou os fatores que mais influenciam na retenção de um aluno.

Por meio da análise dessa tabela os gestores podem tomar medidas para tentar diminuir o índice de reprovação em sua instituição de ensino.

Também foi criado um *data frame* com a posição dos coeficientes de cada um dos modelos treinados para análise como mostra a Tabela 2.

Tabela 2 - Data Frame de Coeficientes

index	order	order1	order2	order3	order4	order5	order6	order7
0	23	126	132	16	20	133	134	133
1	83	87	92	61	83	89	87	86
2	110	105	98	94	105	68	78	111
3	38	42	64	37	40	88	63	41
4	141	144	136	147	131	137	131	147
5	73	80	86	69	72	80	69	78
6	78	74	83	75	73	77	73	72
7	132	76	84	74	140	14	7	80
8	85	59	96	84	101	74	81	87
9	84	95	69	64	62	73	80	59

Fonte: autores

Esse *data frame* possui índices para se relacionar com o nome de coluna de cada coeficiente e as colunas com nome “order” para cada uma das vezes que o modelo foi treinado. Ao analisar o *data frame* pode se ter melhor ideia de qual é o comportamento de cada uma das colunas durante os teste e saber se existe consistência nas posições. Por exemplo a linha de index 1 na Tabela 2 possui posição mínima de 61, máxima de 92, média de 83,5 e moda de 83 e 87 o que mostra grande consistência para esta coluna e dá maior credibilidade para a análise.

Enfim, com as tabelas criadas a partir dessa análise pode-se obter informações valiosas para a tomada de decisão. Há ainda espaço para muita análise e muita coisa ainda pode ser feita, essa é apenas uma amostra do quanto pode se encontrar ao analisar uma base de dados simples e o quanto isso pode ser útil para as organizações educacionais.

5 CONSIDERAÇÕES FINAIS

Ao final da pesquisa foram encontradas informações relevantes para a gestão de soluções em relação às disciplinas que mais influenciam a retenção do aluno de ADS da Fatec Indaiatuba e a descoberta de novos tipos de análises que podem ser realizadas para dar suporte ao negócio o que demonstra a grande oportunidade que o Big Data traz para o ramo educacional.

Considera-se que, com esse conjunto de procedimentos e resultados, foram alcançados os objetivos desse projeto, uma vez que foi desenvolvida uma estratégia de aplicação de ferramentas de análise de Big Data em dados da área educacional, por meio de um projeto piloto na Fatec Indaiatuba, que pode prover amplo conhecimento a respeito do cenário educacional das Fatecs e, desta forma, auxiliar na proposta de ações assertivas na gestão das diversas unidades.

6 REFERÊNCIAS

BIRJALI, M. BENI-HSSANEA, A. ERRITALIB, M. **Analyzing Social Média through Big Data using InfoSphere BigInsights and Apache Flume**. Procedia Computer Science, v.113, p. 280–285, 2017.

BUSSABAN, K. WARAPORN, P. **Preparing undergraduate students majoring in Computer Science and Mathematics with Data Science perspectives and awareness in the age of Big Data**. Social and Behavioral Sciences, v. 197, p. 1443 – 1446, 2015.

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

FLATH, M. C. STEIN, N. **Towards a data science toolbox for industrial analytics applications**. Computers in Industry, v.94, p. 16–25, 2018.

GANDOMI, T.; HAIDER, M. **Beyond the hype: Big data concepts, methods, and analytics**. Jornal Information Fusion v. 35, p. 137–144, 2015.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

JIN, X. et al. **Significance and Challenges of Big Data Research**. Journal Big Data Research, Hong Kong, 26 Fev, 2015.

KUMAR, V. et al. T**heory-Guided Data Science: A New Paradigm for Scientific Discovery from Data**. IEEE Transactions on knowledge and data engineering, Out. 2017, vol. 29, p. 2318-2331.

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

MANYIKA, J., M. CHUI, B. BROWN, J. BUGHIN, R. DOBBS et al. **Big data: The Next Frontier for Innovation, Competition, and Productivity**. McKinsey Global Institute. 2011.

RAMIREZ-GALLEGO S., FERNÁNDEZ A., GARCIA S. et al. **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, 2016.

RUSSOM P., **Big Data Analytics, TDWI best practices report,** The Data Warehousing Institute (TDWI) Research (2011).

SIDDIQA, A. et al. **A survey of big data management: Taxonomy and state-of-the-art**. Journal of Network and Computer Applications, v. 71, p. 151-166, 2016.

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.

SIVARAJAH, U. et al. **Critical analysis of Big Data challenges and analytical methods.** Journal of Business Research, v70 p. 263–286, 2017.

SCAICO P. D., QUEIROZ R. J. G. B. DE, SCAICO A. (2014). **O conceito de Big Data na Educação**. In: WORKSHOP DE INFORMÁTICA NA ESCOLA. 3º Congresso Brasileiro de Informática na Educação (CBIE 2014).

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.