DOI: 10.5748/16CONTECSI/DSC-6154

AN ANALYSIS OF THE SCHOOL DESERTION AND FAILURE RATE OF FATEC SOROCABA STUDENTS, USING BIG DATA

UMA ANÁLISE DO ÍNDICE DE EVASÃO E REPROVA DOS ALUNOS DA FATEC SOROCABA UTILIZANDO BIG DATA

RESUMO: Cada dia mais e mais dispositivos se conectam, seja de forma privada ou pela internet, esses dispositivos geram uma quantidade imensa de dados todos os dias, dados de todas as formas, como estruturados, semiestruturados e não estruturados. Em contrapartida as ferramentas convencionais se mostram ineficazes para armazenar e gerir todos esses dados. Na área educacional não é diferente, existe uma vasta quantidade de dados educacionais gerados, mas que, em geral, não estão disponíveis aos gestores, no tempo e formato adequados. Assim, este trabalho tem como objetivo armazenar, gerenciar e analisar dados de índices de evasão e reprova provenientes da FATEC Sorocaba. Na fundamentação teórica, são apresentados conceitos chave que sustentam a pesquisa como as características do Big Data, ferramentas utilizadas para coleta, armazenamento, modelagem e análise de dados, os quais são ancorados em práticas de pesquisas apresentadas em uma revisão da literatura. Para alcançar os objetivos propostos, foi realizada uma pesquisa experimental, que constituiu em utilizar ferramentas de Big Data de forma a extrair informações que possam ser utilizadas para encontrar um padrão e prever os índices de reprovas e evasão dentro da FATEC Sorocaba. Para o experimento foram selecionadas ferramentas de gerenciamento e armazenamento de grandes volumes de dados como, Hadoop, Spark e ferramentas de análise como Jupter Notebook. Os resultados obtidos com essa analise mostraram que em sua maioria os alunos que ingressam nos cursos noturnos da FATEC Sorocaba, vêm de escolas particulares, e o curso de Análise e Desenvolvimento de Sistemas é a única exceção, pois conta com maioria de alunos vindos de escolas particulares, de um modo geral a FATEC tem altos índices de evasão, dos cursos analisados dois apresentaram índice de evasão superior a 60% e os outros dois de conclusão superior a 50%. Os resultados obtidos também mostraram que em relação ao curso de Análise e Desenvolvimento de Sistemas as disciplinas técnicas são a que mais reprovam alunos.

PALAVRAS-CHAVE: Big Data. Análise. Educação

ABSTRACT: Every day more and more devices are connected, these devices generate a huge amount of data every day, data of all shapes, such as structured, semistructured and unstructured. In contrast, conventional tools prove inefficient to store and manage all of this data. In the educational area there is no difference, there is a vast amount of educational data generated, but in general, these data is not available to managers, in the appropriate time and format. Thus, this academic work aims to store, manage and analyze data from evasion and failure rates from FATEC Sorocaba. In the theoretical basis, key concepts are presented that support the research as the characteristics of Big Data, tools used for collection, storage, modeling and data analysis, which are anchored in research practices presented in a literature review. In order to reach the proposed objectives, an experimental research was carried out, which consisted in using Big Data tools in order to extract information that can be used to a pattern find and predict the failure and evasion rates within FATEC Sorocaba. For the experiment we selected tools for managing and storing large volumes of data such as Hadoop, Spark and analysis tools such as Jupyter Notebook. The results obtained with this analysis showed that the most part of the students in the FATEC Sorocaba's night courses degree come from private schools, and the Systems Analysis and Development course is the only exception because it has a majority of students coming from private schools, in general FATEC has high dropout rates, of the courses analyzed, two had a dropout rate of over 60% and the other two had a dropout rate of over 50%. The obtained results also showed that in relation to the course of Analysis and Development of Systems the technical disciplines are the one that reprove more students.

KEYWORDS: Big data. Analyze. Education. Spark.

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).

As Faculdades de tecnologia do Centro Paula Souza também tem que lidar com todos esses desafios, pois, como escola pública, precisa prestar conta à sociedade a respeito de seus resultados. Assim, as ações e as atividades de planejamento, tanto de Centro Paula Souza, como das unidades de ensino superior, as Fatecs, devem ser subsidiadas por informações precisas para atingir suas metas acadêmicas.

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.

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 *status quo.* 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.

A partir desse contexto, o objetivo geral desta pesquisa foi desenvolver uma estratégia de aplicação de ferramentas de análise de *Big Data* em dados da área educacional (por exemplo, evasão, índice de reprovação) visando a dar um profundo e amplo conhecimento a respeito do cenário educacional das Fatecs e, desta forma, propor ações assertivas na gestão das diversas unidades. Foram propostos os seguintes objetivos específicos para a realização dessa investigação, que se trata de um projeto piloto:

1. Avaliar e identificar quais ferramentas de análise de *Big Data* são adequadas para a realização desta pesquisa.
2. Aplicar, analisar e avaliar o potencial das técnicas de análise de *Big Data* na gestão acadêmica da Fatec Sorocaba.

2 METODOLOGIA

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.

Neste campo, a referência é a abordagem **experimental**, que consiste em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto (GIL, 2002). Para o desenvolvimento desta pesquisa, as seguintes variáveis de controle foram definidas:

1. Momento do processo de integralização que ocorre a evasão;
2. Especificidades das disciplinas que mais reprovam: básicas ou técnicas;

3. Perfil dos alunos que evadem.

A partir disso, as seguintes etapas definidas para a execução deste trabalho:

1. Revisão Bibliográfica: verificação do “estado da arte” na área de *Big Data*;

- Revisão da Literatura utilizando artigos de diversas bases de dados, tais como : *Science direct, IEEE Xplore, Web of Science, Scopus, Compendex;*

2. Pesquisa e avaliação de ferramentas para análise de *Big Data*: ferramentas não pagas disponibilizadas pela *Google*, ferramentas de *Big Data* disponibilizadas na plataforma *Microsoft Azure*, ferramentas não pagas disponibilizadas pela *Apache Foundation*.

- Modelagem, coleta, limpeza e estruturação dos dados, coletados do sistema acadêmico das Fatecs – SIGA

3. Aplicação de ferramentas de análise de *Big Data* (*Big Data Analytics*) nos dados coletados;

4. Análise e avaliação dos resultados obtidos no processo;

2 DESENVOLVIMENTO

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.

Big Data é um campo emergente de pesquisa que usa análise de dados para apoiar as decisões. 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 (SIN e MUTHU,2015; SIDDIQA, 2016).

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; SIVARAJAH, 2017).

Geralmente, *Big Data* é identificado por um conjunto de características fundamentais (RUSSOM,2011;DANIEL, 2015; GANDOMI e HAIDER, 2015; FLATH e STEIN, 2018):

- Volume - grande quantidade de informações, muitas vezes desafiadora para armazenar, processar e transferir, analisar e apresentar;
- 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. Também diz respeito a como os dados são armazenados e extraídos de forma significativa para o problema que está sendo analisado. Abrange ainda questões de confiança e incerteza;
- Variedade - referindo-se a dados em diversos formatos estruturados e desestruturados;
- Verificação - refere-se à verificação e segurança de dados;
- Valor – a característica mais importante, pois diz respeito aos resultados do processo de *Big Data*, ou seja, os dados foram utilizados para gerar valor nos processos de negócios dentro de uma organização?

Além dessas propriedades, há três etapas necessárias para que o *Big Data* agregue valor às atividades de gestão (DANIEL, 2015):

- 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.

É importante notar que há duas entidades técnicas. Primeiro, há o 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 Inteligence*) da atualidade.

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.

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.

A partir da exploração do conceitos teóricos sobre Big Data, foi realizada a análise de ferramentas tecnológicas específicas para esse propósito. Na próxima sessão são apresentados os resultados obtidos no processo experimental.

4 RESULTADOS OBTIDOS

Para a obtenção de conhecimento, na etapa de interpretação de dados, foi utilizada a ferramenta do ecossistema Spark. Os dados foram importados no Jupter Notbook, por meio de comando no console Web da ferramenta.

Os dados para a realização deste projeto foram concedidos pela Fatec Sorocaba em forma de planilha eletrônica (formato xlsx). A planilha foi enviada pelos desenvolvedores do SIGA (Sistema Integrado de Gestão Acadêmica) com a autorização do diretor da Fatec Sorocaba.

A planilha possuía um total de 275.247 linhas. Por se tratar de ferramentas de Big Data, que são preparadas para trabalhar com dados desestruturados, optou-se por transformar a planilha (formato XLSX) em um arquivo texto (TXT) para a manipulação dos dados. Primeiro os dados foram filtrados, selecionando o que seria analisado, em seguida foram convertidos para o formato CSV (*Comma-Separated Values*) e depois copiados em um arquivo TXT para o carregamento dos dados nas ferramentas de análise.

Após o carregamento dos dados e execução de um processo de *map reduce* foi possível realizar as diversas consultas estabelecidas para o projeto piloto e gerar diferentes gráficos. No Quadro 1 são apresentadas as atividades realizadas no projeto e as respectivas ferramentas.

Quadro 1 – Atividades realizadas pelas ferramentas

Ferramenta	Atividade
Hadoop	Responsável pelo armazenamento dos dados
Spark	Ambiente que permitiu a manipulação dos dados. Possui ferramentas para utilização de comantos SQL e programação em Python (pyspark).
Anaconda	Fornece o core da linguagem python e IDE/Editor Jupter
Jupter Notebbok	Aplicação Web que permitiu a manipulação e análise dos dados.

Os atributos escolhidos, visando a dar suporte às variáveis de controle definidas foram: STATUS_ALUNO, que contém informação a respeito do *status* da matrícula. Para análise de evasão, ESCOLA_PUBLICA, que contém a informação se o aluno veio de escola pública ou não, e para a análise de evasão o atributo escolhido foi o campo CONCEITO que pode possuir os valores: concluído, transferido ou cancelado. Além disso a planilha contém os campos CURSO que contém o nome do curso; TURNO com o período que o curso é oferecido; RA que mostra o registro do aluno; NOME com o nome do aluno; RAÇA que pode ser negra, parda, branca, amarela ou não declarada conforme preenchida pelo aluno; NOTA_VESTIBULAR com a nota que o aluno obteve no vestibular; DATA_NASCIMENTO, a data de nascimento do aluno; SEMESTRE_ANO contendo o ano e o semestre que o aluno cursou cada disciplina, SIGLA que contém a sigla das disciplinas; DISCIPLINA; que traz o nome da disciplina e NOTA com a nota que o aluno tirou na respectiva disciplina no semestre em questão.

Após análise de documentação e execução de testes, as ferramentas para a realização do projeto piloto foram escolhidas. Foram utilizadas as ferramentas da Arquitetura Spark como Jupyter Notebook, Spark SQL, GraphX e Anaconda, que contém o core da linguagem Python.

No projeto piloto realizado, salienta-se que, nesse primeiro momento, ainda não foram utilizados dados desestruturados, como por exemplo dados de redes sociais, porque o projeto está em fase inicial e de domínio das tecnologias envolvidas. Salienta-se também que não foi possível fazer qualquer análise pelo campo raça porque o valor desse campo não foi declarado pela maior parte dos alunos.

Foram analisados os dados de todos os cursos noturnos da Fatec Sorocaba, para os alunos matriculados no primeiro semestre de 2012, optou-se por esses dados

pois são consolidados, uma vez que o aluno possui um limite de doze semestres para se formar (incluindo o trancamento de matricula que pode ocorrer duas vezes). Além disso, foi realizada análise das aprovações, semestre a semestre, do curso de Análise e Desenvolvimento de Sistemas.

Dos alunos que iniciaram um curso na Fatec Sorocaba, em 2012, observa-se que a maior parte não conseguiu concluí-lo, isto é, 55,8% do total de alunos. Pôde-se perceber também que a Fatec Sorocaba possui maioria de alunos vindos de escolas particulares. Observa-se ainda que mais da metade dos 76 alunos vindos de escolas públicas não concluíram a graduação, e que dos 123 alunos vindos de escolas particulares, pouco menos da metade, 70 alunos, concluíram o curso que realizava. A Figura 1 apresenta a quantidade de alunos que se formaram e de alunos que não se formaram, dividido por tipo de escola, isto é, pública ou particular. O Gráfico mostra que, proporcionalmente, os alunos que vieram de escolas públicas tendem a evadir mais que os alunos que vieram de escolas particulares.

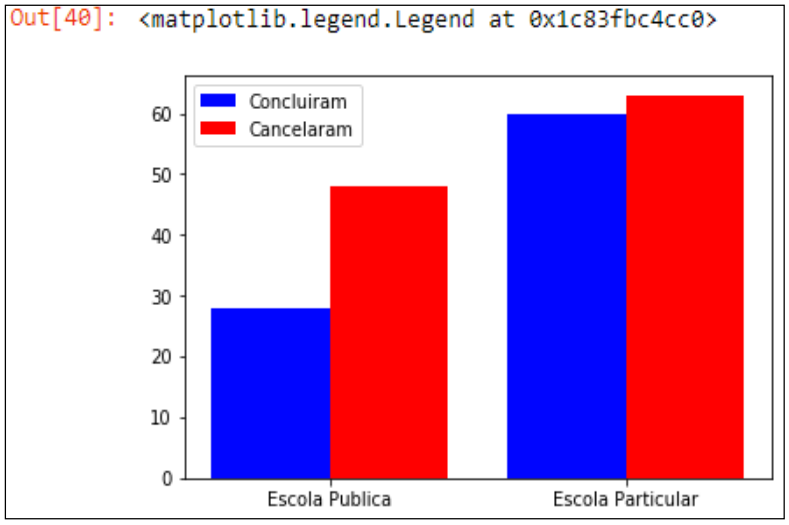

Figura 1 - Concluintes x Escola

O curso de Polímeros (PL) foi o que teve o maior índice de desistência, dos quarenta alunos que iniciaram o curso no primeiro semestre de 2012 somente 12, ou 30%, concluíram o curso. Percebe-se também que os 12 alunos que concluíram são alunos que vieram de escolas particulares. Desses alunos nenhum que veio de escola pública se formou, como apresentado na Figura 2.

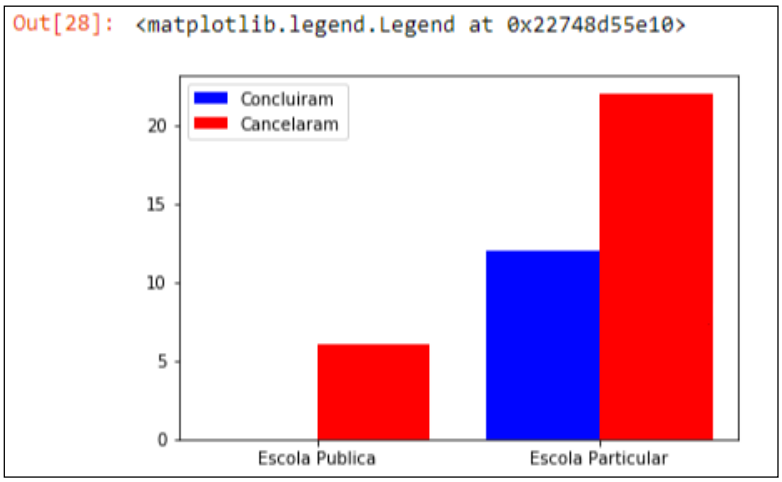

Figura 2 – Alunos de PL X Escola X concluintes

No curso de Projetos Mecânicos (PM) nota-se o mesmo padrão do curso de Polímeros, grande maioria dos alunos não se formou, um total de 67,5% dos alunos. Esse curso possui também maioria de alunos vinda de escolas particulares, dos 28 que iniciaram o curso, 11 se formaram, e dos 12 que vieram de escolas públicas apenas 2 concluíram o curso como apresentado na Figura 3.

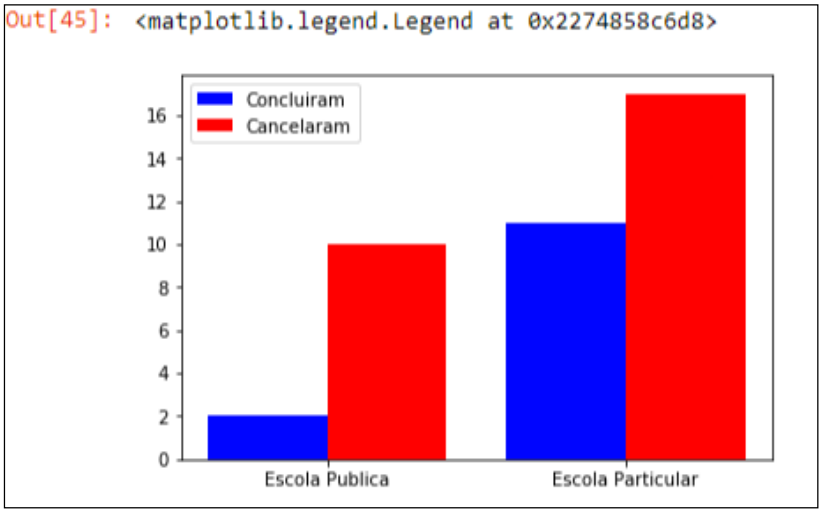

Figura 3 – Alunos de PM X Escola X concluintes

O curso de Fabricação Mecânica (FM) é o único curso noturno da FATEC Sorocaba que oferece 80 vagas noturnas. O curso tem um índice de aprovação superior ao de reprovação. A maior parte dos seus alunos também vêm de escolas particulares. Dos 25 alunos vindos de escolas públicas apenas 6 se formaram, dos que vieram de escolas particulares, 35 se formaram e 20 cancelaram, conforme Figura 4.

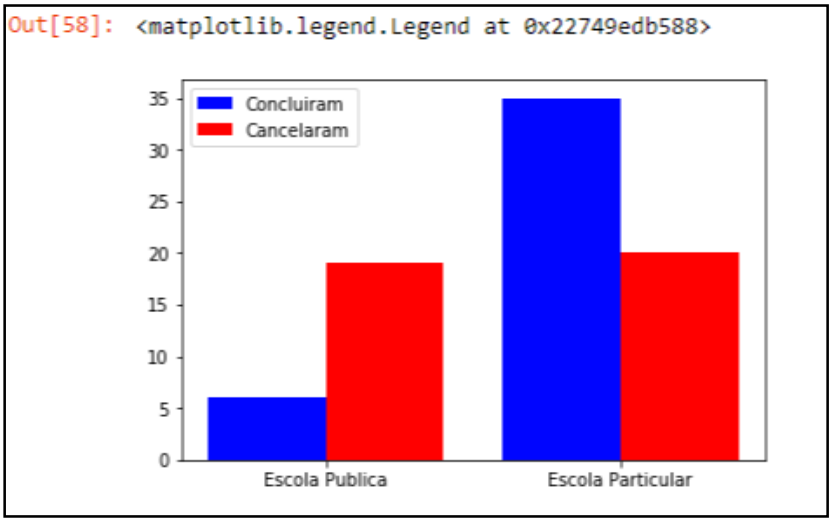

Figura 4 – Alunos de Fabricação Mecânica X Escola X concluintes

O curso de Análise e Desenvolvimento de Sistemas (ADS) é o único curso noturno em que a maioria dos alunos vêm de escola pública, tem um índice de aprovados superior ao de reprovados. No primeiro semestre de 2012 das 40 vagas oferecidas, 39 foram preenchidas, dos 33 alunos que vieram de escolas públicas, 20 se formaram, dos 6 alunos provenientes de escolas particulares, 2 se formaram, conforme o gráfico da Figura 5.

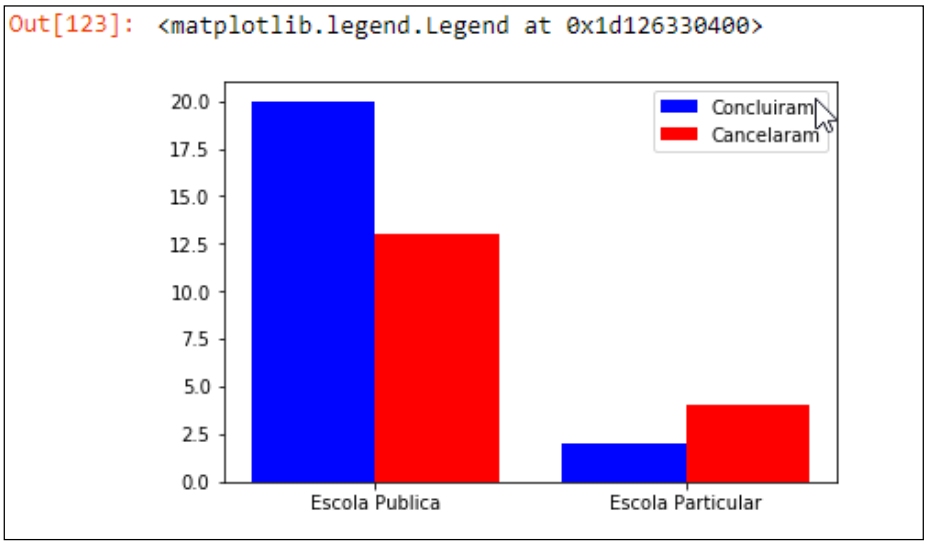

Figura 5 – Alunos de ADS X Escola X concluintes

Foi realizada também análise do índice de reprova das disciplinas do curso de Análise e Desenvolvimento de Sistemas, para cada um dos oito semestres do curso noturno, em cada uma das disciplinas, para os alunos que iniciaram o curso em 2012 ou estavam matriculados em alguma disciplina no período analisado.

As disciplinas que tiveram maior índice de reprova foram as matérias de programação. Uma das disciplinas de programação do terceiro semestre registrou apenas 14 aprovações dos 47 alunos que estavam matriculados.

Foi possível perceber uma diminuição gradativa de matriculados à medida que o aluno avança no curso. O primeiro semestre uma das disciplinas tem 53 alunos matriculados, no último semestre a disciplina com maior número de matriculados tem 27 alunos. Proporcionalmente é o curso que têm o melhor desempenho, pois 56,4% dos alunos concluíram o curso.

5 CONSIDERAÇÕES FINAIS

Os resultados obtidos dão uma dimensão do índice de reprova dos cursos noturnos em geral, mostrando que a maior parte dos alunos não se formam. Também foi possível identificar que, com exceção ao curso de Análise e Desenvolvimento de Sistemas, a maioria dos alunos é proveniente de escolas particulares. Nesse curso as disciplinas técnicas são as que mais reprovam, além disso há uma diminuição gradativa de alunos pelos semestres.

Para finalizar salienta-se que este projeto piloto será a base para outros experimentos que fazem parte da pesquisa sobre análise de big data para dados educacionais. Esse projeto se preocupou em caracterizar a evasão e reprova na Fate Sorocaba, entretanto pretende-se um projeto mais amplo, que contemple diversas Fatecs e que faça uso de dados desestruturado, provenientes de redes sociais, visando a caracterizar:

- Diagnóstico do perfil do aluno;
- Recomendação de novos cursos;
- Diagnóstico do mercado de trabalho e empregabilidade dos cursos tecnológicos;
- Identificação de causas que originam altos índices de evasão e reprova.

REFERÊNCIAS

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.

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.

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