# Minerando Dados do Micro Censo Escolar de 2014 do Brasil

**David de P. Gonçalves, Heron S. Ferreira, Nanci de B. Bonfim**

Departamento de Ciência da Computação – Universidade Federal da Bahia (UFBA)
Av. Adhemar de Barros, s/n – Campus Ondina, 40.170-110, Salvador – BA – Brasil

{davidpinho08,hr.sanches, nancibonfim}@gmail.com

***Resumo.*** *O* Instituto Nacional de Estudos e Pesquisas realiza anualmente um censo escolar no qual são aquisitados dados sobre as diferentes etapas e modalidades de ensino no Brasil. Sabe-se que o *Brasil é um país imenso e possui uma das maiores populações do mundo. Essa grande densidade populacional faz com que os dados do censo escolar sejam considerados como um bigdata. Informações valiosas estão contidas nesse bigdata, mas é necessário um processamento para a extração dessas informações. Neste contexto, faz-se necessário a utilização de conceitos e técnicas de Mineração de Dados, assim como algumas ferramentas para processamento de dados, para que se possa realizar esta tarefa.*

## 1. Introdução

O Censo Escolar é um levantamento de dados estatísticos educacionais de âmbito nacional realizado todos os anos no Brasil e coordenado pelo Inep (Instituto Nacional de Estudos e Pesquisas). Trata-se do principal instrumento de coleta de informações da educação básica, que abrange as diferentes etapas e modalidades: ensino regular (educação infantil e ensinos fundamental e médio), educação especial e educação de jovens e adultos (EJA). Essas informações são valiosas e podem ser utilizadas para traçar um panorama nacional da educação básica e servem de referência para a formulação de políticas públicas e execução de programas na área da educação, incluindo os de transferência de recursos públicos como merenda e transporte escolar, distribuição de livros e uniformes, dentre outros. [1]

Os dados do censo escolar são armazenados e disponibilizados em diferentes arquivos, de diferentes tamanhos, sempre em grande quantidade de dados, e com diversos conteúdos. A tarefa de processar, analisar e fazer correlações entre esses arquivos com enorme quantidade de dados é impraticável sem a utilização de técnicas de mineração de dados. Neste cenário, este trabalho utiliza técnicas de mineração de dados para tentar extrair informações relevantes do Micro Censo Escolar 2014, que possam ajudar a traçar a visão nacional da educação básica brasileira.

O resto deste artigo está organizado da seguinte forma. A seção 2 descreve o pré-processamento dos dados do Micro Censo Escolar de 2014. A seção 3 mostra a mineração dos dados em R, usando o algoritmo Apriori. A seção 4 discute as regras geradas e avaliadas. Por fim, a seção 5 apresenta as considerações finais.

## 2. Pré-processamento dos Dados do Micro Censo Escolar 2014

Os dados do Micro Censo Escolar 2014 [2] abrangem todas as cinco regiões do Brasil (Norte, Nordeste, Centro-oeste, Sudeste e Sul). A base completa possui um tamanho de 12.4 Gigabytes em disco e, por meio dessas informações é possível obter um amplo panorama da educação brasileira. Nesta base de dados estão presentes cerca de 350 variáveis, mais de 50 milhões de matrículas de alunos, mais de 2 milhões de registros de docentes e mais de 190 mil estabelecimentos de ensino. Indubitavelmente, um rico acervo sobre a educação básica brasileira e uma fonte segura e eficaz de obtenção de dados, acessíveis a todos.

Este trabalho, é focado no estudo e no processamento dos dados referentes apenas às regiões Sul e Centro-Oeste do Brasil, totalizando assim 4.1 Gigabytes.

### 2.1. Juntando e Separando os Dados com Spark

Para que seja possível extrair informações relevantes da nossa base de dados, foi necessária a união dos arquivos, que estavam antes separados, em uma base unificada. A forma mais robusta e rápida para juntar estes dados foi a utilização do Spark, uma tecnologia para processamento de dados em larga escala [2].

Cada operação efetuada para juntar as tabelas, gerou uma tabela muito maior, devido ao cruzamento de informações. Devido a algumas características do Spark, foi possível usar o poder computacional de três máquinas distintas.

Para trabalhar de forma distribuída, foi definido o tipo de arquivo no formato Parquet, um formato de colunas, suportado por vários sistemas de processamento de dados. O Spark SQL provê um excelente suporte de leitura e escrita para arquivos parquet, preservando o esquema dos dados originais. Ao unir as tabelas turma, escola e docentes_sul, percebemos que houve um aumento significativo no tamanho do arquivo e, para saber se conseguiríamos trabalhar com toda a base, começamos a trabalhar apenas com essa tabela resultante. Após a realização de alguns testes e a dificuldade de trabalhar com essa base, não unimos as tabelas restantes, matricula_sul, docentes_co e matricula_co.

Através do Spark, além de unir as colunas, foram removidas colunas que tinham apenas um valor, as colunas de nome de professor, escola e códigos únicos, como código da entidade.

### 2.2. Seleção de atributos utilizando o WEKA (CorrelationAttributeEval - WEKA)

Com a base selecionada, foram executados alguns algoritmos de seleção de atributos no Weka. Foi decidido utilizar o CorrelationAttributeEval, que recebe um atributo de referência e calcula para todos os outros um valor de correlação. No nosso caso, o atributo usado como referência foi o código de entidade. Com base nesse valor, a maior correlação foi de pouco mais de 0.68 e com um ponto de corte de 0.15, a base foi diminuída em 40%.

### 2.3. Discretização e binarização dos dados para executar

Para que a execução dos algoritmos de regras de associação, Apriori e Fp-Growth, que serão explicados e usados mais a frente, acontecesse da maneira correta foi necessária uma série de tratamentos nos dados, pois estes algoritmos trabalham com valores binários nos atributos [3]. Para que esses tratamentos acontecessem foram necessários: estudar os casos de valores nulos e perceber que seria válido atribuir o valor zero para todos esses casos; remover linhas iguais entre si; discretização das colunas usando o algoritmo do WEKA; binarizar os dados, pois dessa forma, todas as colunas, já discretizadas, só teriam valores 0 e 1; remover as colunas com valores não significativos após o processo de atribuição de 0, discretização e binarização.

Os processos acima foram executados para 108 atributos, que geraram mais de 900 colunas.

## 3. Mineração dos Dados: R e Apriori

Com a base preparada, iniciou-se o processo para identificar regras de associação utilizando os algoritmos Apriori e FP-Growth. Diversas ferramentas como o WEKA, Spark, R e o Rapidminer foram testadas para a execução desta tarefa. Encontrou-se uma certa dificuldade para escrever os algoritmos no Spark e por conta disso focou-se nas outras ferramentas. O Weka no modo Explorer, o R e o Rapidminer precisam da base em memória e trabalham com um único computador.

O weka não suportou trabalhar com a base, logo não se obteu sucesso. O rapidminer não suportou trabalhar com todas as linhas, então foram gerados subsets para conseguir extrair resultados. Nele rodou-se o algoritmo FP-Growth, que gerou regras que ficaram confusas de serem entendidas. Algumas regras relacionavam o mesmo atributo com valores diferentes.

Por fim, o R conseguiu analisar as mais de 4 milhões de linhas utilizando o algoritmo Apriori com um tempo excelente. O carregamento do arquivo demorou cerca de 15 minutos e o Apriori foi executado em 1 minuto e 30 segundos. Foram geradas 2057 regras que estão disponíveis em [4].

Na próxima sessão são apresentadas as regras as quais conseguimos extrair semântica.

## 4. Avaliação dos Resultados Encontrados

### 4.1. Regras Geradas e Avaliadas

Na seção anterior, foi comentado que 2057 foram geradas pelo R após a execução do Apriori. Estas regras foram estudadas e analisadas, e baseado nos valores de Support, Confidence e Lift, 10 regras foram selecionadas. A tabela 1 mostra as regras e os valores das medidas de avaliação.

**Tabela 1. Regras Selecionadas**

| ID | Regra | Support | Confidence | Lift |
|---|---|---|---|---|
| 1 | {fk_cod_tipo_turma_0,id_dia_semana_sexta_1,id_educacao_fisica_1,id_historia_1,id_lingua_literat_portuguesa_1,id_matematica_1} => {id_geografia_1} | 0.40092 | 0.9771625 | 2.2399 |
| 2 | {fk_cod_tipo_turma_0,id_dia_semana_sexta_1,id_educacao_fisica_1,id_geografia_1,id_lingua_literat_portuguesa_1,id_matematica_1} => {id_historia_1} | 0.40092 | 0.9841149 | 2.2378 |
| 3 | {fk_cod_tipo_turma_0,id_dia_semana_sexta_1,id_educacao_fisica_1,id_geografia_1,id_historia_1,id_lingua_literat_portuguesa_1} => {id_matematica_1} | 0.40092 | 0.995414 | 2.0107 |
| 4 | {fk_cod_mod_ensino_1,id_cozinha_1,id_dia_semana_sexta_1,id_educacao_fisica_1,id_matematica_1} => {id_lingua_literat_portuguesa_1} | 0.40819 | 0.9938599 | 2.0130 |
| 5 | {fk_cod_mod_ensino_1,id_cozinha_1,id_dia_semana_sexta_1,id_lingua_literat_portuguesa_1,id_matematica_1} => {id_educacao_fisica_1} | 0.4082 | 0.9812279 | 2.0808 |
| 6 | {fk_cod_tipo_turma_0,id_dia_semana_sexta_1,id_educacao_fisica_1,id_laboratorio_informatica_1,id_lingua_literat_portuguesa_1,id_matematica_1} => {fk_cod_mod_ensino_1} | 0.40055 | 0.930650 | 1.5866 |
| 7 | {id_alimentacao_1,id_cozinha_1,id_mod_ativ_complementar_1,id_sala_atendimento_especial_1} => {id_aee_1} | 0.41464 | 0.9464141 | 1.6690 |
| 8 | {fk_cod_mod_ensino_1,id_alimentacao_1,id_patio_descoberto_1} => {fk_cod_tipo_turma_0} | 0.40638 | 0.9998266 | 1.2828 |
| 9 | {id_dependencias_pne_1,id_laboratorio_informatica_1,id_patio_descoberto_1} => {id_sanitario_pne_1} | 0.41385 | 0.896638 | 1.3041 |
| 10 | {fk_cod_mod_ensino_1,id_educacao_fisica_1,id_laboratorio_informatica_1,id_lingua_literat_portuguesa_1} => {id_dia_semana_sexta_1} | 0.40790 | 0.99975 | 1.2828 |

A extração semântica das regras listadas na Tabela 1 é descrita abaixo na tabela 2, de acordo com o id de cada regra.

**Tabela 2. Extração Semântica das Regras**

| ID | Semântica |
|---|---|
| 1 | Turmas do tipo de atendimento normal que têm na sexta-feira as disciplinas Educação Física, História, Literatura portuguesa e Matemática, também têm Geografia. |
| 2 | Turmas do tipo de atendimento normal que têm na sexta-feira as disciplinas Educação Física, Geografia, Literatura portuguesa e Matemática, também têm História. |
| 3 | Turmas do tipo de atendimento normal que têm na sexta-feira as disciplinas Educação Física, Geografia, História e Literatura Portuguesa, também têm Matemática. |
| 4 | Turmas do Ensino Regular que têm na sexta-feira as disciplinas Educação Física e Matemática e, têm cozinha, também têm Literatura Portuguesa. |
| 5 | Turmas do Ensino Regular que têm na sexta-feira as disciplinas Literatura Portuguesa e Matemática e, têm cozinha, também têm Educação Física. |
| 6 | Turmas do tipo de atendimento normal que têm aulas na sexta-feira e têm as disciplinas de Educação Física, Literatura Portuguesa e Matemática, em escolas que possuem laboratório de informática, também possuem um modelo de ensino regular. |
| 7 | Escolas que possuem cozinha e sala de atendimento especial e oferecem alimentação para seus alunos e o modelo das atividades complementares não é exclusiva, também possuem atendimento escolar especializado e normal. |
| 8 | Turmas com modelo de ensino regular em escolas que possuem patio descoberto e alimentação para os alunos, também tem um tipo de atendimento normal. |
| 9 | Escolas que possuem depêndencias para alunos portadores de necessidades especiais e laboratórios de informática e patio descoberto, também apresentam sanitário especial para portadores de necessidades especiais. |
| 10 | Turmas das disciplinas de Educação Física e Literatura Portuguesa com modelo de ensino regular em escolas que possuem laboratório de informática, também têm aulas na sexta-feira. |

## 5. Considerações Finais

A seleção de atributos foi feita com um atributo que tem uma distribuição muito grande e a equipe percebeu isso muito tarde. Não foi possível no andamento que estávamos gerar outro subconjunto para aplicar todos os passos que já tinhamos feitos. Como esse conjunto era 40% da base, decidimos continuar trabalhando por acreditar ser suficiente para encontrar relações interessantes o suficiente.

Usamos várias ferramentas diferentes durante este trabalho por cada uma ter um algo melhor para ser usado em um processo específico. Acreditamos que o processamento de big data ainda tem muito a crescer e que as ferramentas devem evoluir para criar um ambiente mais homogêneo. O weka, por exemplo, ferramenta de referência, tem alguns algoritmos que trabalham usando múltiplas threads. Contudo, não são todos e isso limitou um pouco o desempenho. A integração com o Spark via WekaDistributed também possui várias restrições, sendo difícil a sua utilização em sistema distribuídos.

## Referências


[1] Portal Microdados do Censo Escolar. Disponível em: http://dados.gov.br/dataset/microdados-do-censo-escolar. Acessado em 23 de Novembro de 2015.

[2] Base Microdados do Censo Escolar 2014. Disponível em: http://download.inep.gov.br/microdados/micro_censo_escolar_2014.zip. Acessado em 23 de Novembro de 2015.

[3] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[4] David de Pinho, Heron Sanches, Nanci Bonfim. Repositório com as regras geradas pelo R em cima dos dados do Micro Censo Escolar 2014. Disponível em: https://github.com/heronsanches/micro_censo_escolar_2014. Acessado em: 24 de Novembro de 2015.