# Identificando Correlações Entre Bases de Dados Educacionais

**Geraldo Cruz Júnior[1,3], Rafaella Nascimento[2], Roberta Macêdo[1], Gabriel Alves[1]**

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco
Recife – PE – Brasil

[2]Escola Politécnica – Universidade de Pernambuco
Recife – PE – Brasil

[3]Instituto SENAI de Inovação para Tecnologias da Informação e Comunicação – SENAI
Recife – PE – Brasil

{geraldoj8,rafaellalsn,robertammg,gaaj1980}@gmail.com

***Abstract.*** *The availability of public data makes it possible to identify relevant information through methods of knowledge discovery. This information can be useful for society to reflect or even create applications for the improvement of services available to citizens. This work uses an application that, through statistical analysis techniques, calculates correlation coefficients and outliers between information extracted from the Brazilian School Census and the ENEM of 2014, in the state of Pernambuco. For the scenarios evaluated, the results obtained indicate a high, moderate or low correlation between the presence of school characteristics and the students' performance in the ENEM tests.*

***Resumo.*** *A disponibilização de dados públicos possibilita que, através de métodos de descoberta de conhecimento, informações relevantes sejam identificadas. Estas informações podem ser úteis para que a sociedade reflita ou até mesmo crie aplicações para o melhoramento dos serviços disponíveis aos cidadãos. Este trabalho utiliza uma aplicação que através de técnicas de análises estatísticas, calcula coeficientes de correlação e outliers entre informações extraídas de bases do Censo Escolar brasileiro e do ENEM de 2014, no âmbito do estado de Pernambuco. Para os cenários avaliados, os resultados obtidos apontam um alto, moderado ou baixo índice de correlação entre a presença de características das escolas e o desempenho dos alunos nas provas do ENEM.*

## 1. Introdução

Dados Abertos são dados que podem ser livremente utilizados, reutilizados e redistribuídos por qualquer pessoa, contendo informações relativas a diferentes áreas do conhecimento [Data 2014]. Diversos governos pelo mundo estão disponibilizando dados relativos ás suas gestões, para que a população possa ter transparência e controle social em torno dos acontecimentos de seu país [Agune et al. 2010] [Palazzi and Tygel 2014]. Análises discretas e contínuas de Dados Abertos Governamentais (DAG) possibilitam a descoberta de informações em diferentes pilares sociais, como saúde e educação, permitindo uma previsibilidade de acontecimentos para tomadas de decisão.

Como citado anteriormente, dentre os dados abertos existem os dados educacionais, que possibilitam descobertas de conhecimentos no cenário da educação através da

DOI: 10.5753/cbie.wcbie.2017.694 

Mineração de Dados Educacionais (MDE) [Nascimento 2016]. Os dados abertos educacionais podem se referir especificamente aos dados abertos oriundos das instituições educacionais. Estes se referem a administração de instituições, dados do curso e gerados por usuários (como análise de aprendizagem e desempenho). A MDE corresponde a uma área multidisciplinar, que utiliza técnicas de Mineração de Dados (MD) em contextos educacionais como, por exemplo, ambientes virtuais de ensino e aprendizagem, sendo possível inferir sobre perfis de estudantes, professores e tutores, assim como relações entre variáveis de aprendizado [Silva and Silva 2014].

Os conjuntos de dados educacionais são de interesse para uma grande variedade de pessoas, incluindo educadores, alunos, instituições, governo, pais e o público em geral [Guy 2016]. Com a evolução das técnicas de MDE, grandes volumes de dados podem ser armazenados e analisados visando à implantação de um fluxo de *Data-Driven Decision Making* (DDDM) eficiente para que professores, estudantes e gestores possam tomar decisões seguras para processos de melhorias no âmbito educacional [Silva 2015].

Tendo em vista os interesses presentes no cenário educacional, a MDE pode ser interpretada como um processo onde o objetivo não é apenas transformar os dados em conhecimento, mas também filtrar o conhecimento para ajudar na tomada de decisões sobre como modificar o ambiente educacional para melhorar a aprendizagem dos alunos [Romero and Ventura 2013]. A MDE busca respostas para perguntas específicas da educação, relacionadas com processos de aprendizagem, desenvolvimento de materiais instrucionais, acompanhamento e previsões, entre outros, a partir da obtenção de informações e padrões de comportamento importantes para apoiar determinadas práticas pedagógicas [Baker et al. 2011].

Uma forma de obter informações relevantes a partir de dados é através de um processo de MD, no caso, dados educacionais, aplicando-os à estatística probabilística, como é o caso de cálculos de coeficientes de correlação. A correlação indica a força e a direção do relacionamento linear, ou não linear, entre duas variáveis aleatórias. Neste sentido geral, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados [Triola Mário et al. 2005]. No contexto da MDE pode-se utilizar a correlação para relacionar acontecimentos de uma escola, por exemplo, buscando formas de melhorar o ensino e prever evasões.

A identificação de *outliers*, que podem ser definidos como valores aberrantes ou valores atípicos, é outro método que pode ser importante na análise de dados educacionais. Considerando-se os quartis superiores (valor a partir do qual se encontram 25% dos valores mais elevados) e quartis inferiores (valor aos 25% da amostra ordenada), definidos pelo *box-plot* (gráfico utilizado para localizar e analisar a variação de uma variável dentre diferentes grupos de dados), é possível apontar dados discrepantes em uma amostra [Gladwell 2008]. Logo pode-se identificar grandes desigualdade educacionais por regiões, por exemplo.

Desta forma, este trabalho tem como objetivo adaptar o *framework* de análise de dados desenvolvido por Cruz-Junior (2016) para o desenvolvimento de uma aplicação de mineração e identificação de correlações e *outliers* entre bases de dados educacionais provenientes do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), especificamente as bases de dados do Exame Nacional do Ensino Médio (ENEM)

e do Censo Escolar. A metodologia aplicada no desenvolvimento deste trabalho consiste nas fases do projeto KDD [Fayyad et al. 1996]. Ao final, a proposta visa a descoberta de possíveis correlações entre variáveis das bases de dados utilizadas, cruzando e identificando características das escolas com o desempenho de seus alunos no ENEM, no âmbito do estado de Pernambuco.

## 2. Trabalhos Relacionados

A MDE tem crescido nos últimos anos, e os conhecimentos podem servir de subsídio para, por exemplo, a melhoria das práticas em educação presencial e a distância, a diminuição da evasão escolar, além de ser uma importante ferramenta para viabilizar a qualificação e melhoria do ensino como um todo.

Nesta temática, o trabalho apresentado por [Baker et al. 2011] fornece uma revisão das pesquisas realizadas na área, dando ênfase aos métodos e aplicações que veem influenciando a pesquisa e a prática da educação em vários países. Também, é abordado o impacto da MDE na melhoria de cursos na modalidade educação à distância (EaD), que vem recebendo incentivo governamental além de ser crescente o número de alunos que procuram esta forma de ensino.

Já o trabalho apresentado por [Gomes 2015], relata os métodos de MD aplicados com a finalidade de descobrir regras de associação em dados estatísticos educacionais provenientes do ENEM nos anos de 2013 e 2014, no âmbito da região Nordeste. Neste trabalho, seguiu-se a metodologia de descoberta de conhecimento em banco de dados e buscou-se verificar se o algoritmo utilizado na mineração aponta correlações entre a renda familiar e o desempenho do estudante no ENEM. As regras encontradas reforçam a forte relação entre a renda da família e o desempenho dos estudantes, sobretudo, provenientes de escolas públicas.

Na área da educação, o trabalho apresentado por [Gardinal and Marturano 2007] tem como objetivo investigar associações entre comportamento e desempenho do aluno, levando em conta o sexo da criança. A escala avaliada levou em conta a relação do aluno com a tarefa, com os colegas e com o professor. O desempenho foi avaliado por meio de uma sondagem de leitura e escrita. Como principais resultados, obteve-se que para ambos os sexos, comportamentos nos três domínios – relação com a tarefa, com os colegas e o professor, correlacionaram-se com medidas de desempenho. O desempenho escolar foi mais fortemente associado aos comportamentos interpessoais no grupo masculino.

As avaliações dos alunos para medir a eficácia da educação dos instrutores são frequentemente aplicadas no ensino superior por muitos anos. O trabalho de [Mardikyan and Badur 2011] investiga os fatores associados à avaliação do desempenho dos professores usando duas técnicas diferentes de MD: Regressão gradual e árvores de decisão. Os resultados mostram que um fator que resume as perguntas relacionadas ao instrutor no formulário de avaliação, o *status* do trabalho do instrutor, a carga de trabalho do curso, a frequência do aluno e a porcentagem de alunos que completam o formulário são dimensões significativas do desempenho do professor.

O trabalho de [Adeodato et al. 2014] aplica MD para avaliar a qualidade da educação secundária privada no Brasil, a partir das bases do ENEM e do Censo Escolar. Utiliza-se as seguintes técnicas de aprendizagem de máquina: regressão logística, árvore

de decisão e indução. A primeira gerou o escore de propensão ao sucesso, enquanto que a segunda expôs a decisão sequencial humana ideal, e a indução gerou as regras para apoiar a decisão baseada no escore. Os autores utilizaram as métricas *auc_roc* e *max_ks* para avaliar o desempenho do escore de propensão, já a cobertura, confiança e *lift*, mediram a qualidade das regras. Os resultados mostraram que a abordagem *domain-driven data mining* teve muito sucesso na resolução do problema e na validação de políticas públicas.

Outra forma de gerar informações que pode ser agregada à MD é a estatística probabilística, como exemplo, pode-se calcular coeficientes de correlação entre dados e identificar pontos discrepantes, os *outliers*. No trabalho apresentado por [Cruz-Júnior 2016] é desenvolvido um *framework* para análise e monitoramento de dados abertos governamentais, que possui como base algoritmos de inferências estatísticas, correlações, *outliers* e análises de sentimentos. Os resultados gerados são disponibilizados através de gráficos, tabelas e mapas interativos. Com o *framework* busca-se encontrar informações relevantes à saúde, infraestrutura, corrupção e comportamentos sociais, entre outros.

Tendo em vista os trabalhos apresentados, é notável o emergente desenvolvimento de trabalhos relacionados a área educacional, com a preocupação em entender e melhorar o processo de ensino-aprendizagem. Desta forma, este trabalho apresenta como contribuição para a comunidade, uma pesquisa focada em entender se características específicas de escolas, como a presença de quadras, laboratórios e bibliotecas, possuem correlação com o desempenho dos alunos no ENEM.

## 3. Metodologia

A Descoberta de Conhecimento em Banco de Dados, do inglês *Knowledge Discovery in Databases* (KDD), é uma técnica que possibilita analisar dados para gerar descobertas que podem ajudar na tomada de decisão, otimizando os processos e retornando de forma eficiente a informação para que se possa definir a estratégia mais adequada para se aplicar em determinado cenário. As fases deste processo são as seguintes: seleção dos dados, pré-processamento dos dados, transformação, MD e análise [Fayyad et al. 1996]. As seguintes subseções mostram o detalhamento de tais fases aplicadas ao trabalho.

### 3.1. Seleção

Busca-se correlacionar aspectos de infraestrutura das escolas com o desempenho dos alunos no ENEM. As variáveis selecionadas para a análise em ambas as bases são mostradas na Tabela 1. As variáveis presentes na base do Censo Escolar consistem na presença ou não de determinada característica, enquanto as variáveis da base do ENEM consistem na nota obtida para determinada área do conhecimento.

### 3.2. Limpeza e pré-processamento

Com as variáveis selecionadas, é realizada a redução de dimensionalidade vertical da base, na qual exclui-se as colunas indesejadas para o estudo. Desta forma, 9 colunas são mantidas para a base do Censo Escolar e 5 colunas para a base do ENEM, como mostra a Tabela 1. Após isto, foi possível observar nas bases valores ausentes para algumas variáveis, que poderiam interferir na posterior aplicação da técnica de MD. Foi realizada a análise das três alternativas básicas para solucionar este problema: imputação de dados, substituir o valor faltante pela média e excluir o registro.

**Tabela 1. Variáveis selecionadas para análises.**

| ENEM | Censo Escolar |
|---|---|
| Ciências Humanas<br>Ciências da Natureza<br>Linguagens e suas tecnologias<br>Matemática<br>Redação | Água filtrada<br>Água de rede pública<br>Energia elétrica<br>Sistema de esgoto<br>Coleta de lixo<br>Laboratório de Informática<br>Laboratório de Ciências<br>Bibliotecas<br>Salas de Leitura |

As técnicas de imputação fazem a previsão dos dados ausentes. Esta solução é eficiente para um conjunto pequeno de dados, onde pode ser feita pelo especialista no domínio de forma manual. Caso contrário, requer *softwares* para fazer o preenchimento automático que podem não ser muito precisos. Por estas questões esta técnica não foi realizada neste trabalho. Para substituir o valor faltante pela média o campo precisa ser numérico, neste problema, os campos selecionados são categóricos na base do Censo Escolar. Desta forma, a solução para os *missing values* consistiu em excluir os registros com dados faltantes. Além disto não consistiam numa quantidade significativa em relação a base original e elimina o risco da análise ser feita com dados não reais.

### 3.3. Parametrização da fase de Mineração de Dados

A MD é a principal fase do processo KDD. É uma forma de explorar e analisar bases de dados, buscando extrair conhecimentos como regras, padrões e desvios [Kampff 2009]. MD é um passo do processo KDD que consiste em aplicar análise de dados e algoritmos de descoberta que produzem uma enumeração particular de padrões (ou modelos) sobre os dados [Fayyad et al. 1996]. A escolha da técnica de mineração mais adequada depende de aspectos como a área do problema e dos dados disponíveis.

Muitas vezes é preciso conhecer a forma como duas ou mais variáveis estão relacionadas. Existem diversos critérios de avaliação dessa relação, alguns próprios para as que seguem uma distribuição normal e outros para as que não seguem uma distribuição teórica conhecida. Basicamente, existem métodos de avaliação da relação para variáveis contínuas e categóricas [Guimarães 2008]. Neste sentido geral, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados. A correlação indica a força e a direção do relacionamento linear, ou não linear, entre duas variáveis aleatórias.

Para estimar a correlação de dois conjuntos que não têm distribuição conjunta normal bivariada, a alternativa mais usual é o coeficiente de correlação de Spearman. A correlação de Spearman (ou $rho$) é um cálculo estatístico baseado em postos e foi introduzido por Spearman em 1904 que exige apenas que as variáveis X e Y sejam medidas pelo menos em escala ordinal. Este coeficiente é uma medida de correlação não-paramétrica, isto é, ele avalia uma função monótona arbitrária que pode ser a descrição da relação entre duas variáveis, sem fazer nenhuma suposição sobre a distribuição de frequência. Ao contrário do coeficiente de correlação de Pearson, não requer a suposição de que a relação entre as variáveis é linear, nem que elas sejam medidas em intervalo de classe, podendo ser usado para os conjuntos medidos no nível ordinal.

De acordo com [Siegel and Castellan Jr 1975], o coeficiente de correlação por postos de Spearman, designado e representado por $rs$ ou $rho$, é uma medida não paramétrica que permite estabelecer a existência de correlação entre duas variáveis. Para o cálculo de $rs$ considera-se um conjunto de $n$ pares de observações. Em correspondência a cada par, consta-se seu posto em relação uma determinada variável, $Ij$, o índice ambiental e seu posto em relação à outra variável, neste caso, $Yij$, que em geral é representada pela produção do $i$-ésimo genótipo no $j$-ésimo ambiente. Em seguida, determinam-se os valores da diferença, denotada por $di$, entre os pares de postos. Eleva-se cada valor de $di$ ao quadrado e soma-se, obtendo-se assim a soma de quadrados da diferença entre os pares de postos, como pode ser visto no somatório a seguir.

$$\sum_{i}^{n=1} d_i^2 := d_1^2 + d_2^2 + ... + d_n^2 \tag{1}$$

O coeficiente de correlação de Spearman é obtido por meio da equação 2:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1))} \tag{2}$$

Para Cohen (1988) [Cohen 1988], escores de correlação entre 0,10 e 0,29 podem ser considerados pequenos; escores entre 0,30 e 0,49 podem ser considerados como médios; e valores entre 0,50 e 1 podem ser interpretados como grandes. Estes intervalos foram utilizados neste trabalho para definir se os resultados obtidos apresentavam altos, médios ou baixos indicativos de correlação.

Outro método que pode ser utilizado em análises de bases de dados para a extração de informações é a identificação de *outliers*. As observações que apresentam um grande afastamento das restantes, ou são inconsistentes com elas, são habitualmente designadas por *outliers*. Pode-se concluir que um *outlier* é caracterizado pela sua relação com as demais observações que fazem parte da amostra. O seu distanciamento em relação a essas observações é fundamental para se fazer a sua caracterização. Estas observações são também designadas por observações "anormais", contaminantes, estranhas ou aberrantes [Figueira 1998].

Considerando que a amostra esteja ordenada e que se tenha os valores Q1 (Primeiro quartil) e Q3 (Terceiro quartil), deve-se calcular o intervalo inter-quartil, ou *Interquartile Range* (IQR), com ele determinar o *Lower Outlier Boundary* (LOB) e o *Upper Outlier Boundary* (UOB), conforme mostram as equações 3, 4 e 5 abaixo.

$$(IQR) = Q3 - Q1 \tag{3}$$

$$LOB = Q1 - (1.5 * IQR) \tag{4}$$

$$UOB = Q3 + (1.5 * IQR) \tag{5}$$

A operação dos cálculos de LOB e de UOB apontam as barreiras de investigação dos *outliers*, ou seja, qualquer valor que for identificado na amostra que não esteja entre os valores encontrados pode ser apontado como um dado discrepante. Quando o fator de

multiplicação utilizado é o 1,5 se encontram as barreiras internas de investigação, caso se deseje aumentar a aceitação a erros pode-se multiplicar a amplitude interquartílica por 3 [McGill et al. 1978], determinando um intervalo onde qualquer valor identificado fora dele é considerado um *outlier* extremo.

Para a elaboração deste trabalho utilizou-se o *framework* multiplataforma de Cruz-Júnior (2016), para análise e monitoramento de dados governamentais. A aplicação desenvolvida com base no *framework* implementa os cálculos estatísticos abordados. Alguns dos resultados obtidos são ilustrados e discutidos na próxima seção. As significâncias dos resultados obtidos foram identificadas pelo cálculo do valor p ou *p-value*.

## 4. Resultados

As saídas geradas a partir da aplicação desenvolvida são resultantes dos cruzamentos entre a base de dados do ENEM de 2014 e a base de dados do Censo Escolar de 2014. Foram obtidas correlações, *outliers* e análises estatísticas. Com as análises, foi possível verificar informações relevantes para o cenário educacional no estado de Pernambuco, os resultados de algumas destas inferências serão discutidos neste capítulo.

Ao se correlacionar as notas obtidas para cada área de conhecimento da base de dados do ENEM percebe-se indicativos de correlação sempre superiores com $rho$ = 0,7, indicando que independente da área de conhecimento, ciências da natureza, ciências humanas, linguagens e suas tecnologias, matemática ou redação, se a nota de um aluno em uma determinada área tende a aumentar, em uma outra área também tende a aumentar. Removendo as notas da redação da análise o coeficiente sobe para $rho$ = 0,9, ambos indicando um alto índice de correlação para esta primeira análise.

Foi identificado que na base do Censo Escolar, as escolas que possuíam água filtrada também possuíam água da rede pública, energia elétrica, sistema de esgoto e coleta de lixo. As análises quanto a estes aspectos foram realizadas em um só agrupamento. Ao se relacionar as escolas que possuíam, ou não, estas características com as médias gerais das notas que os alunos destas obtiveram no ENEM, obteve-se um $rho$ = 0,276 com um *p-value* = 0,000148. Percebe-se que a análise aponta que estatisticamente há um fraco indício de que a presença ou não destas características de infraestrutura influenciem no desempenho dos alunos no ENEM.

Relacionando-se as escolas que contam com laboratórios de informática com o desempenho geral de seus alunos no ENEM, percebe-se um $rho$ = 0,478. Já ao relacionar laboratório de informática com o resultado na prova de Ciências da Natureza e Matemática, obtém-se, respectivamente, $rho$ = 0,453 e $rho$ = 0,432. Todos os resultados apontam para um indicativo moderado de correlação entre a presença de laboratórios de informática e a melhoria nas notas.

Uma análise semelhante foi realizada analisando-se a presença de laboratório de ciências com a média geral da prova, em Ciências da Natureza e em Matemática. Respectivamente obteve-se $rho$ = 0,553 (*p-value*= 0,000526), $rho$ = 0,504 (*p-value*=0,005442) e $rho$ = 0,507 (*p-value*=0,005575). Ambos os resultados indicam um alto índice de correlação entre a presença de laboratórios de ciência nas escolas e a melhoria das notas dos alunos no ENEM nestas 3 áreas.

A análise da presença de bibliotecas relacionada com a nota geral e a nota de

Linguagens e Tecnologias, resulta em um indicativo de correlação baixo de $rho$ = 0,244 (*p-value*=0,000853) e $rho$ = 0,289 (*p-value*=0,000316), respectivamente. Porém, quando a presença de bibliotecas nas escolas é comparada com a nota da redação, percebe-se uma correlação moderada com $rho$ = 0,362 (*p-value*=0,000213).

Ao se analisar a presença de salas de leitura com as média geral da prova do ENEM, a nota de Redação e a de Linguagens e Tecnologias, respectivamente, obteve-se altos indicativos de correlação, $rho$ = 0,615 (*p-value*=0,000013), $rho$ = 0,557 (*p-value*=0,000467) e $rho$ = 0,557 (p-value=0,000423), respectivamente. Percebe-se que a presença de salas de leitura nas escolas indica possíveis melhorias no desempenho dos alunos nas provas do ENEM para estas 3 áreas.

Por fim, buscou-se identificar *outliers*, ou seja, discrepâncias entre as notas e as características de infraestrutura levando em consideração as escolas, as suas regiões e as áreas de conhecimento do ENEM. Para o cenário analisado neste trabalho não foram identificados *outliers*, ou seja, considerando uma análise de variância, as escolas e notas inferidas apresentam características semelhantes entre si.

Maiores estudos são necessários para determinar a exata causa destes relacionamentos identificados. Também é importante frisar que os altos valores de *p-value* não indicam necessariamente baixa significância quando se trata da correlação de Spearman, como pode ser visto em [Siegel and Castellan Jr 1975], [Bauer 2007] e [Pontes 2010].

## 5. Conclusões

Este estudo contribui para a área da Educação no que diz respeito a entender características de infraestrutura presentes nas escolas que podem estar relacionadas com o desempenho dos alunos nas provas do ENEM. Para medir esta relação, são utilizados modelos estatísticos como cálculo de correlação entre variáveis e identificação de *outliers* presentes nas bases de dados. Tais modelos tem como objetivo oferecer subsídios que ajudem a entender o cenário educacional.

Pode-se verificar índices baixos, moderados ou altos de correlações através de um mapeamento entre as bases do Censo Escolar e do ENEM. Como principais resultados pode-se observar que a presença de atributos como laboratórios de computação e ciências, salas de leitura e biblioteca nas escolas indicam um melhor desempenho dos alunos nas provas de Ciências da Natureza e Matemática, Linguagens e Tecnologias e Redação, respectivamente. O incentivo à melhoria na estrutura das escolas com estas estruturas pode ser um caminho para que seja fornecido aos estudantes uma melhor forma de aprendizagem, através de atividades mais interativas e práticas.

Utilizar a MDE possibilita a identificação prévia de aspectos que podem precisar de melhorias e investimentos mais adequados, melhorando assim o ensino-aprendizagem, mitigando problemas e melhorando índices de desempenho dos alunos, como o medido nas provas do ENEM. Medir a correlação entre variáveis não é o bastante para determinar as dependências no mundo real, uma vez que outros fatores influenciam estes índices e precisam ser estudados. No entanto, este estudo abre caminhos para trabalhos seguintes, uma vez que revela indícios de relações entre variáveis.

Tendo em vista a representatividade destes resultados, pretende-se expandir o agrupamento da pesquisa para cidades de todo o Brasil, e não só para as cidades do

estado de Pernambuco. Outras variáveis como a identificação de localização (zona rural ou urbana) e dependência administrativa da escola (municipal, privada, estadual ou federal) podem ser incluídas no estudo para enriquecer os resultados da pesquisa. Estes dados podem ser obtidos do mesmo Portal de Dados Abertos utilizado neste trabalho, o do INEP. Com a expansão do cenário analisado, espera-se que a identificação de *outliers* possa fornecer resultados interessantes assim como as correlações, trazendo coeficientes embasados em escala nacional.

## Referências

Adeodato, P. J., Santos Filho, M. M., and Rodrigues, R. L. (2014). Predição de desempenho de escolas privadas usando o enem como indicador de qualidade escolar. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 25, page 891.

Agune, R. M., Gregorio Filho, A. S., and Bolliger, S. P. (2010). Governo aberto sp: disponibilização de bases de dados e informações em formato aberto. in: Congresso consad de gestão pública.

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o brasil. *Brazilian Journal of Computers in Education*, 19(02):03.

Bauer, L. (2007). Estimação do coeficiente de correlação de spearman ponderado. *Dissertação de Mestrado em Epidemiologia. Faculdade de Medicina, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil*.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. *2nd ed.Hillsdale, New Jersey: L. [S.l.]: Erlbaum*.

Cruz-Júnior, G. G. (2016). Um framework multiplataforma para análise e monitoramento de dados governamentais.monografia, universidade federal rural de pernambuco.

Data, O. G. (2014). Eight principles of open government data. *Available in: http://resource.org/8_ principles.html. Accessed in: May 15, 2017*.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

Figueira, M. M. C. (1998). Identificação de outliers. *Millenium*.

Gardinal, E. C. and Marturano, E. M. (2007). Meninos e meninas na educação infantil: Associação entre comportamento e desempenho. *Psicologia em estudo*, 12(3).

Gladwell, M. (2008). Fora de série: Outliers. *Rio de Janeiro: Sextante*.

Gomes, T. C. S. (2015). Descoberta de conhecimento utilizando mineração de dados educacionais abertos. monografia, universidade federal rural de pernambuco.

Guimarães, P. R. B. (2008). Métodos quantitativos estatísticos.

Guy, M. (2016). The open education working group: Bringing people, projects and data together. In *Open Data for Education*, pages 166–187. Springer.

Kampff, A. J. C. (2009). Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente.tese de doutorado, universidade federal do rio grande do sul, porto alegre – rs.

Mardikyan, S. and Badur, B. (2011). Analyzing teaching performance of instructors using data mining techniques. *Informatics in Education*, 10(2):245.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, 32(1):12–16.

Nascimento, R. L. S. (2016). Mineração de dados educacionais e visualização de informações geográficas utilizando mapas de calor. monografia, universidade federal rural de pernambuco.

Palazzi, D. and Tygel, A. (2014). Visualização de dados estatísticos representados como dados abertos ligados.

Pontes, A. C. F. (2010). *Ensino da correlação de postos no ensino médio*. Simpósio Nacional de Probabilidade e Estatística (SINAPE), v. 19, p. 26–30.

Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.

Siegel, S. and Castellan Jr, N. J. (1975). *Estatística não-paramétrica para ciências do comportamento*. Artmed Editora.

Silva, L. (2015). Tomada de decisão baseada em dados (dddm) e aplicações em informática em educação. *Jornada de Atualização em Informática na Educação*.

Silva, L. A. and Silva, L. (2014). Fundamentos de mineração de dados educacionais. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 3.

Triola Mário, F. et al. (2005). Introdução à estatística. *Rio de Janeiro: LTC*.