Disponível em: http://periodicos.ufes.br/BJPE/index

ISSN: 2447-5580

Brazilian Journal of Production Engineering

BJPE - Revista Brasileira de Engenharia de Produção

Campus São Mateus
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

ARTIGO ORIGINAL OPEN ACCESS

MINERAÇÃO DE DADOS EDUCACIONAIS NA BASE DE DADOS DO ENEM 2015

EDUCATIONAL DATA MINING ON ENEM 2015 DATABASE

Guilherme Ferrari Bravin[1], Luciana Lee[2] & Silvia das Dores Rissino[3]

[1 2 3] Centro Universitário Norte do Espírito Santo da Universidade Federal do Espírito Santo, Rodovia BR 101 Norte, Km. 60, Bairro Litorâneo, CEP 29932-540, São Mateus.

[1] guilhermefbravin@gmail.com [2] luciana.lee@ufes.br [3] silvia.rissino@ufes.br

ARTIGO INFO.

Recebido em: 27.08.2019

Aprovado em: 15.09.2019

Disponibilizado em: 20.09.2019

PALAVRAS-CHAVE:

Descoberta de Conhecimento; ENEM 2015; Mineração de Dados; Classificação; Regressão Linear.

KEYWORDS:

Knowledge Discovery; ENEM 2015; Data Mining; Classification; Linear Regression

***Autor Correspondente**: Rissino, S. das D.

RESUMO

Este trabalho aplica o processo de descoberta de conhecimento em base de dados (KDD) no conjunto de dados abertos do ENEM por escola no ano de 2015, com o objetivo de encontrar relações entre os indicadores contextuais presentes na base de dados e as notas médias nas diferentes áreas de conhecimento avaliadas pelo exame. No pré-processamento os dados são adequados e filtrados, com o Microsoft Excel e o software R, para serem utilizados na etapa seguinte. Na fase de mineração de dados utiliza-se o software R para a aplicação de algoritmos de classificação e de regressão linear.
Os resultados obtidos através das técnicas de mineração de dados são transformados em conhecimento útil e apresentado através de gráficos. A regressão linear indica uma grande eficiência na previsão da nota de língua portuguesa, mostrando forte influência dos indicadores contextuais para sua determinação.

ABSTRACT

This work applies the steps of Knowledge Discovery in Databases (KDD) in the ENEM open data set, by school, in the year 2015, with the objective of finding relationships between the contextual indicators present in the database and the average scores in the different areas of knowledge assessed by the exam. In pre-processing the data is appropriate and filtered, with Microsoft Excel and R, to be used in the next step. In the data mining phase, R is used for the application of classification and linear regression algorithms. The results obtained through the techniques of data mining are transformed into useful knowledge and presented through graph plots. Linear regression indicates great efficiency in predicting the Portuguese language note, showing strong influence of contextual indicators for its determination.

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

INTRODUÇÃO

Dados vem sendo coletados e acumulados em um ritmo acelerado em uma ampla variedade de domínios. O volume de dados produzidos ultrapassa a capacidade humana de analisá-los sem algum tipo de auxílio computacional. Por isso, é necessário o uso de ferramentas e teorias que auxiliem na extração de informação útil (conhecimento). Tais teorias e ferramentas compõem o que chamamos de descoberta de conhecimento em base de dados, ou KDD (do inglês, "*Knowledge Discovery in Databases*") (Fayyad, et al., 1996).

Mineração de dados ou Data Mining é uma etapa do KDD, nesse sentido, o conhecimento a ser descoberto é o produto final do KDD. Data Mining consiste na aplicação de algoritmos específicos para extrair padrões dos dados. Outros passos da descoberta de conhecimento incluem preparação, seleção e limpeza dos dados e interpretação apropriada dos resultados da mineração.

Mineração de Dados, ou DM (do inglês, "Data Mining"), pode ser também entendido como uma área interdisciplinar, mobilizando principalmente conhecimentos de análise estatística de dados, aprendizagem de máquina, reconhecimento de padrões e visualização de dados (Cabena, et al., 1998).

Alguns autores consideram Data Mining como sinônimo de KDD (Klösgen, et al., 2002), referindo-se a ambas como uma disciplina que objetiva a extração automática de padrões interessantes e implícitos de grandes coleções de dados.

A mineração de dados educacionais, ou EDM (do inglês, "*Educational Data Mining*"), é definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais (Baker, et al., 2011). Através da análise desses dados é possível determinar fatores que influenciam a aprendizagem e melhorá-la de forma eficaz.

O Exame Nacional do Ensino Médio (ENEM), realizado anualmente pelo INEP desde 1998, tem como objetivo avaliar o desempenho escolar ao final da Educação Básica. Atualmente o ENEM permite aos estudantes ingressar no Ensino Superior, através de programas como o SISU, PROUNI e convênios com instituições portuguesas, e em programas de financiamento e apoio estudantil (INEP, 2019c).

Os dados do ENEM 2015 serão utilizados neste trabalho, em função de que 2015 foi o último ano em que o INEP disponibilizou, de forma pública, os dados do ENEM das instituições e seus respectivos indicadores socioeconômicos.

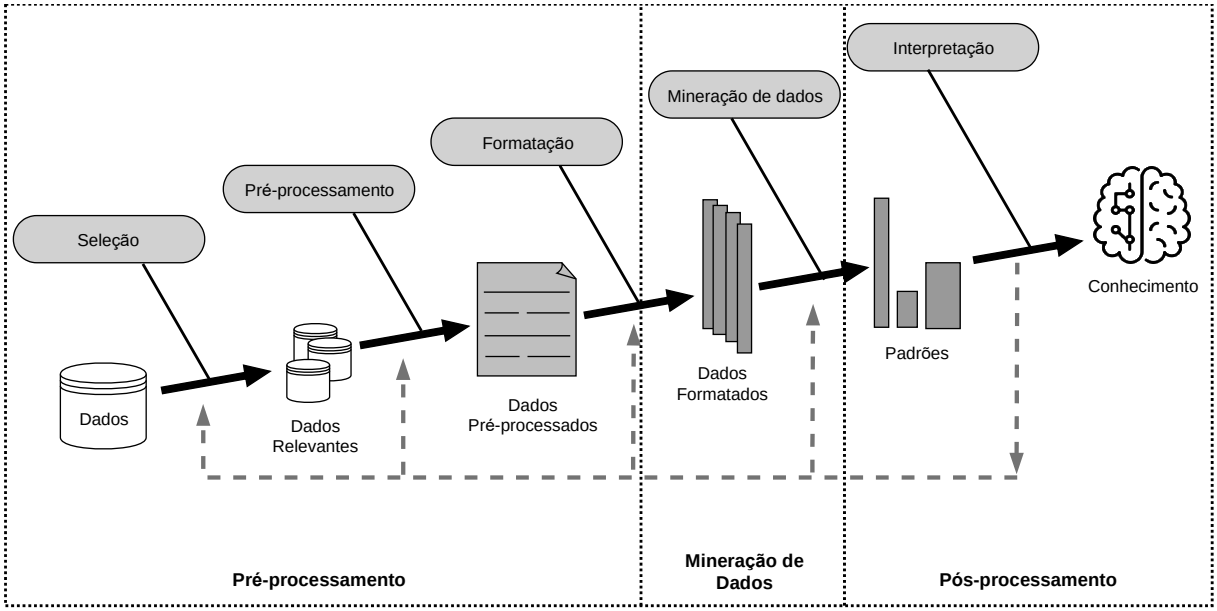
O objetivo deste trabalho é utilizar os dados do ENEM 2015 para avaliar o desempenho das escolas públicas e privadas, que participaram dessa edição. Neste caso, será utilizado um algoritmo de regressão linear, para que se possa determinar se as notas de uma determinada disciplina têm relação com os indicadores contextuais da base de dados de 2015, com ênfase no nível socioeconômico das instituições.

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

2. DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

O processo de KDD, tem como objetivo filtrar, e identificar padrões em conjuntos de dados que analisados gerem informações válidas para estratégias e tomadas de decisões (Fayyad, et al., 1996). A Figura 1 apresenta as etapas do KDD.

Figura 1. Etapas do KDD



Fonte – Adaptado de Fayyad, et al., 1996.

O processo de KDD é composto por três etapas operacionais: Pré-processamento, Mineração de Dados e Pós-processamento. A primeira etapa compreende as funções relacionadas a captação, à organização e ao tratamento dos dados e tem como objetivo a preparação dos dados para os algoritmos para a etapa seguinte.

Na etapa de Mineração de Dados, é realizada a busca efetiva por conhecimentos úteis e, são definidas as técnicas e os algoritmos a serem utilizados no problema em questão. A última etapa abrange o tratamento do conhecimento obtido com o objetivo de viabilizar o conhecimento descoberto (Goldshmidt & Passos, 2005). As etapas operacionais são descritas a seguir:

a) Pré-processamento: é a fase de seleção e preparação dos dados (Han, et al., 2012).

b) Mineração de Dados: é o processo de busca de conhecimento através de algoritmos inteligentes (Goldshmidt & Passos, 2005).

c) Pós-processamento: Esta etapa do KDD envolve análise, interpretação e visualização do modelo de conhecimentos gerado pela etapa de Mineração de Dados (Goldshmidt & Passos, 2005).

2.1. MINERAÇÃO DE DADOS EDUCACIONAIS

A Mineração de Dados Educacionais (EDM) vem ganhando destaque atualmente. Após uma sequência de workshops relacionados ao tema e realizados anualmente desde 2004, criou-se,

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

em 2008, a Conferência Internacional sobre Mineração de Dados (Baker, et al., 2011). Em 2009 foi publicado o primeiro volume da Revista de Mineração de Dados Educacionais (*Journal of Educational Data Mining*).

A EDM busca utilizar ou adaptar métodos e algoritmos de mineração de dados já existentes, de forma a compreender melhor dados produzidos por estudantes e professores. A Mineração de Dados Educacionais pode, entre outras coisas, auxiliar a entender o estudante no seu processo de aprendizagem. Há a necessidade de adequar os algoritmos de mineração de dados existentes para lidar com especificidades dos dados educacionais, como a não independência estatística e a hierarquia dos dados (Costa, et al., 2012).

Muitas das linhas de pesquisa na área de EDM são derivadas diretamente da mineração de dados. Alguns dos tópicos mais interessantes da área são: predição, agrupamento, mineração de relações, destilação de dados para facilitar decisões humanas e descobrimento com modelos. A seguir é exposto uma breve descrição de cada um desses tópicos (Baker, et al., 2011).

- **Predição:** métodos de predição são utilizados para determinar quais características de um modelo são relevantes para a sua predição;
- **Agrupamento:** o objetivo é classificar os dados em grupos de acordo com suas características;
- **Mineração de relações:** esta tarefa envolve descobrir quais variáveis são mais fortemente associadas com uma variável específica. O R, especificamente, gera muito facilmente matriz de correlação, que pode ser usada tanto para ver a relação entre as variáveis como com alguns outros algoritmos, a depender do objetivo;
- **Destilação de dados para facilitar decisões humanas:** a meta aqui é tornar possível a visualização dos dados de forma gráfica e relevante;
- **Descobrimento com modelos:** a partir de um modelo já definido por uma técnica de predição ou agrupamento, será feita uma segunda análise com outra técnica de MDE.

3. CONTEXTUALIZAÇÃO DO AMBIENTE DE DADOS

O Governo é um principal contribuinte na distribuição de dados abertos. “No Brasil, o direito de cada cidadão ter acesso aos dados está previsto na Lei Federal 12.527/2011, conhecida como Lei de Acesso à Informação” (Brasil, 2011).

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) é uma autarquia federal vinculada ao Ministério da Educação (MEC), visa subsidiar a formulação de políticas educacionais dos diferentes níveis de governo com intuito de contribuir para o desenvolvimento econômico e social do país. Dentro deste contexto, são gerados pelo INEP dados referentes ao desempenho dos estudantes de instituições de ensino fundamental, médio e superior, públicas e privadas.

Os dados, utilizados neste trabalho, estão disponíveis no site do INEP, mas o conjunto de dados escolhidos foram os microdados do ENEM por escola (2005 a 2015). Para esta análise,

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

foram filtrados somente os dados correspondentes ao ano de 2015. Os dados estão em um formato CSV, dentro do pacote *microdados_enem_por_escola* que contém também o dicionário de dados.

O ano de 2015 foi escolhido por ser o primeiro a possuir o indicador de nível socioeconômico (INSE) das escolas, além do que, os dados do ano de 2015 foram os últimos com o INSE disponibilizado de forma aberta pelo INEP. Segundo a nota técnica do ENEM de 2015 (INEP, 2019a):

> O INSE possibilita, de modo geral, situar o publico atendido pela escola em um estrato social, apontando o padrão de vida referente a cada um de seus níveis ou estratos. Esse indicador é calculado a partir do nível de escolaridade dos pais e da posse de bens e contratação de serviços pela família dos alunos.

Os dados do INSE de todas as escolas do país podem ser obtidos no site do INEP. O pacote contendo os dados do INSE possui também uma nota técnica que explica o cálculo desse índice de forma detalhada. Os dados presentes nesse pacote foram também utilizados para corrigir os nomes das instituições presentes no conjunto de dados do ENEM, que estavam com problemas de formatação.

Como os dados socioeconômicos serão mencionados com certa frequência no decorrer deste texto, cabe fazer uma breve explicação sobre quais são esses grupos e o que eles representam. As informações aqui mencionadas podem ser encontradas de forma mais aprofundada no INEP (2019b).

Em um primeiro momento o INEP classifica os estudantes de uma escola em um nível socioeconômico que varia de I a VIII. Quanto menor o nível socioeconômico, piores as condições socioeconômicas daquele estudante, e quanto maior o nível socioeconômico, melhores são as condições socioeconômicas.

Os dados utilizados nesse trabalho são do ENEM de 2015 por escola, ou seja, apresenta o índice socioeconômico da escola, não do aluno. As escolas são classificadas em grupos de 1 a 6, sendo que o grupo 1 representa uma maior quantidade de estudantes de níveis socioeconômicos menores, e o grupo 6 representa uma maior quantidade de estudantes de níveis socioeconômicos maiores.

4. FERRAMENTAS E ALGORITMOS

Os dados selecionados através do portal do INEP, por estarem no formato de planilhas *.csv*, puderam ser rapidamente visualizados no Microsoft Excel.

Com exceção da etapa de seleção de dados, todas as demais etapas do KDD foram realizadas com o RStudio, que é um ambiente de desenvolvimento integrado, do inglês *integrated development enviroment* (IDE). O RStudio foi criado para facilitar a utilização da linguagem de programação R, de forma similar ao que ocorre com o popular Eclipse, nesse caso para a linguagem de programação Java. Por esta razão, todas as demais menções em relação a bibliotecas, algoritmos e técnicas serão com relação ao R, pois estes funcionam de forma independente ao RStudio. Para o desenvolvimento do trabalho foram utilizadas as bibliotecas gglopt2, caret, rpart, rpart.plot, corrr, stats e stringr. As versões e os softwares utilizados são:

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

- Excel, versão 16.27, com a licença do Office 365;
- R, versão 3.5.1;
- Bibliotecas – ggplot2 (3.1.0), caret (6.0), rpart (4.1), rpart.plot (3.0.6), corr (0.3.2), stats (3.5.1), stringr (1.3.1);
- RStudio, versão 1.1.463.

4.1. CLASSIFICAÇÃO E REGRESSÃO LINEAR

Por ser inédito na versão de 2015, o indicador de nível socioeconômico, ou INSE, é um dos atributos mais importantes. Em Silva, et al. (2014) os autores criaram um questionário para fazer um levantamento de dados socioeconômicos de alunos de escolas das capitais da região sudeste do Brasil, no ano de 2010. Esses dados foram relacionados com os resultados obtidos no exame e a conclusão consistiu em fatores que influenciaram o desempenho.

Apesar de ser relevante para o contexto que foi realizado o trabalho, a amostra com a qual os autores trabalharam não contemplava toda a extensão do território nacional e seu questionário socioeconômico difere do questionário realizado pelo INEP. A vantagem de utilizar os dados diretos do INEP é justamente por haver uma padronização na coleta desses dados, que podem ser comparados de um ano para o outro, por exemplo.

Influenciado pelo trabalho de Silva et al. (2014), a motivação deste trabalhou consiste em realizar a classificação das escolas considerando o seu INSE (INEP, 2019a). A ideia de usar a classificação veio de Simon & Cazella (2017) que trabalharam também com os dados do ENEM de 2015. A classificação foi feita com a biblioteca rpart. A biblioteca é capaz de gerar modelos de classificação e regressão.

5. METODOLOGIA

5.1. PRÉ-PROCESSAMENTO

5.1.1. PRIMEIRA ETAPA – SELEÇÃO DOS DADOS

Em um primeiro momento os dados do ENEM por escola de 2005 a 2015 foram abertos no Microsoft Excel para melhor visualização dos seus atributos. Através do dicionário de dados foi definido que somente o ano de 2015 apresentaria relevância para este trabalho, então, ainda utilizando o Excel foi feita a filtragem. Após filtragem restaram 15.598 registros com 27 colunas.

5.1.2. SEGUNDA ETAPA – PRÉ-PROCESSAMENTO

De todos os 15.598 registros, somente 101 possuíam algum tipo de dado em branco. Após pesquisa no portal do INEP, constatou-se que esses dados realmente estavam incompletos. Logo, para melhor eficiência dos algoritmos e análises, esses registros foram removidos.

Para auxiliar nas análises, alguns atributos foram decodificados, por exemplo para a dependência administrativa que pode assumir o valor 1, 2, 3 ou 4, que representam, respectivamente, dependência administrativa Estadual, Federal, Municipal ou Privada.

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

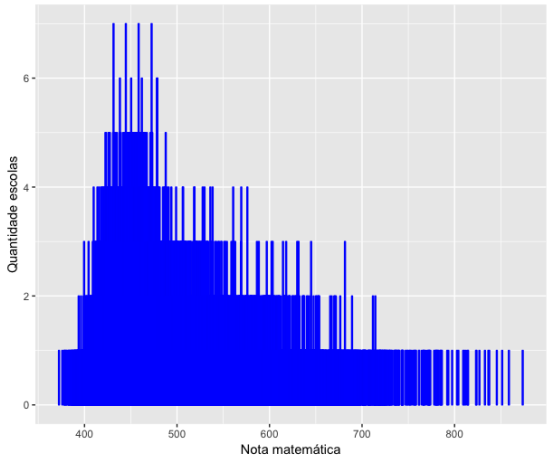
Outros dados que passaram pelo processo de codificação/decodificação incluem: PORTE_ESCOLA, TP_LOCALIZACAO_ESCOLA e INSE. Ainda para exclusivo uso das análises no pré-processamento, alguns atributos foram distribuídos em faixas, são eles:

- As cinco notas nas áreas de conhecimento – faixas de 5;
- PC_FORMACAO_DOCENTE, NU_TAXA_APROVACAO, NU_TAXA_PARTICIPACAO – faixas de 10;

A criação de faixas se faz necessária pois os valores brutos estão distribuídos de forma contínua, em números reais. Ao criar um gráfico com os dados brutos, há uma dificuldade para observar alguns comportamentos, como explicado a seguir.

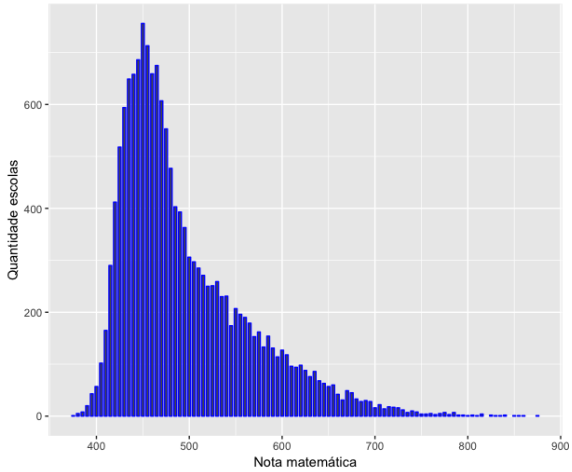
O gráfico da Figura 2 não consegue mostrar a realidade da distribuição das notas, pois há uma sobreposição dos registros diferentes devido a limitação do gráfico de barras, e existem poucos registros iguais, já que há uma precisão decimal com relação à nota bruta. A distribuição das notas por escola é melhor visualizada na Figura 3.

Figura 2. Distribuição das notas brutas de matemática, por escola



Fonte - Autores, 2019

Figura 3. Distribuição das notas em faixas de matemática, por escola



Fonte - Autores, 2019

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

As notas brutas ainda serão utilizadas na mineração de dados, mas a distribuição em faixas garante uma melhor visualização a depender do contexto.

Dados como PC_FORMACAO_DOCENTE e TAXA_APROVACAO encontram-se em porcentagens. Visando ainda o uso de algoritmos de mineração de dados, foi criado mais cinco atributos que colocam as notas das áreas de conhecimento em porcentagens também, já que alguns algoritmos podem dar maior relevância para as notas visto que estas se encontram originalmente numa escala de 0 a 1000, e os dados que estão em porcentagem, intuitivamente, estão numa escala de 0 a 100. Esses atributos foram nomeados como CN_PERCENT, CH_PERCENT, LP_PERCENT, MT_PERCENT e RED_PERCENT.

Ao final do pré-processamento, restaram 15.497 registros e 42 colunas (ou, atributos). O aumento de colunas se deu devido a codificação/decodificação de alguns atributos já presentes na base de dados e também devido a criação de dez novos atributos que são formas diferentes de representação das notas médias brutas nas cinco áreas de conhecimento.

5.2. MINERAÇÃO DE DADOS

A literatura disponibiliza diversos algoritmos e tarefas relacionadas a mineração, mas este são utilizadas de acordo com o objetivo da análise dos dados. Para este trabalho, é importante ressaltar as duas categorias de técnicas de aprendizagem (Camilo & Silva, 2009):

- Algoritmos de aprendizado supervisionado: o conjunto de dados possui uma variável pré-definida, a classe, e os registros são categorizados ou rotulados em relação a esta classe;
- Algoritmos de aprendizado não supervisionado: o conjunto de dados não precisa de uma pré-categorização, ou seja, não é necessário determinar uma variável alvo.

Como este trabalho utilizou a classificação e regressão linear, ambos fazem uso do aprendizado supervisionado. Tanto na classificação como na regressão o conjunto de dados foi dividido em dois. O primeiro conjunto, consiste em 80% dos dados presentes no banco de dados original e é denominado dataTrain. Os 20% restantes foram chamados de dataTest. O dataTrain representa o conjunto de dados que o algoritmo irá utilizar para ser treinado. Com os modelos de classificação e regressão treinados, estes serão submetidos à base de dados dataTest para que sejam validados e sua eficiência seja verificada.

A divisão dos dados é realizada através da biblioteca caret, amplamente documentada e disponível em (Kuhn, 2019). A divisão dos dados pode ser configurada pelo usuário, no caso foi escolhida a divisão 80% e 20% por ser uma divisão padrão da biblioteca a qual faz referência ao princípio de Pareto (Louis & Conway, 2009).

5.2.1. CLASSIFICAÇÃO – ÁRVORE DE DECISÃO

As análises do pré-processamento indicaram a relevância de 17 atributos para a determinação do INSE. São eles: NU_TAXA_PARTICIPACAO, CO_UF_ESCOLA, TP_LOCALIZACAO_ESCOLA, NU_MATRICULAS, NU_PARTICIPANTES_NEC_ESP, NU_PARTICIPANTES, NU_TAXA_ABANDONO, NU_TAXA_REPROVACAO, PORTE_ESCOLA_NUM, PC_FORMACAO_DOCENTE, NU_TAXA_APROVACAO,

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

DEPENDENCIA_ADMINISTRATIVA_NUM, CN_PERCENT, CH_PERCENT, LP_PERCENT, MT_PERCENT e RED_PERCENT.

Os atributos CO_UF_ESCOLA e DEPENDENCIA_ADMINISTRATIVA_NUM representam, respectivamente, os estados do Brasil e o número referente a dependência administrativa das escolas (estadual, municipal, federal e privada).

Os atributos identificados anteriormente foram utilizados para classificar as escolas em determinados grupos socioeconômicos, devidamente contextualizados na seção 3 deste trabalho, o resultado final determina a eficiência do algoritmo para essa tarefa, no qual foi utilizado o aprendizado supervisionado. A classificação foi realizada através de um algoritmo de árvore de decisão, com a biblioteca rpart.

O Algoritmo 1 apresenta o trecho mais importante do script que gera a árvore de decisão. A função rpart() recebe como parâmetro o atributo a ser classificado, e em seguida os atributos que serão utilizados para a classificação. A função predict() recebe o classificador criado como parâmetro, a base de dados de teste e o tipo, que é um classificador.

Algoritmo 1. Trecho do script para geração da árvore de decisão

```
classificador = rpart(formula = inse.num ~ NU_TAXA_PARTICIPACAO +
                      CO_UF_ESCOLA +
                      TP_LOCALIZACAO_ESCOLA +
                      NU_MATRICULAS +
                      NU_PARTICIPANTES_NEC_ESP +
                      NU_PARTICIPANTES +
                      NU_TAXA_ABANDONO +
                      NU_TAXA_REPROVACAO +
                      PORTE_ESCOLA_NUM +
                      PC_FORMACAO_DOCENTE +
                      NU_TAXA_APROVACAO +
                      DEPENDENCIA_ADMINISTRATIVA_NUM +
                      FAIXA_CN +
                      FAIXA_CH +
                      FAIXA_LP +
                      FAIXA_MT +
                      FAIXA_RED,
                    data = dataTrain)
previsoes = predict(classificador, newdata = dataTest, type = 'class')
```

Fonte - Autores, 2019

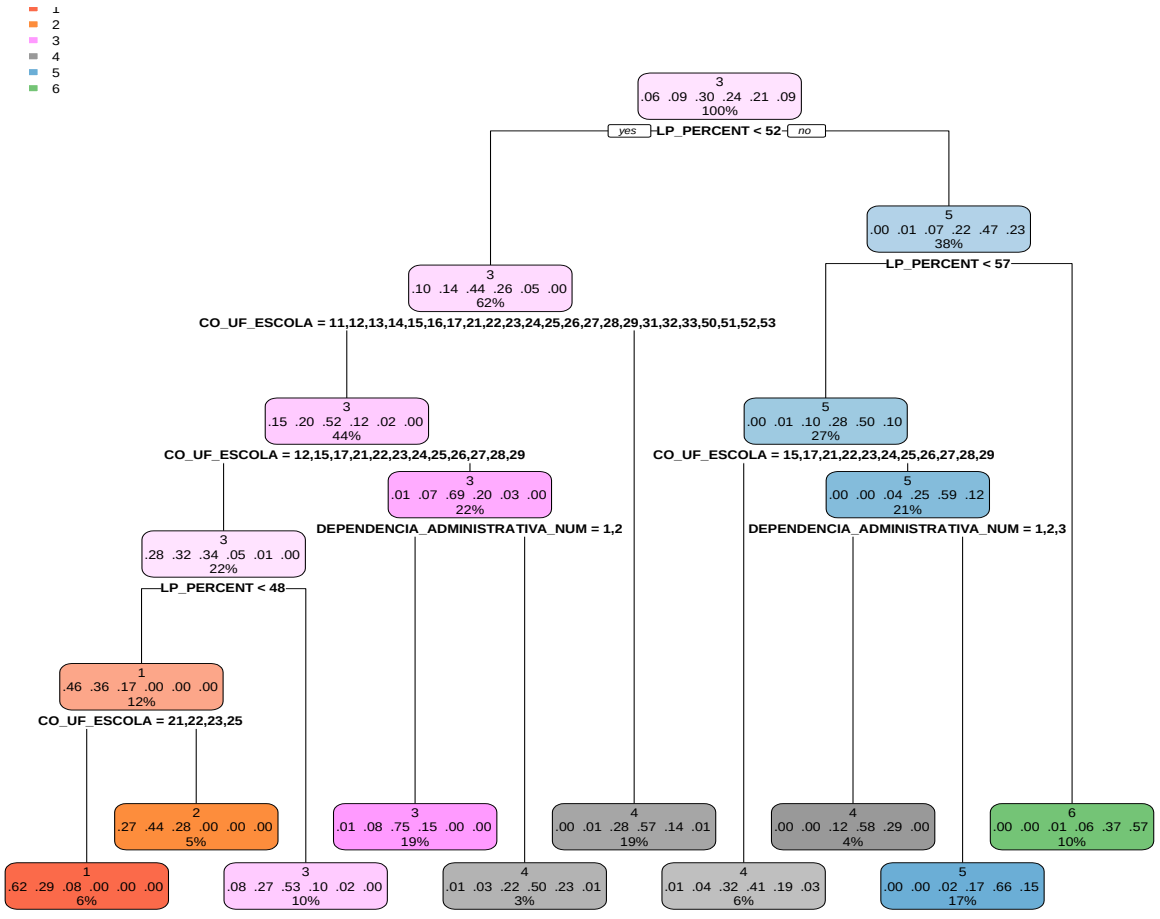
Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

Como pode ser visto na Figura 4, os dados que aparecem na árvore mostram especificamente quais estados ou dependências administrativas são considerados pelo modelo, isso acontece porque os dados estão no formato categórico.

Esse classificador, construído com uma árvore de decisão, possibilitou associar o grupo socioeconômico em função dos atributos CO_UF_ESCOLA e DEPENDENCIA_ADMINISTRATIVA_NUM. Ao analisar a árvore de decisão, fica evidente que o algoritmo considera a nota na área de conhecimento de língua portuguesa muito importante para realizar a classificação, o que motivou o segundo estudo baseado em regressão linear.

O resultado apresentado pela classificação não demonstra relevância devido ao seu baixo índice de acertos. Seu resultado foi exibido nesse trabalho por duas razões. A primeira é para gerar comparações em trabalhos relacionados, já que a inclusão de mais indicadores contextuais podem aumentar significativamente a precisão da árvore. A segunda razão é a motivação do uso da regressão linear, a ser apresentado na próxima seção, já que através da árvore de decisão é possível verificar que as notas em língua portuguesa, tem influência na determinação de certas características dos dados, e a regressão linear será utilizada para avaliar a influência da nota e da disciplina.

Figura 4. Árvore de decisão com o resultado da classificação



Fonte - Autores, 2019

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

5.2.2. REGRESSÃO LINEAR

A utilização da regressão linear consiste em determinar o nota média de língua portuguesa das escolas com base nos mesmos atributos utilizados na classificação, porém aqui os dados de grupo INSE serão utilizados para ajudar a realizar a regressão, e os dados referentes às outras notas serão removidos, já que não faz muito sentido realizar a regressão linear sabendo previamente as notas em outras áreas de conhecimento.

A regressão linear foi realizada com a biblioteca stats, que já está presente no RStudio por padrão. Para avaliar as hipóteses levantadas na etapa de classificação, foi realizada a regressão de duas maneiras.

O Algoritmo 2 mostra um trecho do script para a geração do modelo de regressão linear.

Algoritmo 2. Trecho do script para a geração do modelo de regressão linear

```
regressor = lm(formula = NU_MEDIA_LP ~ ., data = dataTrain)

previsoes = predict(regressor, newdata = dataTest)

ggplot() + geom_point(aes(x = dataTest$PC_FORMACAO_DOCENTE, y  = dataTest$NU_MEDIA_LP), colour = 'blue') +

 geom_point(aes(x = dataTest$PC_FORMACAO_DOCENTE, y = previsoes), colour = 'red')
```

Fonte - Autores, 2019

O Algoritmo 2 apresenta as funções lm(), a função predict(), a função ggplot() e a função geom_point(), as quais são descritas a seguir:

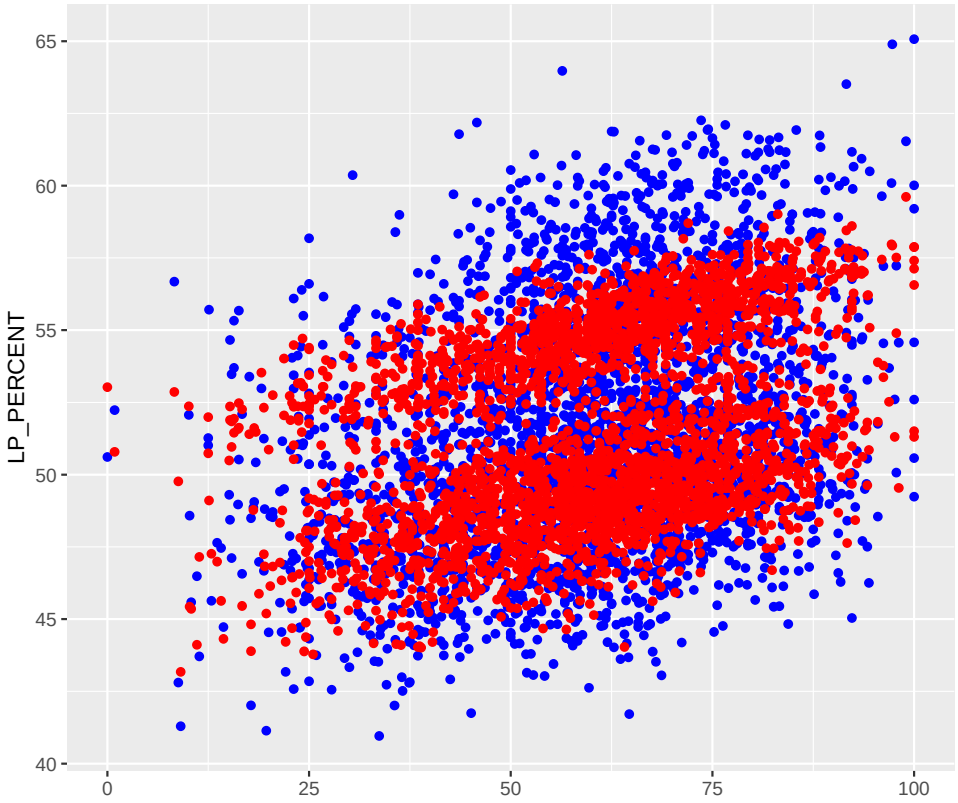
1) A função lm() cria o modelo, recebe como parâmetro a variável a ser estimada e os indicadores contextuais aos quais deve se basear para criar o modelo, representados por '.', o que mostra que todos os indicadores serão utilizados, e por fim, o conjunto de dados;
2) A função predict() recebe o modelo criado e o conjunto de dados para fazer a previsão;
3) A função ggplot() cria um gráfico, o sinal de '+' representa a adição de mais informações nesse gráfico;
4) A função geom_point() renderiza os gráficos de pontos.

Neste caso serão construídos dois gráficos de pontos através da função geom_point(), um com pontos azuis e outro com pontos vermelhos. Os pontos em azul vão representar as notas reais das escolas, enquanto os pontos em vermelho representarão as notas previstas pelo modelo.

No eixo X utilizamos PC_FORMACAO_DOCENTE de forma arbitrária, pois o interesse aqui é comparar a distribuição das notas previstas com as notas reais. A Figura 5 apresenta o primeiro modelo, que não considera os grupos socioeconômicos. Já o segundo modelo, apresentado na Figura 6, considera os grupos socioeconômicos, observando que, o LP_PERCENT indica as notas de língua portuguesa escaladas de 0 a 100.

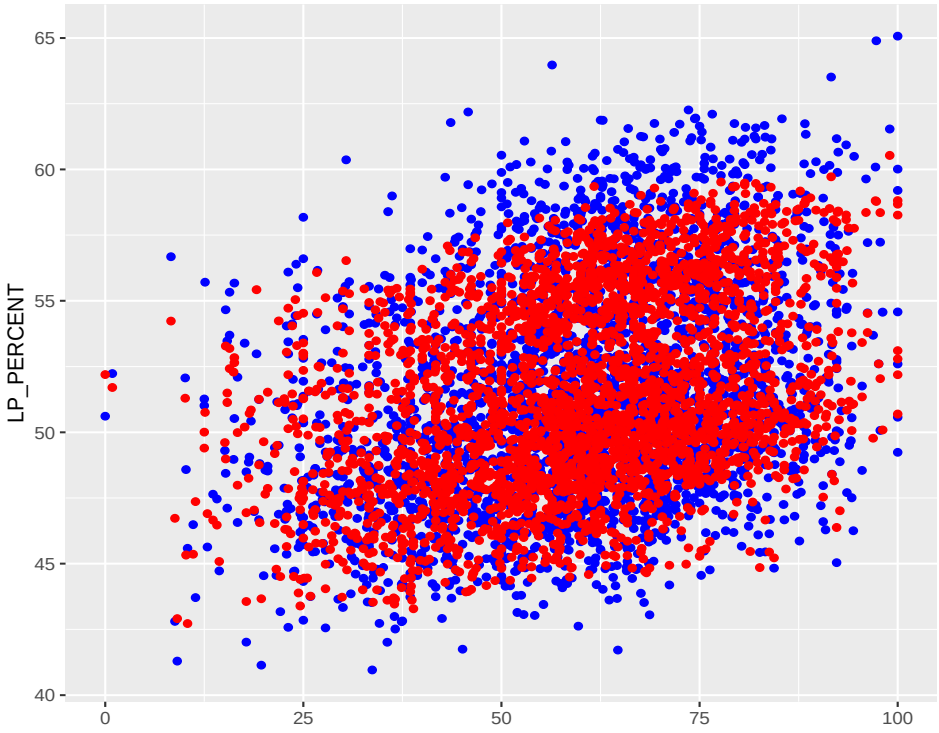
Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

Figura 5. Resultado da regressão linear sem o uso do INSE



Fonte - Autores, 2019

Figura 6. Resultado da regressão linear com o uso do INSE



Fonte - Autores, 2019

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

Através da Tabela 1 e da Tabela 2 podemos fazer uma comparação entre os valores previstos e os reais. As notas no ENEM variam de 0 a 1000, nas tabelas as notas estão em outra escala, de 0 a 100.

Tabela 1. Sumário dos resultados da regressão sem INSE

Menor diferença	Maior diferença	Mediana	Média
0.00167	9.82585	1.58437	1.93200

Fonte - Autores, 2019

Tabela 2. Sumário dos resultados da regressão com INSE

Menor diferença	Maior diferença	Mediana	Média
0.000006	7.993451	1.281874	1.572777

Fonte - Autores, 2019

Tomando com exemplo a Tabela 1 e a coluna Maior Diferença, temos o valor de 9.82585. Esse valor indica que para o modelo treinado na regressão, a maior diferença de nota entre a nota prevista e a nota real foi de 98,2585 pontos. Por exemplo, se uma instituição obteve nota 500 através da regressão linear, a nota real é 598,2585. A comparação dos valores apresentados na Tabela 1 e na Tabela 2 mostram como o desempenho do algoritmo melhora consideravelmente quando o INSE é utilizado para a criação do modelo de regressão.

5.2.3. PÓS-PROCESSAMENTO

Com o objetivo de apresentar os resultados obtidos na mineração de dados e transformá-los em conhecimento, foram gerados gráficos que auxiliam na compreensão do contexto dos dados utilizados. Os gráficos foram gerados utilizando o RStudio juntamente com a biblioteca ggplot, amplamente documentada e utilizada na geração de gráficos. A Figura 7 apresenta o gráfico de densidade de nota.

O Algoritmo 3 mostra o trecho específico do script em R que gera o gráfico da Figura 7. Todos os gráficos apresentados nesse trabalho são gerados através do ggplot. Como o ggplot é uma função, esta recebe o dataset que será utilizado (*full_data*). Os parâmetros utilizados na geração dos gráficos são inseridos dentro da função aes(). Nesse caso, o eixo X apresenta as notas de língua portuguesa distribuídas em faixas. E o eixo Y apresenta a densidade de instituições por grupo. Outras informações podem ser adicionadas para a geração dos gráficos através de um '+', a função geom_density() indica que será um gráfico de densidade, e a variável alpha indica a transparência. Os demais parâmetros adicionados são referentes às legendas do conteúdo, do eixo Y e do eixo X. Para gráficos de barras utilizamos geom_bar() no lugar de geom_density().

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

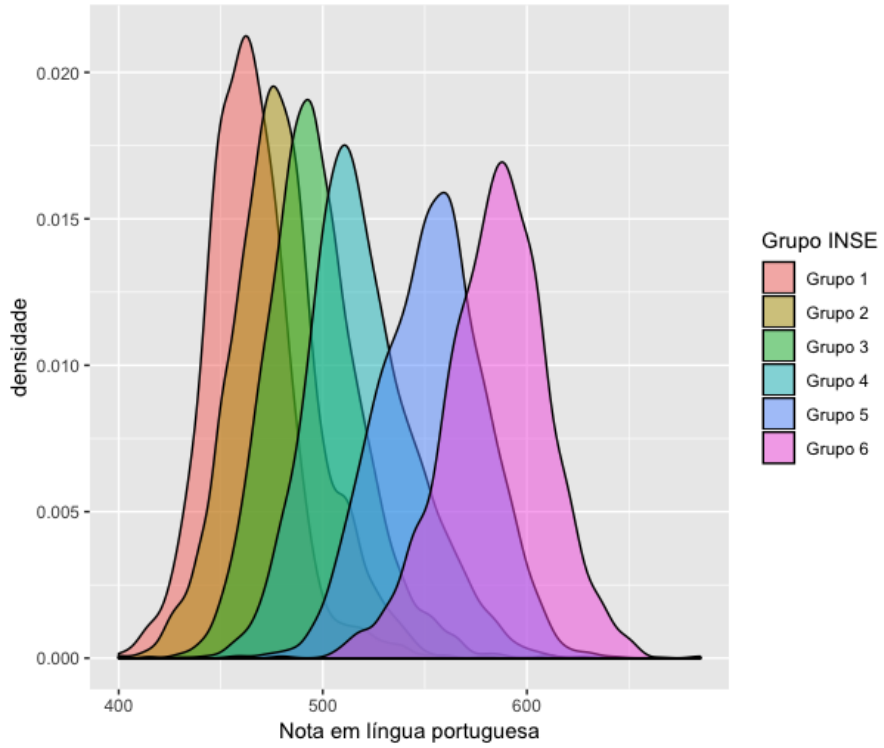
Algoritmo 3. Trecho específico do script em R para geração de gráfico

```
ggplot(full_data, aes(x = full_data$FAIXA_LP, fill = full_data$INSE)) +
  geom_density(alpha = 0.5) +
  labs(fill = "Grupo INSE") +
  ylab("densidade") +
  xlab("Nota em língua portuguesa")
```

Fonte - Autores, 2019

Observa-se que a Figura 7 apresenta a ocorrência de notas maiores conforme troca-se de nível socioeconômico, destacando sua importância para o desempenho no exame.

Figura 7. Gráfico de densidade da nota em língua portuguesa e grupo socioeconômico

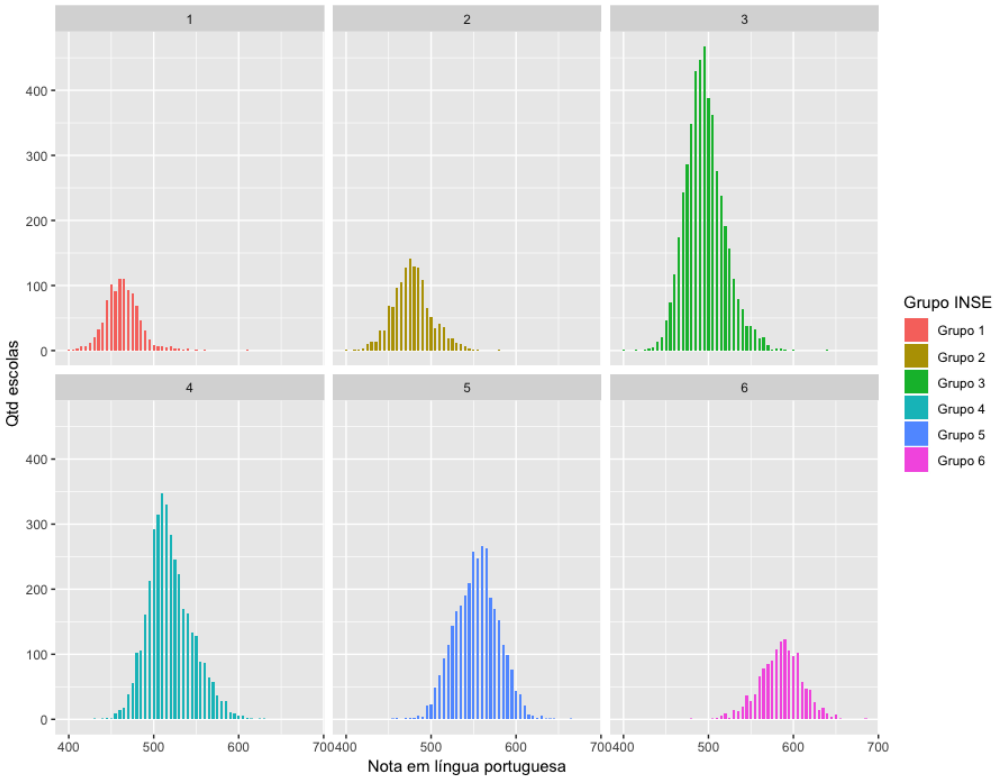


Fonte - Autores, 2019

A Figura 8 apresenta a distribuição por escola em cada um dos grupos socioeconômicos (descritos na seção de contextualização dos dados), o que oferece uma boa perspectiva da realidade socioeconômica nacional de forma geral, observando que o grupo 1 representa um menor poder socioeconômico e o grupo 6 um maior poder socioeconômico.

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

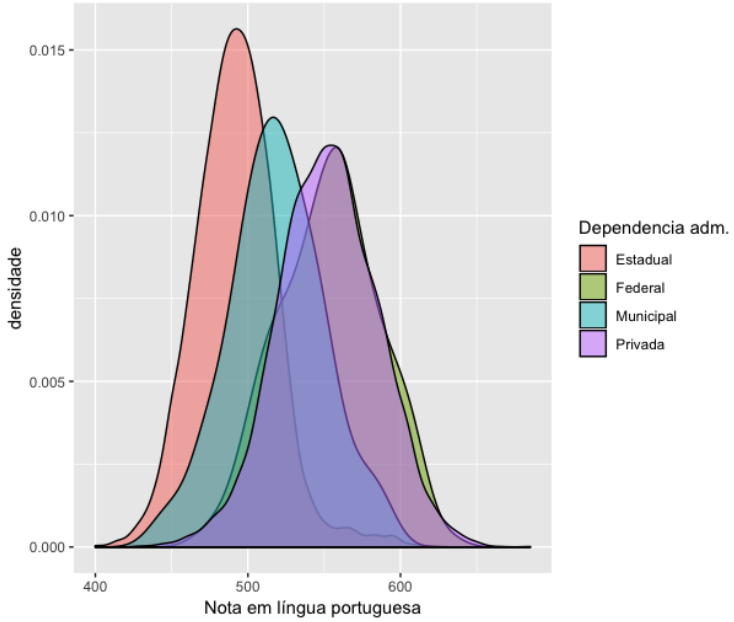
Figura 8. Distribuição de acordo com a nota em língua portuguesa e grupo socioeconômico



Fonte - Autores, 2019

Para o gráfico da Figura 9 é interessante notar que, apesar do desempenho bom para instituições federais, a quantidade de amostras é bem baixa, o que dificulta a comparação, mas mostra uma tendência já esperada de um desempenho superior para essas instituições.

Figura 9. Gráfico de densidade da nota em língua portuguesa e dependência administrativa



Fonte - Autores, 2019

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

6. RESULTADOS

A aplicação do processo de KDD na base de dados do ENEM de 2015 proporcionou o entendimento de cada fase desse processo. Na fase de seleção e pré-processamento o Excel foi utilizado para ver os atributos presentes na tabela e no dicionário de dados; a limpeza e adequação dos dados foi feita utilizando o R.

A fase de mineração constitui a aplicação de duas tarefas, a classificação e a regressão linear. Na classificação utilizou-se árvore de decisão com o objetivo de determinar os grupos socioeconômicos das escolas. A árvore de decisão apresentou baixa eficiência, mas indicou que existe uma relação entre as notas de língua portuguesa e os grupos socioeconômicos. Esse resultado motivou a aplicação do algoritmo de regressão linear, que visa estimar as notas de língua portuguesa utilizando os indicadores contextuais da base de dados do ENEM de 2015.

Na fase de pós-processamento as informações obtidas da mineração de dados, são apresentadas no formato de gráficos, gerados através do software R e a biblioteca ggplot2.

O resultado da árvore de decisão motivou o treinamento de um modelo de regressão linear, com o objetivo de estimar as notas em língua portuguesa no ENEM de 2015. Descobriu-se que a utilização do indicador socioeconômico melhora a precisão do modelo treinado para a determinação das notas, o que indica uma relação entre esses dois atributos.

Este trabalho encontrou relação entre os indicadores contextuais e as notas médias das escolas na disciplina de língua portuguesa. O indicador contextual do grupo socioeconômico das escolas demonstrou maior relevância para o treinamento do modelo de regressão linear, demonstrando que com a presença de um indicador socioeconômico, a precisão na determinação das notas médias é maior do que quando esse indicador não está presente.

7. REFERÊNCIAS

Baker, R.S.J.de & Carvalho, A.M.J.B.de. (2011). Mineração de Dados Educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*. *19(02)*.

Brasil. *Lei de Acesso a Informação – LAI (Lei 12527/2011)*. Retrieved Julho 9, 2019, from http://www2.camara.leg.br/transparencia/acesso-a-informacao.

Cabena, P. & Hadjinian, P. & Stadler, R. & Verhees, J. & Zanasi, A. (1998). Discovering data mining: from concept to implementation. *Upper Saddle River, NJ, USA: Prentice-Hall, Inc.*

Camilo, C. O. & Silva, J. C. DA. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, 1–29.

Costa, E. & Baker R.S.J. d. & Amorim, L. & Magalhães, J. & Marinho, T. (2012). Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. *Jornada de Atualização em Informática na Educação – JAIE*, 1-29.

Fayyad, U. & Shapiro, G.P. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17(3*): 37-54.

Citação (APA): Bravin, G.F., Lee, L. & Rissino, S. das D. (2019). Mineração de dados educacionais na base de dados do ENEM 2015. *Brazilian Journal of Production Engineering*, *5*(*4*), 186-201.

Goldschmidt, R. & Passos, E. (2005). *Data Mining um guia prático*. Elsevier Editora Ltda. Rio de Janeiro. ISBN: 85-352-1877-7.

Han, J. & Kamber, M. & Pei, J. (2012). Data Mining Concepts and Techniques. *Elsevier Editora Ltda*. USA.

Inep (2019a). *Microdados do Enem por Escola*. Retrieved Maio 29, 2019 from http://portal.inep.gov.br/web/guest/microdados.

Inep. *Indicador de Nível Socioeconômico das Escolas de Educação Básica*. (2019b). Retrieved Junho 01, 2019, from: http://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2015/nota_tecnica/nota_tecnica_inep_inse_2015.pdf.

Inep. *ENEM.* (2019c). Retrieved Maio 29, 2019, from http://portal.inep.gov.br/web/guest/enem.

Louis, A.L.B. & Conway, T.R. (2009). Data mining of university philanthropic giving: Cluster-discriminant analysis and Pareto effects. *International Journal of Educational Advancement*.

Klösgen, W. & Zytkow, J.M. (2002). *Handbook of data mining and knowledge discovery*. Oxford University Press, Inc., New York, NY, USA.

Kuhn, M. *The Caret Package* (2019). Retrieved Junho 01, 2019 from https://topepo.github.io/caret/.

Silva, L.A. & Morino, A.H. & Sato, T.M.C. (2014). Prática de Mineração de Dados no Exame Nacional do Ensino Médio. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação.*

Simon, A. & Cazella, S.C. Mineração de Dados Educacionais nos Resultados do ENEM de 2015 (2017*). Anais dos Workshops do VI Congresso Brasileiro de Informática na Educação.*