Relatório do Projeto de Ciencia dos Dados

Gabriel Henrique Daniel da Silva, Anderson Mota

Recife, Pernambuco

Abstract

Nosso projeto visa investigar dados provenientes da área de educacão, visando entender melhor os atributos que fazem uma escola ser melhor do que outra, utilizando como métrica os resultados do ENEM (Exame Nacional do Ensino Médio).

1. Introducão

A educação é um ponto cada vez mais mencionado como imprescindível para a evolucão tanto individual como no contexto da sociedade de modo geral. Diante disto, desejamos trabalhar com dados voltados a educação. Por questão de escopo, optamos por trabalhar com dados voltados ao ensino médio, principalmente por conta das recentes polêmicas envolvendo o Exame Nacional do Ensino Médio (ENEM), nas quais muitos dos estudantes que prestaram o exame clamam que o nível das questões estava incompativel com o ensino das escolas públicas e somente acessível aos estudantes da rede particular.

2. Obtencao de Dados

Os dados que utilizamos foram obtidos basicamente através de uma API chamada App Cívico (http://mobile-aceite.tcu.gov.br:80). Nesssa API podemos encontrar informacoes sobre os equipamentos disponiveis na escola, como por exemplo biblioteca, internet, quantidade de computadores. Alem disso tambem temos informacoes mais admnistrativas sobre a escola como por exemplo o seu

tipo, se publica ou privada, a quantidade de funcionarios, quantidade de alunos, quantidade de salas entre outros.

Complementamos esses dados ja obtidos, com dados provenientes do MEC sobre o desempenho das escolas no ENEM no ano de 2015. Nesses dados podemos observar a quantidade de alunos da escola que participaram, juntamente com suas notas nas 5 diferentes areas de conhecimento avaliadas pelo exame.

3. Análise Exploratória

Uma vez que tinhamos os dados em mãos, começamos a explorá-los para tentar entender melhor o problema que estamos abordando.

Inicialmente nos preocupamos em descobrir a quantidade de escolas publicas e privadas existentes, para então começar a fazer comparações proporcionais entre o tipo das escolas e a quantidade de equipamentos existentes nas mesmas. Diversos resultados chegaram ate mesmo a serem considerados surpreendentes, afinal atualmente acabamos partindo de uma premissa que as escolas públicas são de baixa qualidade e mal equipadas ao contrário das particulares que ao menos em teoria deveriam ser de ótima qualidade e muito bem equipadas.

Uma das principais surpresas encontradas, se deu ao investigarmos o atributo relacionado a quantidade de computadores existentes nas escolas e compararmos com o tipo da escola, como podemos ver abaixo:

Outra surpresa bastante significativa foi encontrada ao investigarmos a relação entre a quantidade de escolas públicas e privadas e a quantidade de bibliotecas existentes nos respectivos tipos de escola.

Em proporções menores, mas ainda sim, de forma bem superior ao esperado, temos a relação entre os laboratórios de ciência e o tipo da escola, onde proporcionalmente, ainda temos um maior número de laboratórios de ciências nas escolas públicas.

Diante dessas análises preliminares começamos a levantar maiores questionamentos sobre a tão falada disparidade de estrutura entre as escolas públicas e particulares, afinal, nos quesitos avaliados pudemos perceber que de modo

geral, as escolas públicas foram tão bem quanto, ou até mesmo superiores proporcionalmente ao conjunto das escolas privadas.

Porém, até este momento, as análises se basearam basicamente em informações estruturais e de equipamento sobre as escolas. Portanto, para enriquecer essa análise, vamos incorporar dados provenientes da prova atualmente mais relevante no contexto atual brasileiro, o Exame Nacional do Ensino Médio (ENEM). Com essa nova métrica, podemos comparar os resultados das escolas públicas e privadas na prática e quem sabe, buscar aprender algo a mais com esse novo conjunto de dados.

4. Aprendizagem

Após essa análise exploratória inicial, partimos para a próxima etapa do projeto. Uma vez que juntamos os dados estruturais com os dados provenientes do ENEM das escolas, falta apenas fazer mais um pequeno pré-processamento nos dados, de forma a converter os atributos booleanos do tipo Sim ou Não, em inteiros 0 e 1, os quais são aceitos nos algoritmos de aprendizagem que utilizamos, o de regressão linear e regressão logística.

Inicialmente, optamos por relacionar alguns atributos estruturais com os resultados que as escolas vem apresentando no Exame Nacional do Ensino Médio. Através de regressão linear, relacionamos o atributo quantidade de computadores com a média geral da escola no ENEM. Como resultado, observamos um ponto interessante, para cada computador disponível na escola, tem se uma média de 1.94 pontos a mais no ENEM.

Outro Atributo que optamos por investigar, foi o de número de salas disponíveis na escola. Para tanto relacionamos o número de salas com a nota geral do ENEM. Chegamos a cerca de 1.82 pontos a mais no ENEM por sala disponível

Além desses, resolvemos investigar também um outro atributo interessante, o número de funcionários na escola. É válido ressaltar que esse número não é restrito apenas a professores, mas a todos os funcionários que trabalham numa determinada escola. Após relacioná-los obtermos como resultado cerca de 0.50

pontos a mais no ENEM por funcionário.

Observando esses dados, torná-se notório que caso uma determinada escola esteja motivada a investir em um projeto de melhoria, uma boa escolha seria investir na compra de computadores, pois certamente dentre as opções citadas apresentou a melhor relação entre custo e benefício.

Mas as análises não cessam por aqui, de agora em diante partimos para análises um pouco mais complexas utilizando regressão logística levando em conta múltiplas variáveis que podem vir a influenciar na nota obtida no Exame Nacional do Ensino Médio.

Na primeira análise desse tipo, optamos por investigar os seguintes fatores com respeito a Média Geral do ENEM: Um valor boleano que indica se a escola possui ou não quadra descoberta, um inteiro que indica a quantidade de salas de uma determinada escola, um inteiro que relata a quantidade de funcionários de uma escola, além desses a quantidade de computadores de uma escola e por fim um indicador se uma escola possui ou não biblioteca. Como resultado, encontramos um coeficiente muito superior no atributo que indicava se uma escola possui ou não biblioteca, seguido a uma certa distância pelo atributo quantidade de computadores.

Seguindo para uma próxima análise, passamos a investigar os seguintes atributos: Quantidade de funcionários, quantidade de salas, quantidade de computadores e um indicador se a escola possui ou não laboratório de ciências. O resultado gerou uma relativa surpresa sobre nós, afinal a vantagem do indicador de laboratório de ciências foi imensa, possuindo um coeficiente quase 50 vezes maior que o atributo que veio em seguida, o qual mensurava a quantidade de computadores.

Após essas duas análises, decidimos fazer uma um pouco diferente, em vez de relacionarmos os atributos com a média geral do ENEM, optamos por relacioná-los com uma parte específica da prova, no caso dessa análise, relacionamos os atributos: Tem biblioteca, tem laboratório de ciências e quantidade de computadores com a Média da prova de Ciências da Natureza do ENEM. Obtemos um valor de coeficiente bastante alto do atributo tem laboratório de ciências em

relação aos demais que tiveram coeficientes bastante inferiores.

Dando prosseguimento, acabamos por analisar utilizando outra parte específica da prova do ENEM, no caso dessa análise, a parte de matemática. Para tanto analisamos os seguintes atributos: Tem quadra descoberta, quantidade de computadores, quantidade de funcionários, quantidade de salas e se tem ou não biblioteca. Dessa vez o maior coeficiente ficou com o atributo que verificava a presença ou não de bibliotecas nas escolas, porém dessa vez não por uma distância tão grande, já que o atributo quantidade de computadores não estava assim tão distante, especialmente se considerarmos a margem de erro, que era especialmente alta no atributo de maior coeficiente.

Finalizando as análises, resolvermos voltar a investigar a média geral do ENEM porém desta vez utilizando apenas os atributos que tiveram os maiores coeficientes nas análises feitas anteriormente: Quantidade de computadores, quantidade de salas, se possui biblioteca ou não, se possui laboratório de ciências ou não e por fim, desta vez, acrescentamos um novo atributo relativo ao tipo da escola se pública ou privada, já que o mesmo não havia sido investigado em nenhum momento anteriormente. O resultado foi bastante interessante e ao mesmo tempo surpreendente, com o maior coeficiente tivemos novamente laboratório de ciências, em seguida quantidade de computadores, em terceiro a quantidade de salas, em quarto se possui biblioteca ou não, e surpreendentemente com o menor coeficiente de todos, o tipo da escola se pública ou privada, fato que nos deixou bastante surpresos.

5. Considerações Finais

Após as diversas análises realizadas, pudemos perceber através dos nosso dados obtidos, que as diferenças estruturais e de equipamentos entre as escolas públicas e privadas não são tão gritantes quanto se imaginava anteriormente. Porém ao compararmos diretamente as médias dos resultados da prova do ENEM entre as escolas públicas e privadas, a diferença acaba sendo notável, especialmente no critério da redação. Mas ainda sim, a diferença não é tão

absurda a ponto de corroborar uma de nossas hipóteses que gostaríamos de investigar, que era se a prova do ENEM estava inacessível para alunos de escolas públicas. Ou seja, provavelmente a diferença entre os tipos das escolas (públicas x privadas) não estejam na parte estrutural ou nos instrumentos utilizados, mas sim em outros fatores que talvez sejam passíveis de uma investigação futura como por exemplo: motivação dos professores ou política interna de avaliação dos estudantes não adequada.

Já observando pela ótica das escolas que visam evoluir, dentre os pontos investigados, certamente a aquisição de computadores, construção de laboratório de ciência e biblioteca, deveriam ser prioridades em termos de fortalecimento estrutural e instrumental da escola,pois tendem a melhorar o rendimento dos alunos das escolas que os possuem.

Por fim, a experiência na disciplina em si, foi extremamente proveitosa do ponto de vista de aprendizagem, especialmente por não focar somente na mecânica de trabalhar com os dados, mas também abordar a questão de senso crítico e interpretação dos resultados obtidos, sem esquecer obviamente, da experiência com uma nova linguagem, num novo ambiente de desenvolvimento. Todo esse conjunto fatores, tem tudo para ser bastante relevante no nosso futuro profissional.