# Aplicando técnicas de regressão para prever resultado da prova SARESP baseado nas informações socioeconômicas dos alunos.

## Introdução ao problema em questão

A fim de descobrir se dados socioeconômicos estão correlacionados com os resultados de proficiência dos alunos na prova de avaliação de ensino escolar do estado de São Paulo, foi feito um estudo com aplicação Machine Learning a fim de tentar prever os resultados da disciplina de matemática com base no dataset de respostas (anônimas) dos alunos à prova do SARESP.

## Objetivo e benefícios do projeto

Sendo identificado correlações fortes das informações obtidas e nas respostas dos alunos com o resultado final podem ser priorizadas medidas de melhoria em deficiências socioeconômicas com mais assertividade a fim de melhorar o desempenho e aprendizado de alunos no Estado.

Sendo identificadas correlações fortes entre as informações obtidas através do questionário socioeconômico e o desempenho nas provas, pode-se reforçar a necessidade de priorização de medidas de combate à desigualdade. Portanto, embora o projeto se destine à predição do desempenho escolar através de dados socioeconômicos dos alunos, vale constatar que sua possibilidade torna evidente o que pode ser feito afim de melhorar o aprendizado de estudantes de escolas estaduais.

## Desenvolvimento do projeto (proposta inicial, problemas encontrados, soluções aplicadas)

Como plano de desenvolvimento, optou-se por seguir as etapas do sistema CRISP-DM que baseia se em 5 etapas que vão desde a análise prévia dos dados até a implantação da modelagem. Sendo um projeto próprio, a primeira etapa, ‘**entendimento do negócio’,** consistiu apenas em entender o que é a prova aplicada pelo SARESP (Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo).

Ao finalizar o entendimento do negócio, iniciou-se a etapa de **entendimento dos dados,** que consistem em uma base com informações dos alunos e todas as alternativas escolhidas pelos alunos na prova socioeconômica. Adicional à base, utilizou-se também um documento com a descrição de cada questão e suas respostas do questionário, informação imprescindível para o entendimento dos dados.

A seguir, na etapa de **Preparação dos Dados** foram realizadas as seguintes etapas:

* analise de registros nulos
* analise de correlação de features
* analise dos tipos de cada variável
* analise de conteúdo de variáveis

Em relação aos nulos, seriam removidas quaisquer colunas com mais de 10% de dados faltantes, entretanto, apenas 5 colunas continham registros faltantes, constituindo mais de 98% dos registros em todas elas, dado isso foram removidas da base.

Pelo problema aqui proposto ser, prever a nota de matemática na prova em questão, as demais notas de outras disciplinas também não estão disponíveis no momento de previsão, portanto serão retiradas para não gerar vieses incorretos no nosso modelo. Partindo dessa premissa, foram removidas features que eram relacionadas com a nossa Target (resultados da prova de matemática) por não fazer sentido estar no conjunto de dados.

Na análise de tipos, foi realizada a conversão de dados de categóricos para numéricos. Não foi utilizada nenhuma técnica muito complexa nessa etapa, apenas uma conversão simples utilizando a biblioteca HotEncoding Outra análise também realizada na etapa de Preparação dos dados, foi um estudo de correlação dos dados com a Target. Porém, nesta etapa, não foi removida nenhuma variável, dado que a média de índice de correlação não foi muito alta. Foi decidido, à princípio, iniciar com um teste de fit com todas as colunas, incluindo as com baixa correlação.

Partindo para a próxima etapa do planejamento, o algoritmo escolhido para iniciar a **modelagem** foi o Random Forest, por normalmente obter resultados melhores que regressões simples, e nos trazer diversas informações adicionais a respeito das nossas variáveis. A base foi separada nos conjuntos de treino e teste para avaliar o modelo com validação cruzada, e aplicou-se o método GridSearch para testar diversos parâmetros no modelo, consistindo em realizar todas as combinações possíveis de parâmetros fornecidos e avaliando seu desempenho.

Após as primeiras previsões, o resultado não foi muito satisfatório, então foi realizada mais uma tentativa de fit após ajustes na base, sendo realizada a remoção de variáveis correlacionadas entre si, deixando apenas uma variável de cada dupla de forte correlação. Apesar da remoção, os resultados não melhoraram significantemente.

Foi feito também um teste com um modelo de regressão simples para comparar **performance** de dois modelos diferentes, provando a hipótese de que o Random Forest inicialmente aplicado e ajustado com GridSearch performa melhor, com um R² 8% melhor e tanto o erro Quadrático médio quanto o erro absoluto médio 1% melhor do que a Regressão Linear Múltipla.

## Conclusão e ganhos do projeto (propostas de melhoria, o que as métricas disseram o que impactou o projeto.)

O Modelo não está representando com grande exatidão a nota de cada aluno.

**R2:0.377**

**MSE:287.467**

**MAE:13.858**

Analisando as métricas de avaliação de modelos R², MSE e MAE, notamos que o modelo está conseguindo representar uma porcentagem baixa dos dados do problema proposto e apresenta uma margem de erro media próximo de 15 pontos percentuais na nota do aluno.

Uma margem de erro de 15% ainda é melhor do que um ‘chute aleatório’, o que abre a possibilidade de melhoria do método de predição, tanto na tentativa de melhorar a regressão, quanto na possibilidade de treinar um classificador para prever as faixas de proficiência dos alunos, sendo a proficiência uma variável categórica atrelada a nota das disciplinas.

No momento ainda não é possível afirmar com convicção que fatores socioeconômicos influenciam no desempenho dos alunos do estado. Porém há expectativa de que em próximos estudos as conclusões melhorem.

Em um projeto próximo será desenvolvido o classificador e novas analises em cima do problema.