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

## Introdução ao problema em questão

Buscando projetos com potencial pra analises utilizando machine learning, me deparei com o seguinte contexto, existe anualmente uma prova de avaliação do ensino estadual do estado de São Paulo, aplicada a todos os alunos da rede estadual.

A fim de descobrir se dados socioeconômicos estão correlacionados com os resultados de proficiência dos alunos na prova, com a tentativa de prever os resultados da disciplina de matemática com base no dataset de respostas (anônimas) dos alunos à prova socioeconômica 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.

Com a análise das métricas de avaliação do modelo de regressão ficou claro que o modelo não performou muito bem, dado isso foi pensada uma nova forma de prever o desempenho dos alunos, não com um valor exato e sim com um intervalo de valores, mesmo o erro médio não sendo muito baixo, ainda é possível classificar corretamente registros com um erro menor. A base que estamos utilizando já tem uma coluna com a classificação da proficiência dos alunos, o que viabiliza tal estudo.

Seguindo para a desenvolvimento da nova forma de previsão, foi necessário a utilização de um modelo de classificação para preverá qual classe o aluno pertencerá. Havia 4 possibilidades de enquadramento, Abaixo do básico, básico, adequado e Avançado.

Ainda na etapa de análise das classes target, percebeu-se um desbalanceamento nas classificações, havia muito mais registros de alunos abaixo do básico do que alunos adequados e avançados, o que trouxe a necessidade de realização de testes de ajuste no modelo com e sem tratamento dessa característica da base de dados.

Dado que as classes estavam desbalanceadas, realizou-se a junção da classe Avançado e Adequado. Pelo objetivo do projeto ser descobrir fatores de problemas, nossa prioridade é classificar com mais detalhes quem está abaixo do adequado, logo a junção das classes não impactará na nossa analise final e ainda ajudará a resolver parcialmente o problema do desbalanceamento.

Na escolha do classificador foram feitos testes com Naive Bayes Gaussiano, e regressão logística multinominal. A regressão logística se adequou melhor ao problema mostrando uma taxa de acerto média 10% melhor que o naive bayes. Apenas a regressão logística teve seus parâmetros ajustados com gridSearch, pelo naive bayes ser um modelo estatístico há poucos parâmetros para ajuste, então decidiu-se utilizar os parâmetros padrão.

Mesmo com a decisão de unificar as classes maiores ou iguais a Adequado foram feitos 3 testes de modelo, um sem a junção de classes e sem a aplicação de undersampling nas classes majoritárias, um com a junção de classes e sem aplicação do undersampling e pro fim um teste com os dois tratamentos.

Os resultados com a junção de classes mostraram uma acurácia melhor do que o teste sem, resultado que já era esperado por motivos simples de probabilidade. O undersampling teve um resultado interessante, no teste onde não foi aplicado, as classes majoritárias estavam com prioridade de previsão, o que seria bom já que queremos prever melhor quem irá mal, porem um modelo desbalanceado, abre espaço para uma maior variância quando aplicado na vida real. O teste com as classes balanceadas gerou resultados constantes de previsão das classes, tendo seus suas métricas registrando aproximadamente 50% a 60% de acurácia e f1 Score.

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

O Modelo de regressão 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 abriu 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 já disponível na base de dados, optou-se por seguir com o desenvolvimento do modelo de classificação para prever faixas de proficiência ao invés de prever a nota com exatidão.

O modelo de Classificação escolhido foi a regressão logística multinominal após ter mostrado uma melhor adequação ao problema do que outros testados, foi treinado com diversas tentativas diferentes de tratamentos nos dados, e obteve seu melhor e mais consistente desempenho com a aplicação de todos os tratamentos idealizados no desenvolvimento do projeto. Os resultados das métricas de avaliação do classificador ficaram com uma acurácia de 58%,métricas como o F1 score, Precision e Recall também foram analisadas.

A acurácia da classificação das faixas de proficiência nos mostra que é possível determinar a nota com um certo grau de erro, o que já nos permite inferir que é possível atuar nas questões abordadas na prova socioeconômica da SARESP a fim de melhorar o desempenho escolar dos alunos.

Os principais pontos de atuação por parte do estado para melhorar o aprendizado dos alunos, podem ser as perguntas das quais as respostas tem uma maior correlação com a nota da prova de conceitos, para isso já temos de output do projeto uma lista ranqueada das questões mais influentes na nota final.

Observa-se ainda mais potencial de descoberta de fatores socioeconômicos correlacionados com a nota do teste, e diversos insights podem ser retirados da analise e dos resultados gerados a partir dos modelos desse projeto porem não serão mencionados nesse relatório de desenvolvimento técnico.