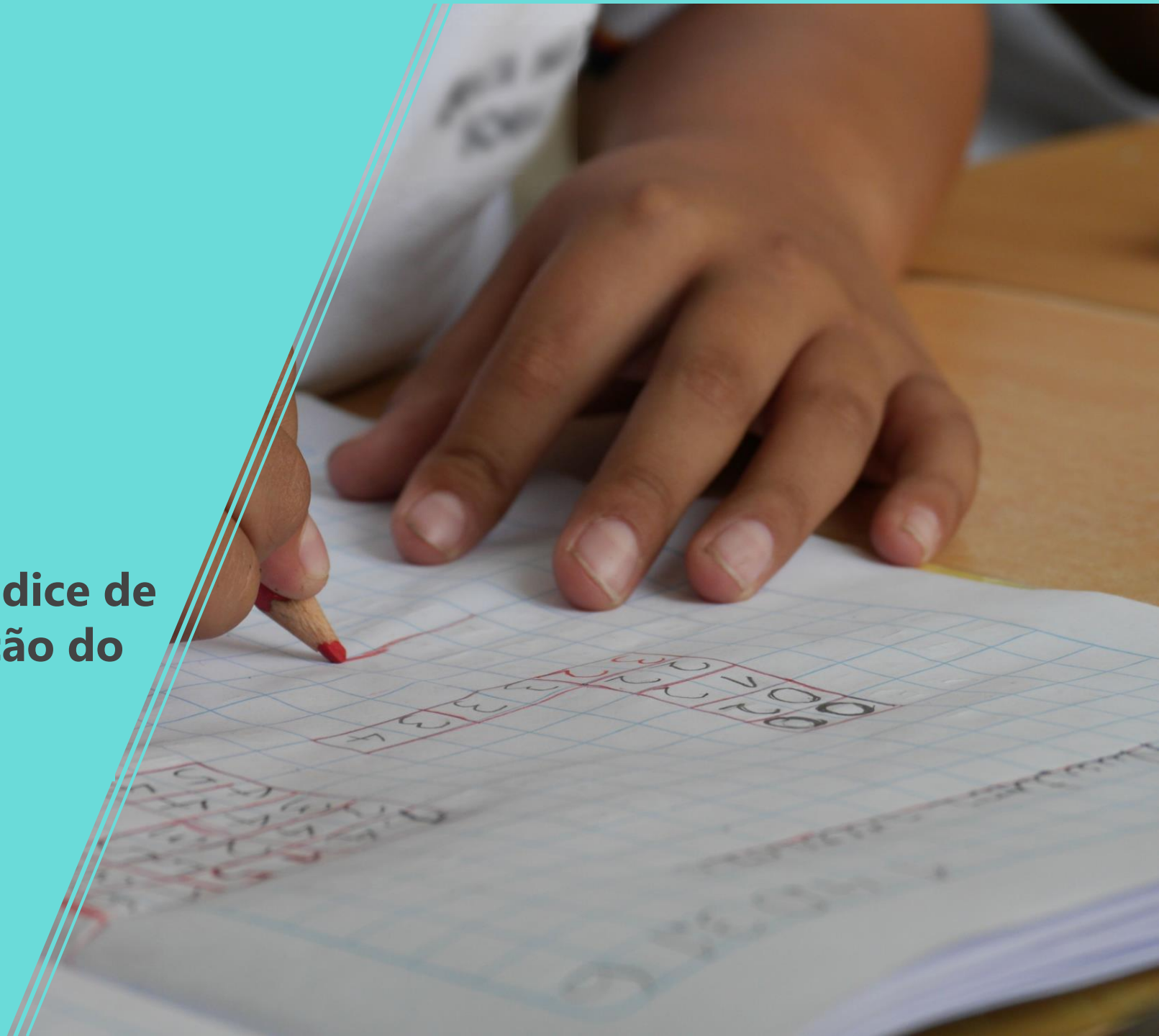




Modelo Preditivo sobre o Índice de Desenvolvimento da Educação do Estado de São Paulo



25/05/2020



Pós-graduação em Análise de Big Data

**Nome do Aluno:**

João Paulo Ribeiro dos Santos

Coordenadores:

Prof.^a Dr.^a Alessandra de Ávila Montini

Prof. Dr. Adolpho Walter Pimazoni Canton



Agenda

- 1. Objetivo do Trabalho
- 2. Contextualização do Problema
- 3. Base de Dados
 - i. Bases originais
 - ii. Filtros
 - iii. Principais variáveis
- 4. Análise Exploratória de Dados
- 5. Modelagem com Estatística Tradicional
- 6. Modelagem com Inteligência Artificial
- 7. Conclusões
- 8. Sugestão para Trabalhos Futuros

Metodologia de análise de dados



Definição do problema

- Objetivos
- Conceitos
- Critérios
- Histórico de dados
- Variáveis

Análise preliminar

- Medidas resumo
- Análise de *outliers*
- Análise de *missings*
- Validação com alguns artigos, e documentos técnicos

Avaliação das técnicas

- Regressão Linear
- Clusterização

Avaliação das técnicas

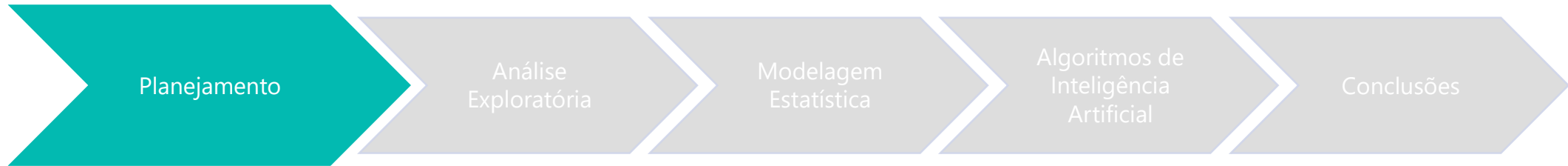
- XGBoost

Definição da técnica

- Validação dos resultados com negócios
- Escolha da técnica que melhor se adequa ao uso e estratégias da área de negócios



Metodologia de análise de dados



Definição do problema

- Objetivos
- Conceitos
- Critérios
- Histórico de dados
- Variáveis

Análise preliminar

- Medidas resumo
- Análise de *outliers*
- Análise de *missings*
- Validação com a área de negócios sobre a consistência das informações

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

Avaliação das técnicas

- Random Forest
- Boosting
- SVM
- Deep Learning

Definição da técnica

- Validação dos resultados com negócios
- Escolha da técnica que melhor se adequa ao uso e estratégias da área de negócios



1. Objetivo do Trabalho

O objetivo do trabalho é **predizer um indicador de qualidade para as escolas do estado de São Paulo, para os próximos 3 anos**, utilizando para isso as informações do IDESP.

Todo o processo de predição bem como de análise será feito utilizando **dados históricos das instituições de ensino**, mais precisamente da **Secretária de Educação do Estado de São Paulo**. Para isso serão utilizados esse dados bem como **modelos estatísticos tradicionais** e **algoritmos de inteligência artificial**, que selecionarão as características mais relevantes que expliquem o índice de qualidade das escolas.

Uma vez tendo essas informações será possível identificar **quais fatores estão relacionados com as escolas com maior qualidade de ensino**, e com tais informações **desenvolver métodos que melhorem a qualidade de ensino nas demais escolas**.





2. Contextualização do Problema

7

Frequentemente nos jornais, artigos, livros, TVs e internet nos deparamos com matérias onde são destacadas as **limitações e deficiências da educação no Brasil**. Matérias recentes apontam inclusive uma **piora nesse índice** tanto que no famoso ranking do PISA o Brasil acabou caindo algumas posições, como aponta matéria do [Globo](#).

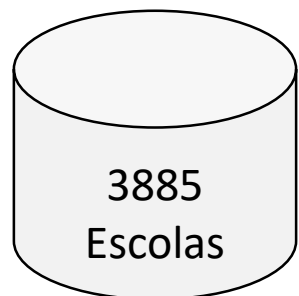
Com o objetivo de melhorar a educação no estado de São Paulo que a Secretária da Educação criou o **PQE (Programa de Qualidade da Escola)** que tem como principal objetivo promover uma melhora na qualidade no sistema de ensino na rede estadual de São Paulo. Para isso foi criado o **IDESP (Índice de Desenvolvimento da Educação do Estado de São Paulo)**, um **indicador** que **mensura a qualidade de ensino** prestada por cada escola na rede estadual.

Com a informação desse índice tanto a secretária de educação quanto as escola podem **adotar medidas** para melhorar esse índice, seja incluindo **atividades extracurriculares** ou **melhorando a grade escolar**.

Diante de tal cenário que será criado um modelo preditivo, com o objetivo de identificar o IDESP para o próximo ano.



3.i. Base original



Visão da base

- Escolas

Filtros de inclusão

- Escolas Ativas

Filtros de exclusão

- Escolas da rede privada, municipal, técnica e federal

Período de Análise

T-5	...	T-2	T-1	T-0			
Histórico					Previsão		

Ano de referência (T0): 2019

Histórico (05 anos): 2014 até 2019

Previsão: 2020

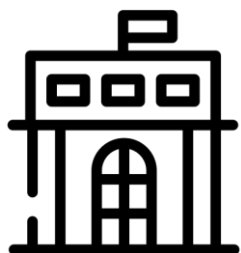


3.ii. Filtros



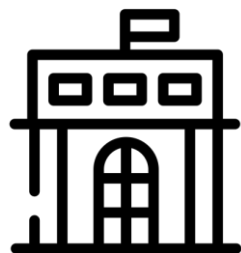
3.iii. Principais variáveis

A base de dados apresenta **20 variáveis explicativas**.



Variáveis cadastrais

- Escola
- Diretoria
- Município
- Tipo Escola
- Zona
- Latitude
- Longitude
- Quantidade de Alunos
- Quantidade de Salas de Aula
- Quantidade de Biblioteca



Variáveis do Ensino

- Indicador IDESP Anual



Variáveis relacionadas a exames de proficiência

- Série
- Ano
- Período
- Disciplina
- Média da proficiência



Variável Resposta

- Target:
- Indicador IDESP 2020





Definição do problema

- Objetivos
- Conceitos
- Critérios
- Histórico de dados
- Variáveis

Análise preliminar

- Medidas resumo
- Análise de *outliers*
- Análise de *missings*
- Validação com a área de negócios sobre a consistência das informações

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

Avaliação das técnicas

- Random Forest
- Boosting
- SVM
- Deep Learning

Definição da técnica

- Validação dos resultados com negócios
- Escolha da técnica que melhor se adequa ao uso e estratégias da área de negócios

