Técnicas de Aprendizado de Máquina para Predição do Desempenho de Escolas no Exame Nacional do Ensino Médio

João Gabriel Viana Hirasawa
Juliana Ferreira Alves
Matheus Malonda dos Santos Macaia
Paulo Henrique Dal Bello
Silvia Cristina de Jesus

Aprendizado de Máquina 1 - 2020.1
Prof. Dr. Diego Furtado Silva

Sumário

- Introdução
- Conjunto de dados
- Métodos de aprendizado
- Análise dos resultados
- Conclusão

Introdução

Motivação

- Aplicação do aprendizado de máquina para fins educacionais
  - Uso de dados para melhorar a qualidade da educação
- Relação de fatores socioeconômicos e infraestrutura das escolas com o desempenho dos estudantes
- Enem
  - Avaliação do desempenho escolar dos estudantes do Ensino Médio
  - Instrumento de ingresso no ensino superior

Objetivo

- Analisar a relação de indicadores socioeconômicos com os resultados no Enem
- Comparar modelos de regressão da nota média das escolas

Código no GitHub

https://github.com/joaohira/am-inep

Conjunto de Dados

Microdados INEP

- **Censo Escolar da Educação Básica**
  - Funcionamento das escolas: infraestrutura, gestores, professores, situação e rendimento escolar dos alunos
  - Alguns atributos: dependência administrativa, indicadores de fornecimento de água filtrada, existência de quadras de esportes, laboratórios etc.
- **Enem por Escola**
  - Desempenho médio por escola na prova do Enem
  - Alguns atributos: nota média de redação, taxa de aprovação no Ensino Médio, taxa de abandono, Indicador de Nível Socioeconômico da escola etc.

Construção do Conjunto

- **Seleção dos dados**
  - Filtragem dos dados do Enem por Escola para o ano de 2015
  - Uso do Censo Escolar de 2015
- **Junção das tabelas**
  - Inner merge na chave identificadora de cada observação (código da escola)
  - Associa os dados do Censo de cada escola a seu desempenho médio no Enem em 2015
- **Após junção**
  - **192** colunas
  - **15598** linhas

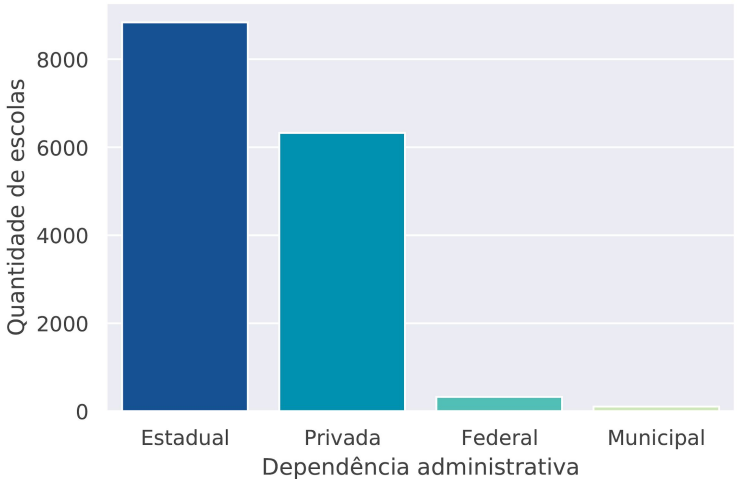
Análise Exploratória

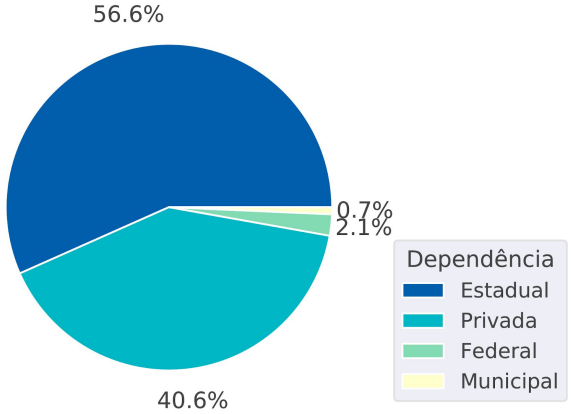
- Criar percepções dos dados, auxiliar o planejamento do processo científico, refinar hipóteses **(Behrens et al., 2012)**
- Uso das bibliotecas *pandas*, *seaborn* e *matplotlib*

seaborn

matplotlib

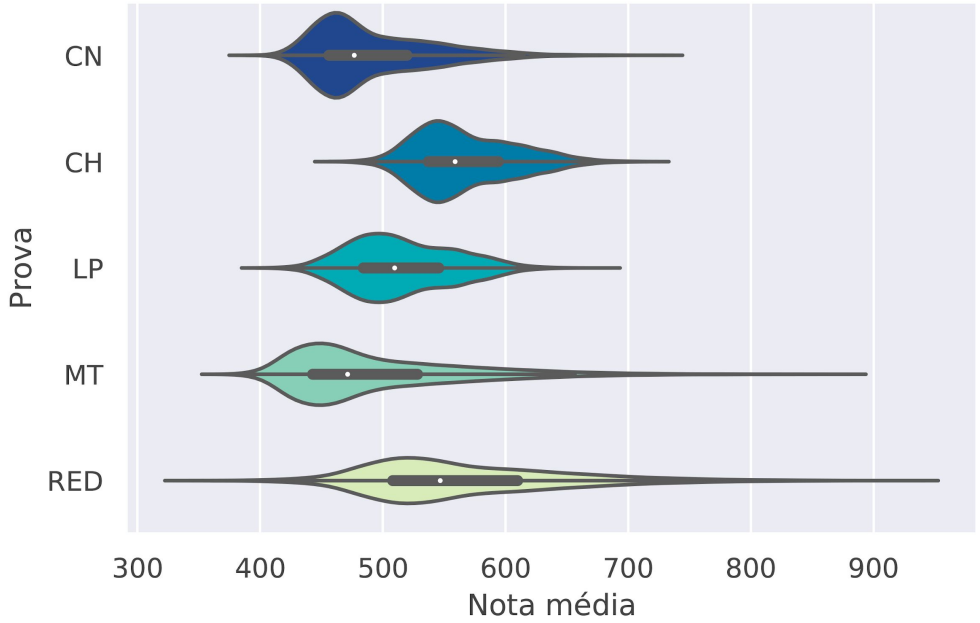
Distribuição do tipo de dependência administrativa



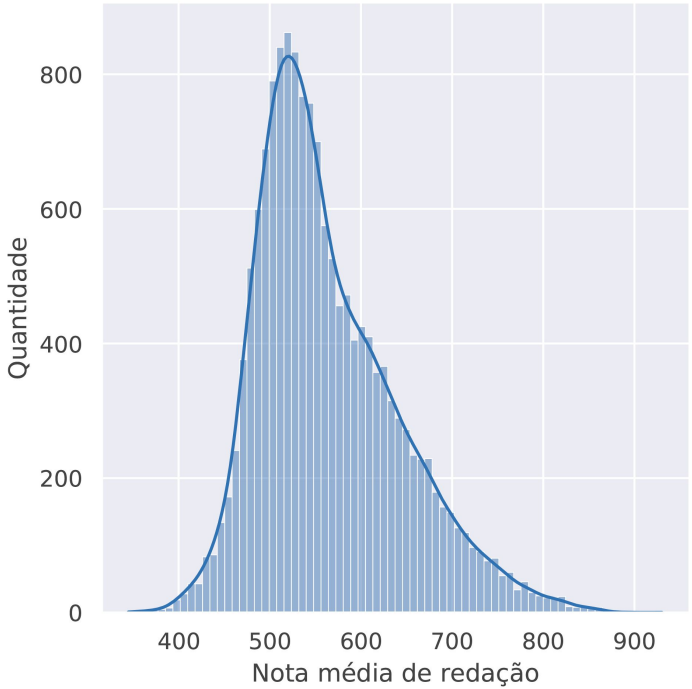


Distribuição da nota média por prova do Enem

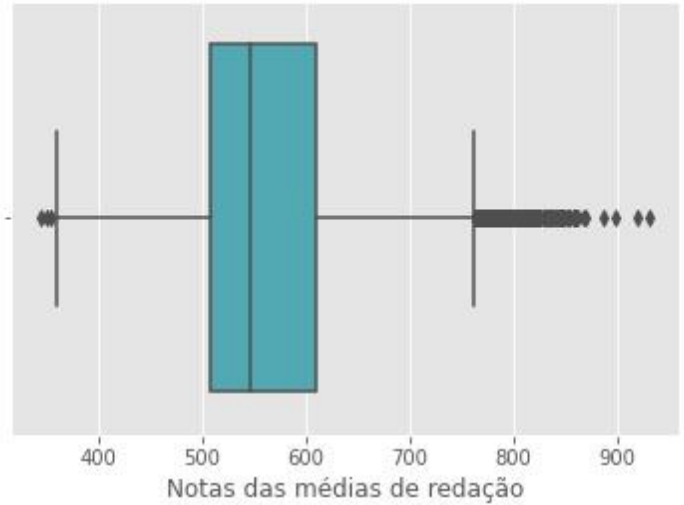
CN: Ciências da Natureza e suas Tecnologias
CH: Ciências Humanas e suas Tecnologias
LP: Linguagens, Códigos e suas Tecnologias
MT: Matemática e suas Tecnologias
RED: Redação



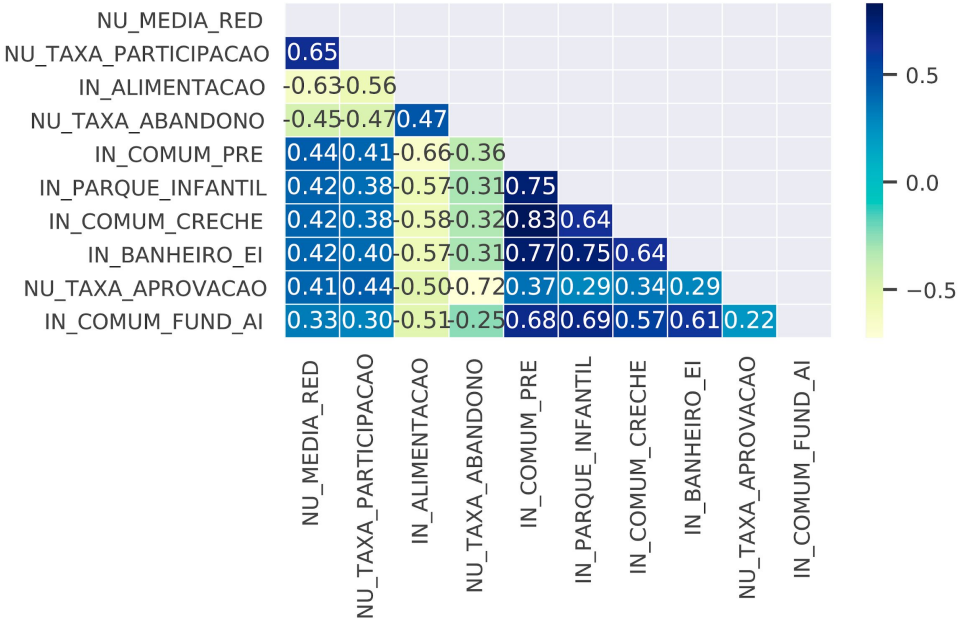
Histograma da nota média de redação



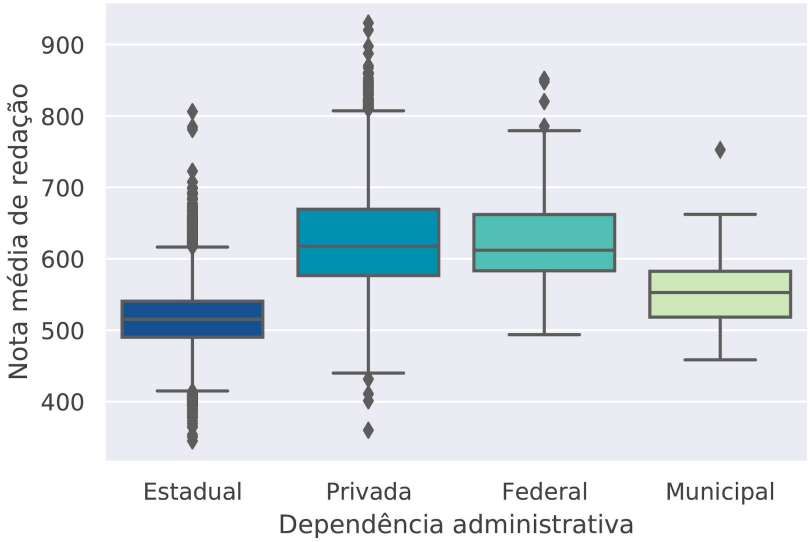
Boxplot da nota média de redação

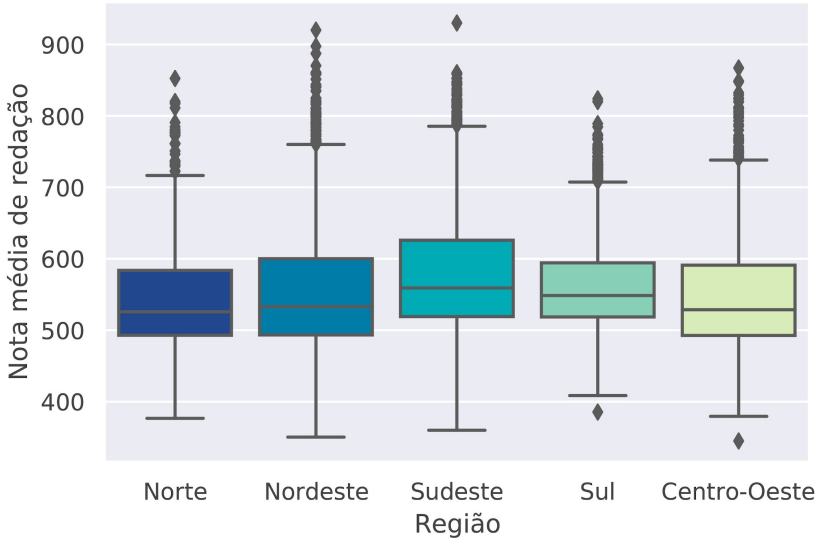


Mapa de calor dos atributos de maior correlação com a nota média de redação

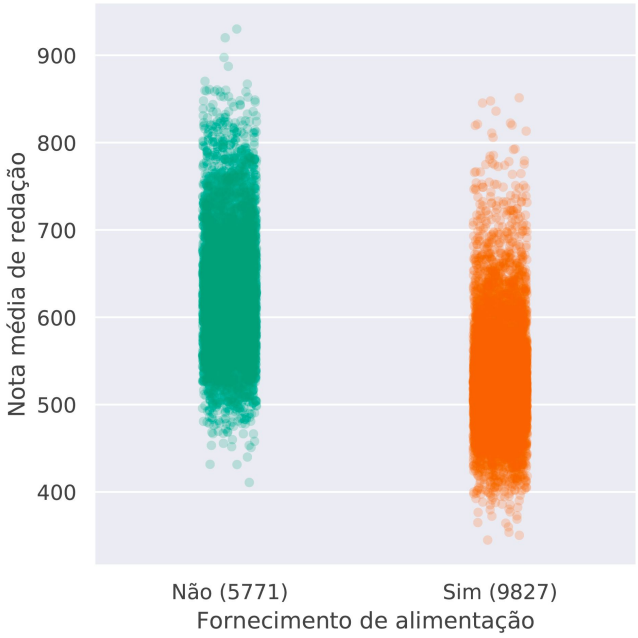


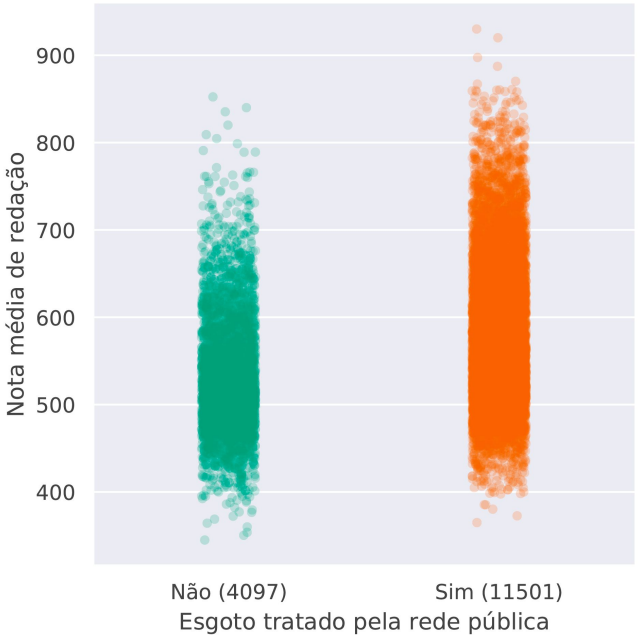
Distribuição da nota média de redação por dependência administrativa e por região





Distribuição da nota média de redação condicionada ao fornecimento de alimentação ou tratamento de esgoto





Limpeza dos Dados

- Detectar e remover erros e inconsistências para melhorar a qualidade dos dados
- Envolve várias fases
  - Inspeção dos dados
  - Definição dos passos de transformação
  - Regras de mapeamento
  - Verificação
  - Transformação
  - Aplicação dos dados

(Rahm, Do, 2000)

Limpeza dos Dados

Remoção de atributos duplicados
A junção das tabelas gerou duplicação de atributos referenciados de forma diferente nas duas tabelas

Remoção de dados faltantes
Atributos com poucos dados (proporção de presença < 4%)

Atributos presentes apenas para escolas privadas

Outras remoções
Atributos com variância (σ^2) nula

Imputação com KNN
O restante dos dados faltantes foi preenchido (baseado na proximidade com outras observações), adição de indicador

Limpeza dos Dados

One-hot encoding
Transformação dos atributos categóricos em atributos binários

n categorias → **n** atributos binários

Padronização dos valores
Para uso com Multi-layer Perceptron (MLP)

Padronização com min-max para o intervalo [0,1]

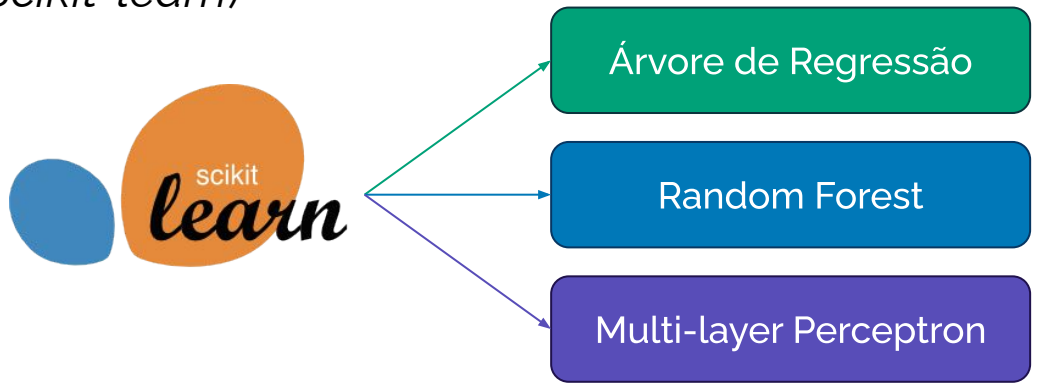
Antes da limpeza
192 colunas
15598 linhas

Após limpeza
200 colunas
15598 linhas

Métodos de Aprendizado

Métodos de Aprendizado

- **Tarefa definida:** regressão da nota média de redação das escolas na prova do Enem
- Uso de três modelos regressores distintos (*Scikit-learn*)



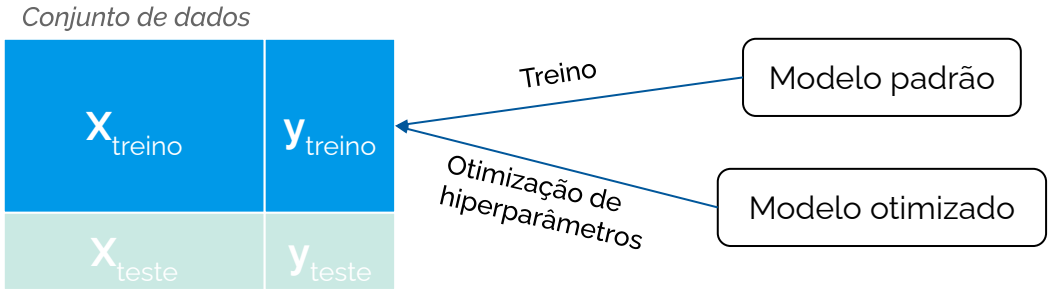
Etapas

1. Divisão dos dados

- Variáveis explicativas (**X**) e variável resposta (**y**)
- Amostragem com *holdout* 80% treino e 20% teste

2. Treinamento dos modelos

- Modelo sem ajuste de hiperparâmetros
- Modelo com *grid search* com validação cruzada 5-fold



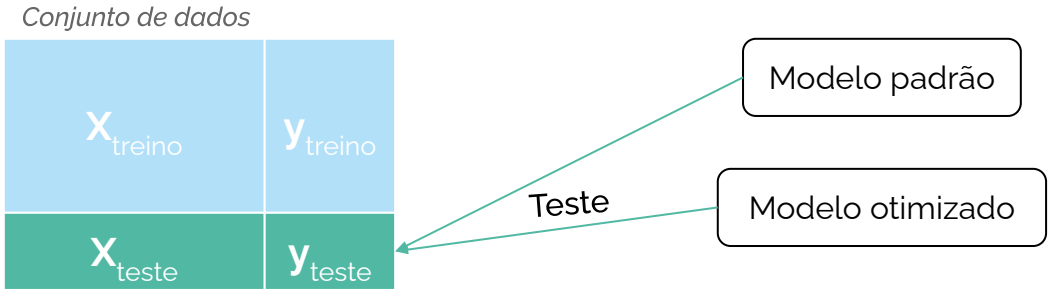
Etapas

1. Divisão dos dados

- Variáveis explicativas (**X**) e variável resposta (**y**)
- Amostragem com *holdout* 80% treino e 20% teste

2. Treinamento dos modelos

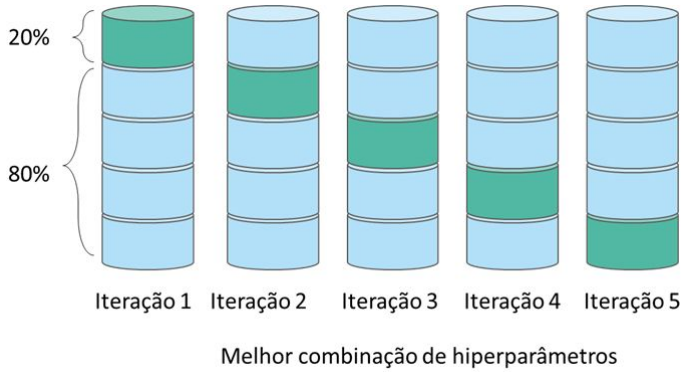
- Modelo sem ajuste de hiperparâmetros
- Modelo com *grid search* com validação cruzada 5-fold



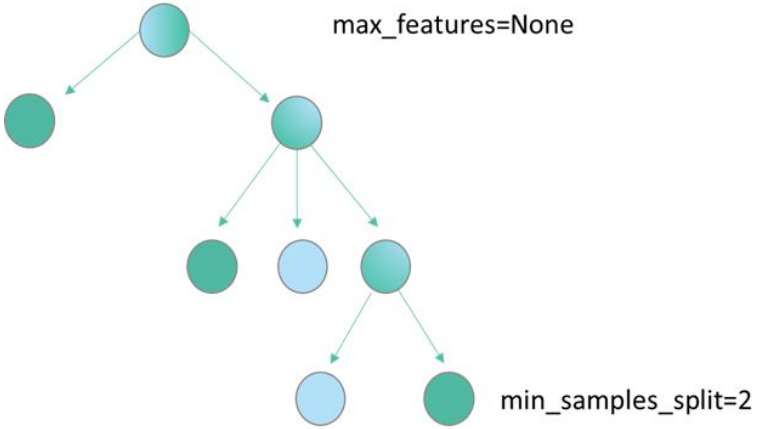
Etapas

2. Treinamento dos modelos

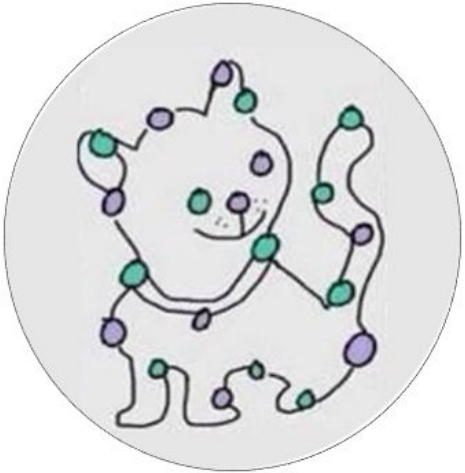
- Modelo com *grid search* com validação cruzada 5-fold



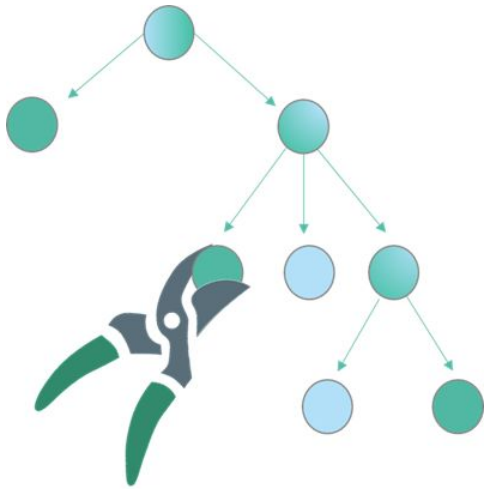
Árvore de Regressão



Overfitting



Árvore de Regressão



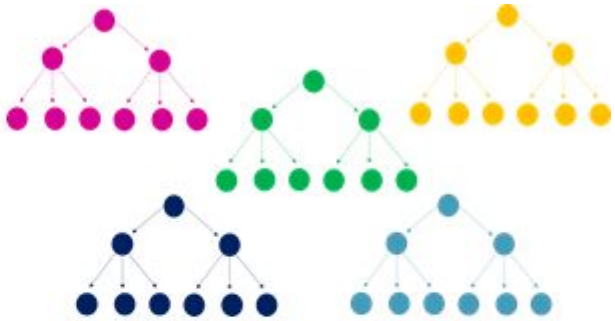
- **Hiperparâmetros otimizados:** altura máxima da árvore (`max_depth`) e número de atributos considerados na divisão dos nós (`max_features`)
- **Altura máxima**
  - Espaço de busca contendo nenhum limite (`None`) e um intervalo de inteiros `[2; 195)`
- **Número de atributos**
  - Espaço de busca contendo o conjunto `{195; 97; 48; 24; 12}`

Random Forest

max_features=None

min_samples_split=2

max_depth=None

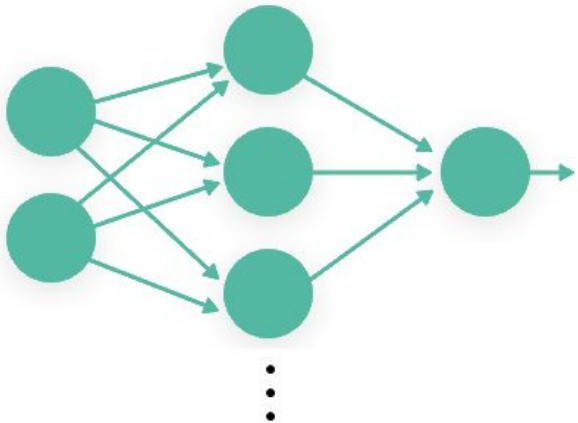


n_estimators=100

Random Forest

- **Hiperparâmetros otimizados:** número de árvores (`n_estimators`) e número de atributos considerados na divisão dos nós (`max_features`)
- **Número de árvores**
  - Espaço de busca contendo o conjunto `{50; 100; 150; 200; 250; 300}`
- **Número de atributos**
  - Espaço de busca contendo o conjunto `{195; 97; 48; 24; 12}`

Multi-layer Perceptron



hidden_layer_sizes=(100,)

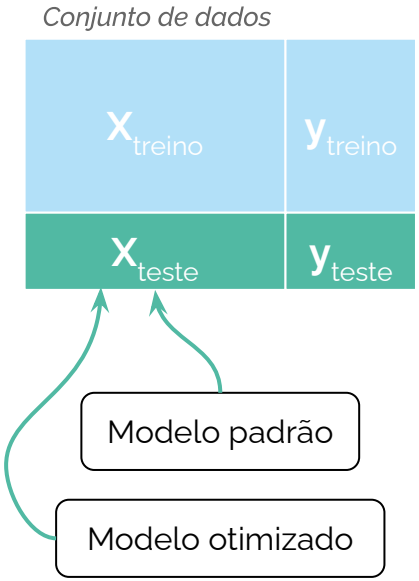
alpha=0.0001

Multi-layer Perceptron

- **Hiperparâmetros otimizados:** camadas ocultas (`hidden_layer_sizes`) e o termo de regularização (`alpha`)
- **Camadas ocultas**
  - Espaço de busca contendo o conjunto de tuplas `{(195,97,); (200,100,); (195,); (100,)}`
- **Termo de regularização**
  - Espaço de busca contendo o conjunto `{0,0001; 0,001; 0,01; 0,1; 1}`

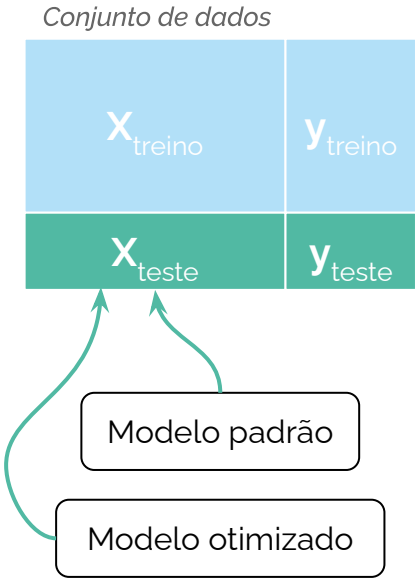
Análise dos Resultados

Método de Análise



- Aplicação de cada modelo no conjunto de teste
- Comparação das predições com os valores verdadeiros

Método de Análise



- Aplicação de cada modelo no conjunto de teste
- Comparação das predições com os valores verdadeiros
- **Métricas aplicadas**
  - Raiz do erro quadrático médio (RMSE)
  - Erro absoluto médio (MAE)

  → Erro médio das predições para os dados verdadeiros

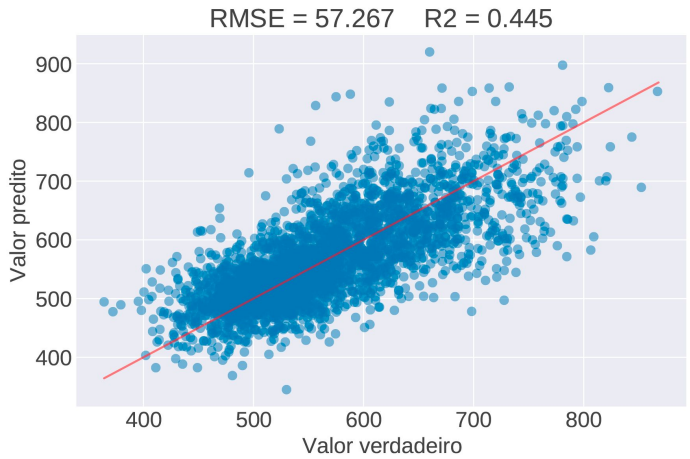
  - Coeficiente de determinação (R^2)

  → Proporção de ajuste do modelo para os dados

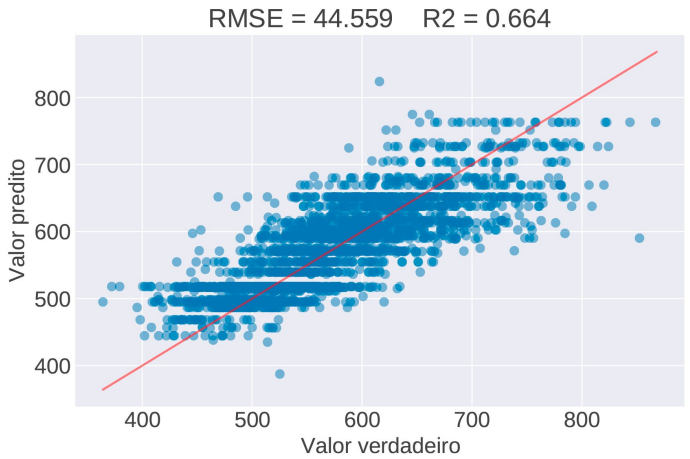
Árvore de Regressão

Método	RMSE	R^2	MAE
Sem ajuste	57,267	0,445	43,254
Grid search	**44,559**	**0,664**	**33,832**

Sem ajuste



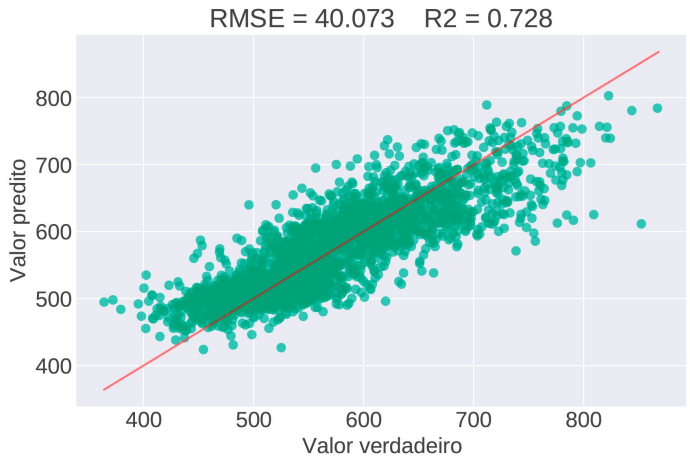
Com grid search



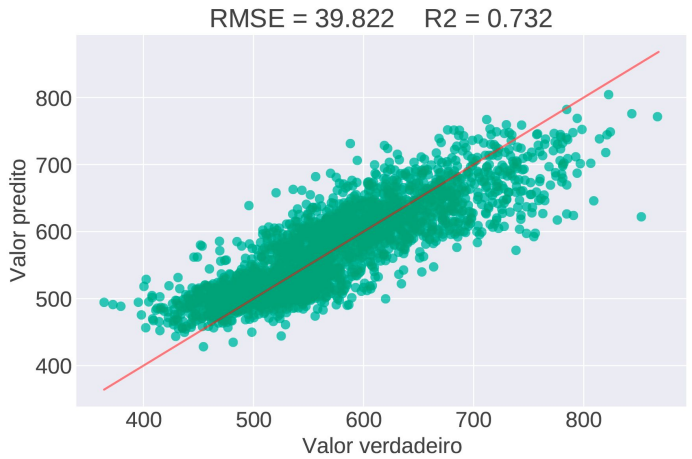
Random Forest

Método	RMSE	R^2	MAE
Sem ajuste	40,073	0,728	30,443
Grid search	**39,822**	**0,732**	**30,276**

Sem ajuste



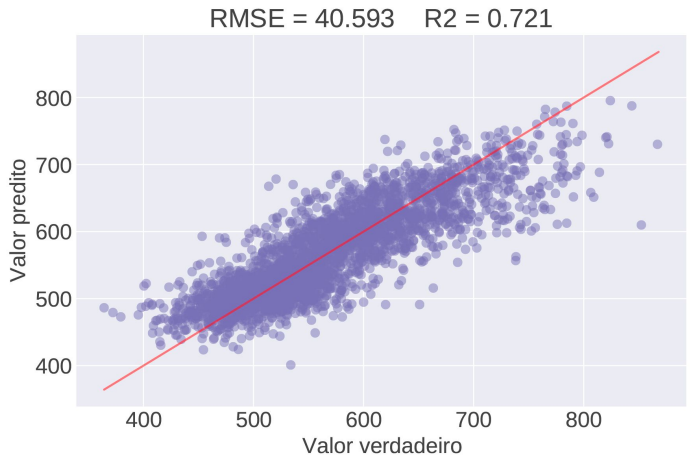
Com grid search



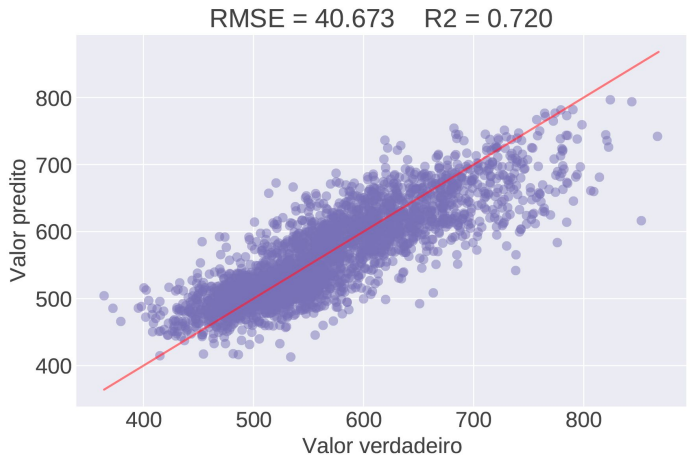
Multi-layer Perceptron

Método	RMSE	R^2	MAE
Sem ajuste	**40,593**	**0,721**	**30,555**
Grid search	40,673	0,720	30,394

Sem ajuste



Com grid search



Análise Comparativa

Modelo	Método	RMSE	R^2	MAE
Árvore de Regressão	Sem ajuste	57,267	0,445	43,254
Árvore de Regressão	**Grid search**	**44,559**	**0,664**	**33,832**
Random Forest	Sem ajuste	40,073	0,728	30,443
Random Forest	**Grid search**	**39,822**	**0,732**	**30,276**
Multi-layer Perceptron	**Sem ajuste**	**40,593**	**0,721**	**30,555**
Multi-layer Perceptron	Grid search	40,673	0,720	30,394

Conclusão

Conclusão

- *Random forest* foi o modelo de melhor desempenho
  - Pode explicar **73%** da variação dos dados
  - Árvore de regressão **66%** e multi-layer perceptron **72%**
- Modelos de aprendizado de máquina podem ser úteis a gestores educacionais
  - Direcionar ações de apoio a escolas de baixo desempenho

Referências

Behrens, John T.; Dicerbo, Kristen E.; Yel, Nedim; Levy, Roy. Exploratory Data Analysis. **Handbook Of Psychology**, [S.L.], v. 2, p. 34-70, 26 set. 2012. John Wiley & Sons, Inc. doi: 10.1002/9781118133880.hop202002.

Rahm, Erhard; Do, Hong Hai. Data Cleaning: problems and current approaches. **IEEE Data Engineering Bulletin**. Leipzig, p. 3-13. jan. 2000.