

# Interpretabilidade em Regressão Simbólica com a representação Interação-Transformação

Guilherme Seidyo Imai Aldeia  
Orientador: Prof. Dr. Fabrício Olivetti de França

Universidade Federal do ABC (UFABC)  
Programa de Pós Graduação em Ciência da Computação

31 de março de 2021

# Índice

- ❶ Introdução
- ❷ Revisão de literatura
- ❸ Interação-Transformação
- ❹ Experimentos preliminares
- ❺ Gerador de relatórios
- ❻ Conclusões parciais

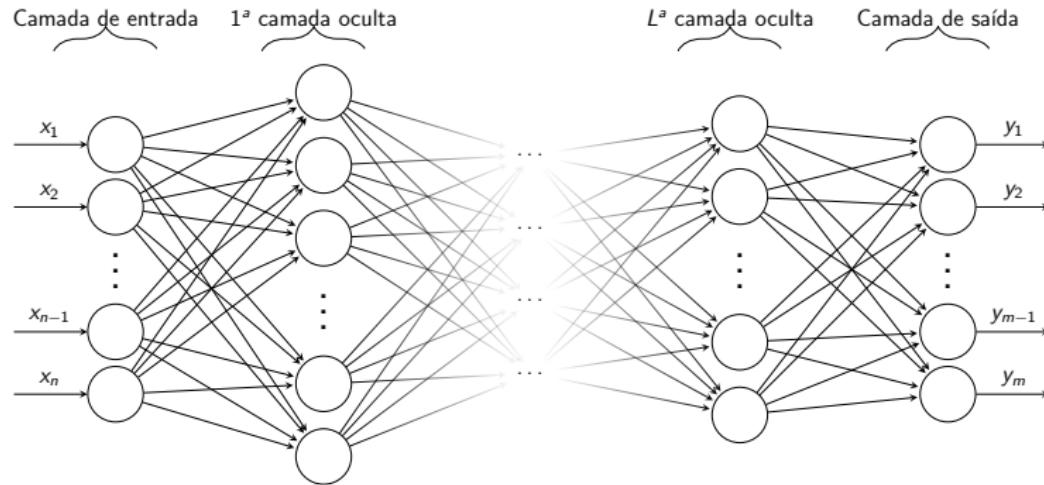
# Introdução à linha de pesquisa

Nas últimas décadas, o aprendizado de máquina teve:

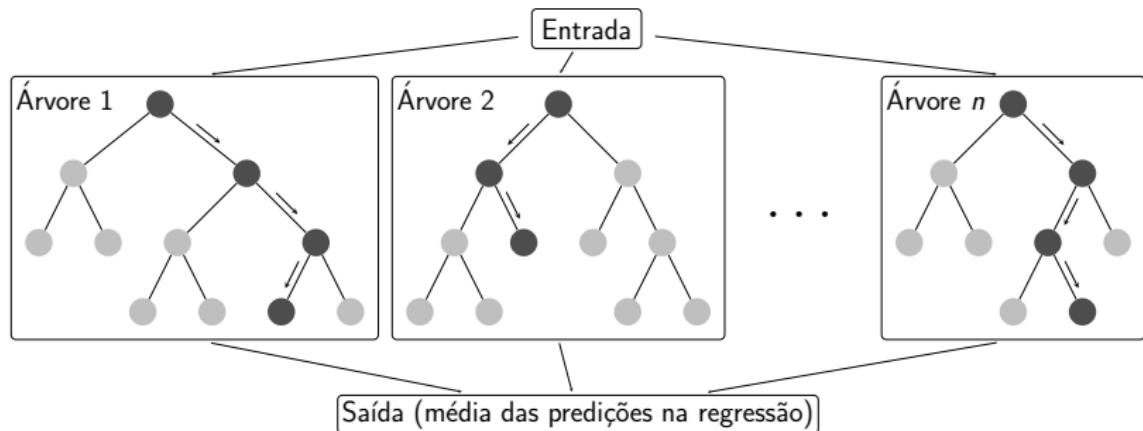
- ✓ Crescimento de uso;
- ✓ Aplicação nas mais diversas áreas;
- ✓ Desempenho competitivo com especialistas.

Surgimento de novos paradigmas para a tarefa.

# Introdução à linha de pesquisa



# Introdução à linha de pesquisa



# Introdução à linha de pesquisa

Porém, essas técnicas possuem desvantagens:

- ✗ Geram modelos complexos;
- ✗ Difícil análise e inspeção;
- ✗ Resultado não é compreendido por humanos.

Classificando-as como **caixas-pretas não interpretáveis**.

# Introdução à linha de pesquisa

Existe um debate sobre o uso de caixas-pretas:

- Caixas-pretas são confiáveis?
- Devemos usar apenas modelos comprehensíveis?
- Todo problema precisa de uma explicação?

# Introdução à linha de pesquisa

Essas questões culminaram no "surgimento" de uma nova área:  
**eXplainable Artificial Intelligence (XAI)**

## Missão da XAI

**Obter explicações** do funcionamento dos modelos, através da criação de **modelos mais interpretáveis** ou do uso de **métodos de explicação de modelos**.

# A Regressão Simbólica

A Regressão Simbólica é uma alternativa caixa-cinza, com:

- ✓ Resultados competitivos;
- ✓ Potencial de obtenção de resultados mais interpretáveis;
- ✗ Alto custo computacional;
- ✗ Amplo espaço de busca.

Pesquisadores da área buscam propor diferentes mecanismos para aliviar esses problemas.

# Objetivos

- Estudar métodos de explicação de modelos;
- Propor uma adaptação ou nova forma para explicar modelos de regressão simbólica;
- Automatizar o processo de obter explicações.

# Índice

① Introdução

② Revisão de literatura

③ Interação-Transformação

④ Experimentos preliminares

⑤ Gerador de relatórios

⑥ Conclusões parciais

# O que é interpretabilidade?

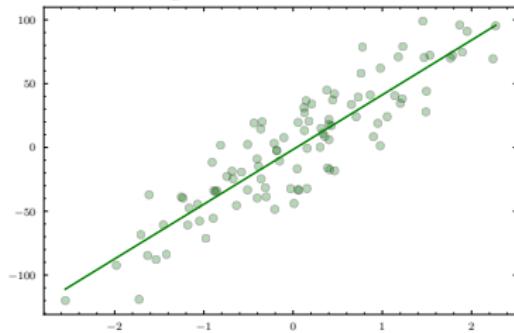
Interpretabilidade possui diferentes definições na literatura.

O que é um modelo interpretável?

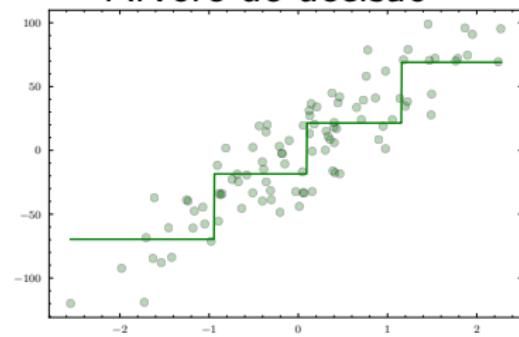
Aquele que pode ser **analisado, simulado, decomposto** em componentes individuais e **entendido** a nível de treinamento.

# O que é um modelo interpretável?

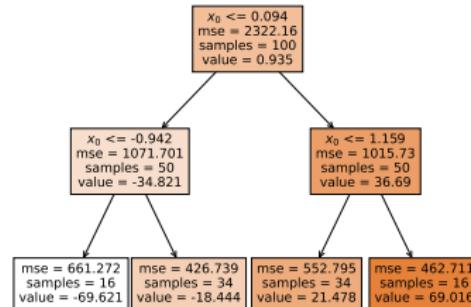
Regressão linear



Árvore de decisão



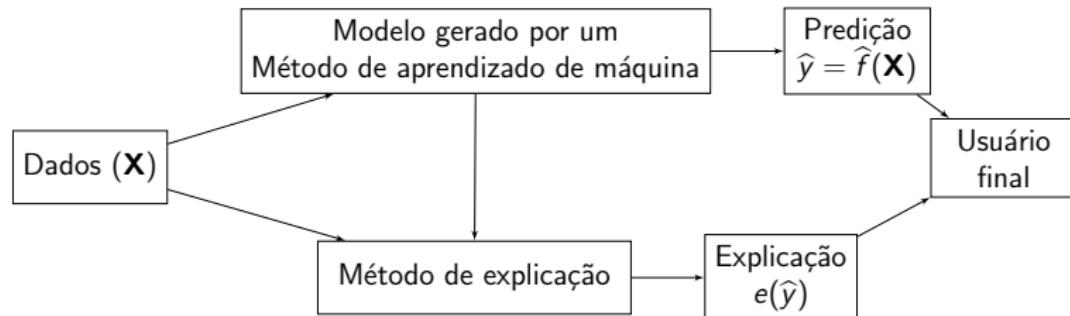
$$\hat{f}(x) = 42.853x_0 + -1.628$$



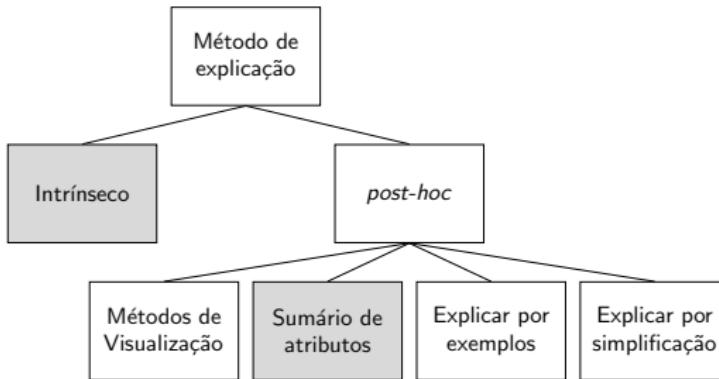
# Lidando com modelos não interpretáveis

Para obter resultados interpretáveis de caixas-pretas/cinzas:

- Substituir por uma caixa-branca;
- Explicar o modelo com um **método de explicação**:



# Métodos de explicação



Característica	Abordagens	
Dependência do modelo	Específico de modelo	Agnóstico de modelo
Interpretabilidade do modelo	Intrínseco	<i>Post-hoc</i>
Escopo da explicação	Local	Global

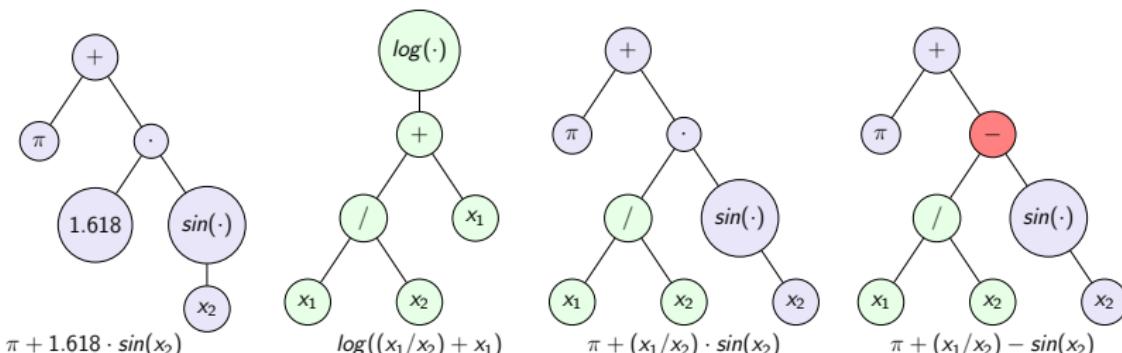
# Regressão Simbólica: uma alternativa caixa-cinza

Pode ser utilizada para obter modelos interpretáveis ou explicar modelos:

- Método que retorna uma expressão simbólica;
- Melhor capacidade de aproximação que a regressão linear.

# Regressão Simbólica: uma alternativa caixa-cinza

Uma **população** de soluções (árvores) passam por um **processo evolutivo** guiado por uma **função de aptidão**, onde **operadores genéticos** criam descendentes que **herdam características** de seus pais.



# Efeito marginal: o explicador da regressão

Popular na econometria para análise de modelos lineares generalizados, mede como uma mudança infinitamente pequena em uma variável afeta a resposta:

$$ME_j = \frac{\partial}{\partial x_j} \hat{f}(\mathbf{x}, \beta). \quad (1)$$

Pode ser utilizado em diferentes cálculos:

- *Average Marginal Effects* (AME);
- *Marginal Effects at the Means* (MEM);
- *Marginal Effects at Representative values* (MER).

# Como avaliar explicações?

Usamos caixas-brancas ou explicadores para obter explicações, mas como avaliar sua qualidade?

- Envolvendo humanos na avaliação:
  - Problemas do mundo real e especialistas;
  - Problemas com menos especificidade e pessoas normais;
  - Sem envolvimento de humanos.
- Utilizando dados de diferentes naturezas:
  - Dados reais *vs* dados sintéticos.

# Problemas em aberto identificados

- ✗ Qualidade de métodos de explication é questionável;
  - (é difícil estabelecer métricas para avaliar explicações);
- ✗ Regressão recebe pouca atenção em relação à classificação;
- ✓ Regressão simbólica tem potencial de encontrar soluções mais simples;
- ✓ O efeito marginal pode ser explorado no contexto da regressão simbólica.

# Índice

- ① Introdução
- ② Revisão de literatura
- ③ **I**nteração-Transformação
- ④ Experimentos preliminares
- ⑤ Gerador de relatórios
- ⑥ Conclusões parciais

# Uma representação alternativa às árvores

A representação Interação-Transformação é uma alternativa à representação de árvores para a regressão simbólica:

- Restringe o espaço de busca;
- Prioriza funções mais simples e interpretáveis.

# A representação Interação-Transformação

Seja uma entrada  $(\mathbf{x}, y)$ , com  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ .

- **A interação** das variáveis é o monômio:

$$p(\mathbf{x}) = \prod_{i=1}^n x_i^{k_i},$$

em que  $\mathbf{k} \in \mathbb{Z}^n$  é um vetor de expoentes para cada variável;

- **A transformação** é a aplicação de uma função  $t$ :

$$t : \mathbb{R} \rightarrow \mathbb{R}.$$

# A representação Interação-Transformação

- **Termo IT** é a tupla  $(t, \mathbf{k})$ , que constrói a composição:

$$TermoIT(\mathbf{x}) = (t \circ p)(\mathbf{x}).$$

- **Expressão IT** é a combinação linear de  $m$  Termos IT:

$$ExprIT(\mathbf{x}) = \sum_{j=1}^m w_j \cdot TermoIT_j(\mathbf{x}) + w_0,$$

em que  $w_j \in \mathbb{R}$  são coeficientes e  $w_0$  é o intercepto, valores obtidos com métodos de otimização.

# Intuições por trás da representação

- Não-linearidade: expoentes e função de transformação:
  - Cria variáveis transformadas, que podem ter relação linear com a variável dependente.
- Coeficientes ajustam a contribuição dos termos;
- Permite soluções simples, polinômios e funções complexas.

# *Interaction-Transformation Evolutionary Algorithm*

- Algoritmo evolutivo recentemente publicado;
- Resultados competitivos e mais simples;
- Se baseia apenas em mutações:
  - *subsInter, adicionar, interPos, interNeg, remover;*
  - Realizam modificações mínimas locais;
  - Bom desempenho em gerar melhorias a cada geração.

# Índice

- ① Introdução
- ② Revisão de literatura
- ③ Interação-Transformação
- ④ Experiments preliminaires
- ⑤ Gerador de relatórios
- ⑥ Conclusões parciais

# Estudo de caso

- Combinar diferentes métodos de explicação e regressão;
- Propor uma metodologia para comparar as combinações;
- Avaliar o efeito marginal com a regressão simbólica.

# Métodos de regressão e explicação avaliados

Métodos de regressão:

- *Interaction-Transformation Evolutionary Algorithm* (ITEA);
- *Random Forest* (RF);
- *eXtreme Gradieng Boosting* (XGB).

Métodos de explicação de **sumário de atributos**:

- *Local Interpretable Model-agnostic Explanations* (LIME);
- *SHapley Additive exPlanations* (SHAP);
- *Explain Like I'm 5* (ELI5);
- *Random Explainer* (RandE);
- *Marginal Effects* (ME).

# Dados utilizados

Foram utilizados dados artificiais de 100 equações da física, em duas partições:

- Treino:
  - 100 observações geradas com distribuição uniforme;
  - Treino dos métodos;
  - Avaliação da explicação global.
- Teste:
  - 100 observações geradas com o *Latin Hypercube Sampling* (LHS);
  - Avaliação da explicação local.

# Metodologia

- Foi feito *grid-search* em todas as execuções do RF e XGB;
- Valores do ITEA definidos de forma sistemática;
- 30 execuções para cada base de dados;
- Gradiente da função original como explicação esperada:
  - Aponta para direção de maior incremento possível.

# Medida de qualidade da explicação

Erro de explicação medido pela equação:

$$ExErr(\mathbf{I}, \widehat{\mathbf{I}}) = \sqrt{\sum_{i=1}^d \frac{(rank(|\mathbf{I}|)_i - rank(|\widehat{\mathbf{I}}|)_i)^2}{d}},$$

com  $\mathbf{I}$  a matriz com a explicação real para cada linha (observação x variável),  $\widehat{\mathbf{I}}$  a importância dada pelos métodos de explicação,  $d$  é o número de observações e  $rank$  uma função que retorna as classificações.

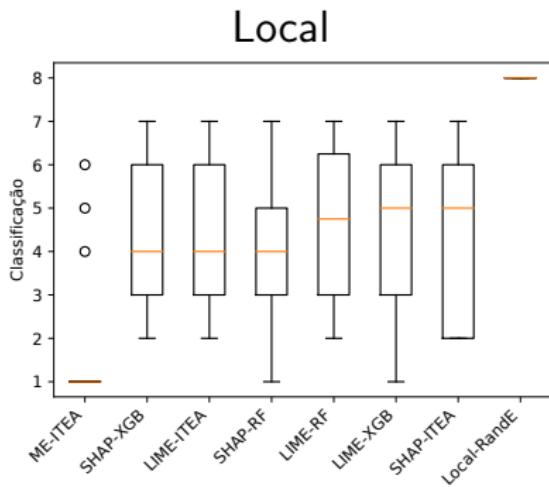
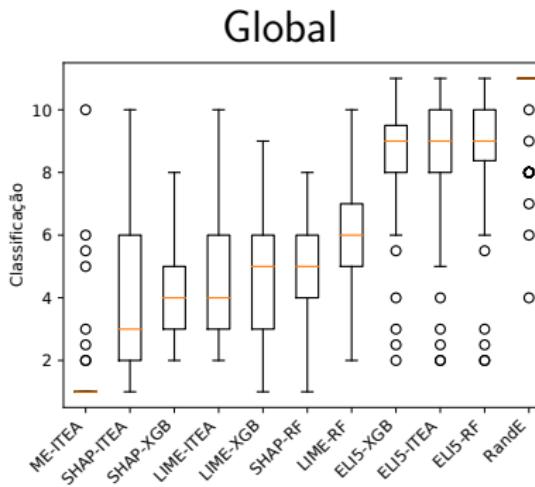
# Resultados e discussão

Todos os experimentos foram executados nas mesmas condições.  
Em relação ao tempo:

	ITEA	XGB	RF
Treino	$108.36 \pm 33.65$	$0.02 \pm 0.00$	$0.06 \pm 0.01$
SHAP	$9.79 \pm 13.61$	$5.64 \pm 6.74$	$7.55 \pm 6.85$
LIME	$358.47 \pm 123.14$	$336.40 \pm 112.87$	$319.35 \pm 101.49$
ELI5	$0.01 \pm 0.01$	$0.01 \pm 0.00$	$0.10 \pm 0.03$
ME	$0.46 \pm 0.24$	-	-

# Resultados e discussão

Classificando as combinações de explicador-regressor pela mediana do erro para cada base de dados:



# Resultados e discussão

Relações entre o ME, o SHAP e o LIME:

- O SHAP busca aproximar os valores Shapley:

$$\phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(F - |S| - 1)!}{F!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)],$$

- O LIME pode ser ajustado para aproximar o SHAP:

$$\xi = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g),$$

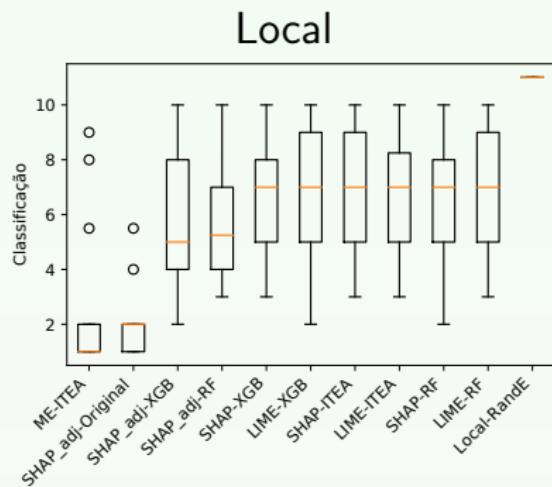
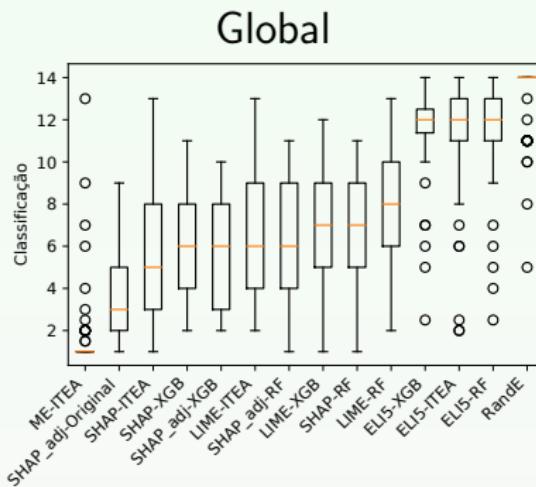
- O SHAP e o ME possuem uma relação aproximada de:

$$\phi_i(f, x) = \frac{\partial}{\partial x_i} f(x) \cdot (x_i - \bar{x}_i).$$

Isso relaciona o gradiente da função e valores Shapley.

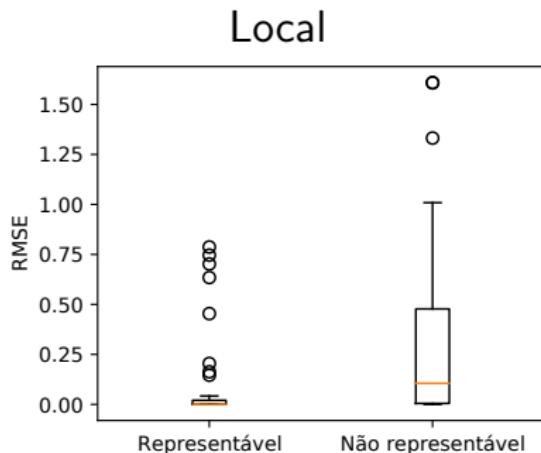
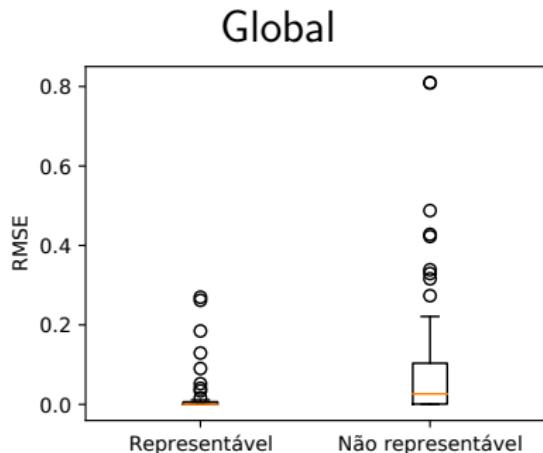
# Resultados e discussão

Podemos incluir o cálculo do SHAP ajustado (comparação mais justa):



# Resultados e discussão

Podemos observar o erro de explicação do ME-ITEA separando equações representáveis e não representáveis:



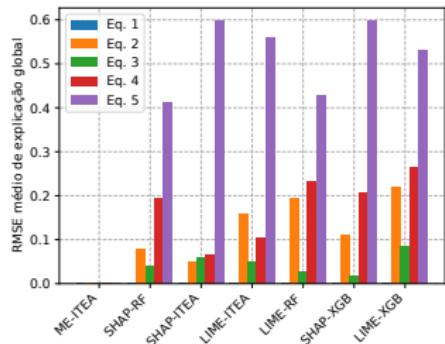
# Resultados e discussão

Seleção de 4 casos com 5 equações cada, de acordo com quando o ME-ITEA apresentou os:

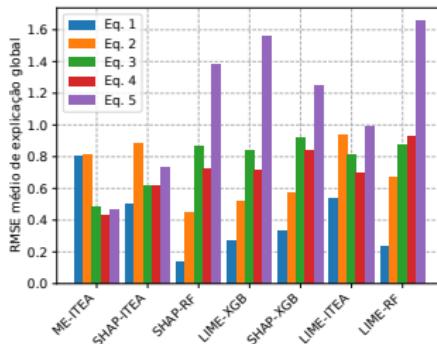
- (a) Melhores resultados para explicações globais;
- (b) Piores resultados para explicações globais;
- (c) Melhores resultados para explicações locais;
- (d) Piores resultados para explicações locais.

# Resultados e discussão

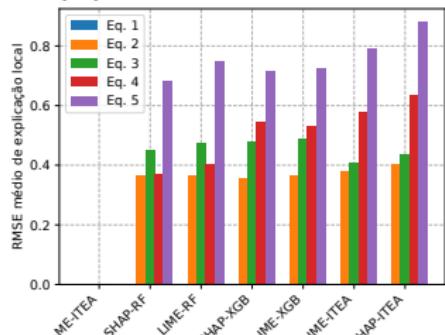
(a) Melhores globais



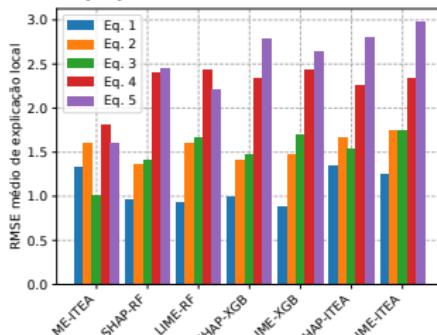
(c) Piores globais



(b) Melhores locais



(d) Piores locais



# Índice

- ① Introdução
- ② Revisão de literatura
- ③ Interação-Transformação
- ④ Experimentos preliminares
- ⑤ Gerador de relatórios**
- ⑥ Conclusões parciais

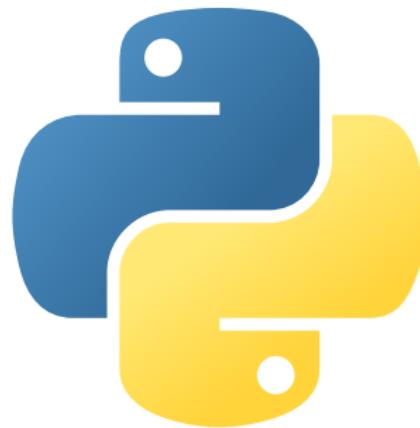
# Automatização do processo de interpretação

O ITEA e a representação IT permitem explicações nos níveis:

- Pré-execução: informações de pré-execução do algoritmo;
- Relacionados à execução: estatísticas, gráficos e equações;
- Pós-execução: uso de métodos *post-hoc*.

# Demonstração

Desenvolvimento de uma biblioteca que automatiza o processo de obtenção de explicações dos resultados do ITEA:



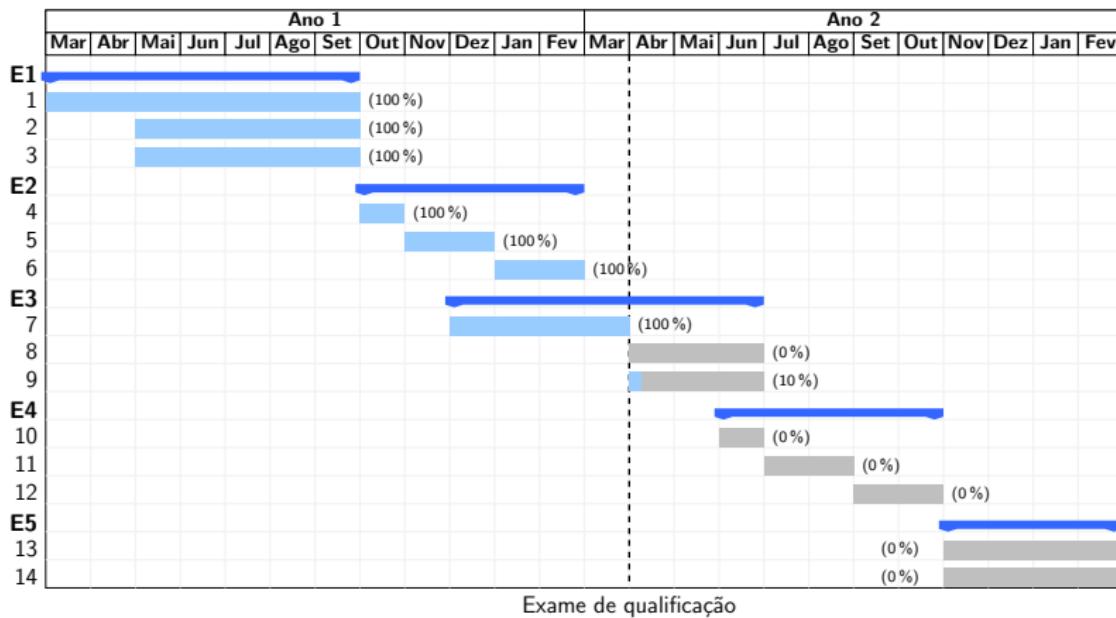
# Índice

- ① Introdução
- ② Revisão de literatura
- ③ Interação-Transformação
- ④ Experimentos preliminares
- ⑤ Gerador de relatórios
- ⑥ Conclusões parciais

# Conclusões parciais

- Modelos caixas-pretas são funções de difícil compreensão;
- A regressão simbólica pode obter resultados mais simples;
- O efeito marginal com o ITEA obteve os melhores resultados no experimento;
- É possível automatizar parcialmente a geração de um relatório de interpretabilidade.

# Cronograma



# Próximas etapas

- E3 Desenvolvimento das **próximas tarefas** e das sugestões levantadas pela banca;
- E4 Realização de experimentos finais;
- E5 Escrita de um artigo científico e dissertação.

# Detalhes das próximas tarefas

As próximas tarefas planejadas para os experimentos finais são:

- Estudo de integração automática de expressões IT;
- Incluir mais explicadores e regressores para comparação;
- Elaboração de mais métricas para comparar as explicações;
- Refinar o relatório automático.

# Resultados aceitos para publicação

Measuring Feature Importance using Marginal Effects with  
Symbolic Regression.

ALDEIA, G. S. I.; de FRANÇA, F. O.

*2021 ACM Genetic and Evolutionary Computation Conference  
(GECCO)*

# Obrigado!

Interpretabilidade em Regressão Simbólica com a representação  
Interação-Transformação

Guilherme Seidyo Imai Aldeia ([guilherme.aldeia@ufabc.edu.br](mailto:guilherme.aldeia@ufabc.edu.br))  
Orientador: Prof. Dr. Fabrício Olivetti de França ([folivetti@ufabc.edu.br](mailto:folivetti@ufabc.edu.br))

# Exemplos de expressões

Equação	Variáveis	Representável
$F = q \cdot (E_f + B \cdot v \cdot \sin(\theta))$	$q, E_f, B, v, \theta$	Não
$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}}$	$m_0, v, c$	Não
$v = \frac{mu_{drift} \cdot q \cdot Volt}{d}$	$mu_{drift}, q, Volt, d$	Sim
$K = \frac{1}{2} \cdot m \cdot (v^2 + u^2 + w^2)$	$m, v, u, w$	Sim

# Representação computacional

$$f(x_1, x_2) = 3.5x_2 + -2.0\cos(x_1) + 15.0$$

```
ExpressaoIT = {
    termos      :: [[Inteiro]]          = [[0, 1], [1, 0]],
    funcoes     :: [nomes de funções]   = ["id", "cos"],
    w           :: [Real]              = [3.5, -2.0],
    intercepto :: Real                = 15.0,
    aptidao    :: Real                = 2.19
}
```

# Pseudo código do ITEA

---

**Algoritmo 1:** Pseudo-código do ITEA.

---

**Entrada:** Variáveis independentes  $\mathbf{X}$  e variável dependente  $\mathbf{y}$

**Saída :** Função simbólica  $\hat{f}$

```
1 pop ← GerarPopAleatoria();
2 enquanto Stop criteria not met faça
3   child ← Ø;
4   para p ∈ pop faça
5     elegibleMut ← {subsInter };
6     se p.nTerms < mt então
7       elegibleMut ← elegibleMut ∪ {add, interPos, interNeg };
8     se p.nTerms > 1 então
9       elegibleMut ← elegibleMut ∪ {drop };
10    f ← Mutacao(p, elegibleMut);
11    f ← RemoveTermosInvalidos(f);
12    se f.nTerms ≥ 1 então
13      child ← child ∪ {f };
14  pop ← Selecao(pop ∪ child);
15 retorna arg max {score (p) for p ∈ pop };
```

---

# Resultados e discussão

Caso	Equação	Caso	Equação
<i>a</i>	$e^{-\frac{\theta^2}{2}} / \sqrt{2\pi}$	<i>b</i>	$\frac{p_d}{4 \cdot \pi \cdot \epsilon} \cdot 3 \cdot z / r^5 \cdot \sqrt{x^2 + y^2}$
	$m_0 / \sqrt{1 - v^2/c^2}$		$\frac{p_d \cdot Ef \cdot t \cdot 2 \cdot \pi}{h} \cdot \frac{\sin((\omega - \omega_0) \cdot t / 2)^2}{((\omega - \omega_0) \cdot t / 2)^2}$
	$\mu \cdot Nn$		$x1 \cdot (\cos(\omega \cdot t) + \alpha \cdot \cos(\omega \cdot t)^2)$
	$q1 \cdot r / (4 \cdot \pi \cdot \epsilon \cdot r^3)$		$Int_0 \cdot \frac{\sin(n \cdot \theta / 2)^2}{\sin(\theta / 2)^2}$
	$q2 \cdot Ef$		$\sin(E_n \cdot t / (h / (2 \cdot \pi)))^2$
<i>c</i>	$e^{-\frac{\theta^2}{2}} / \sqrt{2\pi}$	<i>d</i>	$\frac{p_d}{4 \cdot \pi \cdot \epsilon} \cdot 3 \cdot z / r^5 \cdot \sqrt{x^2 + y^2}$
	$m_0 / \sqrt{1 - v^2/c^2}$		$\frac{p_d \cdot Ef \cdot t \cdot 2 \cdot \pi}{h} \cdot \frac{\sin((\omega - \omega_0) \cdot t / 2)^2}{((\omega - \omega_0) \cdot t / 2)^2}$
	$\mu \cdot Nn$		$x1 \cdot (\cos(\omega \cdot t) + \alpha \cdot \cos(\omega \cdot t)^2)$
	$q1 \cdot r / (4 \cdot \pi \cdot \epsilon \cdot r^3)$		$\frac{G \cdot m1 \cdot m2}{((x2 - x1)^2 + (y2 - y1)^2 + (z2 - z1)^2)}$
	$q2 \cdot Ef$		$q \cdot (Ef + B \cdot v \cdot \sin(\theta))$