



# UNIVERSIDADE DE ITAÚNA

## Aprendizado de Máquina Aplicado à Valoração de Redações

---

**Graduando:** Eugênio Cunha

**Orientador:** Dr. Marco Túlio Alves N Rodrigues

19 de Junho de 2017

Departamento de Ciência da Computação  
Bacharelado em Ciência da Computação

## 1. Introdução

Problema de Pesquisa

Objetivos

Motivação

## 2. Trabalhos Relacionados

## 3. Metodologia

## 4. Resultados Preliminares

## 5. Plano de Trabalho

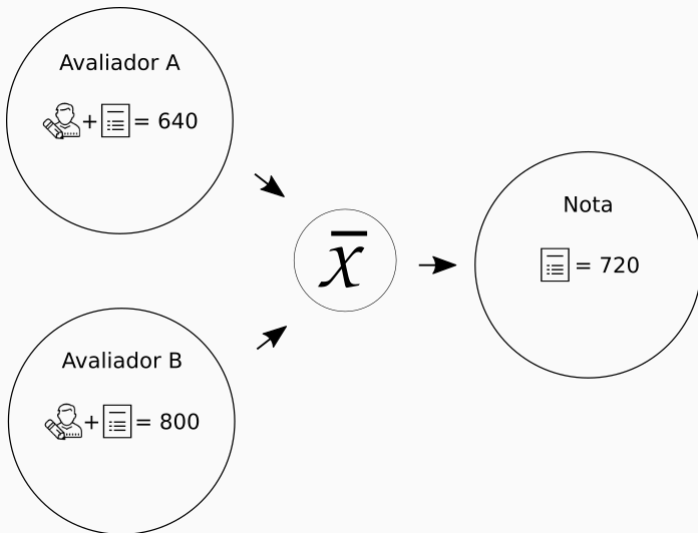
# Introdução

---

O decreto 79.298, de 24 de Fevereiro de 1977 definiu a “inclusão obrigatória da prova ou questão de redação em língua portuguesa” nos concursos e vestibulares (Art. 1º, alínea d).

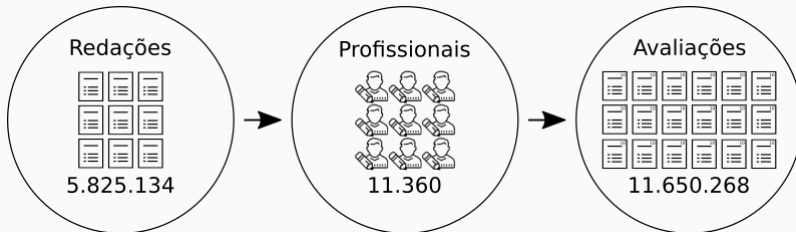
# Problema de Pesquisa

No ENEM cada redação é avaliada por, pelo menos, dois avaliadores, de forma independente [4].



# Problema de Pesquisa

Dados da avaliação de redações do ENEM 2016 [1].



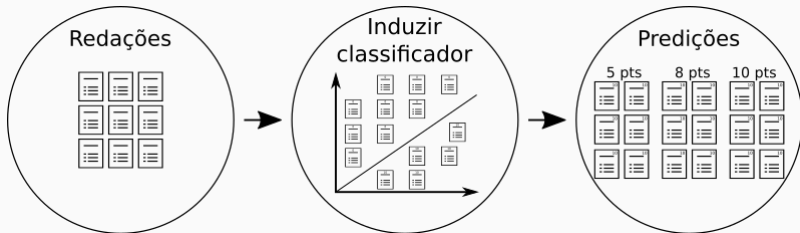
# Problema de Pesquisa

Dado um corpus de redações, classificar as competências exigidas em um texto de redação.



# Objetivos

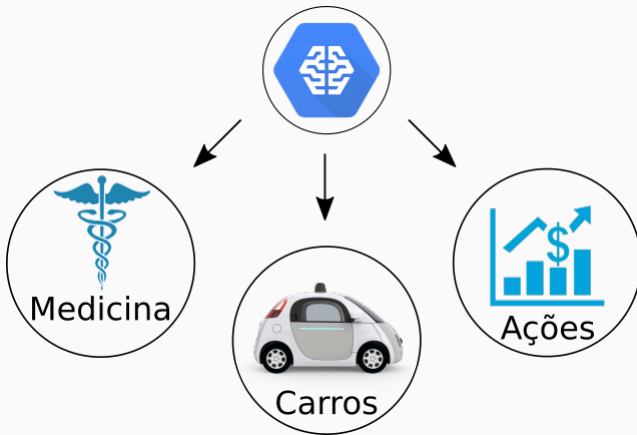
Induzir um modelo de Aprendizado de Máquina a classificar as competências exigidas em um texto de redação.





# Motivação

Aprendizado de Máquina está no centro de muitos avanços tecnológicos, alcançando áreas antes exclusivas de seres humanos.



# Trabalhos Relacionados

---

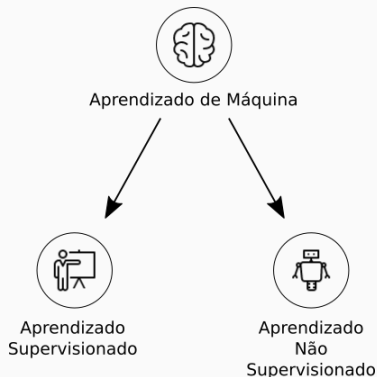
# Matriz de Competências

Silva (2017) cita em seu estudo, que à prova de redação do ENEM é avaliada levando em conta uma matriz de referência elaborada pelo INEP [2].

I	Demonstrar domínio da norma padrão da língua escrita.	200
II	Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.	200
III	Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.	200
IV	Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.	200
V	Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.	200

# Aprendizado de Máquina

Segundo Monard (2003), de uma forma geral o aprendizado indutivo pode ser dividido em supervisionado e não-supervisionado [5].



Freitas (2005), cita que o aprendizado supervisionado exige como entrada um corpus de treino, com exemplos corretamente rotulados [3].

## Supervisionado ou não-supervisionado?

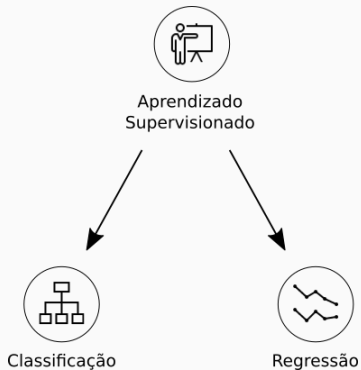
<b>Tema</b>	<i>Direitos em conflito: liberdade de expressão e ...</i>
<b>Título</b>	<i>Os limites da informação</i>
<b>Texto</b>	<i>Analizando todo um conjunto de fatos importantes ...</i>
<b>Competência I</b>	50
<b>Competência II</b>	100
<b>Competência III</b>	50
<b>Competência IV</b>	50
<b>Competência V</b>	50
<b>Nota Total</b>	<b>300</b>

Tabela 1: Propriedades de uma redação.

**Aprendizado supervisionado.**

# Aprendizado de Máquina

De acordo com o estudo de Monard (2003) o aprendizado supervisionado pode ser induzido a resolver problemas de classificação ou regressão [5].



O autor ainda cita, “Para rótulos de classe discretas, esse problema é conhecido como classificação e para valores contínuos como regressão.”

## Classificação ou Regressão?

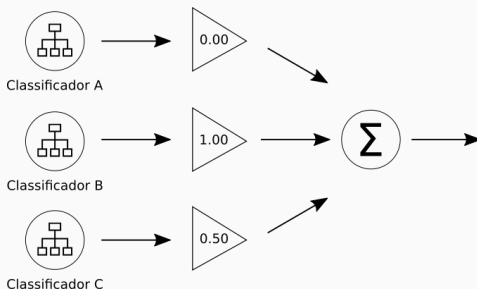
Competência		Valor	Subitens	Valor/Classes
			0	0
			1	50
I	<i>Demonstrar domínio da norma padrão da língua escrita.</i>	200	2	100
			3	150
			4	200

**Tabela 2:** Competência I da matriz de referência.

**Problema de classificação.**

## Boost -> AdaBoost

O AdaBoost é um algoritmo de aprendizado supervisionado do tipo *Boost*, que combina um conjunto de funções simples de classificação, denominadas classificadores fracos para formar um classificador forte.



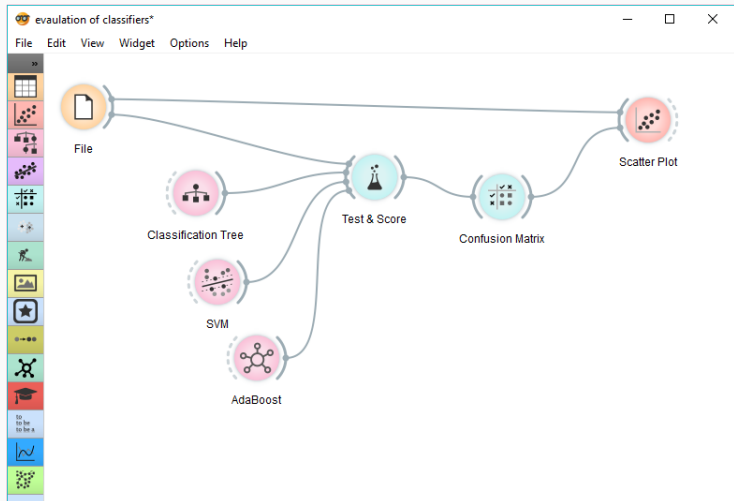
Em cada iteração, o procedimento de atualização aumenta os pesos das amostras classificadas incorretamente, fornecendo desta forma a característica adaptativa do AdaBoost.



Wahbeh et al. (2011) realizou um estudo comparativo entre quatro ferramentas para mineração de dados: KMine, Orange, Tanagra e Weka [6].



Segundo seu trabalho, ferramenta Weka apresentou o melhor desempenho, seguido pela *Orange*.



**Figura 1:** *Orange Data Mining*

# Metodologia

---

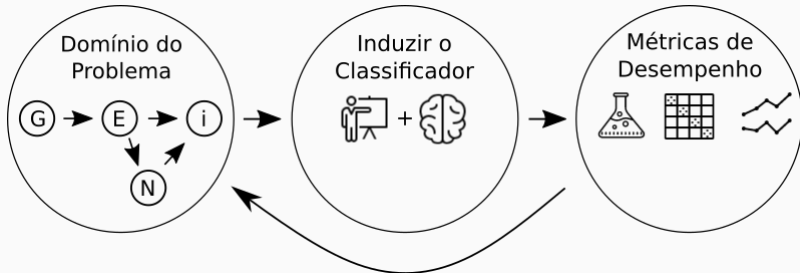
# Coleta de Dados

Coletar textos de redações avaliadas segundo a matriz de referência do INEP, normalizar os textos sem alterar o seu valor textual e armazená-lo de forma estruturada, separando o tema, título, texto e nota.



# Orange Data Mining

Desenvolver uma representação do domínio do problema com auxílio da ferramenta *Orange Data Mining*, induzir o classificador *AdaBoost* sobre a base de conhecimento rotulada, avaliar as métricas de desempenho e repetir o ciclo se necessário.

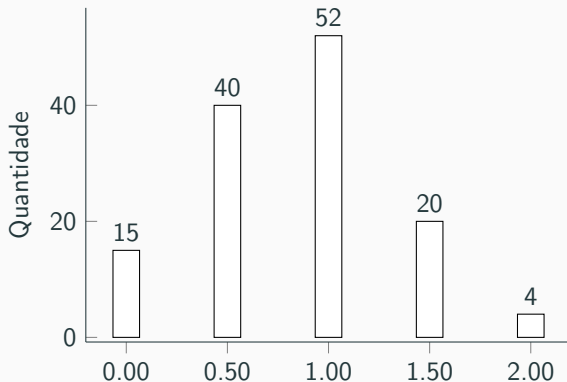


# Resultados Preliminares

---

# Adversidades

Trabalhar com dados desbalanceados tende à produzir regras de classificação que beneficiam as classes majoritárias, resultando em uma baixa taxa de predição para o grupo minoritário.



**Gráfico 1:** Amostra de 30% dos dados no *dataset*.

# Métricas de Desempenho

A Tabela 3 exibe os resultados das principais métricas de desempenho para classificadores e a média geral de cada métrica.

	Resultado da avaliação				
Classes	ROC	Acurácia	<i>F-Score</i>	<i>Precision</i>	<i>Recall</i>
<b>0.00</b>	0.498	0.828	0.096	0.845	0.828
<b>0.50</b>	0.552	0.640	0.349	0.653	0.640
<b>1.00</b>	0.499	0.509	0.422	0.506	0.509
<b>1.50</b>	0.549	0.579	0.222	0.755	0.759
<b>2.00</b>	0.541	0.915	0.140	0.899	0.915
<b>Média</b>	<b>0.529</b>	<b>0.694</b>	<b>0.246</b>	<b>0.737</b>	<b>0.730</b>

**Tabela 3:** Resultado das métricas de desempenho do classificador AdaBoost.



# Matriz de Confusão

A Tabela 4 exibe ao longo da diagonal em tons de cinza as decisões corretas: número de verdadeiros positivos TP e verdadeiros negativos TN;

		Predição					
		0.00	0.50	1.00	1.50	2.00	$\Sigma$
Atual	0.00	4	18	13	2	0	37
	0.50	13	42	44	14	1	114
	1.00	18	51	78	32	11	190
	1.50	7	14	31	15	2	69
	2.00	4	2	14	3	3	26
	$\Sigma$	46	127	180	66	17	436

**Tabela 4:** Matriz de confusão resultante da indução do classificador *AdaBoost*.

# Plano de Trabalho

---

# Plano de Atividades

	Atividade	Fevereiro - 2017	Março - 2017	Abril - 2017	Maiio - 2017	Junho - 2017	Julho - 2017	Agosto - 2017	Setembro - 2017	Outubro - 2017	Novembro - 2017
1	Revisão Bibliográfica	✓	✓	✓							
2	Coleta de Dados				✓						
3	Tratamento dos Dados				✓						
4	Domínio do Problema					✓	✓	✓	✓		
5	Indução do Classificador					✓	✓	✓	✓		
6	Métricas de Desempenho					✓	✓	✓	✓	✓	
7	Escrita da Monografia	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
8	Entrega da Monografia										✓



CEBRASPE.

**CENTRO BRASILEIRO DE PESQUISA EM AVALIAÇÃO E SELEÇÃO E DE PROMOÇÃO DE EVENTOS (CEBRASPE) PROGRAMA DE ATUALIZAÇÃO, QUALIFICAÇÃO E SELEÇÃO DE AVALIADORES DAS REDAÇÕES DO ENEM 2016 , 2016.**

Online; acessado 07 Abril 2017.



S. R. da Silva and T. L. Carvalho.

**Produção de texto escrito no ensino médio: Competências requeridas pela avaliação de redação do enem em (des)uso no livro didático de português.**

*Caminhos em linguística aplicada*, 16(1):1–25, 2017.

## References II



M. C. de Freitas, M. Uzeda-Garrão, C. Oliveira, C. N. dos Santos, and M. C. Silveira.

**A anotação de um corpus para o aprendizado supervisionado de um modelo de sn.**

*In Proceedings of the III TIL/XXV Congresso da SBC, 2005.*



INEP.

**Edital no 10, de 14 de abril de 2016, 2016.**

Online; acessado 05 Junho 2017.



M. C. Monard and J. A. Baranauskas.

**Conceitos sobre aprendizado de máquina.**

*Sistemas Inteligentes-Fundamentos e Aplicações, 1(1), 2003.*



A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi, and E. M. Al-Shawakfa.

**A comparison study between data mining tools over some classification methods.**

*International Journal of Advanced Computer Science and Applications*, 8(2):18–26, 2011.