Classificação da Base do ENEM 2012 com Naive Bayes

Ramon D. Costa, Roberto Caldeira, Mateus Cordeiro
Departamento de Ciência da Computação
Universidade Federal da Bahia
20 de outubro de 2016

COMPUTAÇÃO
UFBA

Sumário

Introdução

- Mineração de dados
- Base ENEM 2012
- Algoritmo Naive Bayes
- Uso da plataforma R

Predição

Status da redação

Identificador	Classificação
P	Presente
B	Em Branco
T	Fuga ao Tema
N	Anulada
I	Texto Insuficiente
A	Não atende ao tipo textual
H	Anulada - Fere Direitos Humanos
C	Cópia de texto motivador
F	Ausente

O Naive Bayes

- Classificador Probabilístico Simples
- Derivado do Teorema de Bayes

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

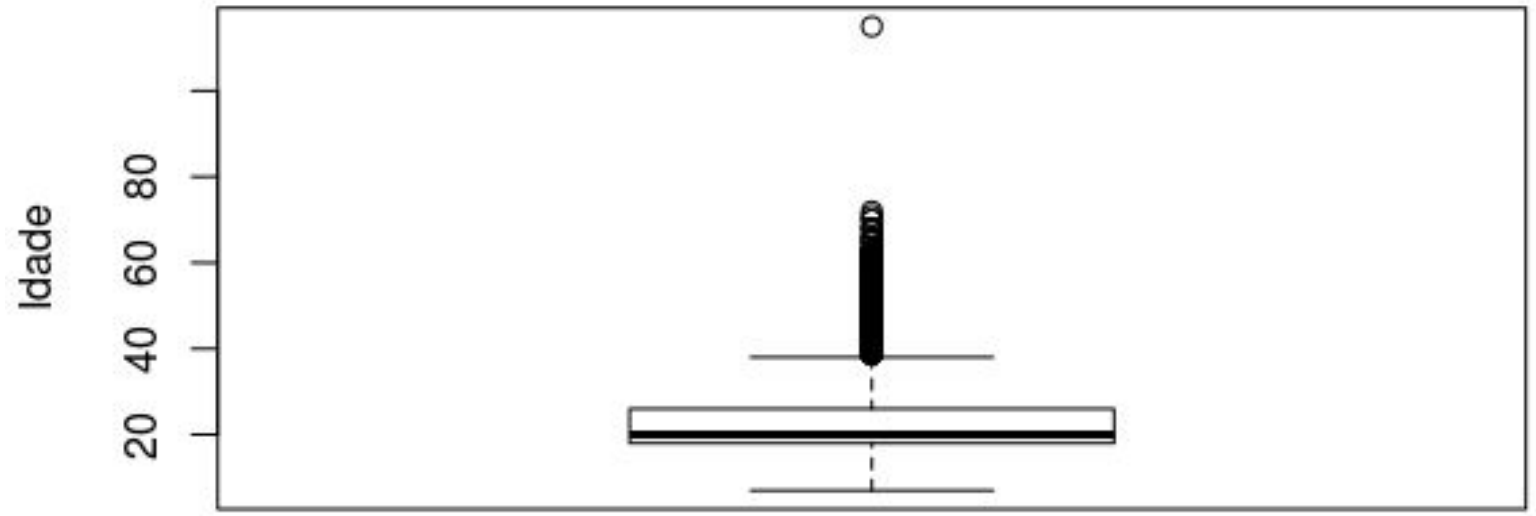
$$p(C_k|x_1, \ldots, x_n)$$

- Dado o atributo a ser predito, assume independência entre os atributos presentes
- Quanto mais consistentes são as evidências, melhor a predição

Análise estatística da base

- Conhecimento sobre:
  - Estrutura da base
  - Correlações entre seus atributos.
- Plataforma R
  - Boxplot
  - Histograma
- Identificação de outliers e inconsistências.
  - Notas
  - Idade

Análise estatística da base



Pré processamento

Uma base de dados pode apresentar certos problemas, como os seguintes:

- Incompletude: Valores Faltando
- Inconsistência: Valores fora do domínio
- Ruído: Variações inexplicáveis
- Dependências

Realização de operações para identificar e corrigir tais problemas.

Pré processamento

- Remoção de colunas
  - Informações redundantes
  - Informações sobre inscrição
  - Informações sobre local de prova
  - Necessidades Especiais
  - Atributos de baixo impacto

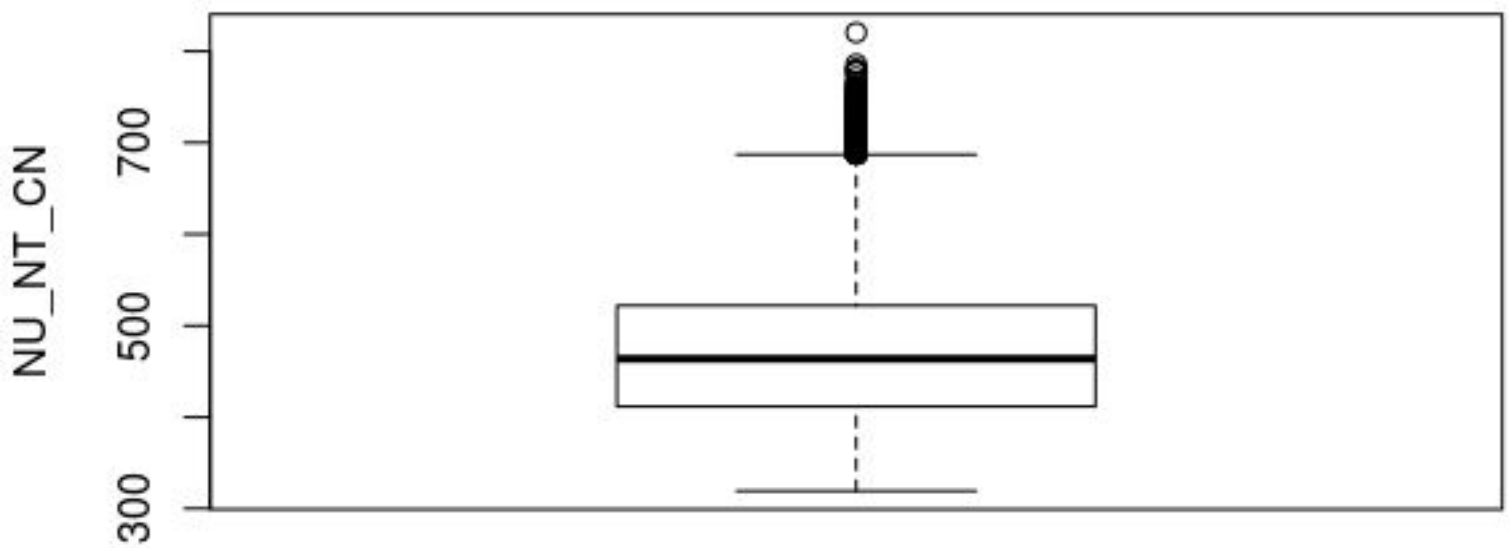
Pré processamento

- Correção de inconsistências
  - Atributos numéricos
  - Idades
  - Atributos condicionados

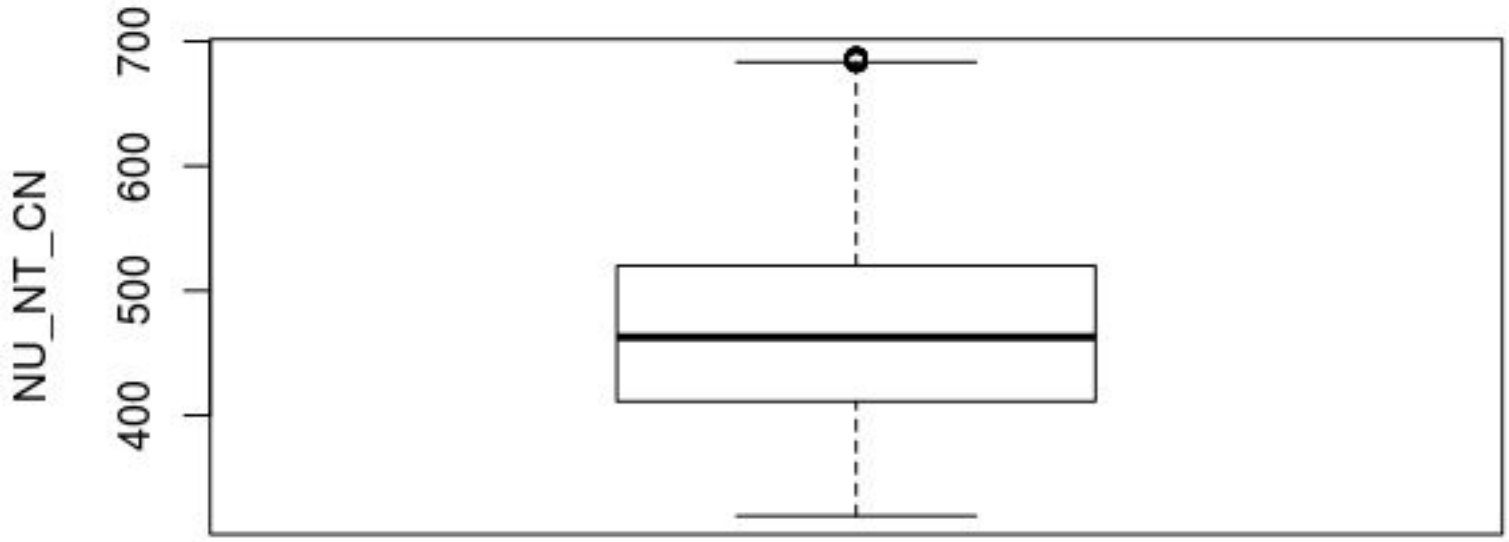
Pré processamento

- Remoção de outliers
  - Notas
  - Idades

Remoção de outliers (notas)



Remoção de outliers (notas)



Classificação e avaliação

A implementação

- Pacote e1071 do R
- Duas etapas

1. Construção do modelo de classificação (conjunto de treinamento)
2. Predição a partir do modelo (conjunto de teste)

Validação

Uso de validação cruzada (k-fold)

- k = 10
- A base é dividida em 10 subconjuntos mutuamente exclusivos
- Dez execuções onde cada um serve como conjunto de teste uma vez e o restante é usado como treinamento

Acurácia

Razão entre as predições corretas e incorretas

- Faz a média entre as acurácias dos k-testes
- Resultado de 97,2%

Acurácia

Nem sempre o suficiente para uma boa avaliação

- Dependendo da distribuição das predições feitas, modelos com menor acurácia ainda podem ser melhores.

Matriz de confusão

Matriz que apresenta os erros e acertos das classificações

- Boa para a avaliação de falsos positivos e falsos negativos em casos de classificação binária
- Para a nossa avaliação, uma pequena adaptação foi feita

Matriz de confusão

Foram criadas duas classes para distribuir os valores possíveis:

- Status Comum = { Presente }
- Status Especial = {Ausente, Em branco, etc}

Matriz de confusão

Distribuição das predições em três conjuntos:

- Classe e valor corretos.
- Apenas classe correta.
- Incorreta

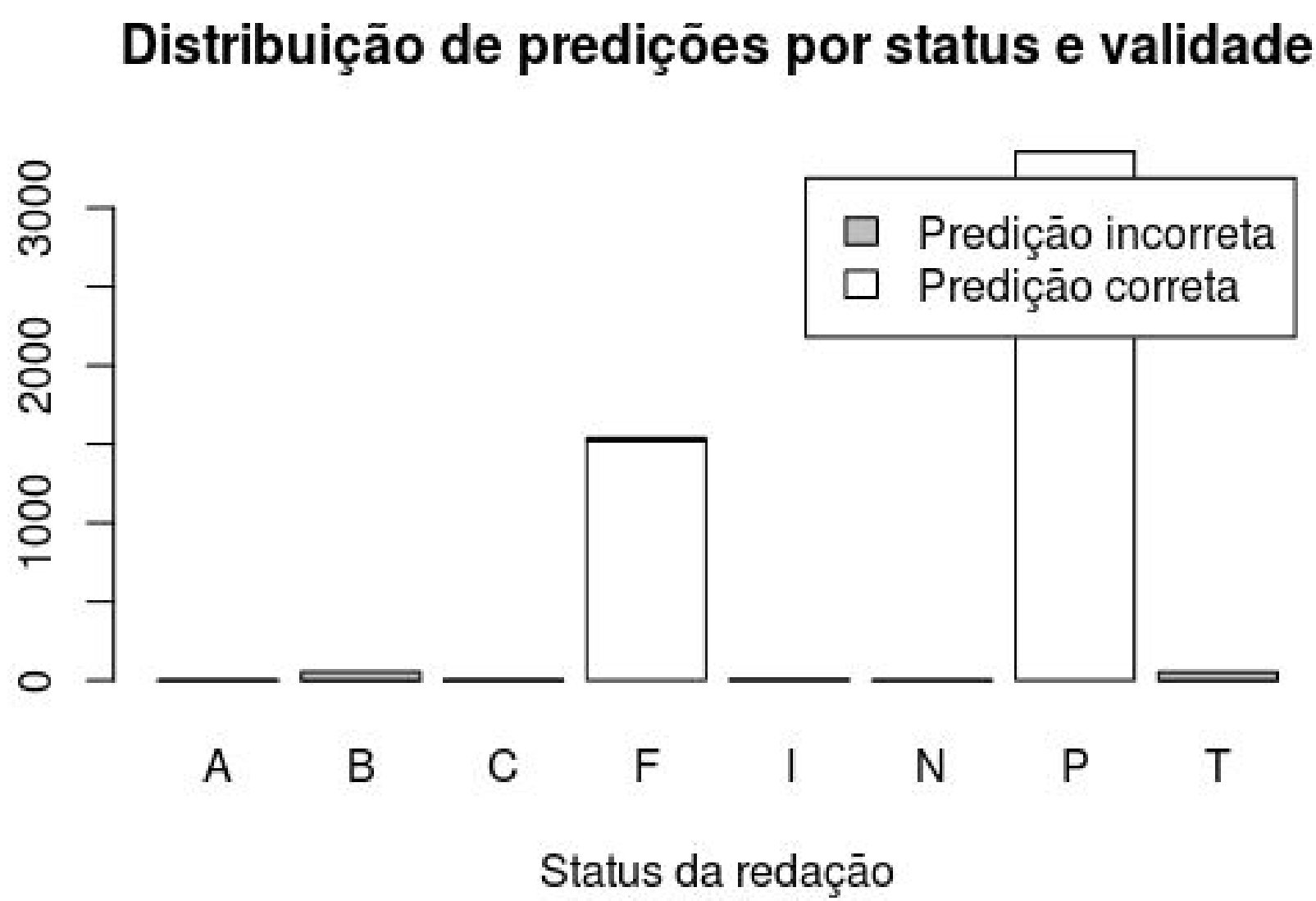
Matriz de confusão

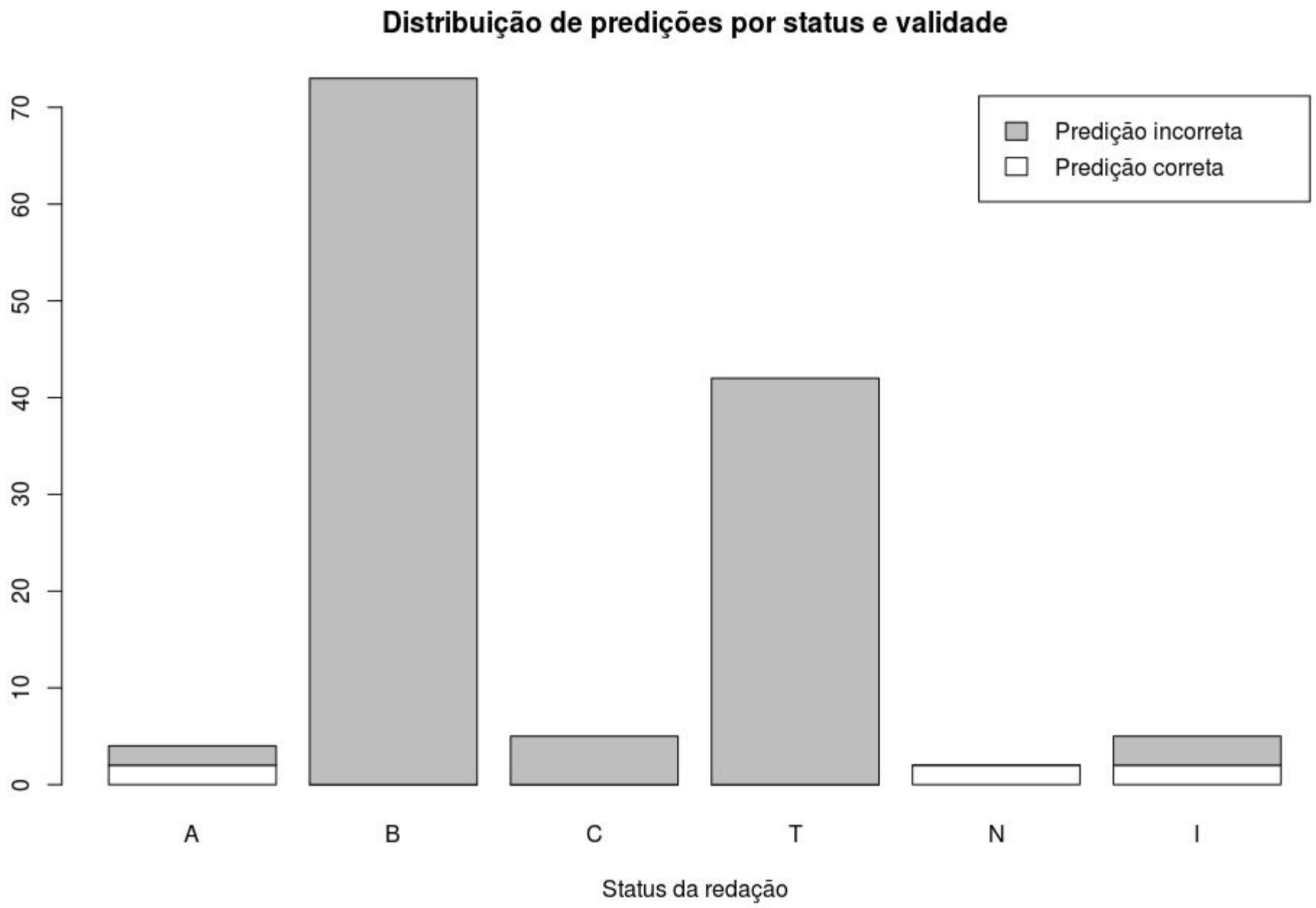
	Status Especial	Status Comum
Classe e valor corretos	1454	3337,3
Apenas classe correta	120,3	-
Incorreta	17,7	0

Matriz de confusão

Nota-se que a maior parte dos erros ocorreram na predição incorreta de status que aconteciam com pouco frequência:

- A escassez desses status provavelmente afetou negativamente a performance do classificador


Distribuição de predições por status e validade
0
1000
2000
3000
Predição incorreta
Predição correta
A
B
C
F
I
N
P
T
Status da redação


Distribuição de predições por status e validade
Predição incorreta
Predição correta
0
10
20
30
40
50
60
70
A
B
C
T
N
I
Status da redação