

## EFC 2

Cláudio Ferreira Carneiro - RA 263796

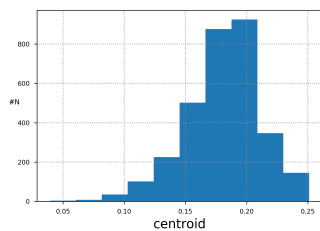
October 17, 2019

## 1 Parte 1 –Classificação binária

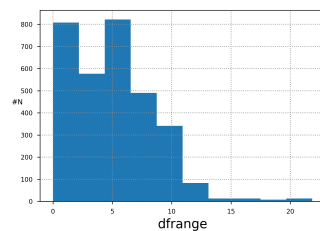
O código referente às atividades se encontra no repositório:  
<https://github.com/carneirofc/IA006.git>

### 1.1 a) Características dos atributos de entrada

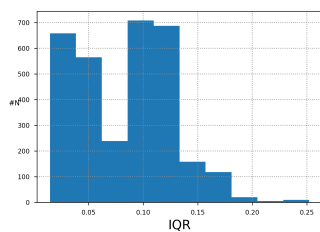
Os histogramas dos atributos em sua forma original são apresentados nas figuras [1], [2] e [3]. A correlação dos atributos é apresentada na forma de um *heatmap* [4] e por gráficos de dispersão [5] (na diagonal principal é exibido o histograma do atributo). Percebe-se que determinados atributos apresentam alto grau de correlação.



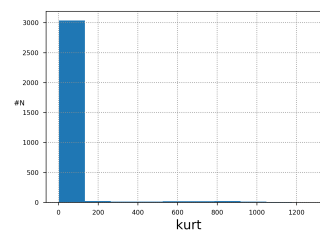
(a) Centroid



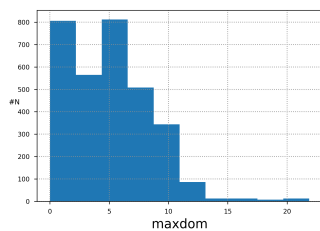
(b) F-Range



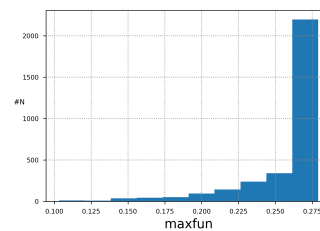
(c) IQR



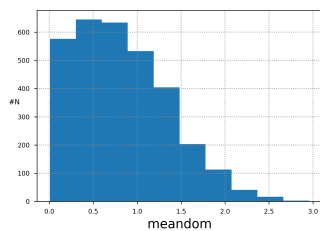
(d) Kurt



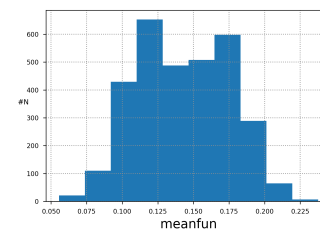
(e) Max dom



(f) Max fun

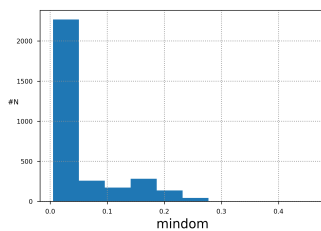


(g) Mean dom

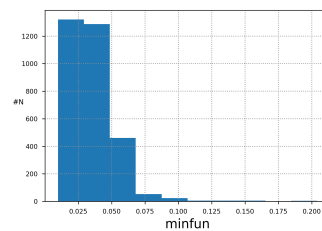


(h) Mean fun

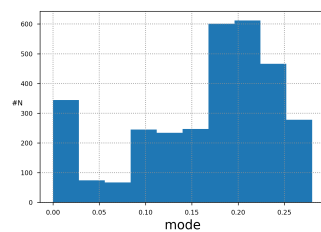
Figure 1: Classificação binária: Histograma dos atributos (1)



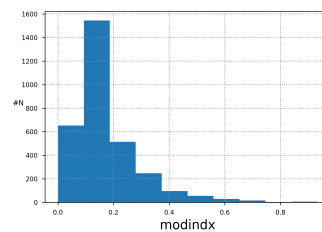
(a) Min dom



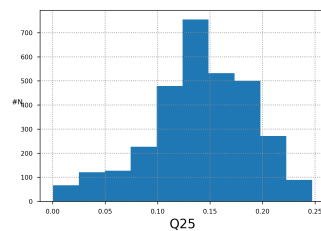
(b) Min fun



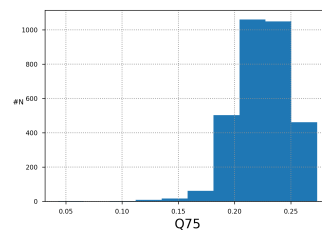
(c) Mode



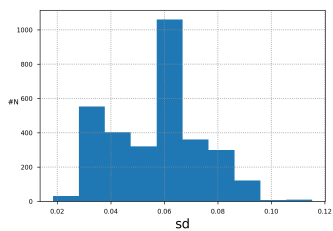
(d) modindx



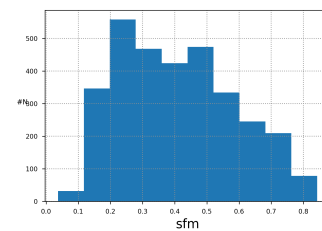
(e) Q25



(f) Q75

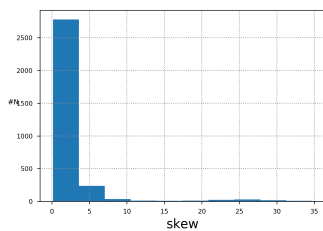


(g) sd

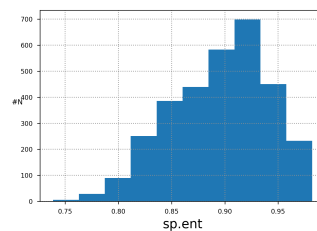


(h) sfm

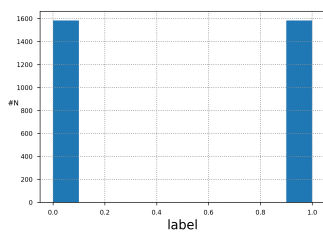
Figure 2: Classificação binária: Histograma dos atributos (2)



(a) skew



(b) histsp.ent



(c) label

Figure 3: Classificação binária: Histograma dos atributos (3)

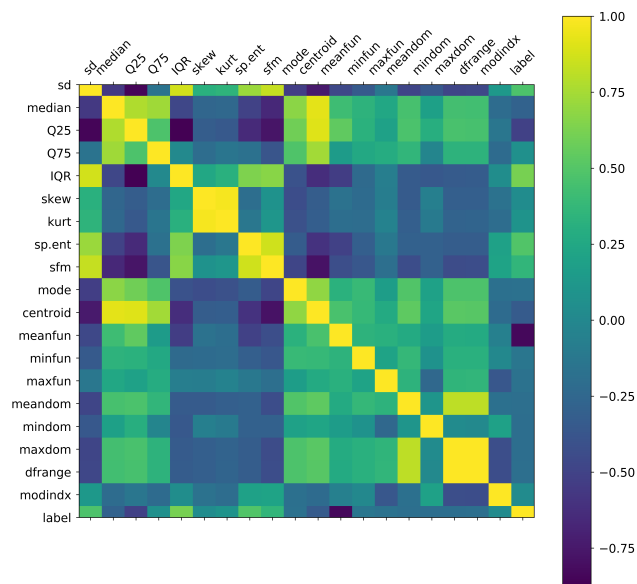


Figure 4: Classificação binária: Mapa de calor da correlação dos atributos

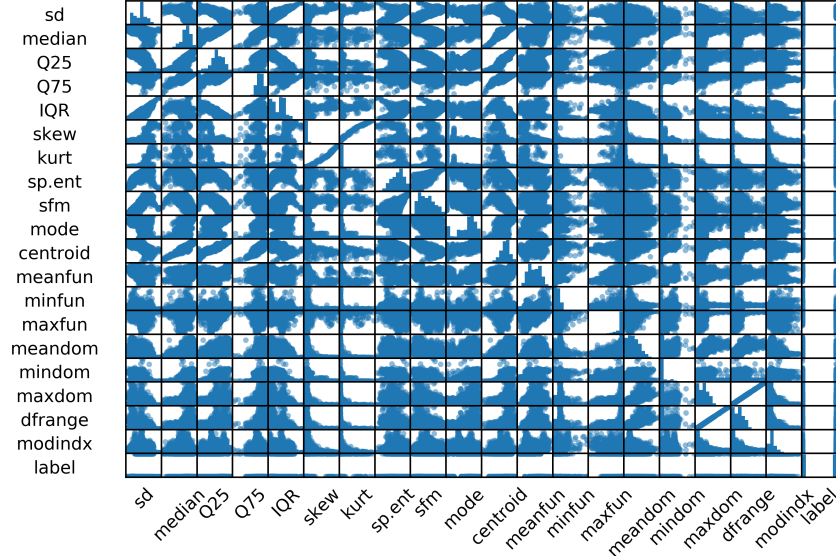


Figure 5: Classificação binária: Correlação dos atributos em gráfico de dispersão

## 1.2 b) Curva ROC e $F_1$ -medida

É utilizado o método *Z-score* para normalização dos dados. Tal método foi escolhido pois favorece o progresso de algoritmos baseados no gradiente descendente, uma vez que deixa as curvas de nível da superfície de erro mais circulares.

O processo de treinamento tem como critério de parada a variação da função de custo. Quando o decréscimo por década do custo for inferior a  $10^{-8}$  é terminado o processo de treinamento.

Parâmetros de treinamento:

$$\eta = 10^{-2}$$

$$tol = 10^{-8}$$

sendo  $\eta$  a taxa de aprendizagem e  $tol$  o limiar para o término do treinamento.

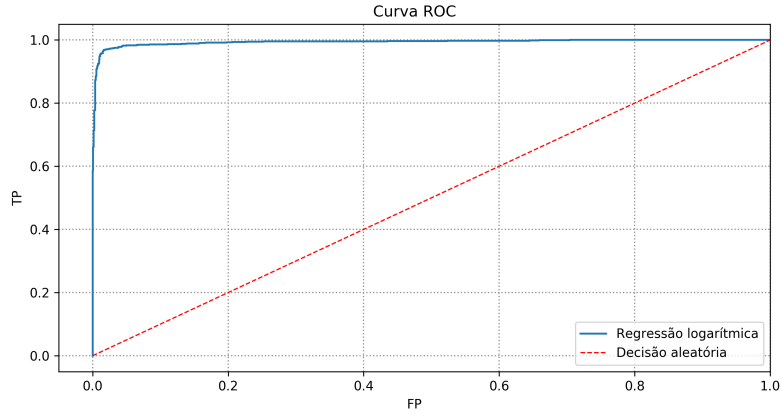


Figure 6: Classificação binária: Curva ROC relativa aos dados de Teste

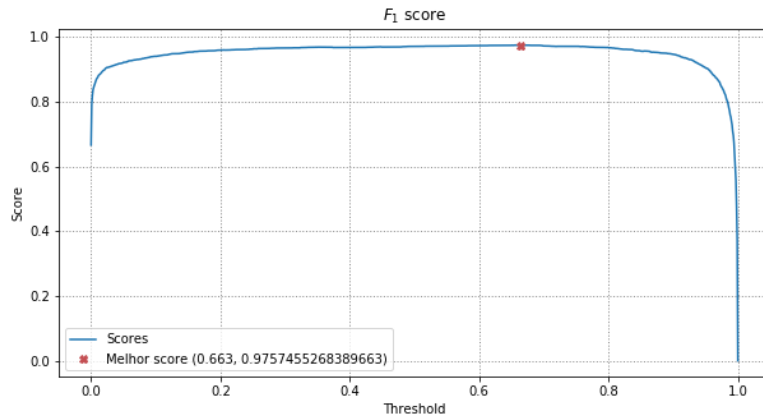


Figure 7: Classificação binária:  $F_1$ -medida relativa aos dados de Teste

### 1.3 c) Melhor *threshold*, matriz de confusão e acurácia

Para a escolha do valor de *threshold* será utilizada a  $F_1$ -medida, de forma que o *recall* e precisão do classificador tenham a mesma importância.

Conforme apresentado na figura [7], o ponto de máxima  $F_1$ -medida é obtido com:

$$\begin{aligned} \text{threshold} &= 0.663 \\ F_1 - \text{medida} &\approx 0.9757 \end{aligned}$$



Utilizando o limiar de máxima  $F_1$ -medida, a classificação do *dataset* de testes é apresentada conforme a matriz de confusão:

			<b>Classe Estimada</b>	
			Masculino	Feminino
			+	-
<b>Classe Verdadeira</b>	Masculino	+	1227	40
	Feminino	-	21	1246

O classificador apresenta acurácia (*acc*), precisão (*prec*) e *recall* de:

$$acc \approx 0.9759$$

$$prec \approx 0.9832$$

$$recall \approx 0.9684$$