EFC 2

Cláudio Ferreira Carneiro - RA 263796

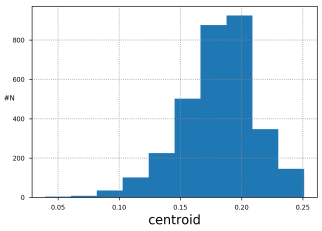
October 20, 2019

1 Parte 1 –Classificação binária

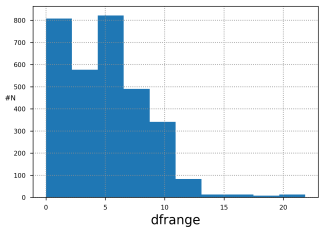
O código referente às atividades se encontra no repositório:
https://github.com/carneirofc/IA006.git

1.1 a) Características dos atributos de entrada

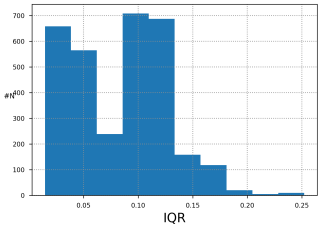
Os histogramas dos atributos em sua forma original são apresentados nas figuras 1, 2 e 3. A correlação dos atributos é apresentada na forma de um *heatmap* 4 e por gráficos de dispersão 5 (na diagonal principal é exibido o histograma do atributo). Percebe-se que determinados atributos apresentam alto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



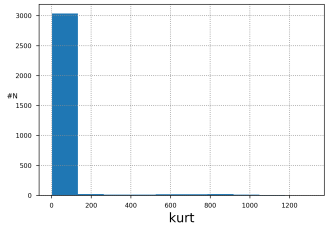
(a) Centroid



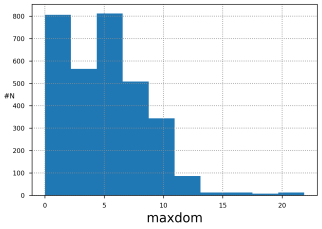
(b) F-Range



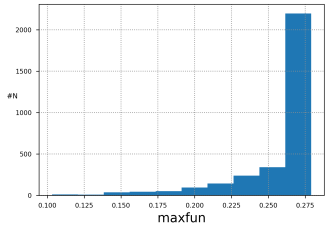
(c) IQR



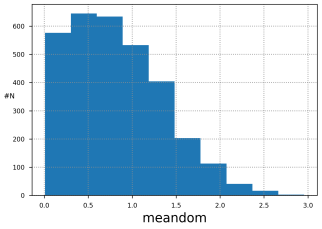
(d) Kurt



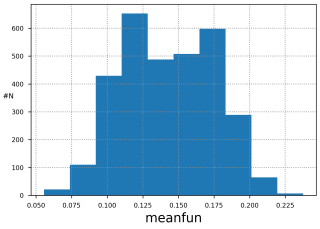
(e) Max dom



(f) Max fun

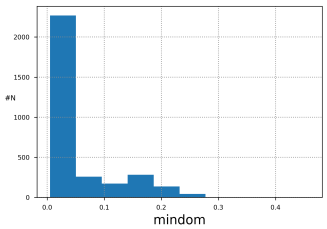


(g) Mean dom

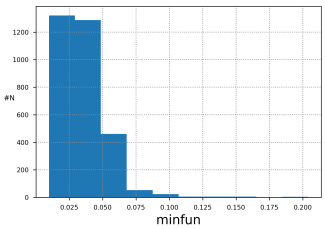


(h) Mean fun

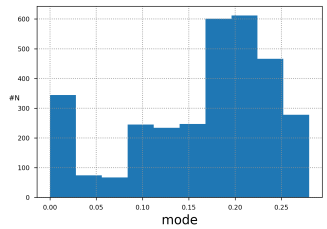
Figure 1: Classificação binária: Histograma dos atributos (1)



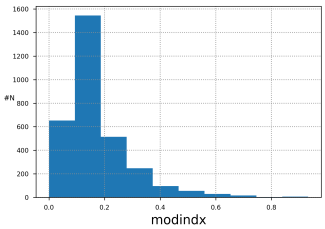
(a) Min dom



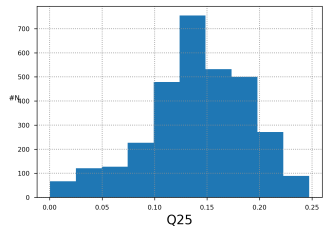
(b) Min fun



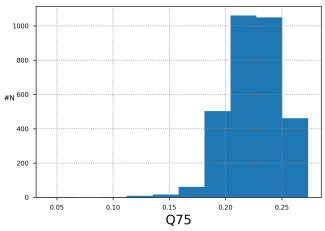
(c) Mode



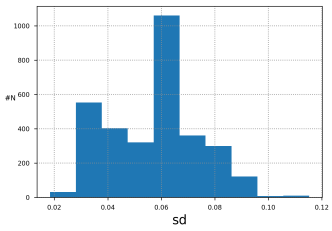
(d) modindx



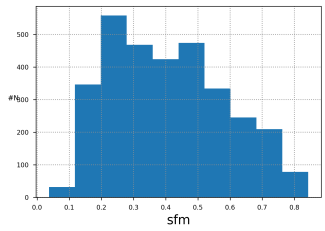
(e) Q25



(f) Q75

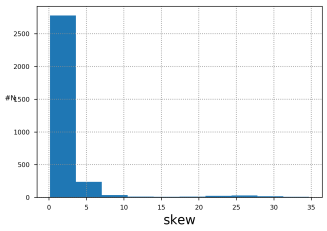


(g) sd

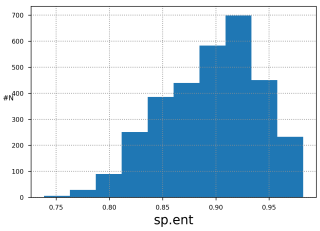


(h) sfm

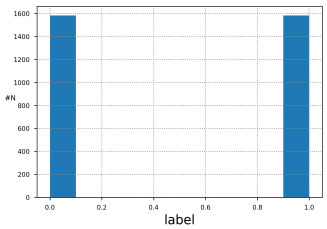
Figure 2: Classificação binária: Histograma dos atributos (2)



(a) skew



(b) histsp.ent



(c) label

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

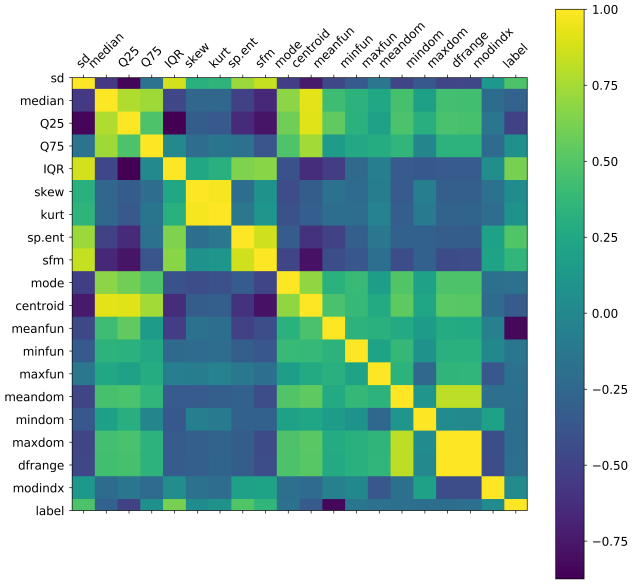


Figure 4: Classificação binária: Mapa de calor da correlaçao dos atributos

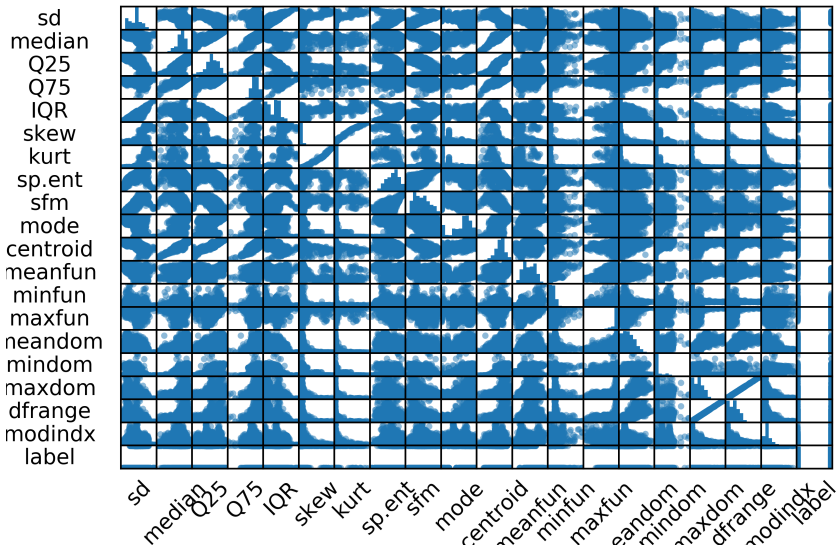


Figure 5: Classificação binária: Correlação dos atributos em gráfico de dispersão

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, sendo η a taxa de aprendizagem e tol o limiar para o término do processo:

$$\eta = 10^{-2}$$
$$tol = 10^{-8}$$

A curva ROC é obtida consideranto o rótulo 1 como classe positiva.

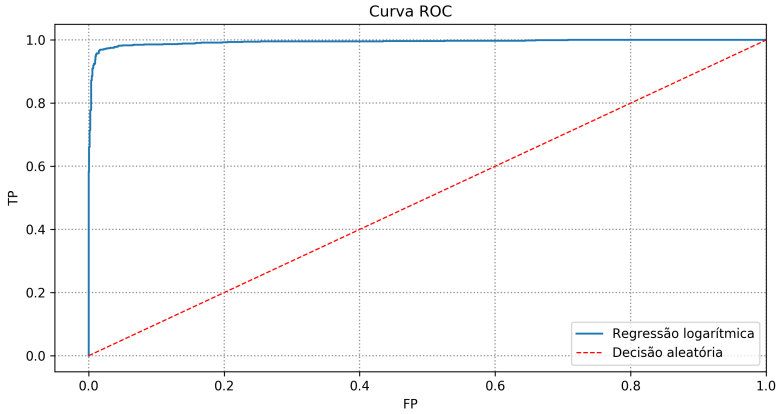


Figure 6: Classificação binária: Curva ROC relativa aos dados de Teste

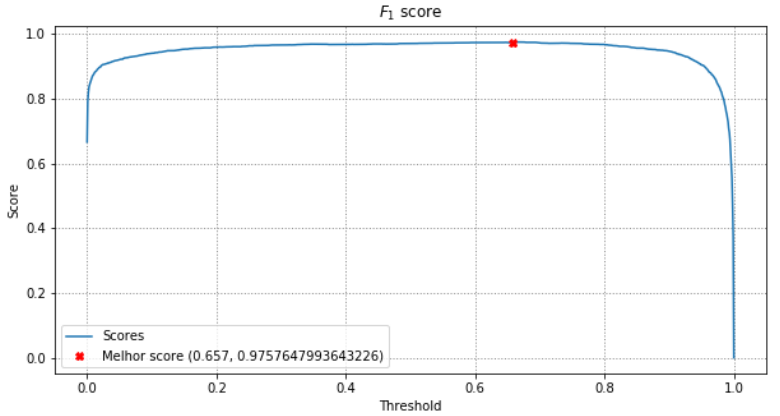


Figure 7: Classificação binária: F_1-medida relativa aos dados de Teste

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 valor de *threshold* para máxima F_1-medida é obtido com:

$$threshold = 0.657$$
$$F_1 - medida \approx 0.9757$$

		Classe Estimada	
		Feminino	Masculino
Classe	Feminino	1245	22
Verdadeira	Masculino	39	1228

Table 1: Matriz de confusão para o *threshold* de 0.657

	Precisão	**Recall**	**F1-medida**	**Amostras**
Feminino	0.969626	0.982636	0.976088	1267
Masculino	0.982400	0.969219	0.975765	1267
Média	0.976013	0.975927	0.975926	2534
Média ponderada	0.976013	0.975927	0.975926	2534

Table 2: Desempenho do classificador para o *threshold* de 0.657

Utilizando o limiar de máxima F_1-medida, a classificação do *dataset* de testes é apresentada conforme a matriz de confusão 1. O classificador apresentou acurácia de aproximadamente 0.975927. Outras medidas de desempenho como precisão, *recall* e F_1-medida são apresentadas na tabela 2. As medidas de desempenho são apresentadas por rótulo (Masculino e Feminino), na forma de uma média e como média ponderada pelo número de amostras de cada classe.

2 Parte 2 – Classificação multi-classe

Será abordado um problema de classificação multi-classe com 6 rótulos, conforme a tabela 3, e 561 atributos.

2.1 a) Regressão logística

Para a classificação multi-classe é adotada a *softmax*, sendo gerado um modelo capaz de produzir Q saídas que representam a probabilidade do padão apresentado pertencer a uma classe específica. Tal modelo apresenta maior robustez que as abordagens um-contra-todos e um-contra-um.

Na etapa de pré-processamento dos dados, o *dataset* de entrada foi normalizado utilizando a $z - score$ e os rótulos (*dataset* de saída) transformados com o processo de *one hot encoding*. O processo de treinamento do modelo foi finalizado com o modelo classificando corretamente $98,89\%$ dos padrões de treinamento.

0	1	2	3	4	**5**
Caminhada	**Subindo Escadas**	**Descendo Escadas**	**Sentado**	**Em pé**	**Deitado**

Table 3: Rótulos

	Caminhada	**Subindo Escadas**	**Descendo Escadas**	**Sentado**	**Em pé**	**Deitado**
Caminhada	479	8	9	0	0	0
Subindo Escadas	8	460	3	0	0	0
Descendo Escadas	11	33	376	0	0	0
Sentado	0	2	0	428	58	3
Em pé	0	0	0	16	516	0
Deitado	0	0	0	0	24	513

Table 4: Matriz de confusão, *dataset* testes com o modelo explorando a função *softmax*

	Precisão	**Recall**	**F1-medida**	**Medidas**
Caminhada	0.961847	0.965726	0.963783	496
Subindo Escadas	0.914513	0.976645	0.944559	471
Descendo Escadas	0.969072	0.895238	0.930693	420
Sentado	0.963964	0.871690	0.915508	491
Em pé	0.862876	0.969925	0.913274	532
Deitado	0.994186	0.955307	0.974359	537
Acurácia			0.940618	2947
Média macro	0.944410	0.939089	0.940363	2947
Média ponderada	0.943691	0.940618	0.940761	2947

Table 5: Métricas de desempenho do classificador

A matriz de confusão apresentada na tabela 4 foi obtida com o teste do modelo.

Será adotada como métrica para avaliação de desempenho a $F_1 - score$ macro, por dar a mesmam importância para a precisão e o *recall* do estimador e pelo tratamento igualitário a todas as classes.

2.2 b) kNN

Para o uso do kNN, os *datasets* de entrada são normalizados utilizando a $z - score$ e o melhor valor para o número de vizinhos k é encontrado com o uso da técnica de validação cruzada $K - Fold$, com 5 folds.

O melhor resultado de classificação na validação cruzada é obtido com $k = 26$ vizinhos, conforme é mostrado no gráfico 8. A figura 8 apresenta o valor médio da quantidade de estimações incorretas obtidos nos 5 folds para k vizinhos.

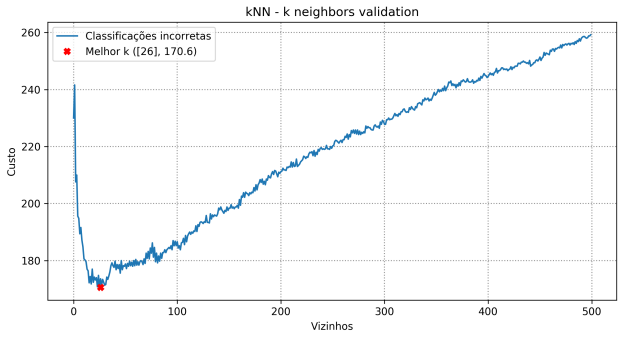


Figure 8: kNN: Média das estimações incorretas dos $K - Folds$ para k vizinhos

Ao término da etapa de testes é obtida a matriz de confusão 6.

	Caminhada	**Subindo Escadas**	**Descendo Escadas**	**Sentado**	**Em pé**	**Deitado**
Caminhada	489	2	5	0	0	0
Subindo Escadas	49	419	3	0	0	0
Descendo Escadas	67	59	294	0	0	0
Sentado	0	2	0	387	100	2
Em pé	0	0	0	21	511	0
Deitado	0	0	0	10	18	509

Table 6: kNN: Matriz de confusão

Os

	Precisão	**Recall**	**F1-medida**	**Medidas**
Caminhada	0.808264	0.985887	0.888283	496
Subindo Escadas	0.869295	0.889597	0.879328	471
Descendo Escadas	0.973510	0.700000	0.814404	420
Sentado	0.925837	0.788187	0.851485	491
Em pé	0.812401	0.960526	0.880276	532
Deitado	0.996086	0.947858	0.971374	537
Acurácia			0.885307	2947
Média macro	0.897566	0.878676	0.880859	2947
Média ponderada	0.896129	0.885307	0.883887	2947

Table 7: kNN: Métricas de desempenho na etapa de testes