Universidade Estadual do Maranhão
Pós-Graduação em Engenharia da Computação e Sistemas
Disciplina de Aprendizagem de Máquina
Prof. Dr. Omar Andres Carmona Cortes
Nome: Danrley Alves dos Santos

Objetivo

Usando as mesmas bases do exercício anterior e os resultados do experimento anterior, compare os algoritmos Naive Bayes, Regressão Logística, MLP e RBF usando K-Fold com k=5. Devem ser testadas pelo menos 2 arquiteturas de MLP e 2 arquiteturas de RBF, totalizando 6 tratamentos. Como a comparação envolve 6 tratamentos é necessário a utilização de ANOVA e teste de Tukey. O teste ANOVA indicará se h0 será aceita ou rejeitada. A diferença é efetivamente demonstrada pelo teste de Tukey.

Data Set 1.0

Mammographic Mass Data Set

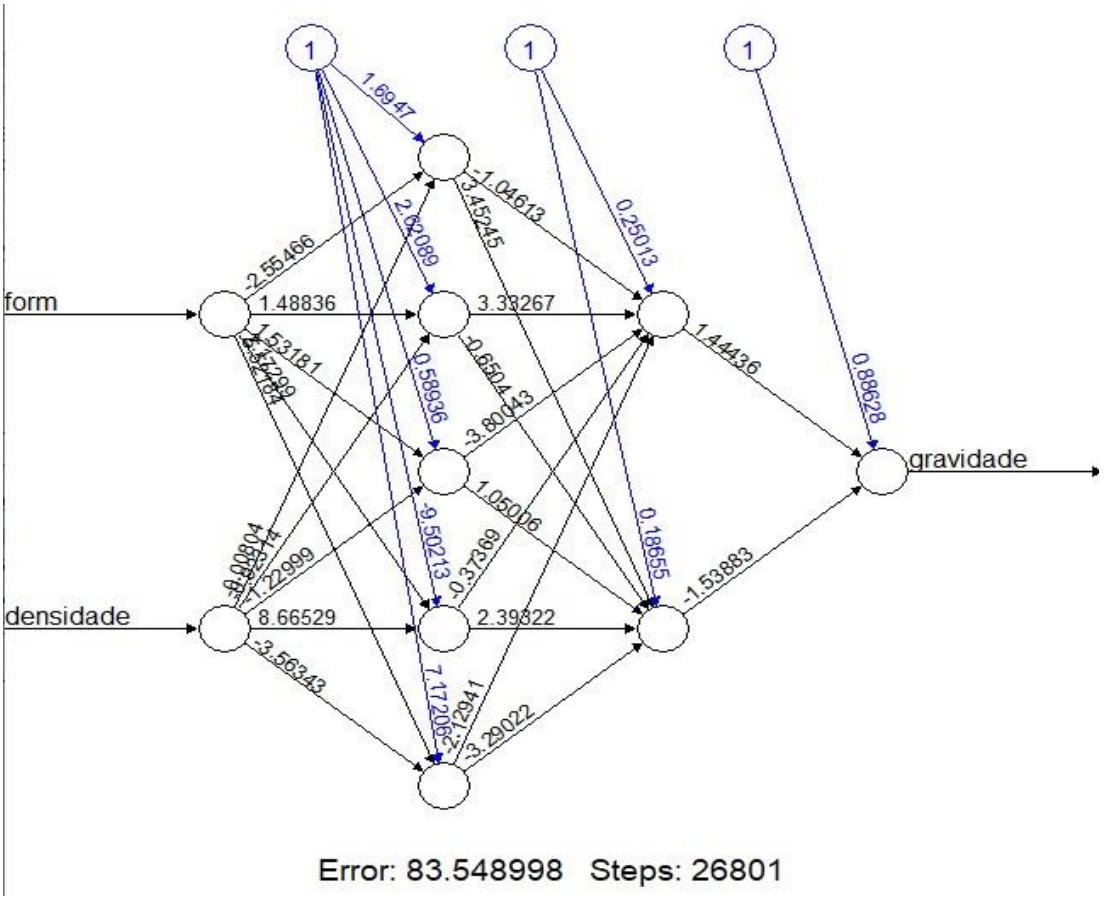
O primeiro data set escolhido se encontra no site da UCI disponível no link https://archive.ics.uci.edu/ml/machine-learning-databases/mammographic-masses/mammographic_masses.data Onde demonstra exames de mamografia para detecção de câncer de mama.
A mamografia é o método mais eficaz para a triagem do câncer de mama atualmente disponível. No entanto, o baixo valor preditivo positivo da biópsia da mama resultante da interpretação da mamografia leva a aproximadamente 70% de biópsias desnecessárias com resultados benignos.
Esse conjunto de dados pode ser usado para prever a gravidade (benigna ou maligna) de uma lesão em massa mamográfica a partir dos atributos BI-RADS e a idade do paciente. Ele contém uma avaliação do BI-RADS, a idade do paciente e três atributos do BI-RADS, juntamente com a verdade básica (o campo de gravidade) de 516 massas benignas e 445 malignas que foram identificadas em mamografias digitais de campo completo coletadas no Institute of Radiology of Universidade Erlangen-Nuremberg entre 2003 e 2006.
Supondo que todos os casos com avaliações BI-RADS maiores ou iguais a um determinado valor (variando de 1 a 5) sejam malignos e os outros casos benignos, sensibilidades e especificidades associadas podem ser calculadas. Isso pode ser uma indicação do desempenho de um sistema CAD em comparação com os radiologistas

Utilizando o primeiro Data Set e aplicando todas as métricas. Foi aplicado os algoritmos para projetar duas arquitetas diferentes de redes neurais no modelo MLP.

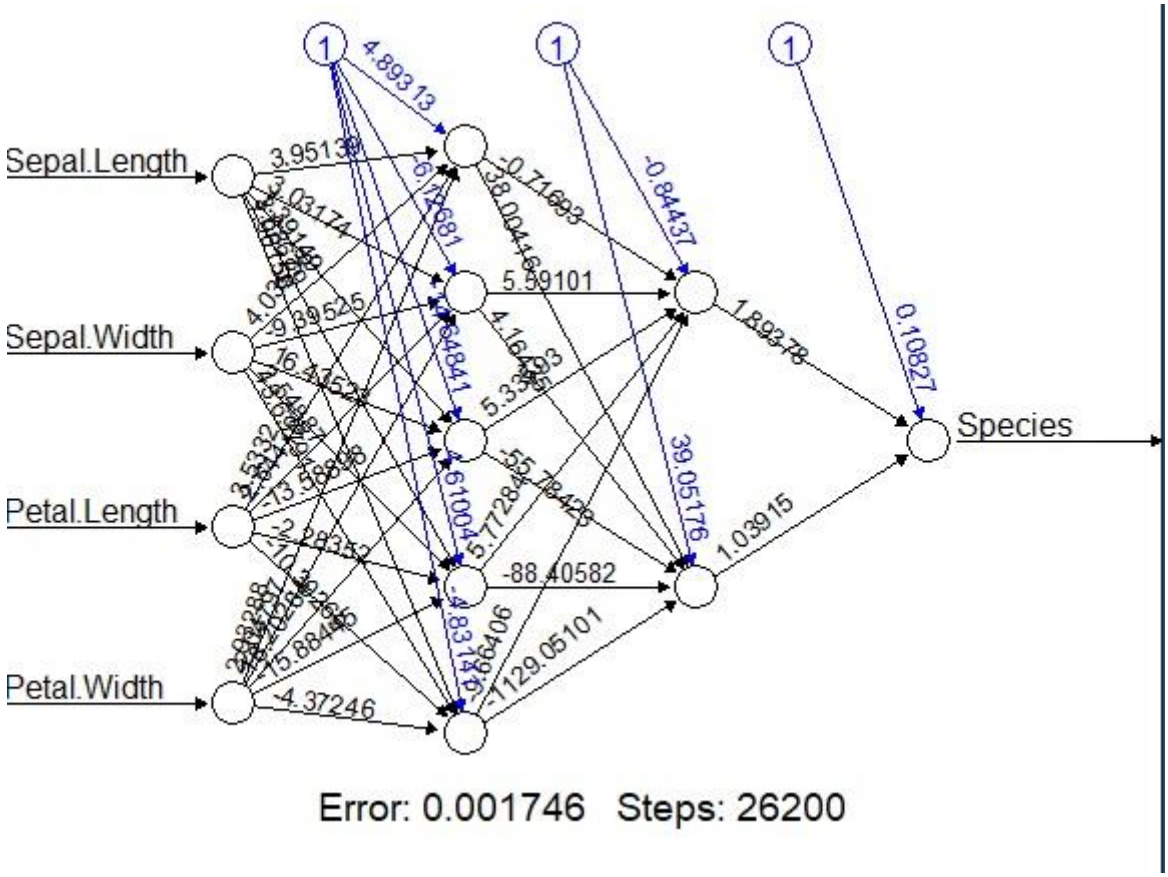
Características

6 atributos no total (1 campo de meta, 1 não preditivo, 4 atributos preditivos)
1. Avaliação BI-RADS: 1 a 5 (ordinal, não preditivo!)
2. Idade: idade do paciente em anos (inteiro)
3. Forma: massa forma: redonda = 1 oval = 2 lobular = 3 irregular = 4 (nominal)
4. Margem: margem de massa: circunscrita = 1 microlobulada = 2 obscurecida = 3 mal definida = 4 especificada = 5 (nominal)
5. Densidade: densidade de massa alta = 1 iso = 2 baixa = 3 contendo gordura = 4 (ordinal)
6. Gravidade: benigno = 0 ou maligno = 1 (binominal, campo de objetivo!).

Redes Neurais MLP (Mammographic Mass Data Set)



ARQUITETURA 01 – MLP (MAMOGRAFIA)



ARQUITETURA 02 – MLP (MAMOGRAFIA)

Teste do K-FOLD dos algoritmos aplicados.

K-FOLD REGRESÃO LOGÍSTICA
> Matriz_Mamografia_K_Fold_RG [,1] [,2] [,3] [,4] [,5] [1,] 0.7854167 0.8041667 0.7895833 0.8020833 0.7895833 [2,] 0.7924528 0.8253968 0.8230453 0.8221344 0.7984791 [3,] 0.7767442 0.7807018 0.7552743 0.7797357 0.7788018
K-FOLD REGRESÃO NAIVE BAYES
> Matriz_Mamografia_K_Fold_NB [,1] [,2] [,3] [,4] [,5] [1,] 0.7770833 0.7875000 0.7895833 0.7958333 0.8083333 [2,] 0.8268398 0.8391304 0.8398268 0.8508772 0.8640351 [3,] 0.7309237 0.7400000 0.7429719 0.7460317 0.7579365 >
K-FOLD ARQUITETURA 01 MLP

```
> metricas
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.6329114 0.8555556 0.8494624 0.8333333 0.8152174
[2,] 0.7083333 0.6875000 0.6458333 0.5989583 0.6302083
[3,] 0.6493506 0.6209677 0.5939850 0.5474453 0.5813953
[4,] 0.7478261 0.8088235 0.7627119 0.7272727 0.7301587
```

K-FOLD ARQUITETURA 02 MLP

```
> metricas2
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.6075949 0.7333333 0.6774194 0.6444444 0.8152174
[2,] 0.4166667 0.5416667 0.4947917 0.4583333 0.6302083
[3,] 0.3720930 0.5076923 0.4846154 0.4461538 0.5813953
[4,] 0.5079365 0.6129032 0.5161290 0.4838710 0.7301587
```

K-FOLD ARQUITETURA 01 RBF

```
> metricas_Mamografia_K_Fold_RBF
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.6075949 0.7222222 0.6774194 0.6444444 0.8152174
[2,] 0.4166667 0.5364583 0.5000000 0.4635417 0.6302083
[3,] 0.3720930 0.5038760 0.4883721 0.4496124 0.5813953
[4,] 0.5079365 0.6031746 0.5238095 0.4920635 0.7301587
```

K-FOLD ARQUITETURA 02 RBF

```
> metricas_Mamografia_K_Fold_RBF2
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.6075949 0.7222222 0.6774194 0.6444444 0.8152174
[2,] 0.4166667 0.5364583 0.5000000 0.4635417 0.6302083
[3,] 0.3720930 0.5038760 0.4883721 0.4496124 0.5813953
[4,] 0.5079365 0.6031746 0.5238095 0.4920635 0.7301587
```

Teste ANOVA & TUKEY

TESTE ANOVA das Acuráricas (Mamographic Mass Data Set)

Anova: fator único

RESUMO

Grupo	*Contagem*	*Soma*	*Média*	*Variância*
K-Fold RG	5	3,580174	0,716035	0,002195
K-Fold NB	5	4,045833	0,809167	0,001481
K-Fold ML	5	3,98648	0,797296	0,008689
K-Fold ML	5	3,478009	0,695602	0,006606
K-Fold RB	5	3,466898	0,69338	0,006421
K-Fold RB	5	3,466898	0,69338	0,006421

ANOVA

te da varia	*SQ*	*gl*	*MQ*	*F*	*valor-P*	*F crítico*
Entre grup	0,073767	5	0,014753	2,782386	0,040427	2,620654
Dentro dos	0,127259	24	0,005302			
Total	0,201026	29				

TESTE TUKEY						
	K-Fold RG	K-Fold NB	K-Fold MLP1	K-Fold MLP2	K-Fold RBF1	K-Fold RBF2
K-Fold RG		0,3596	0,5056	0,9975	0,996	0,996
K-Fold NB	2,86		0,9998	0,1742	0,1595	0,1595
K-Fold MLP1	2,495	0,3645		0,2706	0,2502	0,2502
K-Fold MLP2	0,6274	3,487	3,123		1	1
K-Fold RBF1	0,6957	3,556	3,191	0,06824		1
K-Fold RBF2	0,6957	3,556	3,191	0,06824	0	

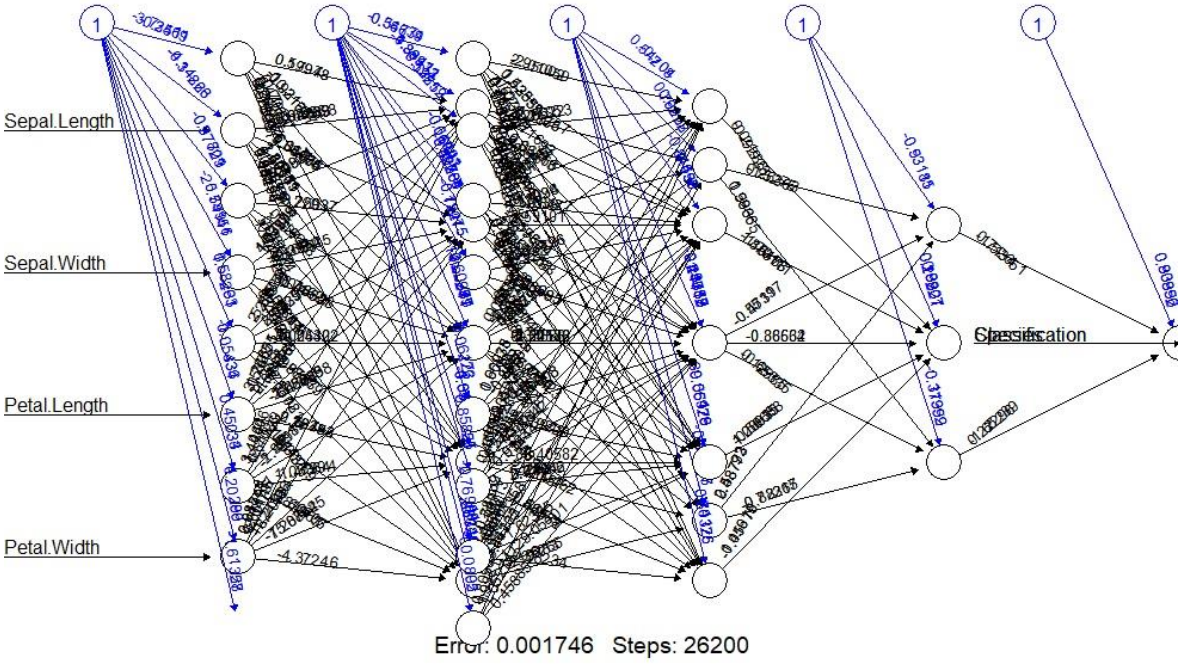
Data Set 2.0 – Breast Cancer Coimbra Data Set

Breast Cancer Coimbra Data Set

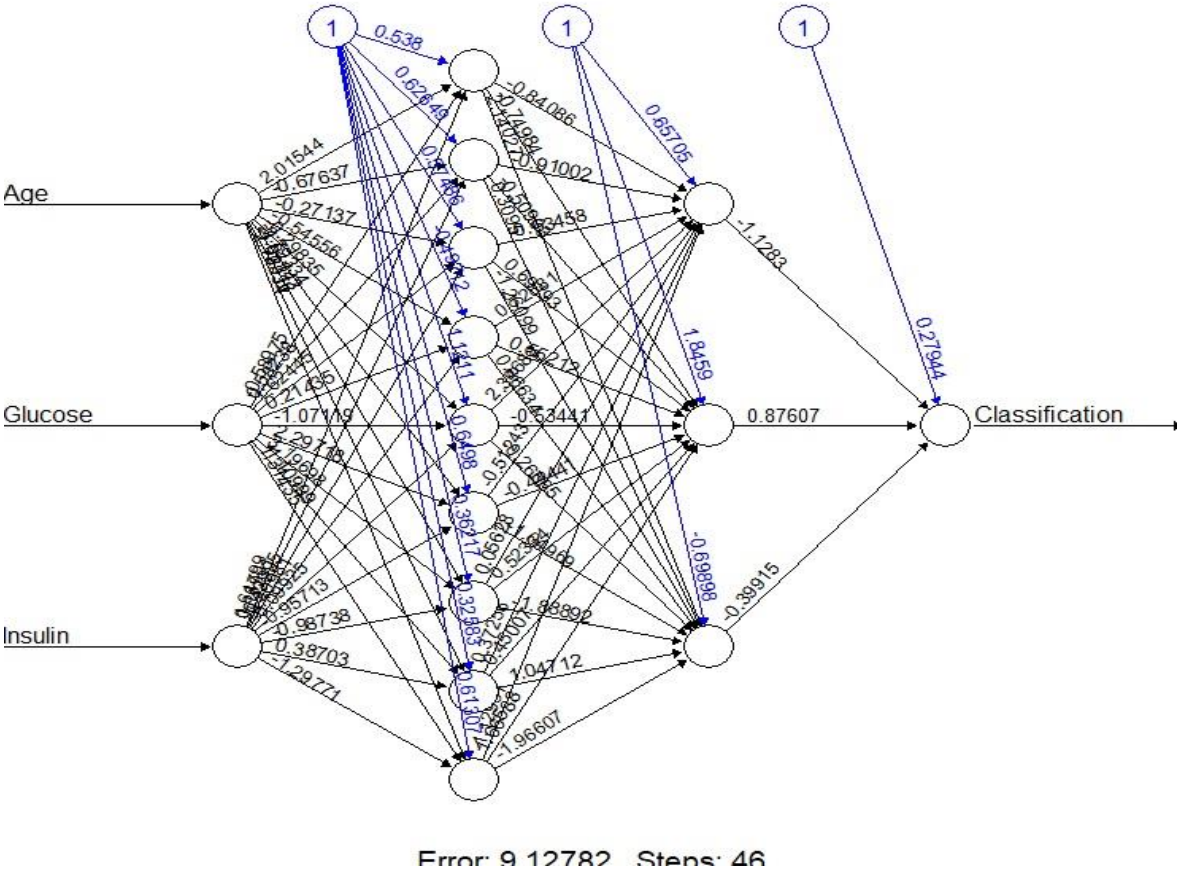
O segundo data set escolhido se encontra no site da UCI disponível no link https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra/dataR2.csv
Onde demonstra à presença de câncer em pacientes.
Existem 10 preditores, todos quantitativos e uma variável dependente binária, indicando a presença ou ausência de câncer de mama. Os preditores são dados e parâmetros antropométricos que podem ser coletados em análises de sangue de rotina. Modelos de previsão baseados nesses preditores, se precisos, podem ser potencialmente usados como um biomarcador de câncer de mama.

Características
Atributos quantitativos:
1-Anos de idade)
2- IMC (kg / m2)
3-Glicose (mg / dL)
4-Insulina (µU / mL)
5-HOMA
6-Leptina (ng / mL)
7-Adiponectina (µg / mL)
8-Resistina (ng / mL)
9-MCP-1 (pg / dL)
10-Classificador (1=controles saudáveis, 2=Pacientes)

Redes Neurais MLP (Breast Cancer Coimbra Data Set)



ARQUITETURA 01 – MLP (CANCER)



ARQUITETURA 02 – MLP (CANCER)

Teste do K-FOLD(Acurácia) dos algoritmos aplicados.

K-FOLD REGRESÃO LOGÍSTICA

```
> Matriz_cancer_K_Fold_RG
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.8103448 0.8448276 0.7758621 0.8275862 0.7413793
[2,] 0.8000000 0.8148148 0.7407407 0.8076923 0.7200000
[3,] 0.8181818 0.8709677 0.8064516 0.8437500 0.7575758
```

K-FOLD REGRESÃO NAIVE BAYES

```
> Matriz_cancer_K_Fold_NB
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.7241379 0.7758621 0.7931034 0.8275862 0.7758621
[2,] 0.6388889 0.6857143 0.7058824 0.7666667 0.7096774
[3,] 0.8636364 0.9130435 0.9166667 0.8928571 0.8518519
```

K-FOLD ARQUITETURA 01 MLP

```
           [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.04545455 0.1363636 0.6000000 0.0000000 0.0000000
[2,] 0.04347826 0.1304348 0.3478261 0.5652174 0.3913043
[3,] 0.50000000 0.7500000 0.1875000 0.0000000 0.0000000
[4,] 0.00000000 0.0000000 0.7142857 0.9285714 0.9000000
```

K-FOLD ARQUITETURA 02 MLP

```
           [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.04545455 0.1363636 0.6000000 0.0000000 0.0000000
[2,] 0.04347826 0.1304348 0.3478261 0.5652174 0.3913043
[3,] 0.50000000 0.7500000 0.1875000 0.0000000 0.0000000
[4,] 0.00000000 0.0000000 0.7142857 0.9285714 0.9000000
```

K-FOLD ARQUITETURA 01 RBF

```
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.5454545 0.5454545 0.4000000 0.0000000 0.0000000
[2,] 0.5217391 0.5217391 0.3913043 0.3913043 0.3913043
[3,] 0.9230769 0.9230769 0.1538462 0.0000000 0.0000000
[4,] 0.0000000 0.0000000 0.7000000 0.9000000 0.9000000
>
```

K-FOLD ARQUITETURA 02 RBF

```
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.5454545 0.5454545 0.4000000 0.0000000 0.0000000
[2,] 0.5217391 0.5217391 0.3913043 0.3913043 0.3913043
[3,] 0.9230769 0.9230769 0.1538462 0.0000000 0.0000000
[4,] 0.0000000 0.0000000 0.7000000 0.9000000 0.9000000
```

TESTE ANOVA das Acuráricas (Breast Cancer Coimbra Data Set)

Anova: fator único

RESUMO

Grupo	Contagem	Soma	Média	Variância
K-Fold RG	5	3,454237	0,690847	0,000226
K-Fold NB	5	3,896552	0,77931	0,001397
K-Fold ML	5	0,781818	0,156364	0,064603
K-Fold ML	5	0,781818	0,156364	0,064603
K-Fold RB	5	1,490909	0,298182	0,07762
K-Fold RB	5	1,490909	0,298182	0,07762

ANOVA

te da varia	SQ	gl	MQ	F	valor-P	F crítico
Entre grup	1,83924	5	0,367848	7,71521	0,000191	2,620654
Dentro dos	1,144279	24	0,047678			
Total	2,983519	29				

TESTE TUKEY

	K-Fold RG	K-Fold NB	K-Fold MLP1	K-Fold MLP2	K-Fold RBF1	K-Fold RBF2
K-Fold RG		0,9866	0,008454	0,008454	0,08419	0,08419
K-Fold NB	0,9059		0,001784	0,001784	0,02089	0,02089
K-Fold MLP1	5,473	6,379		1	0,9042	0,9042
K-Fold MLP2	5,473	6,379	0		0,9042	0,9042
K-Fold RBF1	4,021	4,927	1,452	1,452		1
K-Fold RBF2	4,021	4,927	1,452	1,452	0	

Data Set 3.0 - Heart Disease Data Set

Heart Disease Data Set

O terceiro data set escolhido se encontra no site da UCI disponível no link https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/. Onde demonstra à presença de doenças cardíacas em pacientes.

Esse banco de dados contém 76 atributos, mas todas as experiências publicadas se referem ao uso de um subconjunto de 14 deles. Em particular, o banco de dados de Cleveland é o único usado pelos pesquisadores de ML para

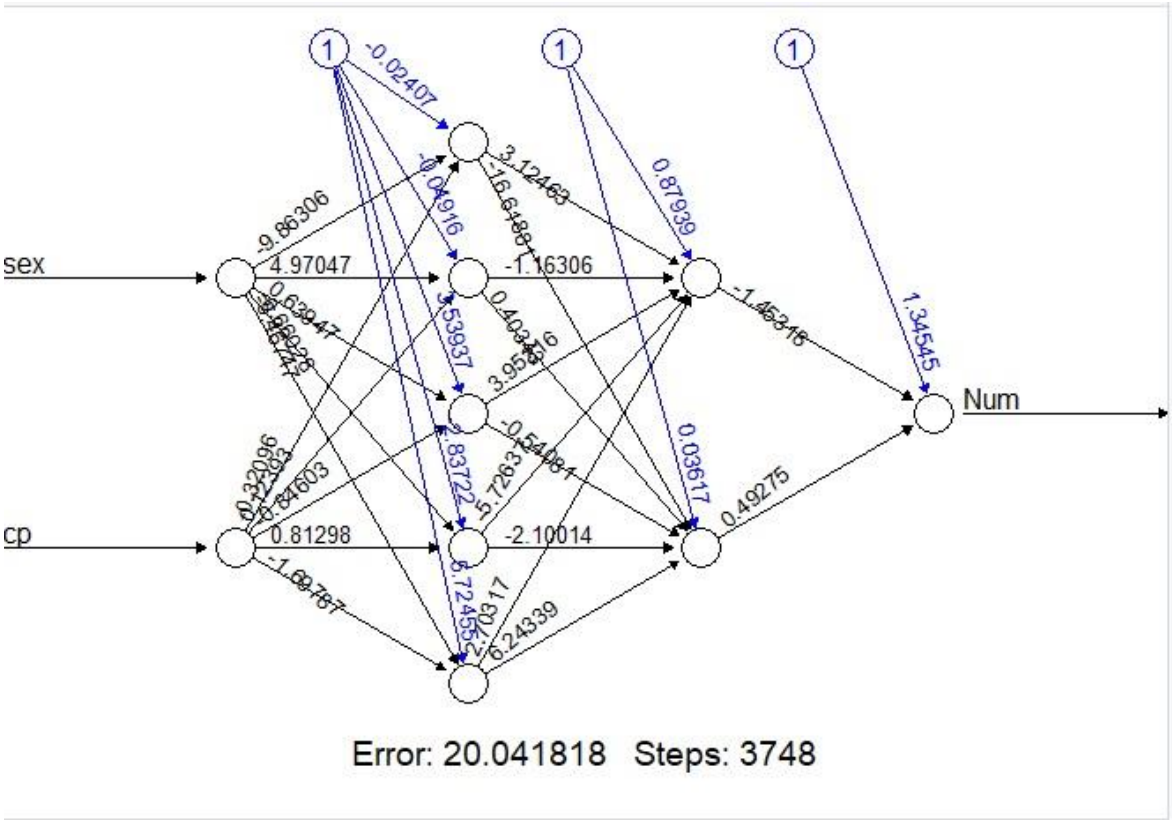
esta data. O campo "objetivo" refere-se à presença de doença cardíaca no paciente. É um número inteiro avaliado de 0 (sem presença) a 4. As experiências com o banco de dados de Cleveland concentraram-se em simplesmente tentar distinguir presença (valores 1,2,3,4) da ausência (valor 0). Os nomes e os números de seguridade social dos pacientes foram recentemente removidos do banco de dados, substituídos por valores fictícios.

Um arquivo foi "processado", aquele que contém o banco de dados de Cleveland. Todos os quatro arquivos não processados também existem nesse diretório.

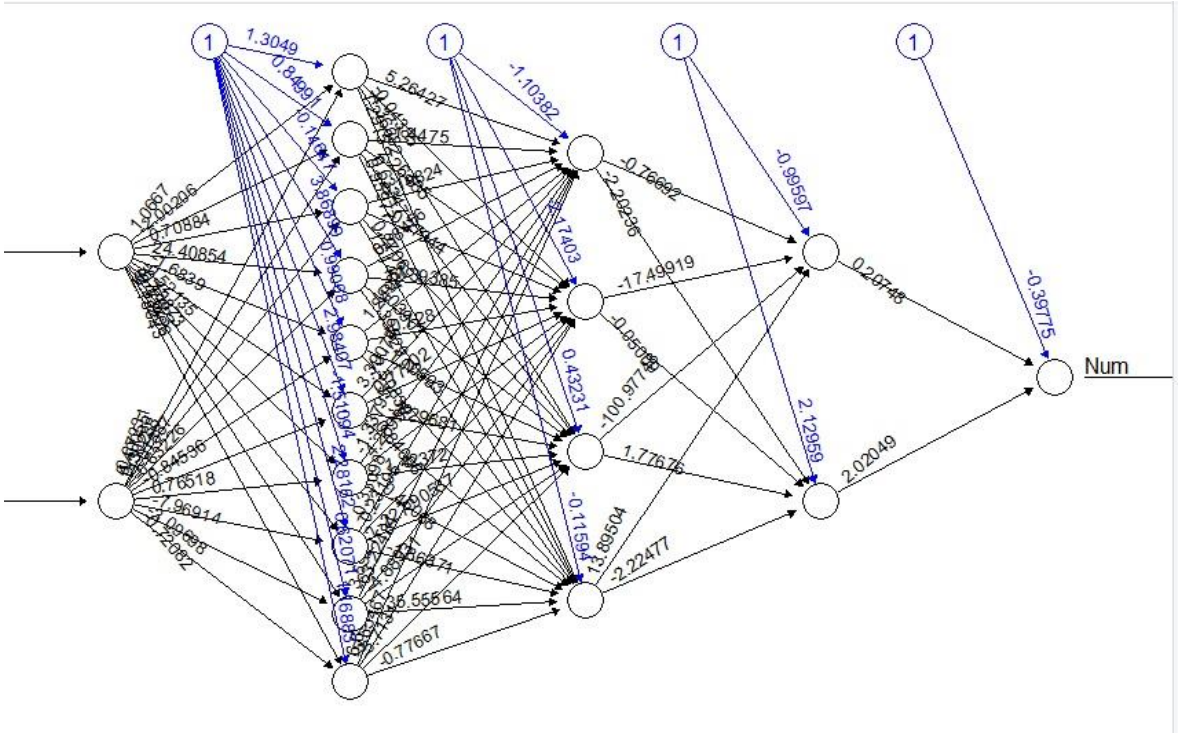
Doador: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

Características
Essa base de dados contém 14 atributos principais que serão trabalhados:
1. (idade)
2. (sexo)
3. (cp)
4. (trestbps)
5. (col)
6. (fbs)
7. (restecg)
8. (thalach)
9. (exang)
10.(pico antigo)
11.(inclinação)
12.(ca)
13.(thal)
14.(num) (o atributo previsto

Redes Neurais MLP (Heart Disease Data Set)



ARQUITETURA 01 – MLP (CARDIACO)



ARQUITETURA 02 – MLP (CARDIACO)

Teste do K-FOLD(Acurácia) dos algoritmos aplicados.

K-FOLD REGRESÃO LOGÍSTICA					
	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.8466667	0.8466667	0.8533333	0.8800000	0.8733333
[2,]	0.8372093	0.8372093	0.8470588	0.9090909	0.8604651
[3,]	0.8593750	0.8593750	0.8615385	0.8493151	0.8906250
>					

K-FOLD REGRESÃO NAIVE BAYES					
	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.8733333	0.8733333	0.8733333	0.8733333	0.8733333
[2,]	0.8604651	0.8604651	0.8604651	0.8604651	0.8604651
[3,]	0.8906250	0.8906250	0.8906250	0.8906250	0.8906250
>					

K-FOLD ARQUITETURA 01 MLP					
	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.7307692	0.7931034	0.7692308	0.7692308	0.7419355
[2,]	0.7666667	0.7666667	0.7666667	0.7666667	0.7333333
[3,]	0.7307692	0.7419355	0.7142857	0.7142857	0.7419355
[4,]	0.7941176	0.7931034	0.8125000	0.8125000	0.7241379
>					

K-FOLD ARQUITETURA 02 MLP

```
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.7307692 0.7931034 0.7692308 0.7692308 0.7419355
[2,] 0.7666667 0.7666667 0.7666667 0.7666667 0.7333333
[3,] 0.7307692 0.7419355 0.7142857 0.7142857 0.7419355
[4,] 0.7941176 0.7931034 0.8125000 0.8125000 0.7241379
```

K-FOLD ARQUITETURA 01 RBF

```
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.7307692 0.7931034 0.7692308 0.7692308 0.7419355
[2,] 0.7666667 0.7666667 0.7666667 0.7666667 0.7333333
[3,] 0.7307692 0.7419355 0.7142857 0.7142857 0.7419355
[4,] 0.7941176 0.7931034 0.8125000 0.8125000 0.7241379
```

K-FOLD ARQUITETURA 02 RBF

```
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.7419355 0.7419355 0.7419355 0.7419355 0.7419355
[2,] 0.7333333 0.7333333 0.7333333 0.7333333 0.7333333
[3,] 0.7419355 0.7419355 0.7419355 0.7419355 0.7419355
[4,] 0.7241379 0.7241379 0.7241379 0.7241379 0.7241379
```

Teste do K-FOLD (Acurácia) dos algoritmos aplicados.

TESTE ANOVA das Acuráricas (Breast Cancer Coimbra Data Set)

Anova: fator único

RESUMO

Grupo	*Contagem*	*Soma*	*Média*	*Variância*
0,846667	4	3,453333	0,863333	0,000252
0,873333	4	3,493333	0,873333	0
0,73077	4	3,073501	0,768375	0,000437
0,730769	4	3,073501	0,768375	0,000437
0,730769	4	3,073501	0,768375	0,000437
0,741936	4	2,967742	0,741936	0

ANOVA

te da varia	*SQ*	*gl*	*MQ*	*F*	*valor-P*	*F critico*
Entre grup	0,062867	5	0,012573	48,2396	8,41E-10	2,772853
Dentro dos	0,004692	18	0,000261			
Total	0,067558	23				

TESTE TUKEY

	K FOLD RL	K FOLD NB	K FOLD MLP1	K FOLD MLP2	K FOLD RBF1	K FOLD RBF2
K FOLD RL		0,0585	0,003219	0,0006503	0,001692	0,002585
K FOLD NB	4,269		0,8072	0,4255	0,659	0,7597
K FOLD MLP1	6,039	1,77		0,9858	0,9998	1
K FOLD MLP2	6,957	2,689	0,9185		0,9987	0,9927
K FOLD RBF1	6,41	2,141	0,371	0,5474		1
K FOLD RBF2	6,166	1,897	0,1269	0,7915	0,2441	

Referências

https://archive.ics.uci.edu/ml/index.php