



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Reconhecimento de Padrões

José Delmondes
Matheus Miranda



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

DATASET

**EACH**Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Sobre o DATASET

DATASET	Reconhecimento de Gênero por Voz
Resumo	DATASET criado para identificar se uma determinada voz é masculina ou feminina, com base nas propriedades acústicas da voz e da fala. O conjunto de dados consiste em 3.168 amostras de voz gravadas. As amostras de voz foram pré-processadas por análise acústica em R usando os pacotes seewave e tuneR, com uma faixa de frequência analisada de 0hz-280hz.
Classes	Masculina (1584) Feminina (1584)
Instâncias	3168
Características	20
Tipo das Características	Continuas e Racionais
Missing Values	0

**EACH**Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Descrição das Características

Nome	Descrição	Tipo
meanfreq	Mean Frequency (in kHz)	Continua e Racional
sd	Standard Deviation of Frequency	Continua e Racional
median	Median Frequency (in kHz)	Continua e Racional
Q25	First Quantile (in kHz)	Continua e Racional
Q75	Third Quantile (in kHz)	Continua e Racional
IQR	Interquantile Range (in kHz)	Continua e Racional
skew	Skewness	Continua e Racional
kurt	Kurtosis	Continua e Racional
sp.ent	Spectral Entropy	Continua e Racional
sfm	Spectral Flatness	Continua e Racional

**EACH**Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Descrição das Características

Nome	Descrição	Tipo
mode	Mode Frequency	Continua e Racional
centroid	Frequency Centroid	Continua e Racional
meanfun	Average of fundamental frequency measured across acoustic signal	Continua e Racional
minfun	Minimum fundamental frequency measured across acoustic signal	Continua e Racional
maxfun	Maximum fundamental frequency measured across acoustic signal	Continua e Racional
meandom	Average of dominant frequency measured across acoustic signal	Continua e Racional
mindom	Minimum of dominant frequency measured across acoustic signal	Continua e Racional
maxdom	Maximum of dominant frequency measured across acoustic signal	Continua e Racional
dfrange	Range of dominant frequency measured across acoustic signal	Continua e Racional
modindx	Modulation Index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range.	Continua e Racional

**EACH**Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Medidas das Características

Nome	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
meanfreq	0.03936	0.16366	0.18484	0.18091	0.19915	0.25112
sd	0.01836	0.04195	0.05916	0.05713	0.06702	0.11527
median	0.01097	0.16959	0.19003	0.18562	0.21062	0.26122
Q25	0.0002288	0.1110865	0.1402864	0.1404556	0.1759388	0.2473469
Q75	0.04295	0.20875	0.22568	0.22476	0.24366	0.27347
IQR	0.01456	0.04256	0.09428	0.08431	0.11418	0.25223
skew	0.1417	16.496	21.971	31.402	29.317	347.255
kurt	2.068	5.670	8.319	36.569	13.649	1.309.613
sp.ent	0.7387	0.8618	0.9018	0.8951	0.9287	0.9820
sfm	0.03688	0.25804	0.39634	0.40822	0.53368	0.84294



Medidas das Características

Nome	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
mode	0.0000	0.1180	0.1866	0.1653	0.2211	0.2800
centroid	0.03936	0.16366	0.18484	0.18091	0.19915	0.25112
meanfun	0.05557	0.11700	0.14052	0.14281	0.16958	0.23764
minfun	0.009775	0.018223	0.046110	0.036802	0.047904	0.204082
maxfun	0.1031	0.2540	0.2712	0.2588	0.2775	0.2791
meandom	0.007812	0.419828	0.765795	0.829211	1.177.166	2.957.682
mindom	0.004883	0.007812	0.023438	0.052647	0.070312	0.458984
maxdom	0.007812	2.070.312	4.992.188	5.047.277	7.007.812	21.867.188
dfrange	0.000	2.045	4.945	4.995	6.992	21.844
modindx	0.00000	0.09977	0.13936	0.17375	0.20918	0.93237



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Análise de Componentes Principais



Métodos e Funções

Função	Descrição	Parâmetros
Cov	Analise de Covariância	Dataset
Cor	Analise de Correlação	Dataset
Summary	Analise de Medidas	Dataset / Lista (Retorno PCA)
Princomp	Execução do PCA	Dataset, cor=TRUE, scores=TRUE
Fviz_eig	Gráfico de variâncias	Lista (Retorno do PCA)
fviz_pca_var	Gráfico Componentes	Lista (Retorno do PCA) col.var="contrib" (Exibe contribuição das variáveis) gradient.cols (Esquema de cores) repel = TRUE (Evita sobreposição de rótulos)



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Matriz de Correlação - Dataset

	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx
meanfreq	1	-0,73904	0,92545	0,91142	0,741	-0,62761	-0,32233	-0,31604	-0,6012	-0,78433	0,68772	1	0,46084	0,38394	0,274	0,53667	0,22926	0,51953	0,51557	-0,21698
sd	-0,73904	1	-0,5626	-0,84693	-0,16108	0,87466	0,3146	0,34624	0,71662	0,83809	-0,52915	-0,73904	-0,46628	-0,34561	-0,12966	-0,48273	-0,35767	-0,48228	-0,476	0,12266
median	0,92545	-0,5626	1	0,77492	0,73185	-0,47735	-0,25741	-0,24338	-0,502	-0,66169	0,67743	0,92545	0,41491	0,3376	0,25133	0,45594	0,19117	0,43892	0,43562	-0,2133
Q25	0,91142	-0,84693	0,77492	1	0,47714	-0,87419	-0,31948	-0,35018	-0,64813	-0,76687	0,59128	0,91142	0,54504	0,32099	0,19984	0,4674	0,30225	0,45968	0,45439	-0,14138
Q75	0,741	-0,16108	0,73185	0,47714	1	0,00964	-0,20634	-0,14888	-0,17491	-0,3782	0,48686	0,741	0,15509	0,258	0,28558	0,35918	-0,02375	0,33511	0,33565	-0,21647
IQR	-0,62761	0,87466	-0,47735	-0,87419	0,00964	1	0,2495	0,31618	0,64081	0,6636	-0,40376	-0,62761	-0,53446	-0,22268	-0,06959	-0,33336	-0,35704	-0,33788	-0,33156	0,04125
skew	-0,32233	0,3146	-0,25741	-0,31948	-0,20634	0,2495	1	0,97702	-0,19546	0,07969	-0,43486	-0,32233	-0,16767	-0,21695	-0,08086	-0,33685	-0,06161	-0,30565	-0,30464	-0,16932
kurt	-0,31604	0,34624	-0,24338	-0,35018	-0,14888	0,31618	0,97702	1	-0,12764	0,10988	-0,40672	-0,31604	-0,19456	-0,2032	-0,04567	-0,30323	-0,10331	-0,2745	-0,27273	-0,20554
sp.ent	-0,6012	0,71662	-0,502	-0,64813	-0,17491	0,64081	-0,19546	-0,12764	1	0,86641	-0,3253	-0,6012	-0,51319	-0,30583	-0,12074	-0,29356	-0,29487	-0,32425	-0,31905	0,19807
sfm	-0,78433	0,83809	-0,66169	-0,76687	-0,3782	0,6636	0,07969	0,10988	0,86641	1	-0,48591	-0,78433	-0,42107	-0,3621	-0,19237	-0,42844	-0,28959	-0,43665	-0,43158	0,21148
mode	0,68772	-0,52915	0,67743	0,59128	0,48686	-0,40376	-0,43486	-0,40672	-0,3253	-0,48591	1	0,68772	0,32477	0,38547	0,17233	0,49148	0,19815	0,47719	0,47377	-0,18234
centroid	1	-0,73904	0,92545	0,91142	0,741	-0,62761	-0,32233	-0,31604	-0,6012	-0,78433	0,68772	1	0,46084	0,38394	0,274	0,53667	0,22926	0,51953	0,51557	-0,21698
meanfun	0,46084	-0,46628	0,41491	0,54504	0,15509	-0,53446	-0,16767	-0,19456	-0,51319	-0,42107	0,32477	0,46084	1	0,33939	0,31195	0,27084	0,16216	0,27798	0,27515	-0,05486
minfun	0,38394	-0,34561	0,3376	0,32099	0,258	-0,22268	-0,21695	-0,2032	-0,30583	-0,3621	0,38547	0,38394	0,33939	1	0,21399	0,37598	0,08202	0,31786	0,31649	0,00204
maxfun	0,274	-0,12966	0,25133	0,19984	0,28558	-0,06959	-0,08086	-0,04567	-0,12074	-0,19237	0,17233	0,274	0,31195	0,21399	1	0,33755	-0,24343	0,35539	0,35988	-0,36303
meandom	0,53667	-0,48273	0,45594	0,4674	0,35918	-0,33336	-0,33685	-0,30323	-0,29356	-0,42844	0,49148	0,53667	0,27084	0,37598	0,33755	1	0,09966	0,81284	0,8113	-0,18095
mindom	0,22926	-0,35767	0,19117	0,30225	-0,02375	-0,35704	-0,06161	-0,10331	-0,29487	-0,28959	0,19815	0,22926	0,16216	0,08202	-0,24343	0,09966	1	0,02664	0,00867	0,20021
maxdom	0,51953	-0,48228	0,43892	0,45968	0,33511	-0,33788	-0,30565	-0,2745	-0,32425	-0,43665	0,47719	0,51953	0,27798	0,31786	0,35539	0,81284	0,02664	1	0,99984	-0,42553
dfrange	0,51557	-0,476	0,43562	0,45439	0,33565	-0,33156	-0,30464	-0,27273	-0,31905	-0,43158	0,47377	0,51557	0,27515	0,31649	0,35988	0,8113	0,00867	0,99984	1	-0,42927
modindx	-0,21698	0,12266	-0,2133	-0,14138	-0,21647	0,04125	-0,16932	-0,20554	0,19807	0,21148	-0,18234	-0,21698	-0,05486	0,00204	-0,36303	-0,18095	0,20021	-0,42553	-0,42927	1

[correlacao] >= 0.7

0.3 < [correlacao] < 0.7

[correlacao] < 0.3



Matriz de Correlação - Componentes

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17	Comp.18	Comp.19	Comp.20
Comp.1	1	-4,5E-17	-2,2E-16	7,3E-16	2E-16	-3,6E-16	-4,3E-16	4,2E-16	-9,4E-17	1,3E-15	-1,8E-16	1E-15	-1,2E-15	1,5E-15	1,2E-15	1,1E-15	9,8E-15	0,11716	-0,01437	0,04915
Comp.2	-4,5E-17	1	-1,7E-15	5,2E-16	-3,6E-15	-1,4E-15	-2,2E-15	3,3E-15	-1,1E-15	-7E-16	-2,4E-15	-1,8E-15	3,7E-16	-1,1E-15	1,8E-15	-4,5E-15	-4,1E-15	-0,00862	-0,0323	0,03194
Comp.3	-2,2E-16	-1,7E-15	1	-5,1E-16	-3,5E-15	-1,1E-15	-1,8E-15	2,4E-15	-1,7E-15	-6,9E-16	-7,6E-16	-1,7E-16	-1,1E-15	2,8E-16	3,6E-16	7,4E-15	-3,8E-16	0,00947	-0,00351	-0,0259
Comp.4	7,3E-16	5,2E-16	-5,1E-16	1	2,5E-16	-1,7E-15	-4,5E-16	2,3E-16	1,8E-18	-1,4E-15	-1,4E-15	1,1E-15	4,4E-15	-4,6E-16	-9,1E-16	-6,2E-15	-1E-14	-0,0816	-0,01299	-0,11217
Comp.5	2E-16	-3,6E-15	-3,5E-15	2,5E-16	1	-5,2E-15	-4,3E-15	2,1E-15	-2,8E-15	-6,1E-15	1,7E-15	3,8E-15	-2,4E-15	1,1E-15	3,7E-15	3,8E-15	-5,3E-15	0,02364	-0,05665	-0,02606
Comp.6	-3,6E-16	-1,4E-15	-1,1E-15	-1,7E-15	-5,2E-15	1	-8,3E-16	3,1E-15	5,4E-16	-1,7E-15	2,2E-15	9,8E-16	-3,9E-17	-5E-16	-4E-15	1,9E-15	2,1E-15	0,00757	0,04452	-0,12768
Comp.7	-4,3E-16	-2,2E-15	-1,8E-15	-4,5E-16	-4,3E-15	-8,3E-16	1	3,1E-15	3,9E-15	-2,4E-15	2,1E-15	2,3E-16	9,7E-16	-3,5E-16	-3,5E-15	-9,3E-15	1,8E-15	-0,01483	0,00799	-0,09788
Comp.8	4,2E-16	3,3E-15	2,4E-15	2,3E-16	2,1E-15	3,1E-15	3,1E-15	1	-6,6E-17	1,9E-15	-2E-15	3,4E-15	-5,4E-16	2E-15	5,6E-15	-5,9E-15	-1,6E-14	0,00252	0,05632	-0,1402
Comp.9	-9,4E-17	-1,1E-15	-1,7E-15	1,8E-18	-2,8E-15	5,4E-16	3,9E-15	-6,6E-17	1	1,8E-15	3,6E-15	1,5E-15	-4,8E-16	1,8E-15	-3,3E-15	-5,6E-15	9,9E-15	-0,0103	0,10939	0,07145
Comp.10	1,3E-15	-7E-16	-6,9E-16	-1,4E-15	-6,1E-15	-1,7E-15	-2,4E-15	1,9E-15	1,8E-15	1	1,1E-15	3,7E-15	-4,4E-15	8E-16	6,7E-15	-1,1E-15	3,1E-14	-0,03362	0,06241	0,35333
Comp.11	-1,8E-16	-2,4E-15	-7,6E-16	-1,4E-15	1,7E-15	2,2E-15	2,1E-15	-2E-15	3,6E-15	1,1E-15	1	-9,9E-16	-3,4E-15	-1,1E-15	-5,8E-15	7,4E-15	5E-15	0,04429	0,24558	-0,08372
Comp.12	1E-15	-1,8E-15	-1,7E-16	1,1E-15	3,8E-15	9,8E-16	2,3E-16	3,4E-15	1,5E-15	3,7E-15	-9,9E-16	1	-8,9E-16	-5,8E-15	-8,8E-16	7,5E-15	4,7E-14	-0,03078	-0,01824	0,63671
Comp.13	-1,2E-15	3,7E-16	-1,1E-15	4,4E-15	-2,4E-15	-3,9E-17	9,7E-16	-5,4E-16	-4,8E-16	-4,4E-15	-3,4E-15	-8,9E-16	1	-3,3E-15	1,9E-14	2,2E-14	3,3E-14	0,0871	0,00925	0,15502
Comp.14	1,5E-15	-1,1E-15	2,8E-16	-4,6E-16	1,1E-15	-5E-16	-3,5E-16	2E-15	1,8E-15	8E-16	-1,1E-15	-5,8E-15	-3,3E-15	1	1,2E-15	1,2E-14	8,2E-14	0,16705	-0,63165	0,14626
Comp.15	1,2E-15	1,8E-15	3,6E-16	-9,1E-16	3,7E-15	-4E-15	-3,5E-15	5,6E-15	-3,3E-15	6,7E-15	-5,8E-15	-8,8E-16	1,9E-14	1,2E-15	1	2,1E-14	1,6E-13	0,32999	-0,18318	0,29003
Comp.16	1,1E-15	-4,5E-15	7,4E-15	-6,2E-15	3,8E-15	1,9E-15	-9,3E-15	-5,9E-15	-5,6E-15	-1,1E-15	7,4E-15	7,5E-15	2,2E-14	1,2E-14	2,1E-14	1	1,5E-13	0,09042	-0,36333	-0,5022
Comp.17	9,8E-15	-4,1E-15	-3,8E-16	-1E-14	-5,3E-15	2,1E-15	1,8E-15	-1,6E-14	9,9E-15	3,1E-14	5E-15	4,7E-14	3,3E-14	8,2E-14	1,6E-13	1,5E-13	1	-0,90366	-0,13895	0,01219
Comp.18	0,11716	-0,00862	0,00947	-0,0816	0,02364	0,00757	-0,01483	0,00252	-0,0103	-0,03362	0,04429	-0,03078	0,0871	0,16705	0,32999	0,09042	-0,90366	1	-0,02509	0,06316
Comp.19	-0,01437	-0,0323	-0,00351	-0,01299	-0,05665	0,04452	0,00799	0,05632	0,10939	0,06241	0,24558	-0,01824	0,00925	-0,63165	-0,18318	-0,36333	-0,13895	-0,02509	1	0,08612
Comp.20	0,04915	0,03194	-0,0259	-0,11217	-0,02606	-0,12768	-0,09788	-0,1402	0,07145	0,35333	-0,08372	0,63671	0,15502	0,14626	0,29003	-0,5022	0,01219	0,06316	0,08612	1

|correlacao| >= 0.7 0.3 < |correlacao| < 0.7 |correlacao| < 0.3

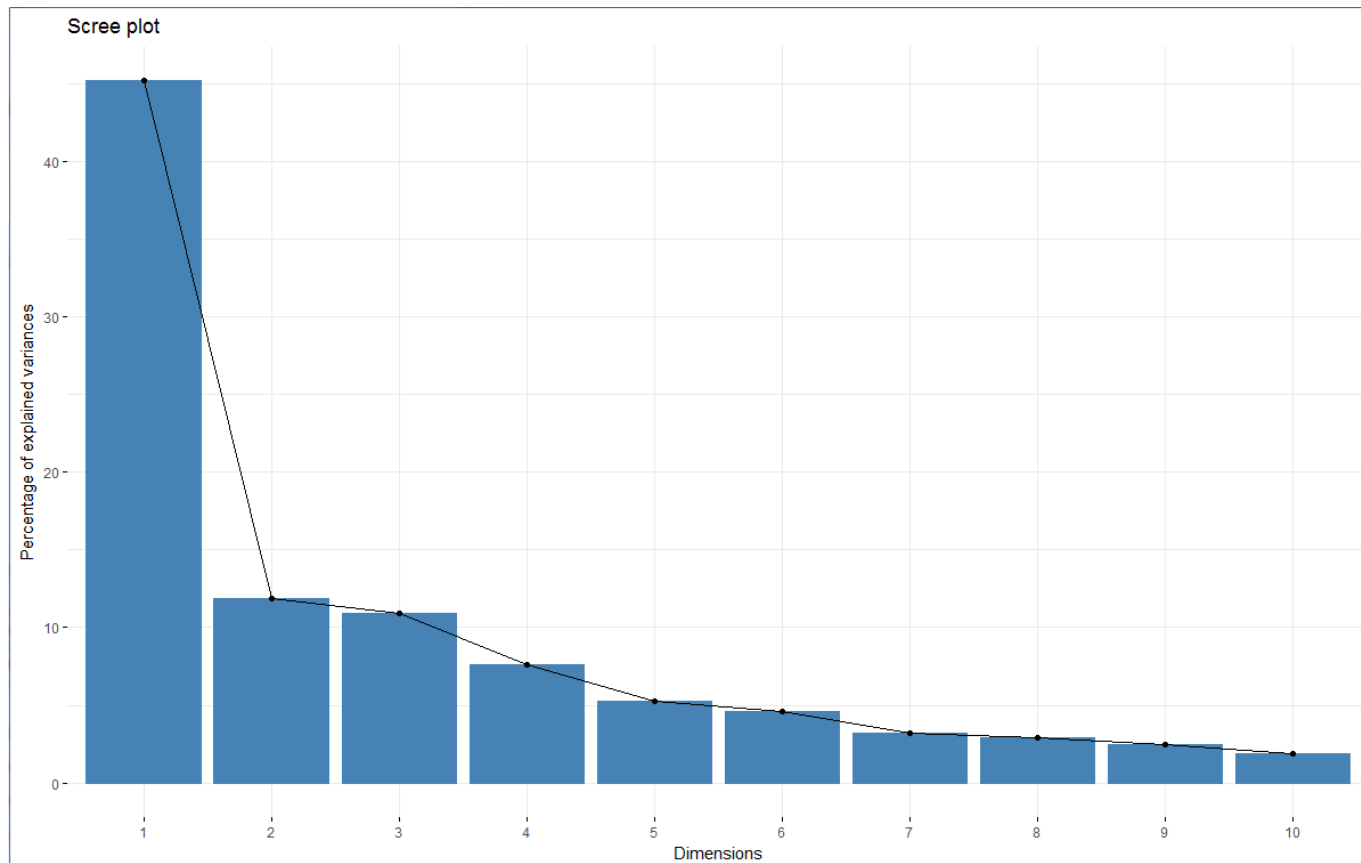


EACH



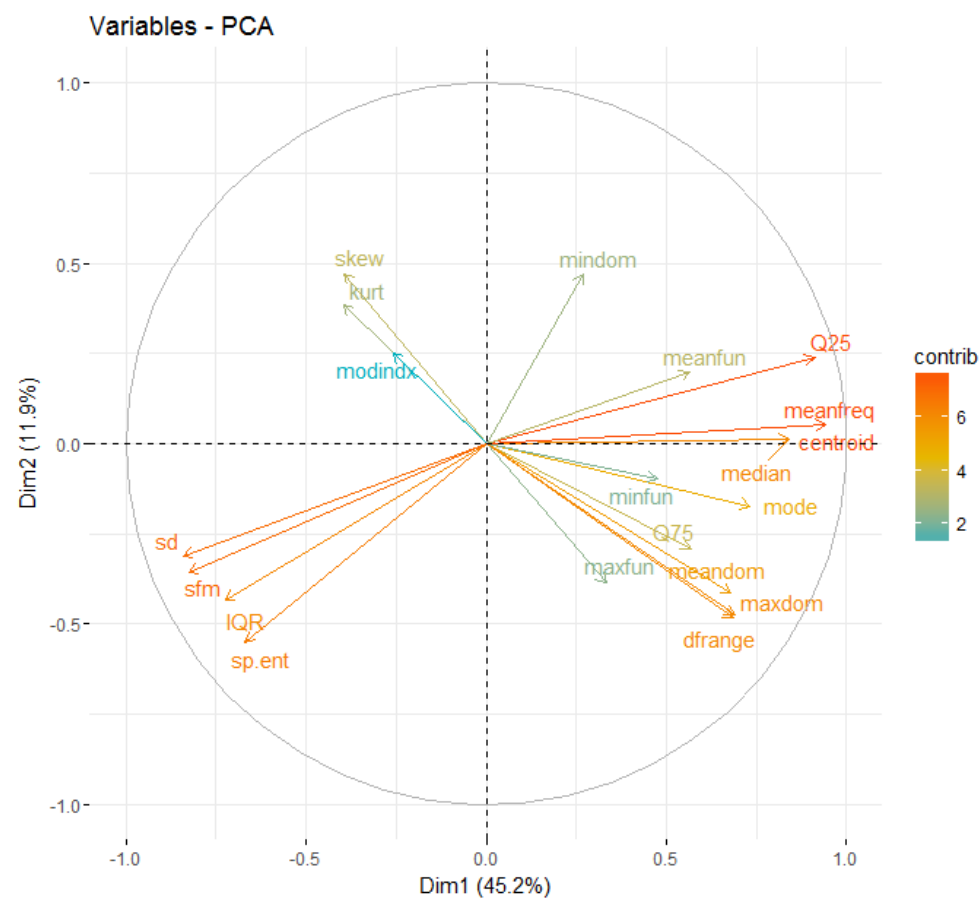
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Resultados



**Percentual de Variância x
Componente Principal**

Resultados



As Variáveis correlacionadas positivamente apontam para o mesmo lado do gráfico.

As Variáveis correlacionadas negativamente apontam para lados opostos do gráfico.

A cor das variáveis mostra o quando ela contribuiu para a formação dos componentes principais.

Resultados

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7
Standard deviation	3.0072044	1.5408186	1.4771553	1.23448476	1.02897402	0.96072539	0.80056008
Proportion of Variance	0.4521639	0.1187061	0.1090994	0.07619763	0.05293938	0.04614966	0.03204482
Cumulative Proportion	0.4521639	0.5708700	0.6799694	0.75616702	0.80910640	0.85525606	0.88730089

	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12	Comp. 13	Comp. 14
Standard deviation	0.76136639	0.70024659	0.61245694	0.57548340	0.423283006	0.371561974	0.306283001
Proportion of Variance	0.02898394	0.02451726	0.01875518	0.01655906	0.008958425	0.006902915	0.004690464
Cumulative Proportion	0.91628482	0.94080209	0.95955726	0.97611632	0.985074747	0.991977662	0.996668126

	Comp. 15	Comp. 16	Comp. 17	Comp. 18	Comp. 19	Comp. 20
Standard deviation	0.213968620	0.1136242763	0.0891315808	2.023668e-08	0	0
Proportion of Variance	0.002289129	0.0006455238	0.0003972219	2.047616e-17	0	0
Cumulative Proportion	0.998957254	0.9996027781	1.0000000000	1.000000e+00	1	1

**EACH**Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Escolha dos Componentes

Variância acumulada: Escolher os n primeiros componentes onde a variância acumula seja o menor valor mais próximo de 80%.

Variância Individual dos componentes principais: Considerar apenas os componentes principais que possuem variância ≥ 1 .

Com base nas regras acima, escolhemos os cinco primeiros componentes:

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Standard deviation	3.0072044	1.5408186	1.4771553	1.23448476	1.02897402
Proportion of Variance	0.4521639	0.1187061	0.1090994	0.07619763	0.05293938
Cumulative Proportion	0.4521639	0.5708700	0.6799694	0.75616702	0.80910640

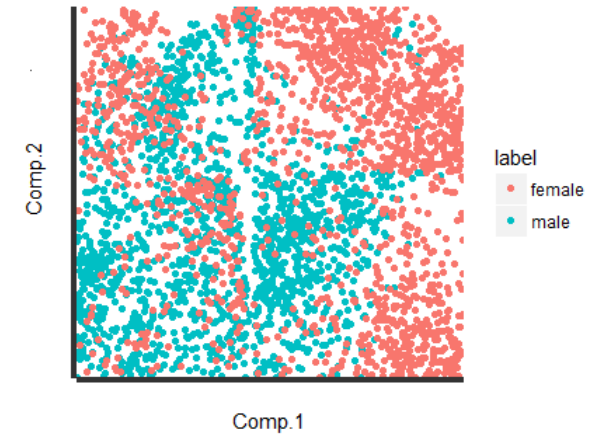
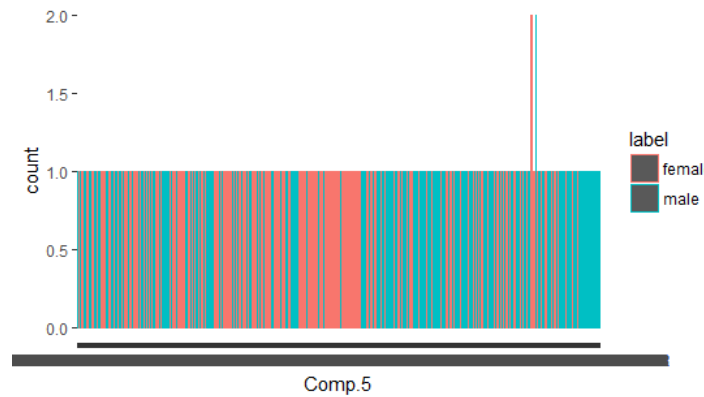
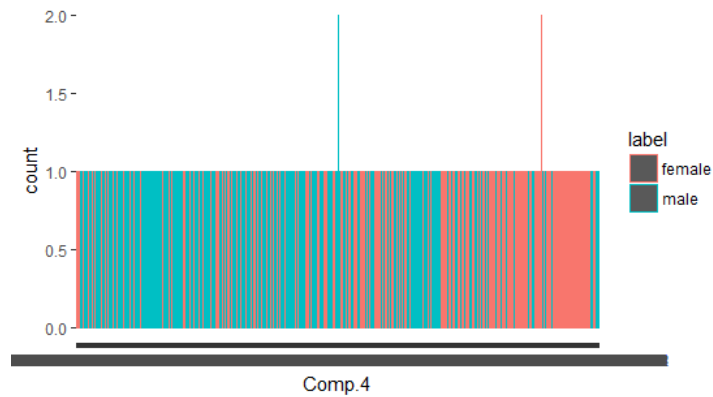
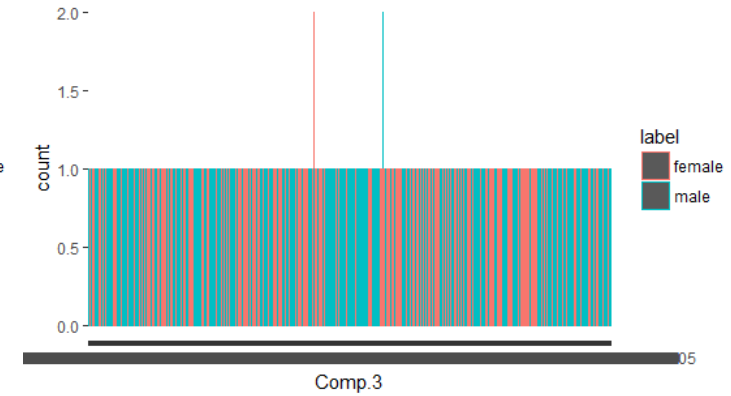
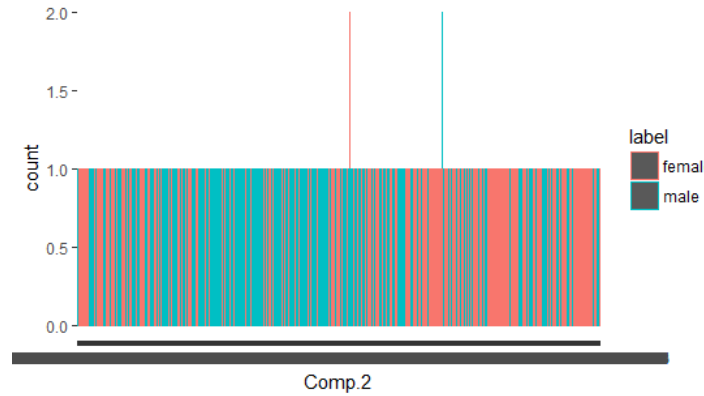
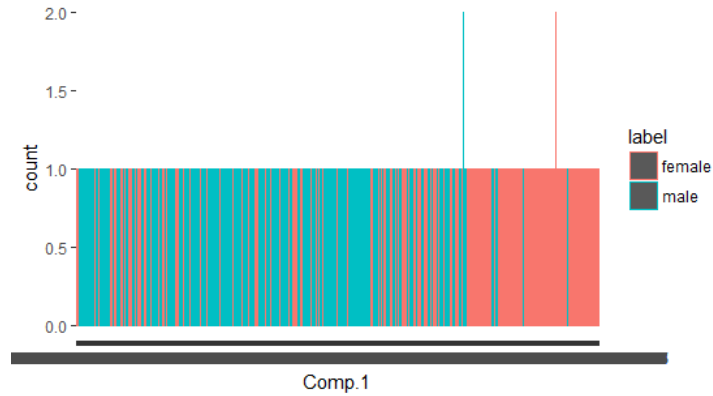


EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Histogramas e Dispersão





EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Discussão

- PCA reduziu a dimensionalidade do dataset em 75%;
- Pequena perda de informação na escolha dos componentes;
- Não foram encontradas diferenças significativas nos resultados entre as funções testadas (princomp, prcomp, utilizando matriz de correlação e covariância);
- Escolha da função princomp devido ao baixo numero de características do dataset;
- PCA transformou variáveis com correlação média/alta em componentes com correlação baixa;
- Variáveis com maior grau de correlação tiveram maior contribuição na formação dos componentes;
- PCA aumentou a mistura entre as classes;



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Seleção de Características



Métodos e Funções

Função	Descrição	Parâmetros
relief	Calculo dos Pesos	dataset, neighbours.count = 5, sample.size = 3168
Cutoff.k	Seleção de Subconjuntos	Array de pesos, Quantidade de características



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Resultados

meanfun	0.136259694	dfrange	0.028709800
IQR	0.091543117	maxdom	0.028632422
Q25	0.073612411	Q75	0.028336438
sd	0.060222125	meanfreq	0.026424285
sp.ent	0.055612398	centroid	0.026424285
mode	0.051762411	maxfun	0.026358006
sfm	0.042358517	modindx	0.025066052
meandom	0.036232346	minfun	0.023221873
median	0.035739156	skew	0.012851451
mindom	0.030321237	kurt	0.005483969



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Características Seleccionadas

- PCA Reduziu a dimensionalidade em 75% (Os cinco primeiros componentes foram seleccionados);
- As cinco primeiras características seleccionadas pelo relief apresentaram um bom peso (49% do peso total);
- Optamos por utilizar o mesmo número de características do PCA, para fins de comparação dos resultados;



Características Seleccionadas

Característica	Relief
meanfun	0.136259694
IQR	0.091543117
Q25	0.073612411
sd	0.060222125
sp.ent	0.055612398

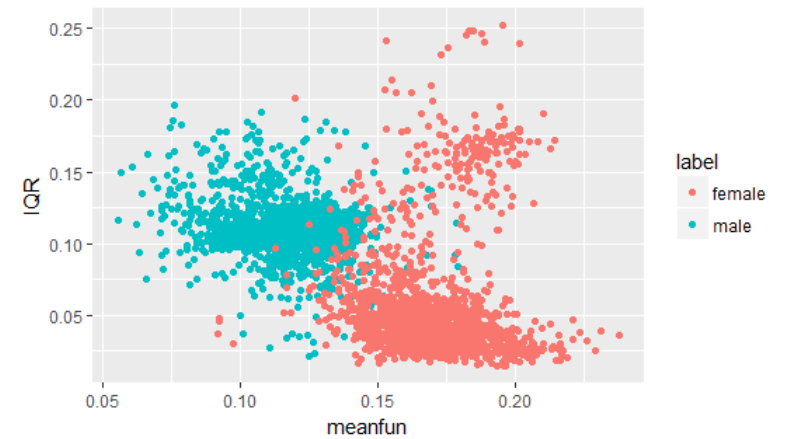
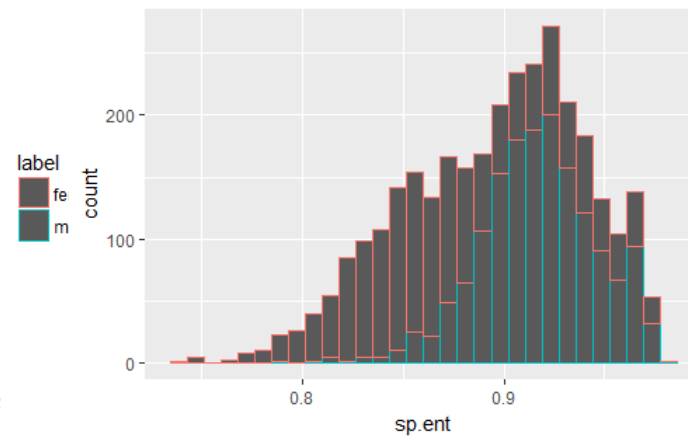
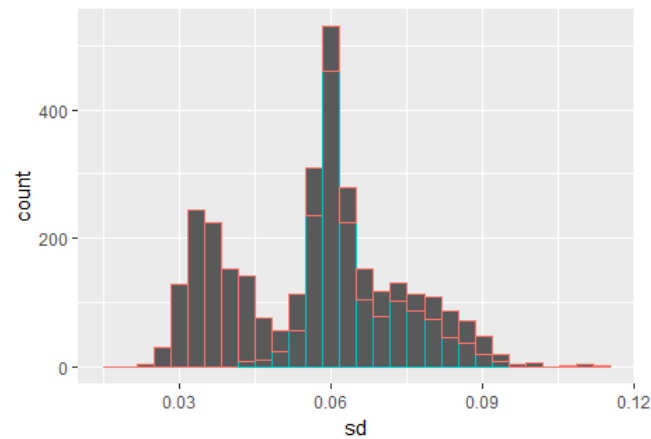
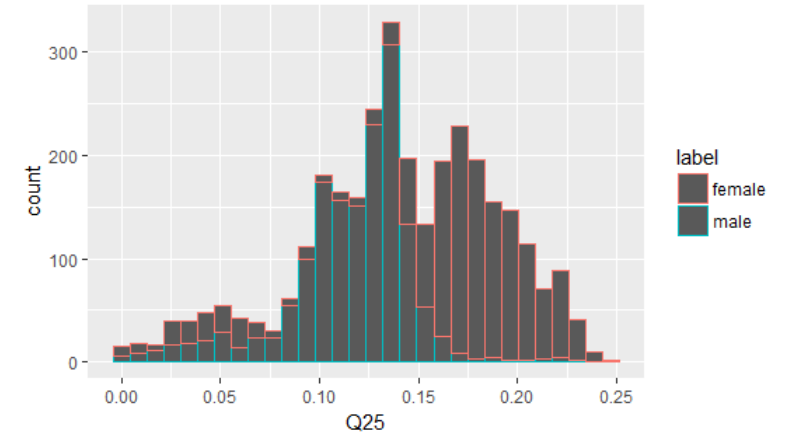
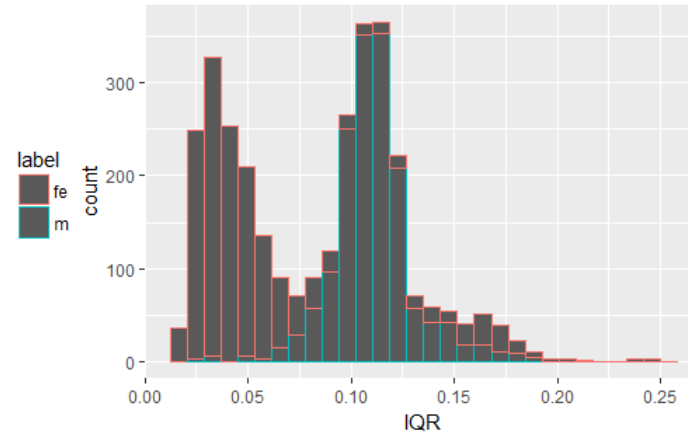
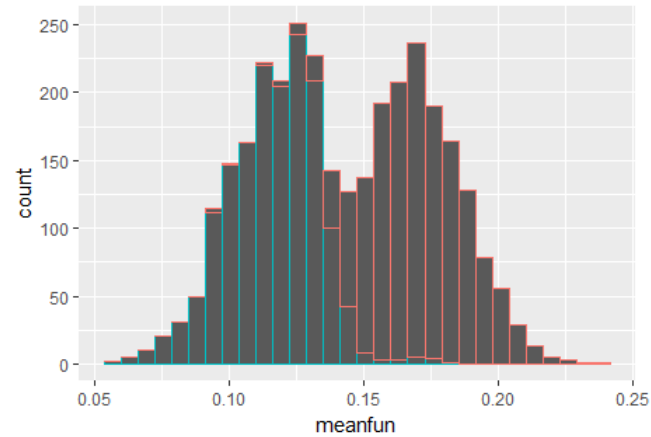
Peso acumulado de todas as características: 0.860287

Peso acumulado das características seleccionadas: 0.4243438

Percentual do peso das características seleccionadas: 49%



Histogramas e Dispersão





EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Procedimentos utilizados para execução dos algoritmos de classificação



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Validação Cruzada e Medidas de Acurácia

- Todas as execuções foram realizadas utilizando validação cruzada, com 10 partições, com diferença máxima de tamanho em 1 elemento, e distribuição proporcional das classes.
 - Como medidas de acurácia foram consideradas:
 - **Erro:** $fp + fn / tot$ (falso positivo + falso negativo / total de elementos)
 - **Precisão:** $tp / tp + fp$ (verdadeiro positivo / verdadeiro positivo + falso positivo)
 - **Sensibilidade:** tp / pos (verdadeiro positivo / total de elementos positivos)
- Obs.:** Para o cálculo das medidas, a classe **male** foi considerada como sendo positiva e a classe **female**, como negativa.
- A Escolha das melhores execuções foram realizadas com base no **score**, dado pela seguinte formula: Sensibilidade + Precisão - Erro.



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Maquinas de Vetores de Suporte (SVM)



Métodos e Funções

Função	Descrição	Parâmetros
svm (e1071)	Executa svm	formula = label; data = (completo, relief, pca) kernel = (linear, polinomial, radial, sigmoid) cost = (0.1, 0.5, 1, 5, 10) gamma = (0.1, 0.5, 1, 5, 10) coef0 = (0.1, 0.5, 1, 5, 10) degree = (0.1, 0.5, 1, 5, 10) type = C
predict (e1071)	Realiza predição	object = modelo retornado pela função svm newdata = amostra de teste Type = "class"



Analise Kernel Linear

- Melhor desempenho com o dataset completo e com as características selecionadas pelo Relief;
- Desempenho inferior com os componentes do PCA.

Melhor Execução com todas as características:

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Todas	5	linear	0	0	0	10	0,024305195	0,975372184	0,976309414	1,927376

Melhor Execução Relief:

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Relief	2	linear	0	0	0	0,5	0,029672963	0,975388106	0,965838174	1,911553

Melhor Execução PCA:

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
PCA	1	linear	0	0	0	0,1	0,113018608	0,931175862	0,855920559	1,674078



Analise Kernel Polinomial

- Melhores Execuções com os parâmetros **degree** = 5 e **gamma** = 0,1
- Melhor desempenho com o dataset completo e com as características selecionadas pelo Relief;
- Desempenho inferior com os componentes do PCA, no entanto superou o kernel linear.

Melhores Execuções

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Todas	391	polynomial	5	0,1	1	0,1	0,019576329	0,980427514	0,980484271	1,941335
Relief	400	polynomial	5	0,1	5	10	0,021462085	0,977266937	0,979832958	1,935638
Todas	388	polynomial	5	0,1	0,5	1	0,021781536	0,977903829	0,978599392	1,934722
Relief	394	polynomial	5	0,1	1	5	0,022409456	0,976641987	0,97853444	1,932767

Melhor Execução PCA

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
PCA	397	polynomial	5	0,1	5	0,5	0,055239987	0,960862989	0,931050057	1,836673



Analise Kernel Radial

- Melhores Execuções com o parâmetro **gamma** = 0,1
- Desempenho semelhante ao Kernel Polinomial.

Melhor Execução com todas as características

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Todas	635	radial	0	0,1	0	10	0,017370123	0,981048483	0,984292932	1,947971

Melhor Execução Relief

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Relief	643	radial	0	1	0	1	0,020522701	0,975364223	0,983542101	1,938384

Melhor Execução PCA

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
PCA	640	radial	0	0,5	0	10	0,055242982	0,96336677	0,929214921	1,837339



Analise Kernel Sigmoid

- Pior desempenho, em comparação ao outros Kernels.
- Melhor desempenho com as características selecionadas pelo Relief e com os componentes principais;
- Desempenho inferior com o dataset completo

Melhores Execuções

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Relief	656	sigmoid	0	0,1	0,1	0,1	0,08963383	0,948228644	0,882367513	1,740962
PCA	675	sigmoid	0	0,1	5	10	0,133838797	0,968442003	0,804554808	1,639158
PCA	656	sigmoid	0	0,1	0,1	0,1	0,134464721	0,931155959	0,823460832	1,620152
Relief	761	sigmoid	0	10	0,5	0,1	0,135083656	0,88005334	0,854347173	1,599317

Melhor Execução com todas as características

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Todas	656	sigmoid	0	0,1	0,1	0,1	0,280937787	0,72977868	0,714628615	1,16347



Analise Parâmetro degree

- Parâmetro utilizado apenas com o Kernel Polinomial.
- Valores Testados: 0.1, 0.5, 1, 5, 10
- Melhores resultados com o valor: 5
- Em geral o PCA precisou de valores maiores para o parâmetro, em comparação as outras abordagens de seleção de características.

Melhores Execuções

Caracteristicas	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Todas	391	polynomial	5	0,1	1	0,1	0,019576329	0,980427514	0,980484271	1,941335
Relief	400	polynomial	5	0,1	5	10	0,021462085	0,977266937	0,979832958	1,935638
Todas	388	polynomial	5	0,1	0,5	1	0,021781536	0,977903829	0,978599392	1,934722
Relief	394	polynomial	5	0,1	1	5	0,022409456	0,976641987	0,97853444	1,932767
Relief	405	polynomial	5	0,1	10	10	0,022416444	0,975376164	0,979805118	1,932765



Analise Parâmetro gamma

- Parâmetro utilizado com os kernels polinomial, radial e sigmoid.
- Valores Testados: 0.1, 0.5, 1, 5, 10
- Melhores resultados com os valores: 0.1

Melhores Execuções

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Todas	635	radial	0	0,1	0	10	0,017370123	0,981048483	0,984292932	1,947971
Todas	634	radial	0	0,1	0	5	0,017358144	0,981072367	0,984218252	1,947932
Todas	632	radial	0	0,1	0	0,5	0,017677595	0,980423533	0,984288779	1,947035
Todas	639	radial	0	0,5	0	5	0,018630955	0,975993153	0,986624096	1,943986
Todas	391	polynomial	5	0,1	1	0,1	0,019576329	0,980427514	0,980484271	1,941335



Analise Parâmetro coef0

- Parâmetro utilizado com os kernels polinomial e sigmoid.
- Valores Testados: 0.1, 0.5, 1, 5, 10
- Valores testados não apresentaram variação nos resultados.

Melhores Execuções

Características ▾	Execucao ▾	kernel ▾	degree ▾	gamma ▾	coef0 ▾	cost ▾	erro ▾	sensibilidade ▾	precisao ▾	score ▾
Todas	391	polynomial	5	0,1	1	0,1	0,019576329	0,980427514	0,980484271	1,941335
Relief	400	polynomial	5	0,1	5	10	0,021462085	0,977266937	0,979832958	1,935638
Todas	388	polynomial	5	0,1	0,5	1	0,021781536	0,977903829	0,978599392	1,934722
Relief	394	polynomial	5	0,1	1	5	0,022409456	0,976641987	0,97853444	1,932767
Relief	405	polynomial	5	0,1	10	10	0,022416444	0,975376164	0,979805118	1,932765



Analise Parâmetro cost

- Parâmetro utilizado com todos os kernels.
- Valores Testados: 0.1, 0.5, 1, 5, 10
- Valores testados não apresentaram variação nos resultados.

Melhores Execuções

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Todas	635	radial	0	0,1	0	10	0,017370123	0,981048483	0,984292932	1,947971
Todas	634	radial	0	0,1	0	5	0,017358144	0,981072367	0,984218252	1,947932
Todas	632	radial	0	0,1	0	0,5	0,017677595	0,980423533	0,984288779	1,947035
Todas	639	radial	0	0,5	0	5	0,018630955	0,975993153	0,986624096	1,943986
Todas	391	polynomial	5	0,1	1	0,1	0,019576329	0,980427514	0,980484271	1,941335



Analise PCA

- Pior desempenho, em comparação com o dataset completo e as características selecionadas pelo Relief;
- Necessitou de mais dimensões para separação das classes;
- Melhor desempenho com os kernels polinomial e radial;
- Desempenho ruim com os kernels linear e sigmoid;

Melhores Execuções

Caracteristicas	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
PCA	640	radial	0	0,5	0	10	0,055242982	0,96336677	0,929214921	1,837339
PCA	397	polynomial	5	0,1	5	0,5	0,055239987	0,960862989	0,931050057	1,836673
PCA	396	polynomial	5	0,1	5	0,1	0,055555445	0,960878911	0,930607703	1,835931
PCA	643	radial	0	1	0	1	0,056181368	0,962753762	0,92791398	1,834486
PCA	402	polynomial	5	0,1	10	0,5	0,057130735	0,960859008	0,927715996	1,831444



Analise Relief

- Desempenho próximo ao dataset completo;
- Melhor desempenho que o PCA, rodando com o mesmo número de características;
- Favorecido devido à separação natural das classes no dataset original;
- Melhor desempenho com os kernels polinomial e radial;
- Bom desempenho com o kernel linear;

Melhores Execuções

Caracteristicas	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Relief	643	radial	0	1	0	1	0,020522701	0,975364223	0,983542101	1,938384
Relief	642	radial	0	1	0	0,5	0,021151619	0,976653929	0,981039157	1,936541
Relief	640	radial	0	0,5	0	10	0,021150621	0,974759175	0,982861016	1,93647
Relief	644	radial	0	1	0	5	0,021466078	0,974743253	0,982449808	1,935727
Relief	400	polynomial	5	0,1	5	10	0,021462085	0,977266937	0,979832958	1,935638



Analise Todas as Características

- Melhor desempenho, em comparação ao PCA e ao Relief;
- Melhor desempenho com os kernels polinomial e radial;
- Bom desempenho com o kernel linear;

Melhores Execuções

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Todas	635	radial	0	0,1	0	10	0,017370123	0,981048483	0,984292932	1,947971
Todas	634	radial	0	0,1	0	5	0,017358144	0,981072367	0,984218252	1,947932
Todas	632	radial	0	0,1	0	0,5	0,017677595	0,980423533	0,984288779	1,947035
Todas	639	radial	0	0,5	0	5	0,018630955	0,975993153	0,986624096	1,943986
Todas	391	polynomial	5	0,1	1	0,1	0,019576329	0,980427514	0,980484271	1,941335



Resultado geral

- Melhores resultados com o dataset completo e com as características selecionadas pelo Relief, com os kernels radial e polinomial

Melhores Execuções

Características	Execucao	kernel	degree	gamma	coef0	cost	erro	sensibilidade	precisao	score
Todas	635	radial	0	0,1	0	10	0,017370123	0,981048483	0,984292932	1,947971
Todas	634	radial	0	0,1	0	5	0,017358144	0,981072367	0,984218252	1,947932
Todas	632	radial	0	0,1	0	0,5	0,017677595	0,980423533	0,984288779	1,947035
Todas	639	radial	0	0,5	0	5	0,018630955	0,975993153	0,986624096	1,943986
Todas	391	polynomial	5	0,1	1	0,1	0,019576329	0,980427514	0,980484271	1,941335
Todas	633	radial	0	0,1	0	1	0,019569341	0,978552663	0,982319441	1,941303
Todas	638	radial	0	0,5	0	1	0,019890788	0,974759175	0,985406239	1,940275
Relief	643	radial	0	1	0	1	0,020522701	0,975364223	0,983542101	1,938384
Todas	640	radial	0	0,5	0	10	0,020518708	0,972844519	0,986033469	1,938359
Relief	642	radial	0	1	0	0,5	0,021151619	0,976653929	0,981039157	1,936541
Relief	640	radial	0	0,5	0	10	0,021150621	0,974759175	0,982861016	1,93647
Relief	644	radial	0	1	0	5	0,021466078	0,974743253	0,982449808	1,935727
Relief	400	polynomial	5	0,1	5	10	0,021462085	0,977266937	0,979832958	1,935638
Relief	645	radial	0	1	0	10	0,021461087	0,975396067	0,98162716	1,935562
Relief	639	radial	0	0,5	0	5	0,021467077	0,974739272	0,98225541	1,935528



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Discussão

- Dataset completo apresentou melhores resultados.
- Redução do número de características (75%), com PCA ou Relief, compensa devido a pequena diferença nos resultados.
- Relief apresentou resultados superiores ao PCA, devido a separação natural das classes no dataset original.
- Kernels Polynomial e Radial apresentaram melhores resultados.



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Redes Neurais



Métodos e Funções

Função	Descrição	Parâmetros
nnet (nnet)	Executa Redes Neurais	x = amostra de treinamento (completo, relief, pca) Y = labels syze = (2,4,6,8,10) rang = (0.2, 0.4, 0.6, 0.8, 1) decay = (0.01, 0.02, 0.04, 0.08, 0.1) maxit = 500
predict (nnet)	Realiza predição	object = modelo retornado pela função nnet newdata = amostra de teste



Analise Parâmetro syze

- Valores Testados: 2, 4, 6, 8, 10
- Melhores resultados com os valores: 10, 8, 6
- Relief precisou de menos neurônios para apresentar bons resultados.

Melhores Execuções

Características	Execução	size	rang	decay	erro	sensibilidade	precisao	score
Todas	114	10	0,6	0,08	0,020835164	0,981056445	0,977471288	1,93769257
Todas	124	10	1	0,08	0,021151619	0,980415572	0,977446899	1,93671085
Todas	72	6	1	0,02	0,021150621	0,978540721	0,979234543	1,93662464
Todas	122	10	1	0,02	0,021462085	0,979165672	0,978009529	1,93571312
Todas	94	8	0,8	0,08	0,02146508	0,979169652	0,977979729	1,9356843



Analise Parâmetro rang

- Valores Testados: 0.2, 0.4, 0.6, 0.8, 1.
- Valores testados não influenciaram nos resultados.

Melhores Execuções

Características ▾	Execução ▾	size ▾	rang ▾	decay ▾	erro ▾	sensibilidade ▾	precisao ▾	score ▾
Todas	114	10	0,6	0,08	0,020835164	0,981056445	0,977471288	1,93769257
Todas	124	10	1	0,08	0,021151619	0,980415572	0,977446899	1,93671085
Todas	72	6	1	0,02	0,021150621	0,978540721	0,979234543	1,93662464
Todas	122	10	1	0,02	0,021462085	0,979165672	0,978009529	1,93571312
Todas	94	8	0,8	0,08	0,02146508	0,979169652	0,977979729	1,9356843



Analise Parâmetro decay

- Valores Testados: 0.01, 0.02, 0.04, 0.08, 0.1.
- Em conjunto com o Relief, os valores 0.01 e 0.02 apresentaram melhores resultados.

Melhores Execuções Relief

Características	Execução	size	rang	decay	erro	sensibilidade	precisao	score
Relief	46	4	1	0,01	0,027782215	0,976001115	0,968878644	1,91709754
Relief	97	8	1	0,02	0,027776225	0,976013056	0,968815338	1,91705217
Relief	16	2	0,8	0,01	0,027782215	0,974719369	0,970108466	1,91704562
Relief	81	8	0,4	0,01	0,027774228	0,975372184	0,969300115	1,91689807
Relief	56	6	0,4	0,01	0,028092681	0,975376164	0,968956773	1,91624026



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Analise PCA

- Pior desempenho, em comparação com o dataset completo e as características selecionadas pelo Relief;
- Apresentou melhores resultados com um numero maior de neurônios comparado ao número de características

Melhores Execuções

Características	Execução	size	rang	decay	erro	sensibilidade	precisao	score
PCA	121	10	1	0,01	0,057776624	0,956412706	0,930291537	1,82892762
PCA	61	6	0,6	0,01	0,058397556	0,958363188	0,928145166	1,8281108
PCA	58	6	0,4	0,04	0,058708022	0,960246	0,925432592	1,82697057
PCA	57	6	0,4	0,02	0,058721998	0,959565321	0,92599953	1,82684285
PCA	125	10	1	0,1	0,05870902	0,958327362	0,927067313	1,82668566



Analise Relief

- Em geral precisou de menos neurônios para apresentar bons resultados, em comparação com o PCA e o dataset completo;
- Apresentou melhores resultados com a taxa de aprendizagem baixa, 0.01 e 0.02.

Melhores Execuções

Características	Execução	size	rang	decay	erro	sensibilidade	precisao	score
Relief	46	4	1	0,01	0,027782215	0,976001115	0,968878644	1,91709754
Relief	97	8	1	0,02	0,027776225	0,976013056	0,968815338	1,91705217
Relief	16	2	0,8	0,01	0,027782215	0,974719369	0,970108466	1,91704562
Relief	81	8	0,4	0,01	0,027774228	0,975372184	0,969300115	1,91689807
Relief	56	6	0,4	0,01	0,028092681	0,975376164	0,968956773	1,91624026



Analise Todas as Características

- Necessitou de mais neurônios, devido ao maior numero de características;

Melhores Execuções

Características ▾	Execução ▾	size ▾	rang ▾	decay ▾	erro ▾	sensibilidade ▾	precisao ▾	score ▾
Todas	114	10	0,6	0,08	0,020835164	0,981056445	0,977471288	1,93769257
Todas	124	10	1	0,08	0,021151619	0,980415572	0,977446899	1,93671085
Todas	72	6	1	0,02	0,021150621	0,978540721	0,979234543	1,93662464
Todas	122	10	1	0,02	0,021462085	0,979165672	0,978009529	1,93571312
Todas	94	8	0,8	0,08	0,02146508	0,979169652	0,977979729	1,9356843



Resultado geral

- Melhores resultados com o dataset completo e número de neurônios entre 6 e 10.

Melhores Execuções

Características	Execução	size	rang	decay	erro	sensibilidade	precisao	score
Todas	114	10	0,6	0,08	0,020835164	0,981056445	0,977471288	1,93769257
Todas	124	10	1	0,08	0,021151619	0,980415572	0,977446899	1,93671085
Todas	72	6	1	0,02	0,021150621	0,978540721	0,979234543	1,93662464
Todas	122	10	1	0,02	0,021462085	0,979165672	0,978009529	1,93571312
Todas	94	8	0,8	0,08	0,02146508	0,979169652	0,977979729	1,9356843
Todas	107	10	0,4	0,02	0,021469073	0,979181594	0,977970884	1,9356834
Todas	58	6	0,4	0,04	0,021782534	0,979794602	0,977274137	1,9352862
Todas	93	8	0,8	0,04	0,021778541	0,979794602	0,976921726	1,93493779
Todas	123	10	1	0,04	0,021776544	0,97977868	0,976872249	1,93487438
Todas	78	8	0,2	0,04	0,021781536	0,979161691	0,977476051	1,93485621
Todas	116	10	0,8	0,01	0,021780537	0,977891888	0,978541968	1,93465332
Todas	84	8	0,4	0,08	0,022092002	0,979806544	0,976448764	1,93416331
Todas	119	10	0,8	0,08	0,022096993	0,979806544	0,976235786	1,93394534
Todas	98	8	1	0,04	0,022094997	0,97853276	0,97750531	1,93394307
Todas	51	6	0,2	0,01	0,022092002	0,976634026	0,97930553	1,93384755



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Discussão

- Dataset completo apresentou melhores resultados.
- Redução do numero de características (75%), com PCA ou Relief, compensa devido a pequena diferença nos resultados.
- Relief apresentou resultados superiores ao PCA.
- Em Geral Melhores resultados foram apresentados com o maior número de neurônios na camada oculta (10, 8), valores testados (2, 4, 6, 8, 10).
- Em geral Piores resultados foram apresentados com a taxa de aprendizagem baixa (0.01, 0.02), valores testados (0.01, 0.02, 0.04, 0.08, 0.1). Limite de épocas:500.
- Relief apresentou bons resultados com menor número de neurônios e taxa de aprendizagem mais baixa.



EACH |



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Naive Bayes



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Métodos e Funções

Função	Descrição	Parâmetros
naiveBayes (e1071)	Executa Naive Bayes	Formula = label data = amostra de treinamento (completo, relief, pca)
predict (e1071)	Realiza predição	object = modelo retornado pela função naiveBayes newdata = amostra de teste Type = "class"



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Resultado Geral

- Relief apresentou o melhor resultado e uma diferença significativa em relação ao PCA e ao dataset completo.

características ▼	execução ▼	erro ▼	sensibilidade ▼	precisao ▼	score ▼
Relief	1	0,045768279	0,974747234	0,936547393	1,865526348
PCA	1	0,10544164	0,924245681	0,872894474	1,691698515
Todas	1	0,107640858	0,900883688	0,88628203	1,67952486



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Discussão

- Maior diferença nos resultados em comparação com redes neurais e svm.
- Único algoritmo em que o dataset contendo todas as características foi superado.
- Características a mais trouxeram erro ao modelo.



EACH |



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Conclusão



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Conclusão

- A separação natural das classes favoreceu o algoritmo de seleção de características Relief, que se mostrou superior ao PCA e ao dataset completo.
- Relief selecionou as características que separavam melhor as classes enquanto a formação dos componentes principais acabou misturando as classes.
- Embora o dataset completo tenha apresentado melhor pontuação comparado ao Relief, nas execuções da SVM e das Redes Neurais, a diferença foi mínima e poderia ser removida facilmente com a adição de novas características ao Relief.



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Conclusão

- PCA necessitou de mais esforço por parte dos algoritmos para separar as classes, devido a mistura nos componentes principais, com SVM foi necessário um numero maior de dimensões para separar as classes, e com as Redes Neurais, um numero maior de neurônios.
- SVM e Redes Neurais apresentaram melhor desempenho comparados ao Nayve Bayes, por utilizarem álgebra linear para separação das classes.
- Concluimos que o Relief em conjunto com a SVM apresentou melhores resultados para o nosso dataset, devido a separação natural das classes através de algumas características.



EACH |



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Trabalhos Futuros



EACH



Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Trabalhos Futuros

- Testar o desempenho do Relief com SVM, outros números de características;
- Testar outras variações de valores para os parâmetros da SVM.