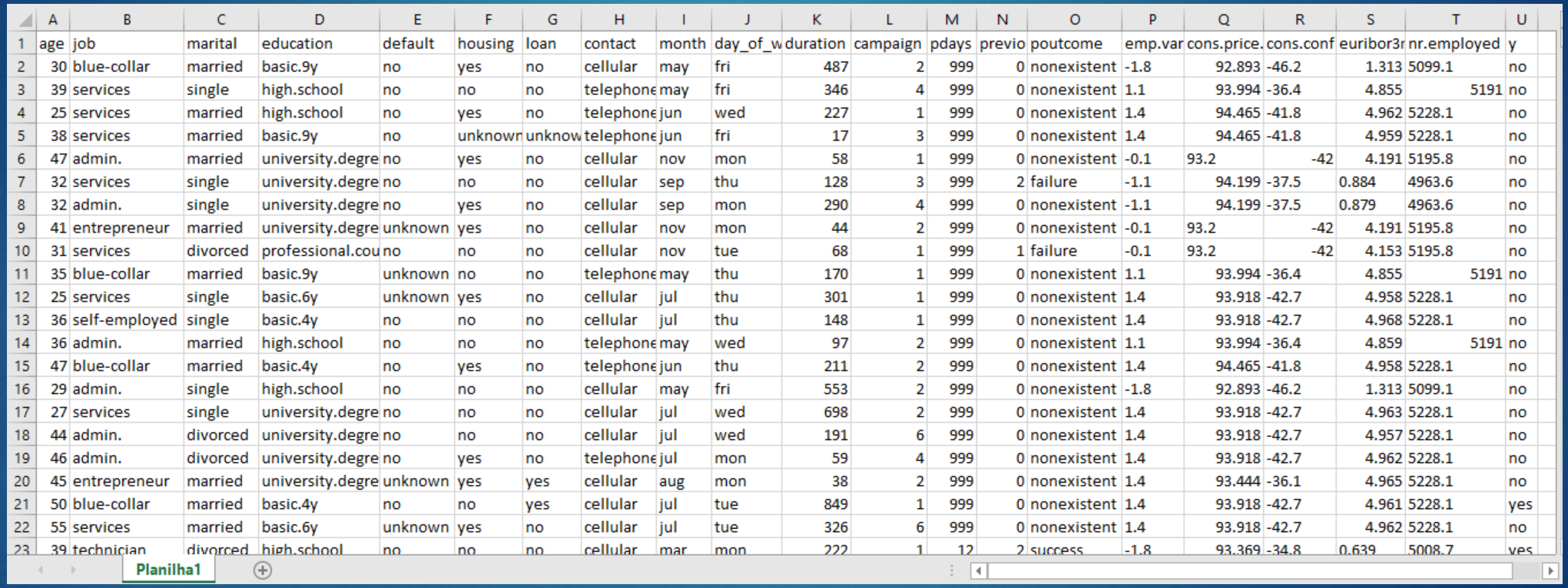
Comparação de classificadores: **SVM**

IGOR GUILHERME BIANCHI – 558400
CÉSAR ABREU DE ANDRADE – 412147
LUCAS SERRA DE ASSIS – 510475
FELIPE CHURUYUKI CHINEN – 496235
LUCAS DA COSTA MARTINS SILVA - 412031

Bank Marketing Data Set

- Prever se um consumidor tem o perfil de compra de um produto do banco
- 20 atributos:
  - **Dados do cliente:** idade, profissão, status relacionamento, educação, padrão de concessão de crédito, aluguel, empréstimo, tipo de contato, mês do último contato, dia do último contato, duração da ligação
  - **Outros:** ligações feitas, dias depois do último contato, número de ligações antes desta campanha, resultado da última campanha
  - **Índices sócio-econômicos:** taxa de variação de emprego, índice de preços ao consumidor, índice de confiança do consumidor, taxa de juros da Zona do Euro, número de empregados atualmente no banco
- **Classe: O cliente comprou o produto? Sim ou não.**

Dataset original



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	age	job	marital	education	default	housing	loan	contact	month	day_of_w	duration	campaign	pdays	previo	poutcome	emp.var	cons.price.	cons.conf	euribor3r	nr.employed	y
2	30	blue-collar	married	basic.9y	no	yes	no	cellular	may	fri	487	2	999	0	nonexistent	-1.8	92.893	-46.2	1.313	5099.1	no
3	39	services	single	high.school	no	no	no	telephone	may	fri	346	4	999	0	nonexistent	1.1	93.994	-36.4	4.855	5191	no
4	25	services	married	high.school	no	yes	no	telephone	jun	wed	227	1	999	0	nonexistent	1.4	94.465	-41.8	4.962	5228.1	no
5	38	services	married	basic.9y	no	unknown	unknow	telephone	jun	fri	17	3	999	0	nonexistent	1.4	94.465	-41.8	4.959	5228.1	no
6	47	admin.	married	university.degre	no	yes	no	cellular	nov	mon	58	1	999	0	nonexistent	-0.1	93.2	-42	4.191	5195.8	no
7	32	services	single	university.degre	no	no	no	cellular	sep	thu	128	3	999	2	failure	-1.1	94.199	-37.5	0.884	4963.6	no
8	32	admin.	single	university.degre	no	yes	no	cellular	sep	mon	290	4	999	0	nonexistent	-1.1	94.199	-37.5	0.879	4963.6	no
9	41	entrepreneur	married	university.degre	unknown	yes	no	cellular	nov	mon	44	2	999	0	nonexistent	-0.1	93.2	-42	4.191	5195.8	no
10	31	services	divorced	professional.cou	no	no	no	cellular	nov	tue	68	1	999	1	failure	-0.1	93.2	-42	4.153	5195.8	no
11	35	blue-collar	married	basic.9y	unknown	no	no	telephone	may	thu	170	1	999	0	nonexistent	1.1	93.994	-36.4	4.855	5191	no
12	25	services	single	basic.6y	unknown	yes	no	cellular	jul	thu	301	1	999	0	nonexistent	1.4	93.918	-42.7	4.958	5228.1	no
13	36	self-employed	single	basic.4y	no	no	no	cellular	jul	thu	148	1	999	0	nonexistent	1.4	93.918	-42.7	4.968	5228.1	no
14	36	admin.	married	high.school	no	no	no	telephone	may	wed	97	2	999	0	nonexistent	1.1	93.994	-36.4	4.859	5191	no
15	47	blue-collar	married	basic.4y	no	yes	no	telephone	jun	thu	211	2	999	0	nonexistent	1.4	94.465	-41.8	4.958	5228.1	no
16	29	admin.	single	high.school	no	no	no	cellular	may	fri	553	2	999	0	nonexistent	-1.8	92.893	-46.2	1.313	5099.1	no
17	27	services	single	university.degre	no	no	no	cellular	jul	wed	698	2	999	0	nonexistent	1.4	93.918	-42.7	4.963	5228.1	no
18	44	admin.	divorced	university.degre	no	no	no	cellular	jul	wed	191	6	999	0	nonexistent	1.4	93.918	-42.7	4.957	5228.1	no
19	46	admin.	divorced	university.degre	no	yes	no	telephone	jul	mon	59	4	999	0	nonexistent	1.4	93.918	-42.7	4.962	5228.1	no
20	45	entrepreneur	married	university.degre	unknown	yes	yes	cellular	aug	mon	38	2	999	0	nonexistent	1.4	93.444	-36.1	4.965	5228.1	no
21	50	blue-collar	married	basic.4y	no	no	yes	cellular	jul	tue	849	1	999	0	nonexistent	1.4	93.918	-42.7	4.961	5228.1	yes
22	55	services	married	basic.6y	unknown	yes	no	cellular	jul	tue	326	6	999	0	nonexistent	1.4	93.918	-42.7	4.962	5228.1	no
23	39	technician	divorced	high.school	no	no	no	cellular	mar	mon	222	1	12	2	success	-1.8	93.369	-34.8	0.639	5008.7	yes

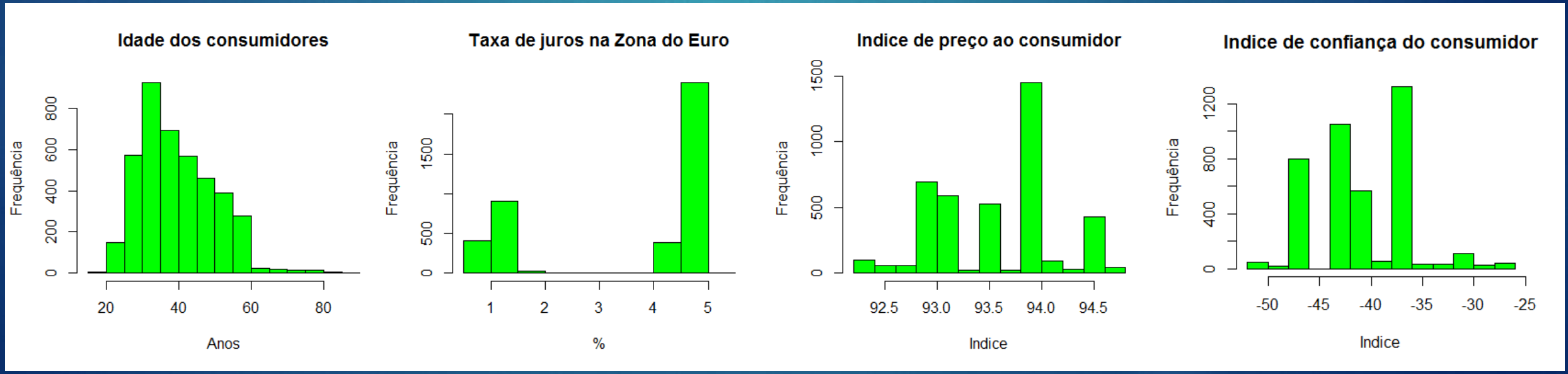
Planilha1

Pré-processamento

- Correlação foi aplicada em cada par de atributo contínuo presente no dataset original
- Descobriu-se correlação fortíssima entre número de empregados e a taxa de juros da zona do Euro
- E também para taxa de variação de empregos e a taxa de juros da zona do Euro
- Optou-se por excluir os dois atributos de empregados

Pré-processamento

- Discretização dos atributos de acordo com os histogramas:
  - Índice de preços ao consumidor – 3 categorias
  - Índice de confiança do consumidor – 5 categorias
  - Idade dos consumidores – 3 categorias
  - Taxa de juros – 2 categorias



Pré-processamento

- Exclusão do atributo que continha a duração da última ligação ao consumidor, induzia sempre ao não quando fosse 0 (recomendação prevista no "dataset information")
- Atributos categóricos foram transformados em numéricos para facilitar a análise
- Ao fim do pré-processamento: 17 atributos

Método de amostragem

- Para este exemplo foi usada o método holdout
- 75% das instâncias para treinamento
- 25% para teste
- Total 4119 instâncias

Resultados

Kernel linear

Tempo	Custo	Taxa de acerto
0,06s	0,1	90,2%
0,09s	1	90,1%
0,14	10	90,1%

Kernel radial

Tempo	Custo	Gama	Taxa de Acerto
0,16s	0,1	0,1	90,4%
0,3s	0,1	0,5	89,6%
0,37s	0,1	1	89,6%
0,17s	1	0,1	90,3%
0,3s	1	0,5	89,9%
0,39s	1	1	89,6%
0,18s	10	0,1	89,2%
0,28s	10	0,5	87,8%
0,38s	10	1	88,9%

Kernel Polinomial

Tempo	Custo	Gama	Grau	Coef.	Taxa de acerto
0,09s	1	1	3	1	83,4%
0,08s	1	0,5	3	1	83,6%
0,09s	1	0,1	3	1	89,6%
0,07s	1	0,1	2	1	90,5%
0,09s	1	0,1	2	10	90,5%
0,08s	0,1	0,1	2	10	90,2%
0,06s	10	0,1	2	10	89,6%
0,08s	1	0,05	2	1	90,4%
0,9s	1	2	4	1	83,8%

Comparação com a árvore de decisão

- Árvore
  - Precisão: 90,2%
  - Tempo de execução: 0,02s
  - Falsos positivos: 43,5%
  - Falsos negativos: 8,2%

- SVM (kernel polinomial)
  - Precisão: 90,5%
  - Tempo de execução: 0,07s
  - Falsos positivos: 36,3%
  - Falsos negativos: 8,6%

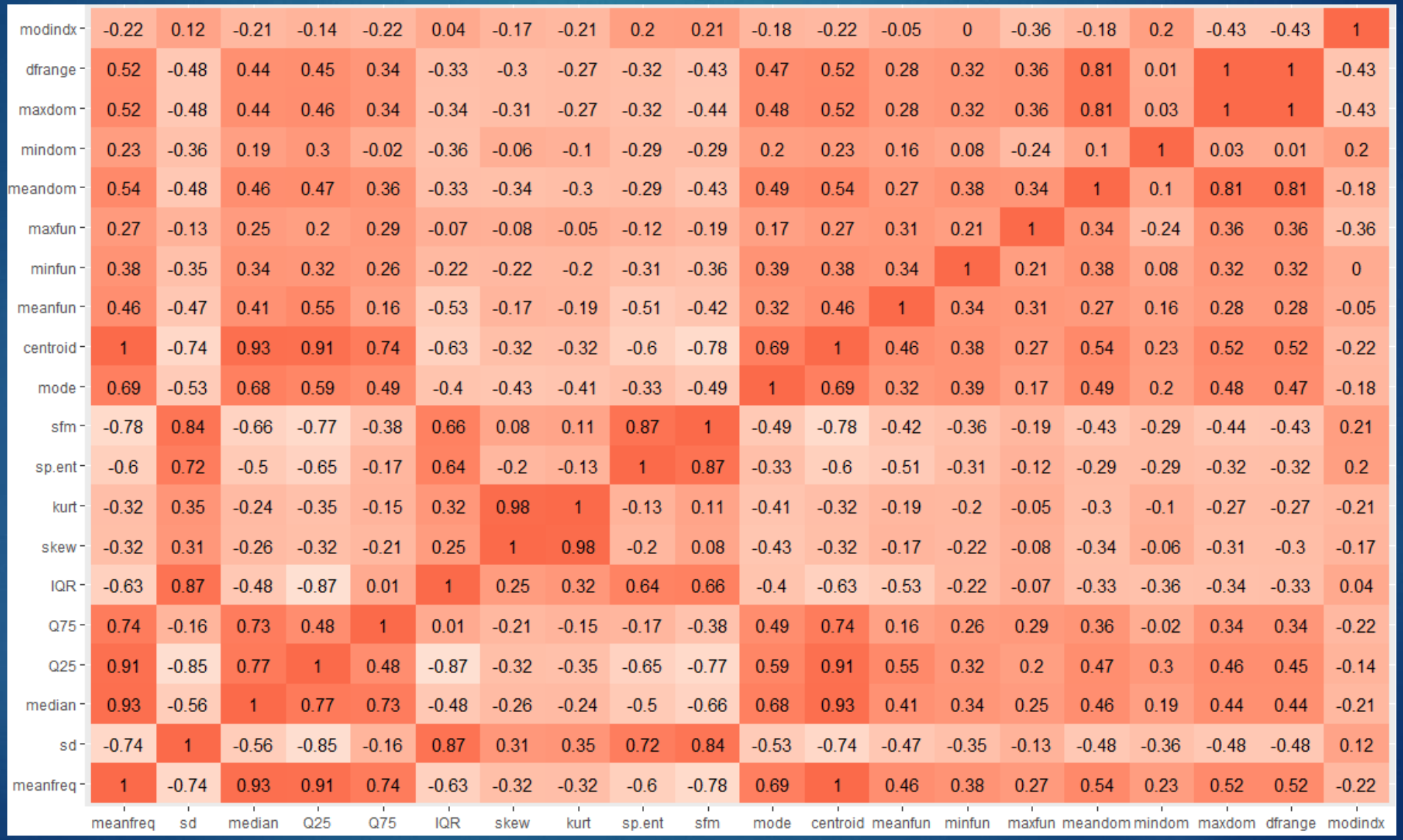
Gender Recognition by Voice

- Objetivo de identificar o gênero do orador através das características acústicas da voz e da fala;
- Consiste em 3.168 instâncias;
  - 1584 homens, 1584 mulheres;
- Composto por 21 atributos;
- **Classe: Sexo (homem ou mulher)**

Dataset original

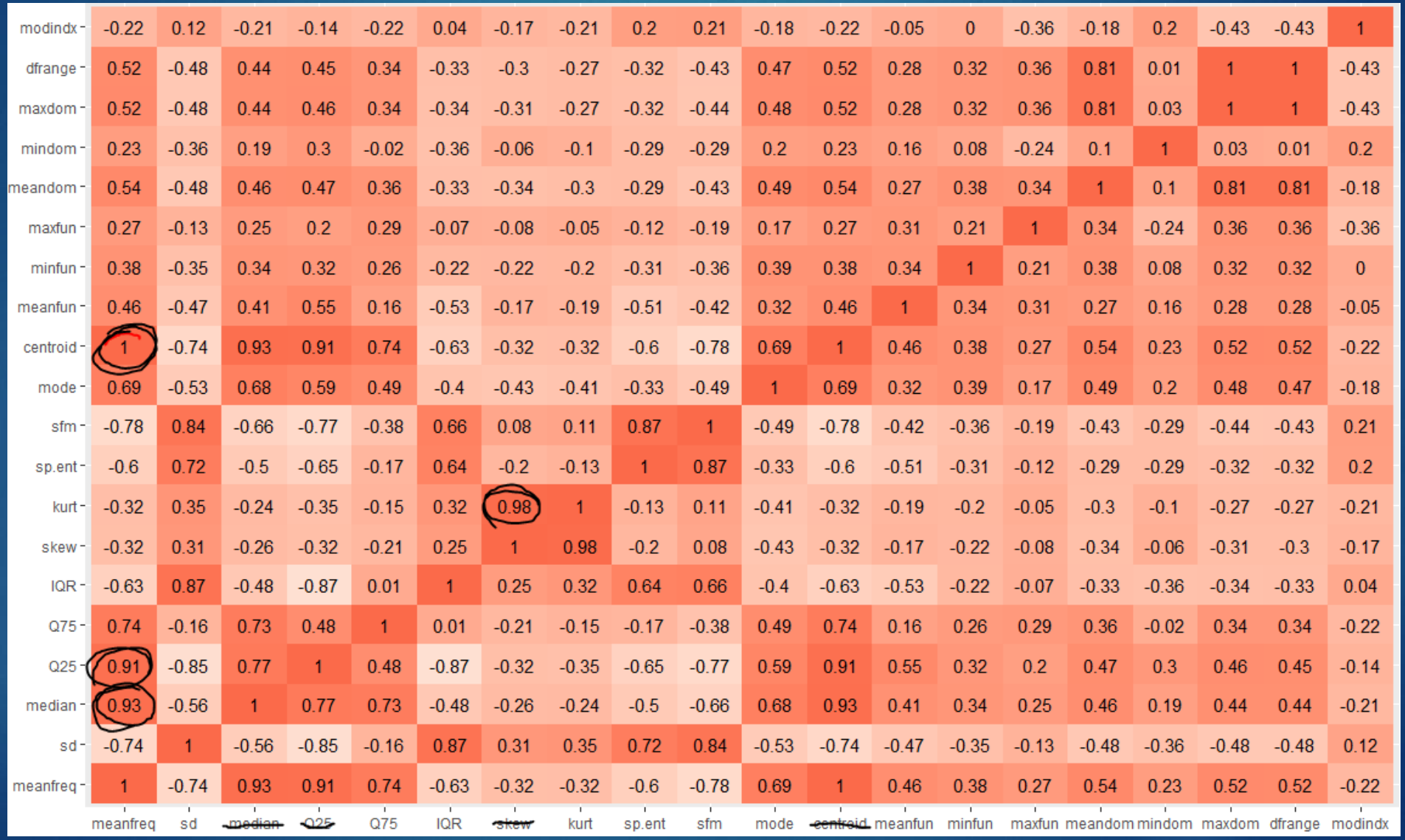
	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx	label
1	0.05978098	0.06424127	0.03202691	0.0150714886	0.09019344	0.07512195	12.8634618	274.402906	0.8933694	0.49191777	0.0000000000	0.05978098	0.08427911	0.01570167	0.2758621	0.007812500	0.0078125	0.0078125	0.0000000	0.00000000	male
2	0.06600874	0.06731003	0.04022873	0.0194138670	0.09266619	0.07325232	22.4232854	634.613855	0.8921932	0.51372384	0.0000000000	0.06600874	0.10793655	0.01582591	0.2500000	0.009014423	0.0078125	0.0546875	0.0468750	0.05263158	male
3	0.07731550	0.08382942	0.03671846	0.0087010566	0.13190802	0.12320696	30.7571546	1024.927705	0.8463891	0.47890498	0.0000000000	0.07731550	0.09870626	0.01565558	0.2711864	0.007990057	0.0078125	0.0156250	0.0078125	0.04651163	male
4	0.15122809	0.07211059	0.15801119	0.0965817278	0.20795525	0.11137352	1.2328313	4.177296	0.9633225	0.72723180	0.0838781852	0.15122809	0.08896485	0.01779755	0.2500000	0.201497396	0.0078125	0.5625000	0.5546875	0.24711908	male
5	0.13512039	0.07914610	0.12465623	0.0787202178	0.20604493	0.12732471	1.1011737	4.333713	0.9719551	0.78356806	0.1042614023	0.13512039	0.10639784	0.01693122	0.2666667	0.712812500	0.0078125	5.4843750	5.4765625	0.20827389	male
6	0.13278641	0.07955687	0.11908985	0.0679579930	0.20959160	0.14163361	1.9325624	8.308895	0.9631813	0.73830700	0.1125554259	0.13278641	0.11013192	0.01711230	0.2539683	0.298221983	0.0078125	2.7265625	2.7187500	0.12515964	male
7	0.15076233	0.07446321	0.16010638	0.0928989362	0.20571809	0.11281915	1.5306432	5.987498	0.9675731	0.76263767	0.0861968085	0.15076233	0.10594452	0.02622951	0.2666667	0.479619565	0.0078125	5.3125000	5.3046875	0.12399186	male
8	0.16051433	0.07676688	0.14433678	0.1105321684	0.23196187	0.12142971	1.3971564	4.766611	0.9592546	0.71985791	0.1283240667	0.16051433	0.09305243	0.01775805	0.1441441	0.301339286	0.0078125	0.5390625	0.5312500	0.28393665	male
9	0.14223942	0.07801846	0.13858744	0.0882062780	0.20858744	0.12038117	1.0997462	4.070284	0.9707229	0.77099205	0.2191031390	0.14223942	0.09672895	0.01795735	0.2500000	0.336476293	0.0078125	2.1640625	2.1562500	0.14827202	male
10	0.13432878	0.08035003	0.12145135	0.0755799890	0.20195712	0.12637713	1.1903684	4.787310	0.9752461	0.80450525	0.0116987356	0.13432878	0.10588093	0.01930036	0.2622951	0.340364583	0.0156250	4.6953125	4.6796875	0.08991998	male
11	0.15702051	0.07194293	0.16816015	0.1014299333	0.21673975	0.11530982	0.9794423	3.974223	0.9652491	0.73369288	0.0963584366	0.15702051	0.08889398	0.02206897	0.1176471	0.460227273	0.0078125	2.8125000	2.8046875	0.20000000	male
12	0.13855052	0.07705399	0.12752660	0.0873138298	0.20273936	0.11542553	1.6267699	6.291365	0.9660038	0.75204201	0.0121010638	0.13855052	0.10419932	0.01913876	0.2622951	0.246093750	0.0078125	2.7187500	2.7109375	0.13235137	male
13	0.13734274	0.08087671	0.12426251	0.0831448993	0.20922677	0.12608187	1.3787282	5.008952	0.9635135	0.73614996	0.1084340481	0.13734274	0.09264402	0.01678909	0.2133333	0.481670673	0.0156250	5.0156250	5.0000000	0.08850000	male
14	0.18122546	0.06004205	0.19095321	0.1288388215	0.22953206	0.10069324	1.3694305	5.475600	0.9374458	0.53707999	0.2198266898	0.18122546	0.13150372	0.02500000	0.2758621	1.277113971	0.0078125	2.8046875	2.7968750	0.41655028	male
15	0.18311528	0.06698235	0.19123253	0.1291486658	0.24015248	0.11100381	3.5681040	35.384748	0.9403326	0.57139420	0.0499872935	0.18311528	0.10279870	0.02083333	0.2758621	1.245738636	0.2031250	6.7421875	6.5390625	0.13933177	male
16	0.17427211	0.06941105	0.19087411	0.1156019791	0.22827927	0.11267730	4.4850384	61.764908	0.9509720	0.63519918	0.0500274876	0.17427211	0.10204599	0.01832761	0.2461538	1.621299342	0.0078125	7.0000000	6.9921875	0.20931099	male
17	0.19084630	0.06579028	0.20795099	0.1322804629	0.24435671	0.11207624	1.5623037	7.834350	0.9385460	0.53880958	0.0501293397	0.19084630	0.11332280	0.01754386	0.2758621	1.434114583	0.0078125	6.3203125	6.3125000	0.25477979	male
18	0.17124697	0.07487157	0.15280665	0.1223908524	0.24361746	0.12122661	3.2071698	25.765565	0.9369535	0.58641951	0.0599584200	0.17124697	0.07971831	0.01567091	0.2622951	0.106279481	0.0078125	0.5703125	0.5625000	0.13835470	male
19	0.16834595	0.07412082	0.14561847	0.1157559098	0.23982408	0.12406817	2.7043347	18.484703	0.9345234	0.55974232	0.0600329852	0.16834595	0.08348402	0.01571709	0.2318841	0.146562500	0.0078125	3.1250000	3.1171875	0.05953660	male

Pré-processamento



	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx
modindx	-0.22	0.12	-0.21	-0.14	-0.22	0.04	-0.17	-0.21	0.2	0.21	-0.18	-0.22	-0.05	0	-0.36	-0.18	0.2	-0.43	-0.43	1
dfrange	0.52	-0.48	0.44	0.45	0.34	-0.33	-0.3	-0.27	-0.32	-0.43	0.47	0.52	0.28	0.32	0.36	0.81	0.01	1	1	-0.43
maxdom	0.52	-0.48	0.44	0.46	0.34	-0.34	-0.31	-0.27	-0.32	-0.44	0.48	0.52	0.28	0.32	0.36	0.81	0.03	1	1	-0.43
mindom	0.23	-0.36	0.19	0.3	-0.02	-0.36	-0.06	-0.1	-0.29	-0.29	0.2	0.23	0.16	0.08	-0.24	0.1	1	0.03	0.01	0.2
meandom	0.54	-0.48	0.46	0.47	0.36	-0.33	-0.34	-0.3	-0.29	-0.43	0.49	0.54	0.27	0.38	0.34	1	0.1	0.81	0.81	-0.18
maxfun	0.27	-0.13	0.25	0.2	0.29	-0.07	-0.08	-0.05	-0.12	-0.19	0.17	0.27	0.31	0.21	1	0.34	-0.24	0.36	0.36	-0.36
minfun	0.38	-0.35	0.34	0.32	0.26	-0.22	-0.22	-0.2	-0.31	-0.36	0.39	0.38	0.34	1	0.21	0.38	0.08	0.32	0.32	0
meanfun	0.46	-0.47	0.41	0.55	0.16	-0.53	-0.17	-0.19	-0.51	-0.42	0.32	0.46	1	0.34	0.31	0.27	0.16	0.28	0.28	-0.05
centroid	1	-0.74	0.93	0.91	0.74	-0.63	-0.32	-0.32	-0.6	-0.78	0.69	1	0.46	0.38	0.27	0.54	0.23	0.52	0.52	-0.22
mode	0.69	-0.53	0.68	0.59	0.49	-0.4	-0.43	-0.41	-0.33	-0.49	1	0.69	0.32	0.39	0.17	0.49	0.2	0.48	0.47	-0.18
sfm	-0.78	0.84	-0.66	-0.77	-0.38	0.66	0.08	0.11	0.87	1	-0.49	-0.78	-0.42	-0.36	-0.19	-0.43	-0.29	-0.44	-0.43	0.21
sp.ent	-0.6	0.72	-0.5	-0.65	-0.17	0.64	-0.2	-0.13	1	0.87	-0.33	-0.6	-0.51	-0.31	-0.12	-0.29	-0.29	-0.32	-0.32	0.2
kurt	-0.32	0.35	-0.24	-0.35	-0.15	0.32	0.98	1	-0.13	0.11	-0.41	-0.32	-0.19	-0.2	-0.05	-0.3	-0.1	-0.27	-0.27	-0.21
skew	-0.32	0.31	-0.26	-0.32	-0.21	0.25	1	0.98	-0.2	0.08	-0.43	-0.32	-0.17	-0.22	-0.08	-0.34	-0.06	-0.31	-0.3	-0.17
IQR	-0.63	0.87	-0.48	-0.87	0.01	1	0.25	0.32	0.64	0.66	-0.4	-0.63	-0.53	-0.22	-0.07	-0.33	-0.36	-0.34	-0.33	0.04
Q75	0.74	-0.16	0.73	0.48	1	0.01	-0.21	-0.15	-0.17	-0.38	0.49	0.74	0.16	0.26	0.29	0.36	-0.02	0.34	0.34	-0.22
Q25	0.91	-0.85	0.77	1	0.48	-0.87	-0.32	-0.35	-0.65	-0.77	0.59	0.91	0.55	0.32	0.2	0.47	0.3	0.46	0.45	-0.14
median	0.93	-0.56	1	0.77	0.73	-0.48	-0.26	-0.24	-0.5	-0.66	0.68	0.93	0.41	0.34	0.25	0.46	0.19	0.44	0.44	-0.21
sd	-0.74	1	-0.56	-0.85	-0.16	0.87	0.31	0.35	0.72	0.84	-0.53	-0.74	-0.47	-0.35	-0.13	-0.48	-0.36	-0.48	-0.48	0.12
meanfreq	1	-0.74	0.93	0.91	0.74	-0.63	-0.32	-0.32	-0.6	-0.78	0.69	1	0.46	0.38	0.27	0.54	0.23	0.52	0.52	-0.22

Pré-processamento



	meanfreq	sd	~~median~~	~~Q25~~	Q75	IQR	~~skew~~	kurt	sp.ent	sfm	mode	~~centroid~~	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx
modindx	-0.22	0.12	-0.21	-0.14	-0.22	0.04	-0.17	-0.21	0.2	0.21	-0.18	-0.22	-0.05	0	-0.36	-0.18	0.2	-0.43	-0.43	1
dfrange	0.52	-0.48	0.44	0.45	0.34	-0.33	-0.3	-0.27	-0.32	-0.43	0.47	0.52	0.28	0.32	0.36	0.81	0.01	1	1	-0.43
maxdom	0.52	-0.48	0.44	0.46	0.34	-0.34	-0.31	-0.27	-0.32	-0.44	0.48	0.52	0.28	0.32	0.36	0.81	0.03	1	1	-0.43
mindom	0.23	-0.36	0.19	0.3	-0.02	-0.36	-0.06	-0.1	-0.29	-0.29	0.2	0.23	0.16	0.08	-0.24	0.1	1	0.03	0.01	0.2
meandom	0.54	-0.48	0.46	0.47	0.36	-0.33	-0.34	-0.3	-0.29	-0.43	0.49	0.54	0.27	0.38	0.34	1	0.1	0.81	0.81	-0.18
maxfun	0.27	-0.13	0.25	0.2	0.29	-0.07	-0.08	-0.05	-0.12	-0.19	0.17	0.27	0.31	0.21	1	0.34	-0.24	0.36	0.36	-0.36
minfun	0.38	-0.35	0.34	0.32	0.26	-0.22	-0.22	-0.2	-0.31	-0.36	0.39	0.38	0.34	1	0.21	0.38	0.08	0.32	0.32	0
meanfun	0.46	-0.47	0.41	0.55	0.16	-0.53	-0.17	-0.19	-0.51	-0.42	0.32	0.46	1	0.34	0.31	0.27	0.16	0.28	0.28	-0.05
centroid	1	-0.74	0.93	0.91	0.74	-0.63	-0.32	-0.32	-0.6	-0.78	0.69	1	0.46	0.38	0.27	0.54	0.23	0.52	0.52	-0.22
mode	0.69	-0.53	0.68	0.59	0.49	-0.4	-0.43	-0.41	-0.33	-0.49	1	0.69	0.32	0.39	0.17	0.49	0.2	0.48	0.47	-0.18
sfm	-0.78	0.84	-0.66	-0.77	-0.38	0.66	0.08	0.11	0.87	1	-0.49	-0.78	-0.42	-0.36	-0.19	-0.43	-0.29	-0.44	-0.43	0.21
sp.ent	-0.6	0.72	-0.5	-0.65	-0.17	0.64	-0.2	-0.13	1	0.87	-0.33	-0.6	-0.51	-0.31	-0.12	-0.29	-0.29	-0.32	-0.32	0.2
kurt	-0.32	0.35	-0.24	-0.35	-0.15	0.32	0.98	1	-0.13	0.11	-0.41	-0.32	-0.19	-0.2	-0.05	-0.3	-0.1	-0.27	-0.27	-0.21
skew	-0.32	0.31	-0.26	-0.32	-0.21	0.25	1	0.98	-0.2	0.08	-0.43	-0.32	-0.17	-0.22	-0.08	-0.34	-0.06	-0.31	-0.3	-0.17
IQR	-0.63	0.87	-0.48	-0.87	0.01	1	0.25	0.32	0.64	0.66	-0.4	-0.63	-0.53	-0.22	-0.07	-0.33	-0.36	-0.34	-0.33	0.04
Q75	0.74	-0.16	0.73	0.48	1	0.01	-0.21	-0.15	-0.17	-0.38	0.49	0.74	0.16	0.26	0.29	0.36	-0.02	0.34	0.34	-0.22
Q25	0.91	-0.85	0.77	1	0.48	-0.87	-0.32	-0.35	-0.65	-0.77	0.59	0.91	0.55	0.32	0.2	0.47	0.3	0.46	0.45	-0.14
median	0.93	-0.56	1	0.77	0.73	-0.48	-0.26	-0.24	-0.5	-0.66	0.68	0.93	0.41	0.34	0.25	0.46	0.19	0.44	0.44	-0.21
sd	-0.74	1	-0.56	-0.85	-0.16	0.87	0.31	0.35	0.72	0.84	-0.53	-0.74	-0.47	-0.35	-0.13	-0.48	-0.36	-0.48	-0.48	0.12
meanfreq	1	-0.74	0.93	0.91	0.74	-0.63	-0.32	-0.32	-0.6	-0.78	0.69	1	0.46	0.38	0.27	0.54	0.23	0.52	0.52	-0.22

Pré-processamento

- Os dados foram normalizados para valores entre 0 e 1;
- Ao retirar atributos com alta correlação, restaram 16 atributos.

Método de amostragem

- Para este exemplo foi usada o método holdout
- 75% aleatórios das instâncias para treinamento
- 25% aleatórios para teste

Resultados

Kernel linear

```
  SVM-Type:  C-classification
SVM-Kernel:  linear
      cost:  10
```

```
           test_y
pred_linear female male
     female    378    7
     male        9  398
```

Acerto: 97.98%

Kernel radial

```
  SVM-Type:  C-classification
SVM-Kernel:  radial
      cost:  2
     gamma:  0.125
```

```
           test_y
pred_radial female male
     female    379    9
     male        8  396
```

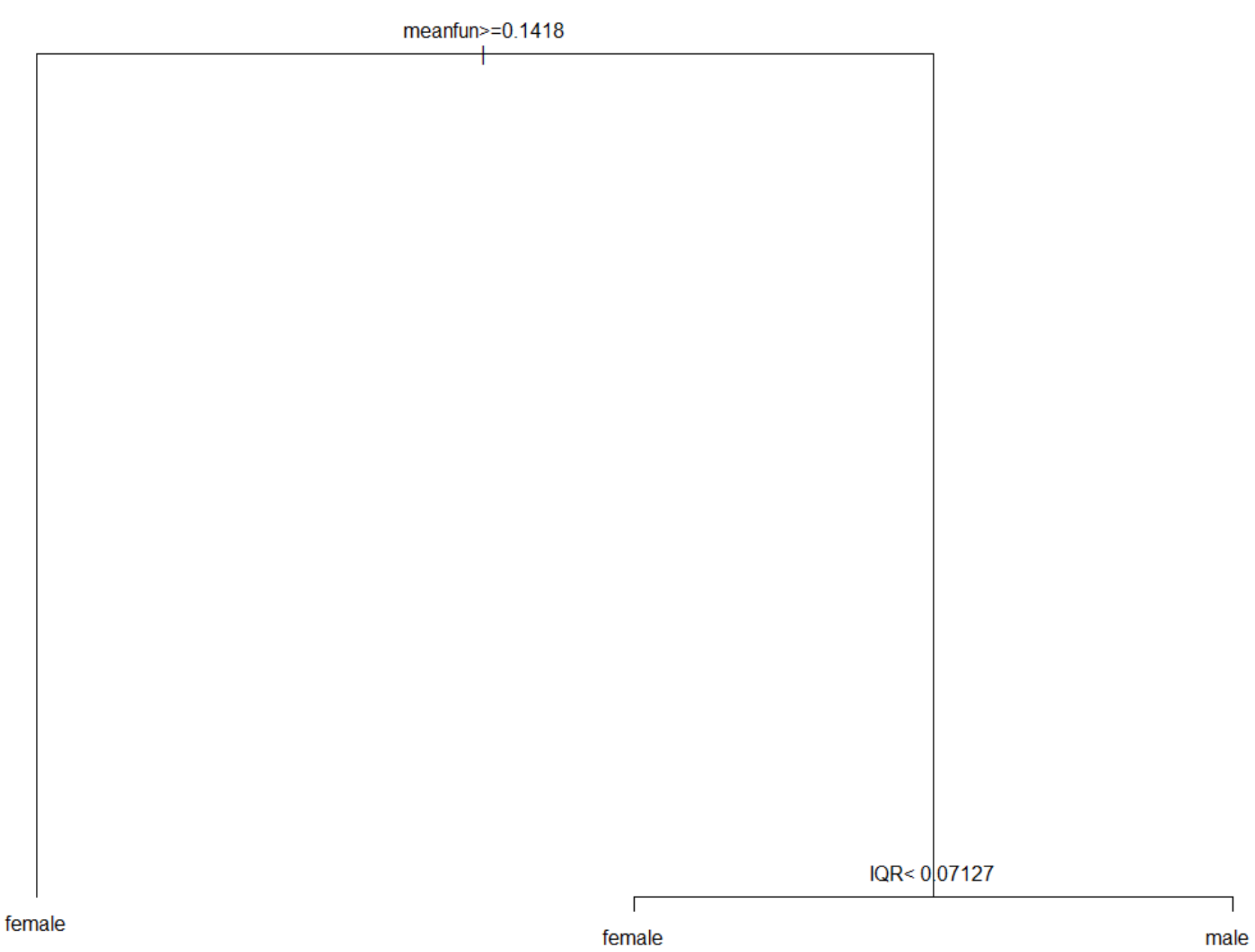
Acerto: 97.85%

Kernel Polinomial

```
  SVM-Type:  C-classification
SVM-Kernel:  polynomial
      cost:  1
    degree:  3
     gamma:  0.125
    coef.0:  1
```

```
           test_y
pred_poly female male
   female    377   10
   male       10  395
```

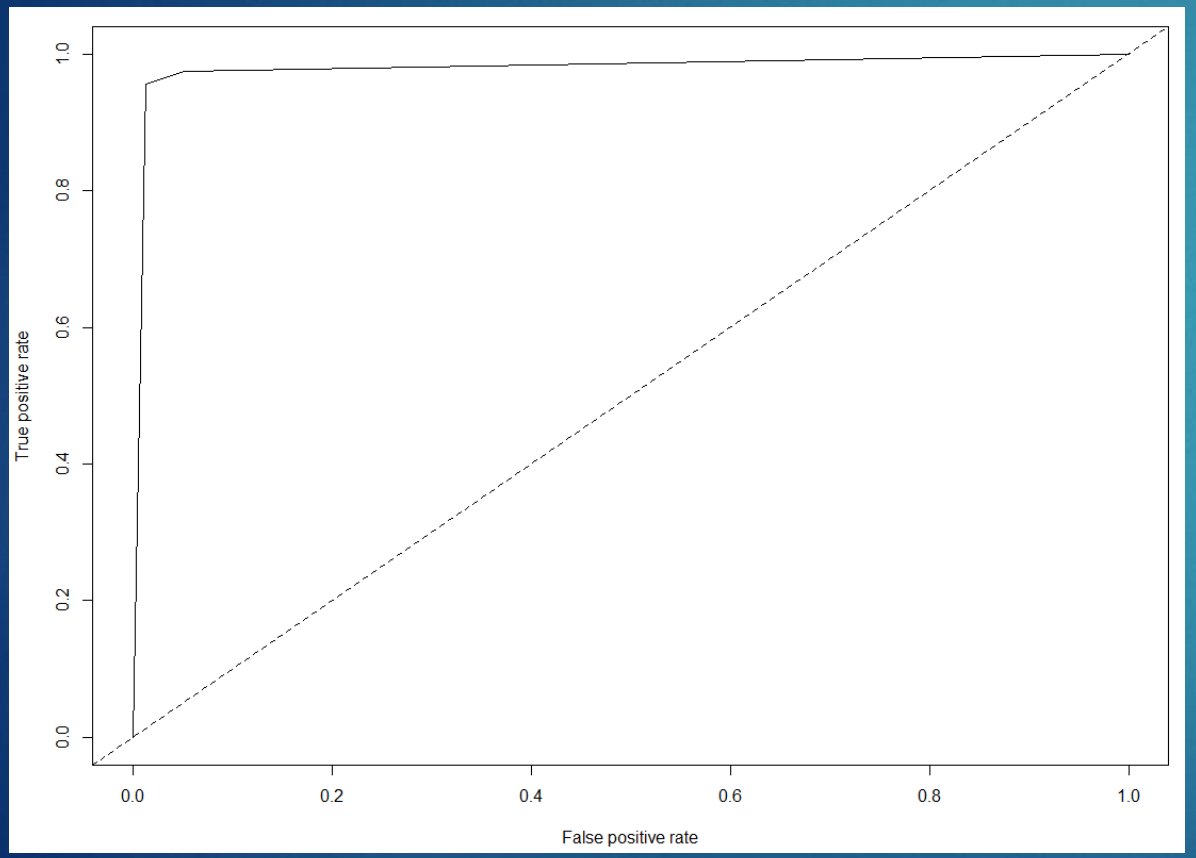
Acerto: 97.47%

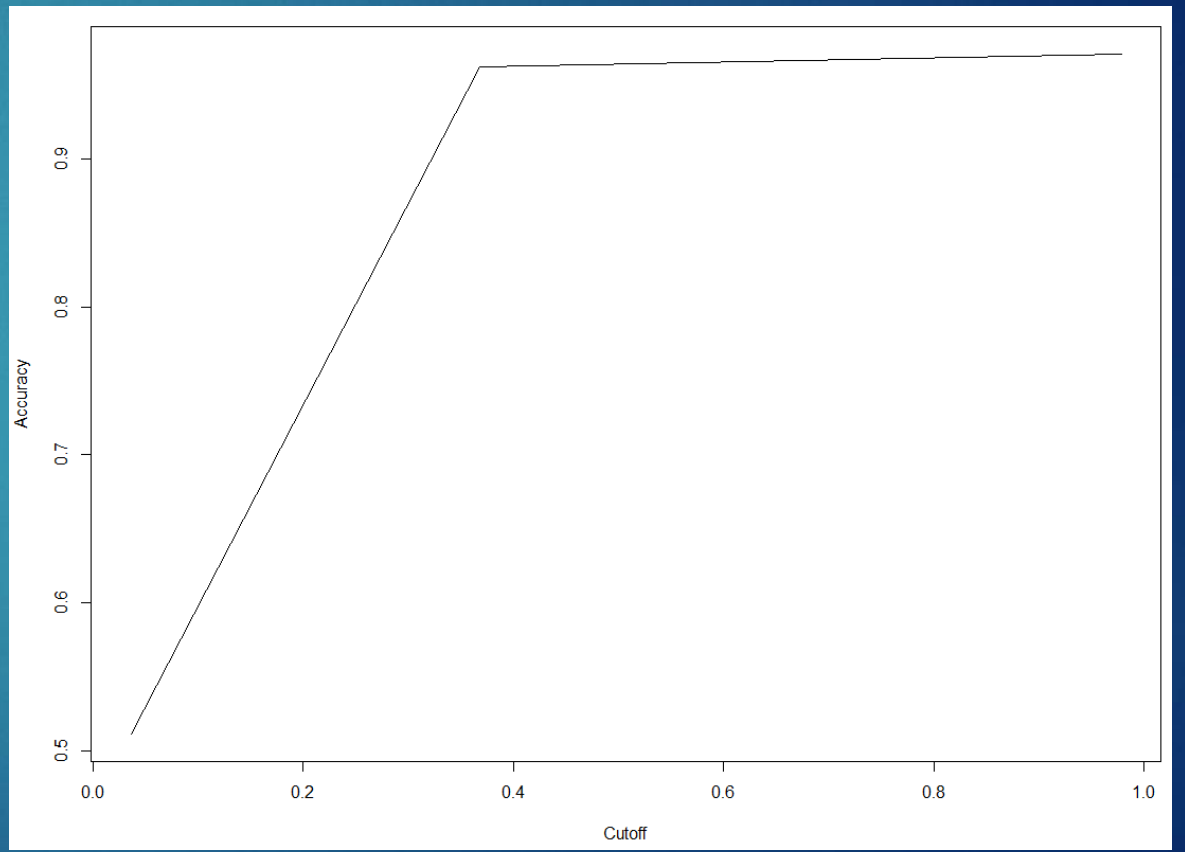

meanfun>=0.1418
female
IQR< 0.07127
female
male

Comparação com a árvore de decisão

- Árvore
  - Precisão: 97,09%
- SVM (kernel linear)
  - Precisão: 97,98%

ROC Árvore


True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


Accuracy
Cutoff
0.5
0.6
0.7
0.8
0.9
0.0
0.2
0.4
0.6
0.8
1.0

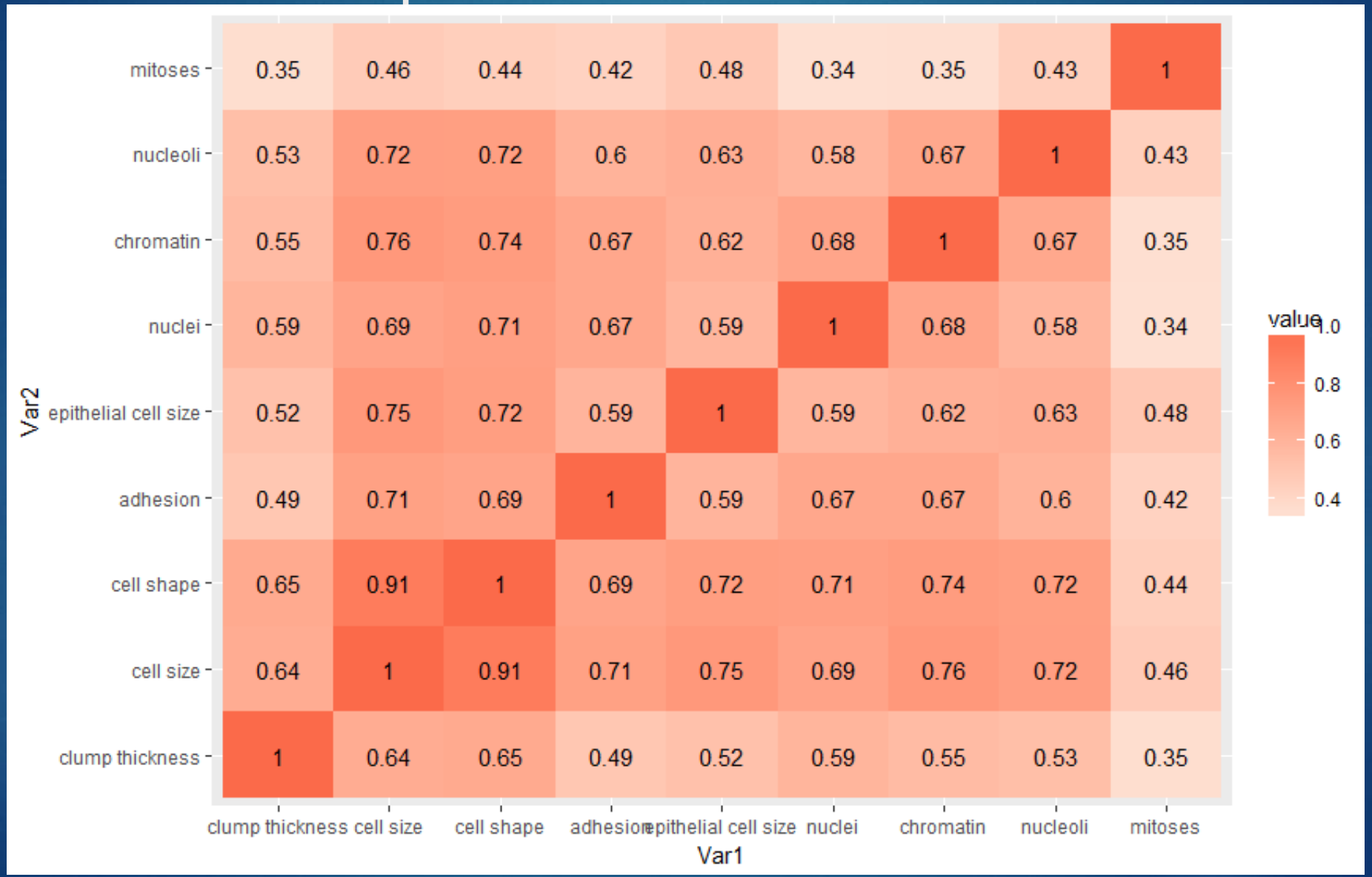
Breast Cancer in Wisconsin

- Objetivo de classificar um tumor como benigno ou maligno
- Consiste em 699 instâncias;
  - 458 benignos, 241 malignos;
- Composto por 10 atributos;
- **Classe: Tumor (maligno ou benigno)**

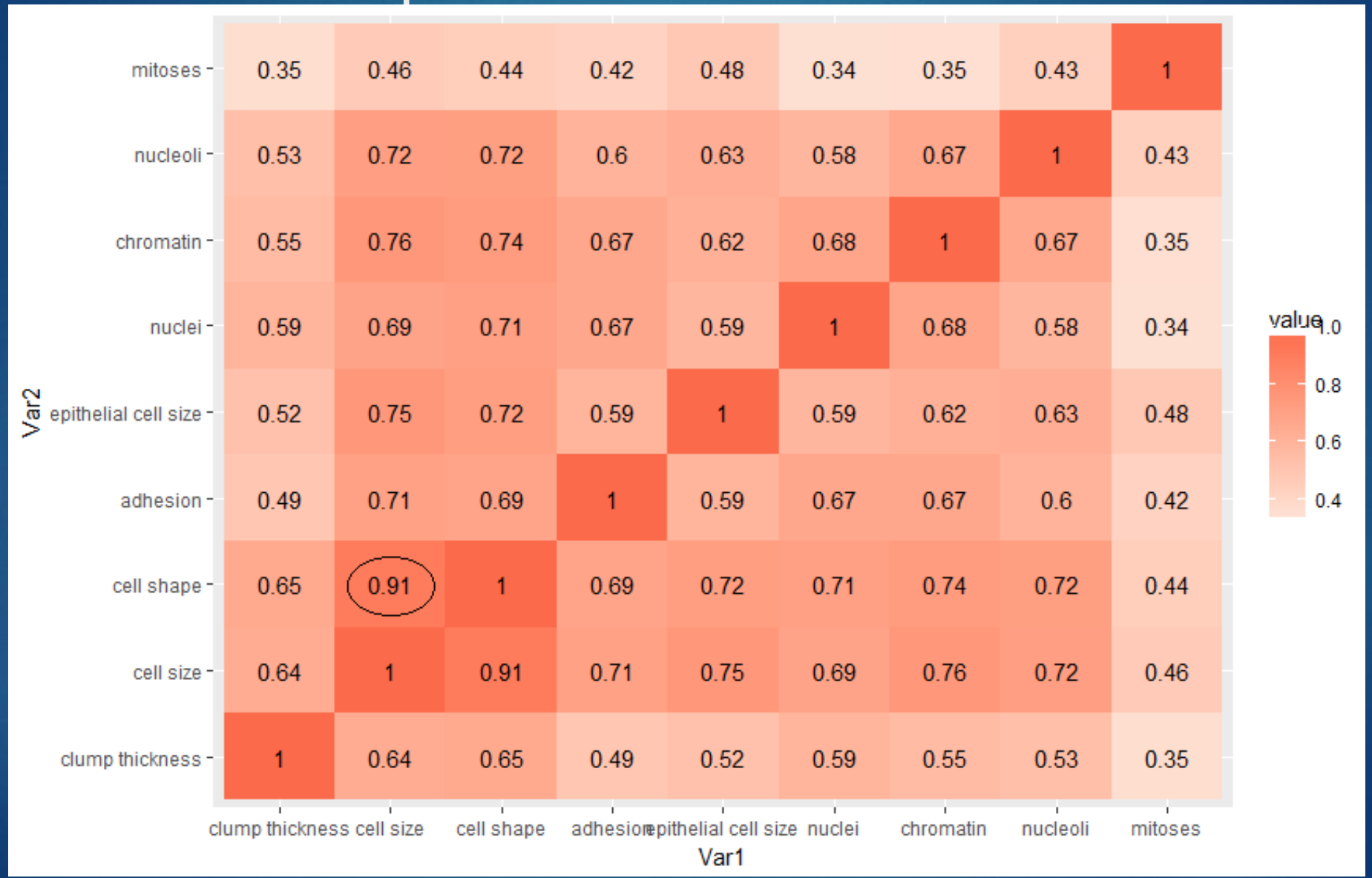
Dataset original

id (integer)	clump thickness (integer)	cell size (integer)	cell shape (integer)	adhesion (integer)	epithelial cell size (integer)	nuclei (integer)	chromatin (integer)	nucleoli (integer)	mitoses (integer)	class (integer)
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4
1048672	4	1	1	1	2	1	2	1	1	2
1049815	4	1	1	1	2	1	3	1	1	2
1050670	10	7	7	6	4	10	4	1	2	4
1050718	6	1	1	1	2	1	3	1	1	2
1054590	7	3	2	10	5	10	5	4	4	4

Pré-processamento



Pré-processamento



Pré-processamento

- O atributo *id* foi ignorado;
- Os dados foram normalizados para valores entre 0 e 1;
- 16 instâncias possuíam dados indisponíveis sobre o atributo *nuclei*, as quais foram retiradas do conjunto;
- Foi retirado 1 atributo com alta correlação (*cell shape*).

Método de amostragem

- Para este exemplo foi usada o método *holdout*
- 75% aleatórios das instâncias para treinamento
- 25% aleatórios para teste

Resultados

Kernel linear

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.125

Number of Support Vectors:  41
```

	test_y	
pred_linear	benign	malign
benign	107	2
malign	4	58

Acerto: 96,49%

Kernel radial

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  2
      gamma:  0.0625

Number of Support Vectors:  60
```

	test_y	
pred_radial	benign	malign
benign	105	1
malign	6	59

Acerto: 95,91%

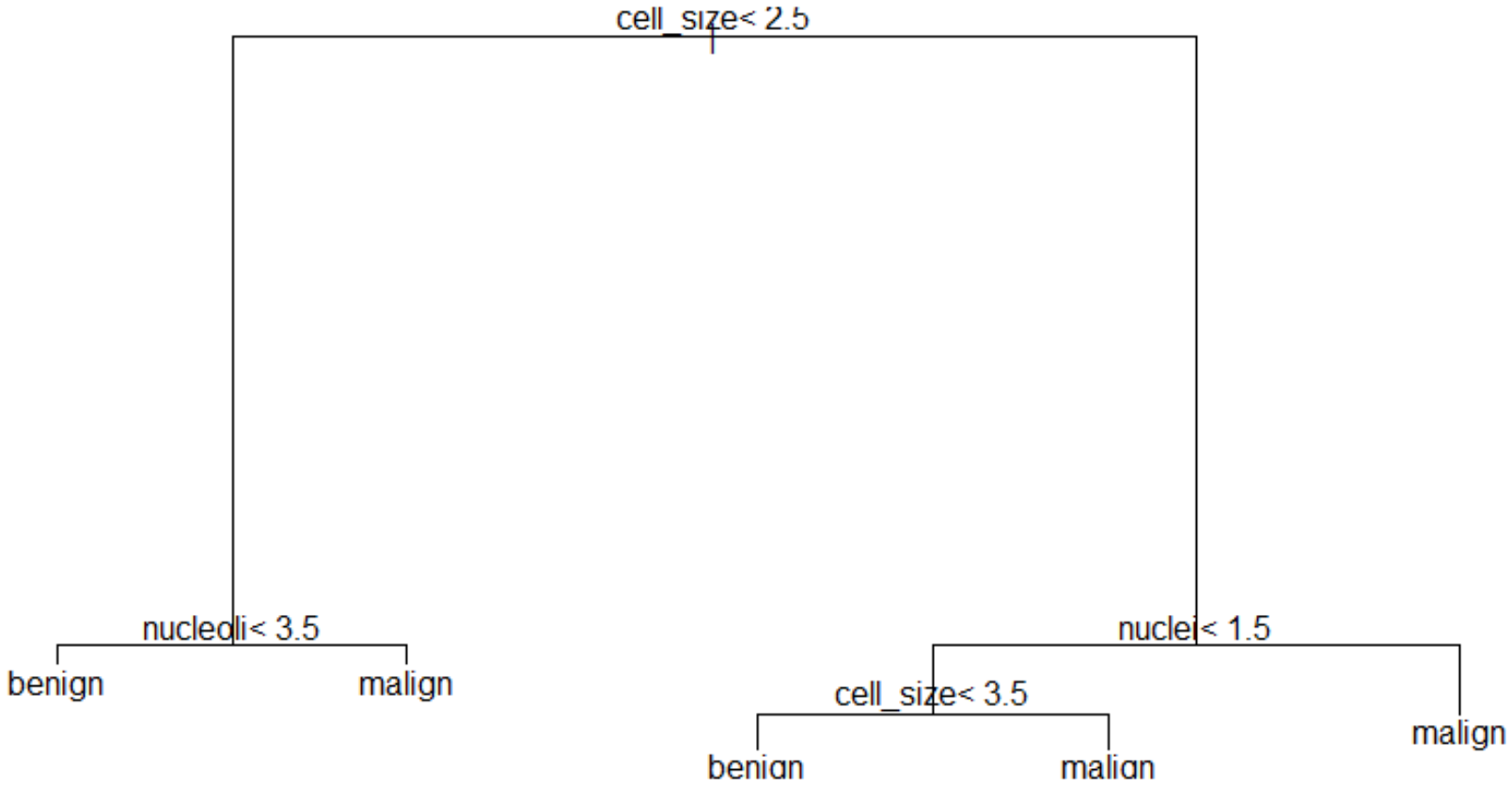
Kernel Polinomial

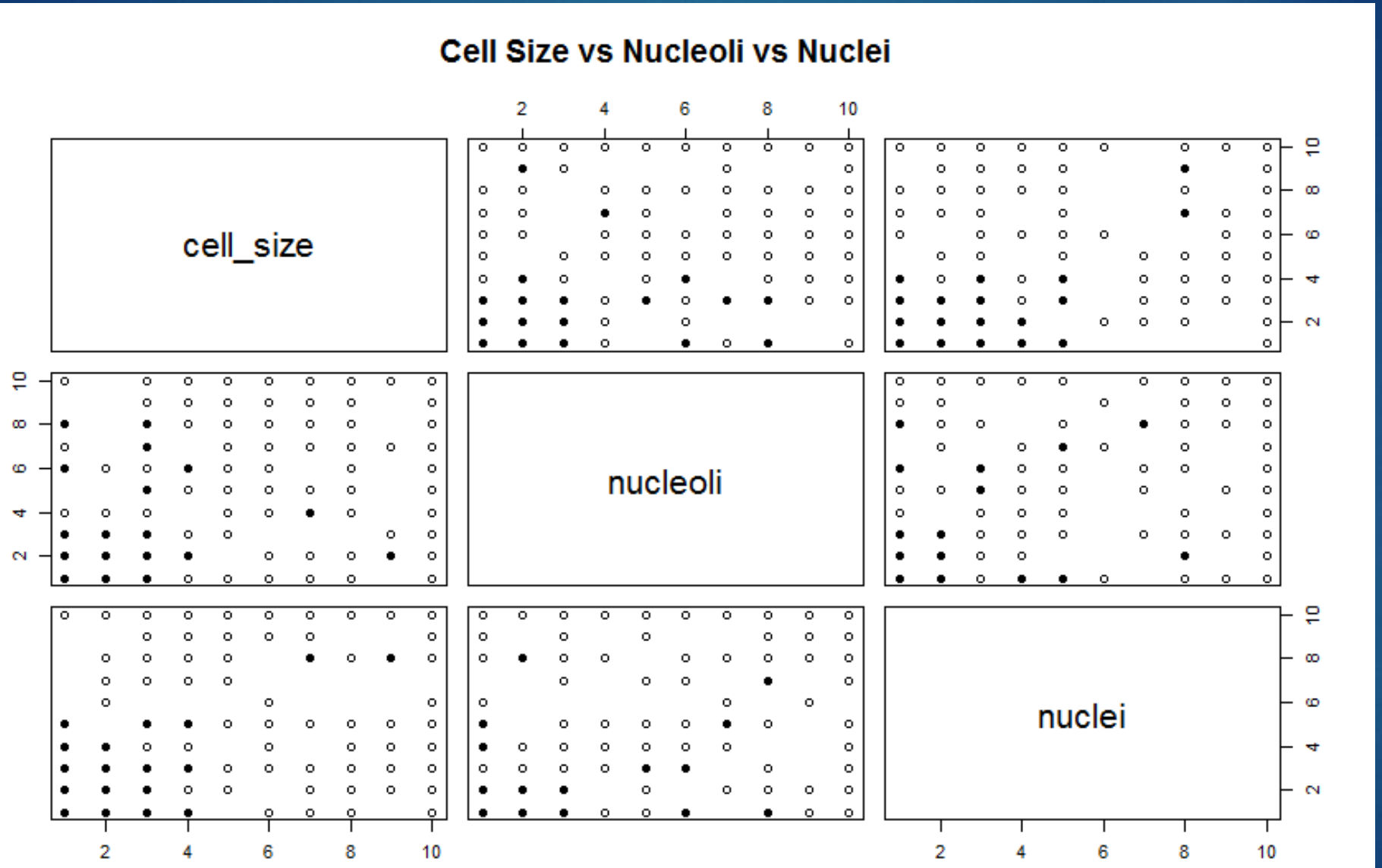
```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  1
     degree:  2
      gamma:  0.125
     coef.0:  1

Number of Support Vectors:  52
```

	test_y	
pred_poly	benign	malign
benign	106	2
malign	5	58

Acerto: 95,91%


cell_size< 2.5
nucleoli< 3.5
benign
malign
nuclei< 1.5
cell_size< 3.5
benign
malign
malign


Cell Size vs Nucleoli vs Nuclei
cell_size
nucleoli
nuclei
2
4
6
8
10

Comparação com a árvore de decisão

- Árvore
  - Precisão: 95,32%
- SVM (kernel linear)
  - Precisão: 96,49%

Referências bibliográficas

- Bank Marketing Data Set. Disponível em: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. Acesso em 04/06/2017.
- NEVES, R. de Cássia David das. *Pré-processamento de Descoberta de Conhecimento em Banco de Dados*. 2003. Dissertação (Pós-Graduação em Computação) – UFRGS, Porto Alegre. Disponível em: http://www.lume.ufrgs.br/bitstream/handle/10183/2701/000375412.pdf?...1. Acesso em: 04/06/2017.
- SVM example with Iris Data in R. Disponível em: http://rischanlab.github.io/SVM.html. Acesso em: 04/06/2017.

Referências bibliográficas

- Breast Cancer Wisconsin Data Set. Disponível em:
- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). Acesso em 06/06/2017
- Using Support Vector Machines Effectively. Disponível em: http://neerajkumar.org/writings/svm/ . Acesso em 06/06/2017