WINE DATASET

Trabalho final disciplina SIN-5007

BRUNO ABREU KEMMER
LUCIANA SILVA
ALAN REBELO

Dataset:

Características:

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

178 instâncias:

Classe 1: 59

Classe 2: 71

~~Classe 3: 48~~

130 instâncias:

Classe 1: 59

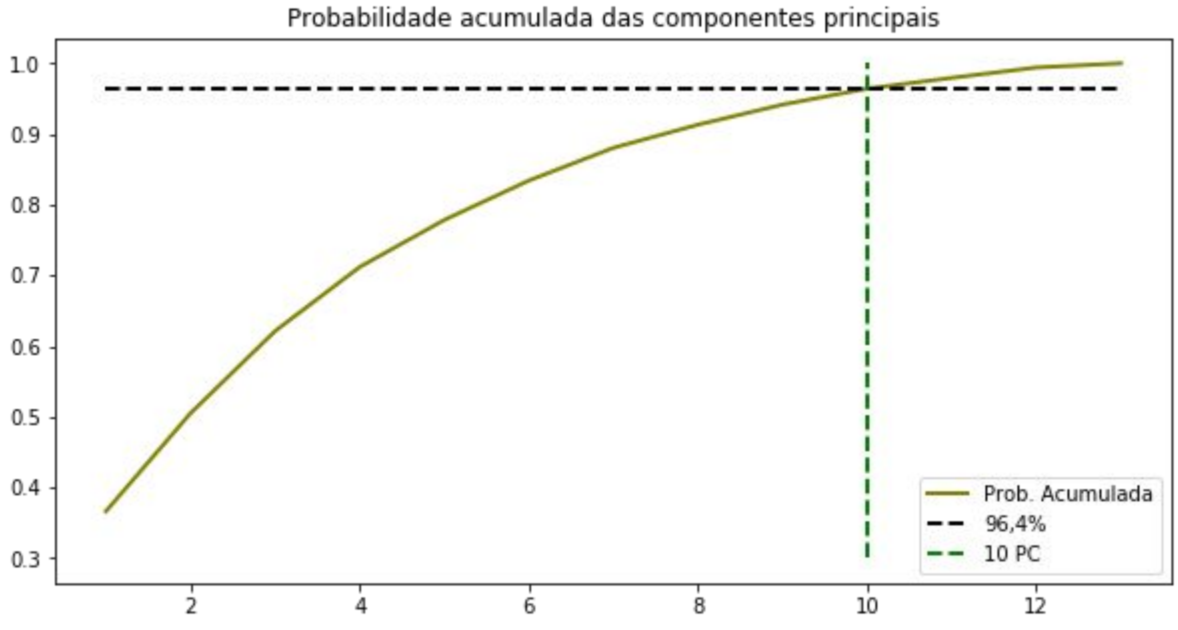
Classe 2: 71

https://archive.ics.uci.edu/ml/datasets/wine

PCA

Sem normalização: 1 componente principal com 99,834% da variabilidade.

Normalizado:



PCA

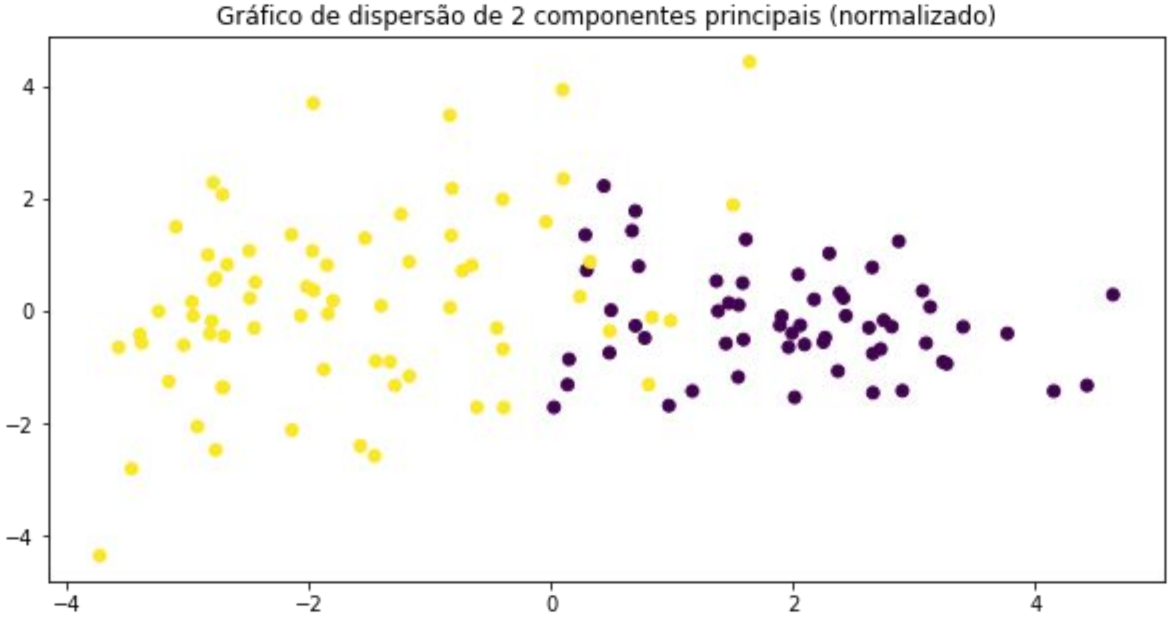


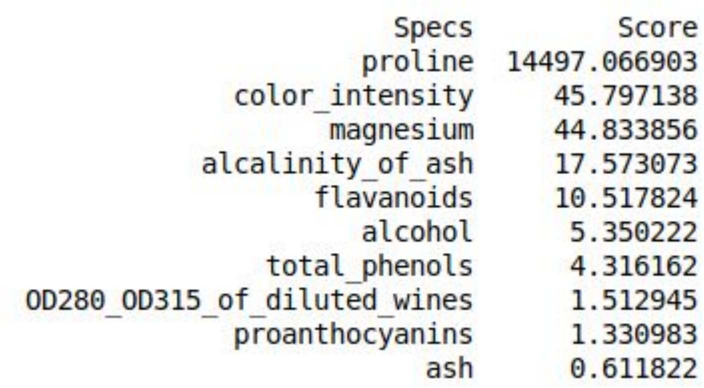
Gráfico de dispersão de 2 componentes principais (normalizado)
4
2
0
−2
−4
−4
−2
0
2
4

SELEÇÃO DE CARACTERÍSTICAS

1. SelectKBest

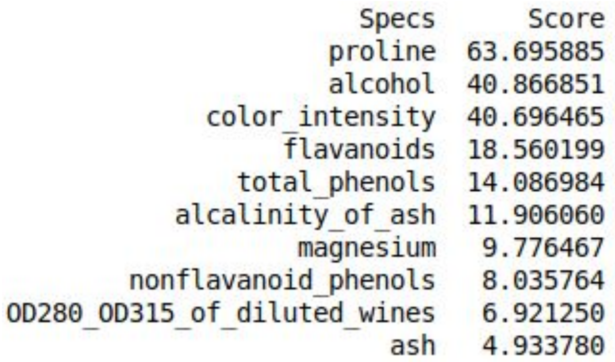
Selecionará as variáveis que tiverem o maior valor para o teste estatístico chi-quadrado, esse teste mede a dependência entre variáveis estocásticas, ou seja a dependência entre as características e a variável que contém a classe.

Sem normalização:



```
                      Specs         Score
                    proline  14497.066903
            color_intensity     45.797138
                  magnesium     44.833856
          alcalinity_of_ash     17.573073
                 flavanoids     10.517824
                    alcohol      5.350222
              total_phenols      4.316162
OD280_OD315_of_diluted_wines     1.512945
            proanthocyanins      1.330983
                        ash      0.611822
```

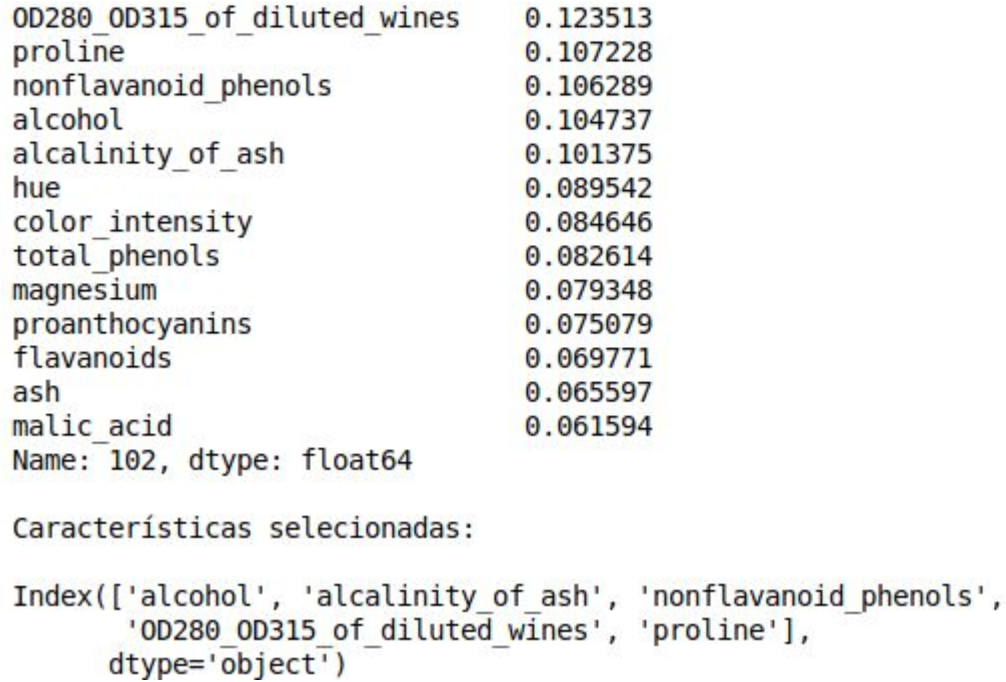
Com normalização:



```
                       Specs      Score
                     proline  63.695885
                     alcohol  40.866851
             color_intensity  40.696465
                  flavanoids  18.560199
               total_phenols  14.086984
           alcalinity_of_ash  11.906060
                   magnesium   9.776467
        nonflavanoid_phenols   8.035764
OD280_OD315_of_diluted_wines   6.921250
                         ash   4.933780
```

SELEÇÃO DE CARACTERÍSTICAS

2. Relief (implementação própria):

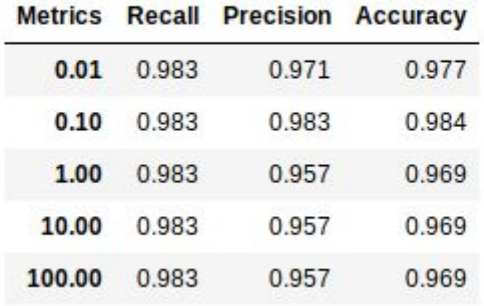


```
OD280_OD315_of_diluted_wines    0.123513
proline                         0.107228
nonflavanoid_phenols            0.106289
alcohol                         0.104737
alcalinity_of_ash               0.101375
hue                             0.089542
color_intensity                 0.084646
total_phenols                   0.082614
magnesium                       0.079348
proanthocyanins                 0.075079
flavanoids                      0.069771
ash                             0.065597
malic_acid                      0.061594
Name: 102, dtype: float64

Características selecionadas:

Index(['alcohol', 'alcalinity_of_ash', 'nonflavanoid_phenols',
       'OD280_OD315_of_diluted_wines', 'proline'],
      dtype='object')
```

NoSample = 30
Threshold = 0.1
Seed = 42

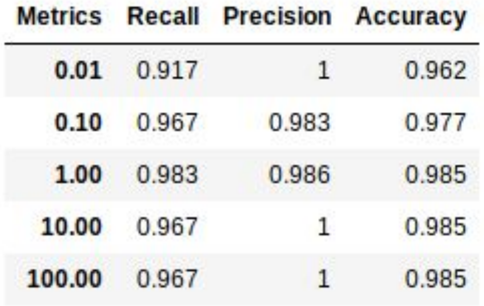
SVM - Kernel Linear (validação cruzada k=10)

Todas as características



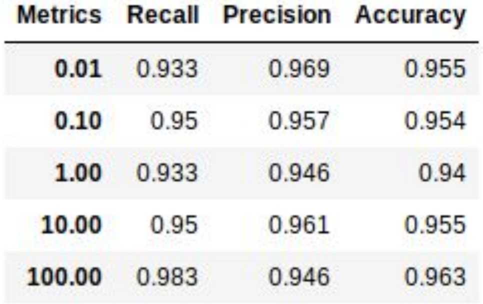
Metrics	Recall	Precision	Accuracy
0.01	0.983	0.971	0.977
0.10	0.983	0.983	0.984
1.00	0.983	0.957	0.969
10.00	0.983	0.957	0.969
100.00	0.983	0.957	0.969

PCA 10



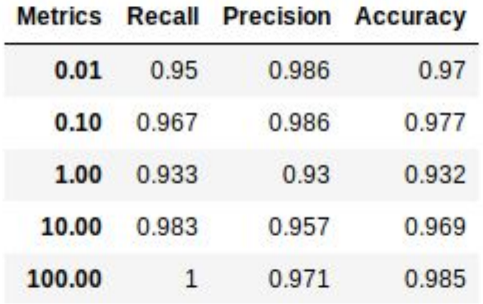
Metrics	Recall	Precision	Accuracy
0.01	0.917	1	0.962
0.10	0.967	0.983	0.977
1.00	0.983	0.986	0.985
10.00	0.967	1	0.985
100.00	0.967	1	0.985

SelectKBest (k=6)

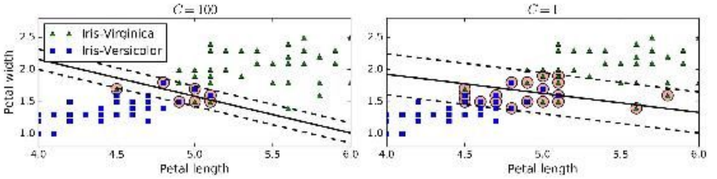


Metrics	Recall	Precision	Accuracy
0.01	0.933	0.969	0.955
0.10	0.95	0.957	0.954
1.00	0.933	0.946	0.94
10.00	0.95	0.961	0.955
100.00	0.983	0.946	0.963

Relief (5 carac.)



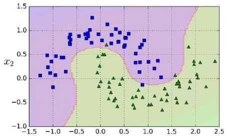
Metrics	Recall	Precision	Accuracy
0.01	0.95	0.986	0.97
0.10	0.967	0.986	0.977
1.00	0.933	0.93	0.932
10.00	0.983	0.957	0.969
100.00	1	0.971	0.985



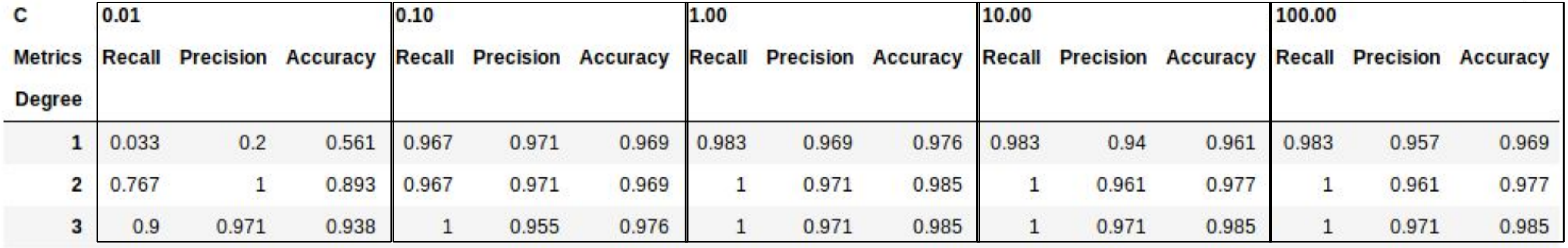
```
*** SVM normalizada,  SelectKBest (k=6) e com kernel Linear ***
Características selecionadas:   'proline', 'alcohol', 'color_intensity',
                                 flavanoids', 'total_phenols', 'alcalinity_of_ash'
```

```
*** SVM normalizada,  Relief (5 carac.) e com kernel Linear ***
Características selecionadas:   'alcohol', 'alcalinity_of_ash', 'nonflavanoid_phenols',
                                'OD280_OD315_of_diluted_wines', 'proline'
```

SVM - Kernel Polinomial (validação cruzada k=10)

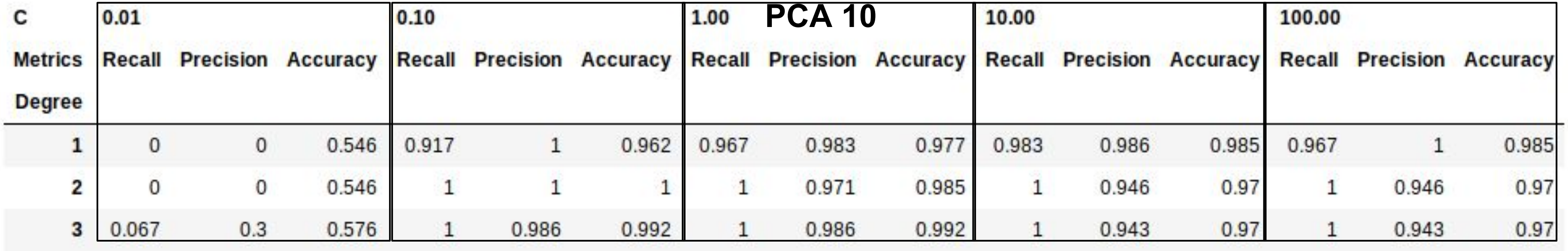


Todas as características



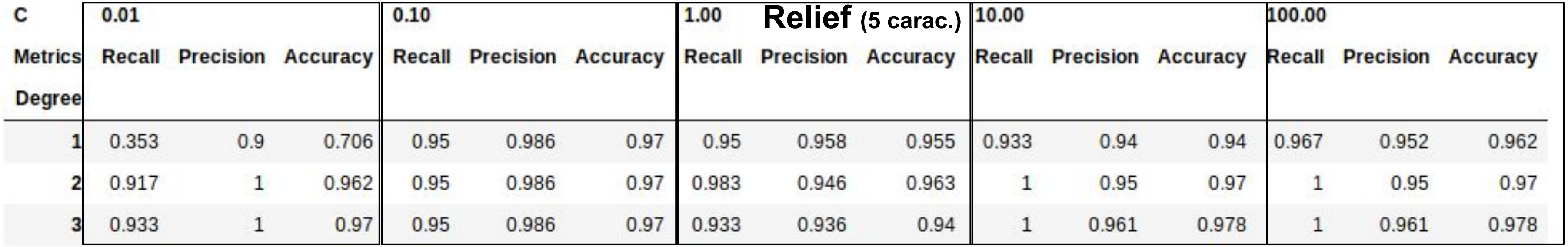
C	0.01			0.10			1.00			10.00			100.00		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Degree															
1	0.033	0.2	0.561	0.967	0.971	0.969	0.983	0.969	0.976	0.983	0.94	0.961	0.983	0.957	0.969
2	0.767	1	0.893	0.967	0.971	0.969	1	0.971	0.985	1	0.961	0.977	1	0.961	0.977
3	0.9	0.971	0.938	1	0.955	0.976	1	0.971	0.985	1	0.971	0.985	1	0.971	0.985

PCA 10



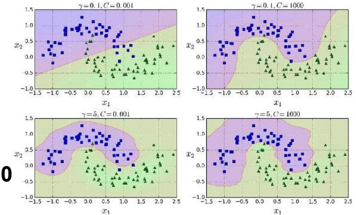
C	0.01			0.10			1.00			10.00			100.00		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Degree															
1	0	0	0.546	0.917	1	0.962	0.967	0.983	0.977	0.983	0.986	0.985	0.967	1	0.985
2	0	0	0.546	1	1	1	1	0.971	0.985	1	0.946	0.97	1	0.946	0.97
3	0.067	0.3	0.576	1	0.986	0.992	1	0.986	0.992	1	0.943	0.97	1	0.943	0.97

Relief (5 carac.)



C	0.01			0.10			1.00			10.00			100.00		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Degree															
1	0.353	0.9	0.706	0.95	0.986	0.97	0.95	0.958	0.955	0.933	0.94	0.94	0.967	0.952	0.962
2	0.917	1	0.962	0.95	0.986	0.97	0.983	0.946	0.963	1	0.95	0.97	1	0.95	0.97
3	0.933	1	0.97	0.95	0.986	0.97	0.933	0.936	0.94	1	0.961	0.978	1	0.961	0.978

SVM - Kernel RBF (validação cruzada k=10)



Todas as características

C	0.01			0.10			1.00			10.00			100		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Gamma															
0.1	0	0	0.546	0.933	1	0.969	0.983	1	0.992	1	1	1	1	1	1
1.0	0	0	0.546	0	0	0.546	0.357	1	0.708	0.407	1	0.731	0.407	1	0.731
10.0	0	0	0.546	0	0	0.546	0	0	0.546	0	0	0.546	0	0	0.546

PCA 10

C	0.01			0.10			1.00			10.00			100.00		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Gamma															
0.1	0	0	0.546	0.443	0.9	0.747	0.983	0.971	0.977	0.983	0.986	0.985	0.983	0.986	0.985
1.0	0	0	0.546	0	0	0.546	0.323	1	0.693	0.423	1	0.739	0.423	1	0.739
10.0	0	0	0.546	0	0	0.546	0	0	0.546	0	0	0.546	0	0	0.546

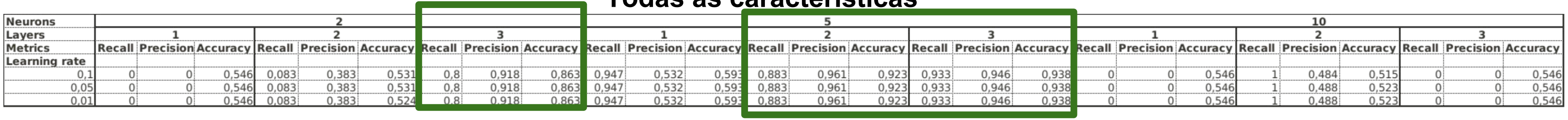
Relief (5 carac.)

C	0.01			0.10			1.00			10.00			100.00		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Gamma															
0.1	0	0	0.546	0.95	1	0.977	0.983	1	0.992	1	0.946	0.97	1	0.961	0.978
1.0	0	0	0.546	0.293	0.9	0.678	0.917	0.986	0.954	0.917	0.969	0.947	0.917	0.969	0.947
10.0	0	0	0.546	0	0	0.546	0.193	0.7	0.633	0.193	0.7	0.633	0.193	0.7	0.633

Redes Neurais

(2,3) (5,2) (5,3)

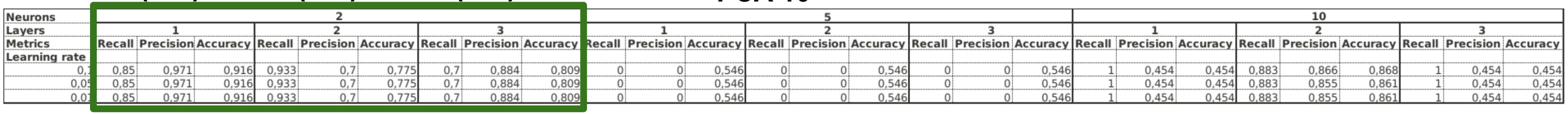
Todas as características



Neurons	2									5									10								
Layers	1			2			3			1			2			3			1			2			3		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Learning rate																											
0,1	0	0	0,546	0,083	0,383	0,531	0,8	0,918	0,863	0,947	0,532	0,593	0,883	0,961	0,923	0,933	0,946	0,938	0	0	0,546	1	0,484	0,515	0	0	0,546
0,05	0	0	0,546	0,083	0,383	0,531	0,8	0,918	0,863	0,947	0,532	0,593	0,883	0,961	0,923	0,933	0,946	0,938	0	0	0,546	1	0,488	0,523	0	0	0,546
0,01	0	0	0,546	0,083	0,383	0,524	0,8	0,918	0,863	0,947	0,532	0,593	0,883	0,961	0,923	0,933	0,946	0,938	0	0	0,546	1	0,488	0,523	0	0	0,546

(2,1) (2,2) (2,3)

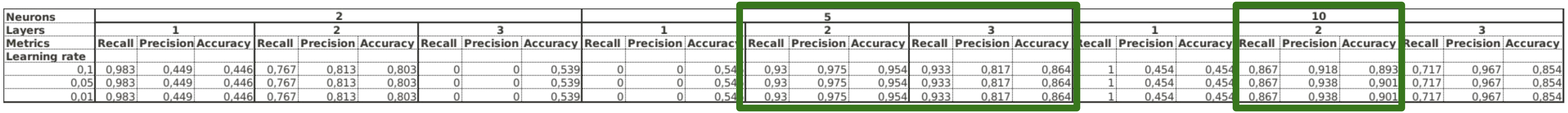
PCA 10



Neurons	2									5									10								
Layers	1			2			3			1			2			3			1			2			3		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Learning rate																											
0,1	0,85	0,971	0,916	0,933	0,7	0,775	0,7	0,884	0,809	0	0	0,546	0	0	0,546	0	0	0,546	1	0,454	0,454	0,883	0,866	0,868	1	0,454	0,454
0,05	0,85	0,971	0,916	0,933	0,7	0,775	0,7	0,884	0,809	0	0	0,546	0	0	0,546	0	0	0,546	1	0,454	0,454	0,883	0,855	0,861	1	0,454	0,454
0,01	0,85	0,971	0,916	0,933	0,7	0,775	0,7	0,884	0,809	0	0	0,546	0	0	0,546	0	0	0,546	1	0,454	0,454	0,883	0,855	0,861	1	0,454	0,454

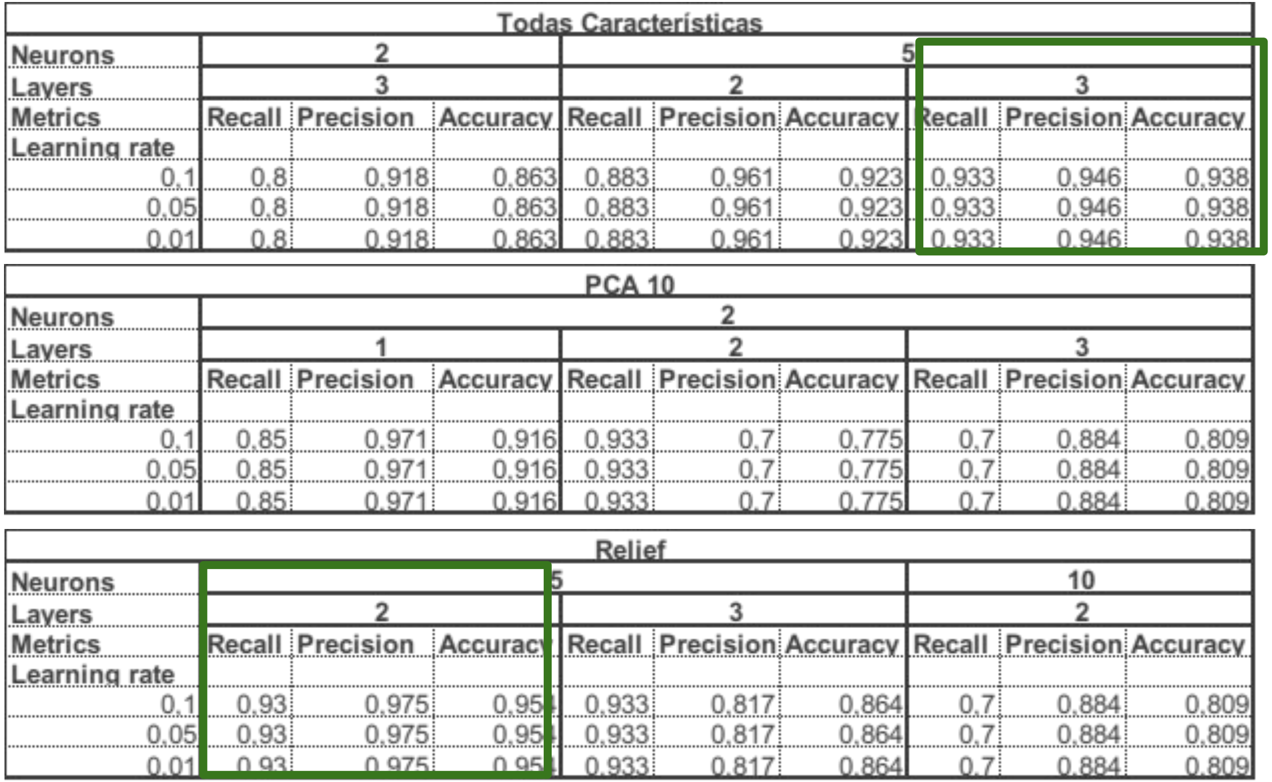
Relief

(5,2) (5,3) (10,2)



Neurons	2									5									10								
Layers	1			2			3			1			2			3			1			2			3		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accurac	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Learning rate																											
0,1	0,983	0,449	0,446	0,767	0,813	0,803	0	0	0,539	0	0	0,54	0,93	0,975	0,954	0,933	0,817	0,864	1	0,454	0,454	0,867	0,918	0,893	0,717	0,967	0,854
0,05	0,983	0,449	0,446	0,767	0,813	0,803	0	0	0,539	0	0	0,54	0,93	0,975	0,954	0,933	0,817	0,864	1	0,454	0,454	0,867	0,938	0,901	0,717	0,967	0,854
0,01	0,983	0,449	0,446	0,767	0,813	0,803	0	0	0,539	0	0	0,54	0,93	0,975	0,954	0,933	0,817	0,864	1	0,454	0,454	0,867	0,938	0,901	0,717	0,967	0,854

Redes Neurais

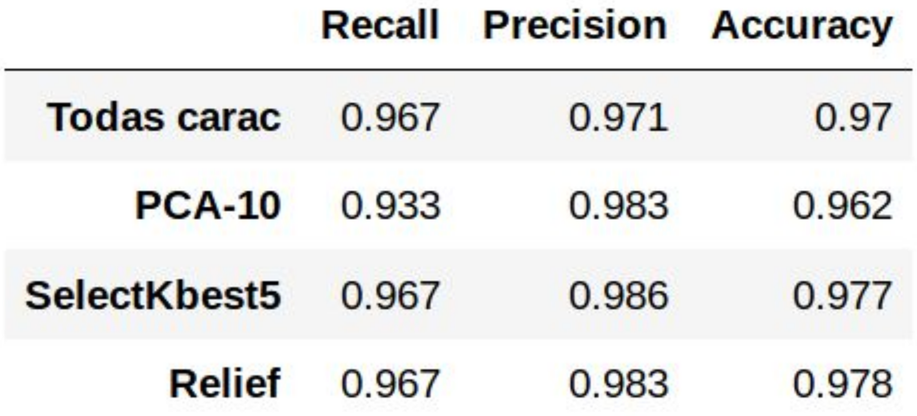


Todas Características									
Neurons	2			5					
Layers	3			2			3		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Learning rate									
0,1	0,8	0,918	0,863	0,883	0,961	0,923	0,933	0,946	0,938
0,05	0,8	0,918	0,863	0,883	0,961	0,923	0,933	0,946	0,938
0,01	0,8	0,918	0,863	0,883	0,961	0,923	0,933	0,946	0,938

PCA 10									
Neurons	2								
Layers	1			2			3		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Learning rate									
0,1	0,85	0,971	0,916	0,933	0,7	0,775	0,7	0,884	0,809
0,05	0,85	0,971	0,916	0,933	0,7	0,775	0,7	0,884	0,809
0,01	0,85	0,971	0,916	0,933	0,7	0,775	0,7	0,884	0,809

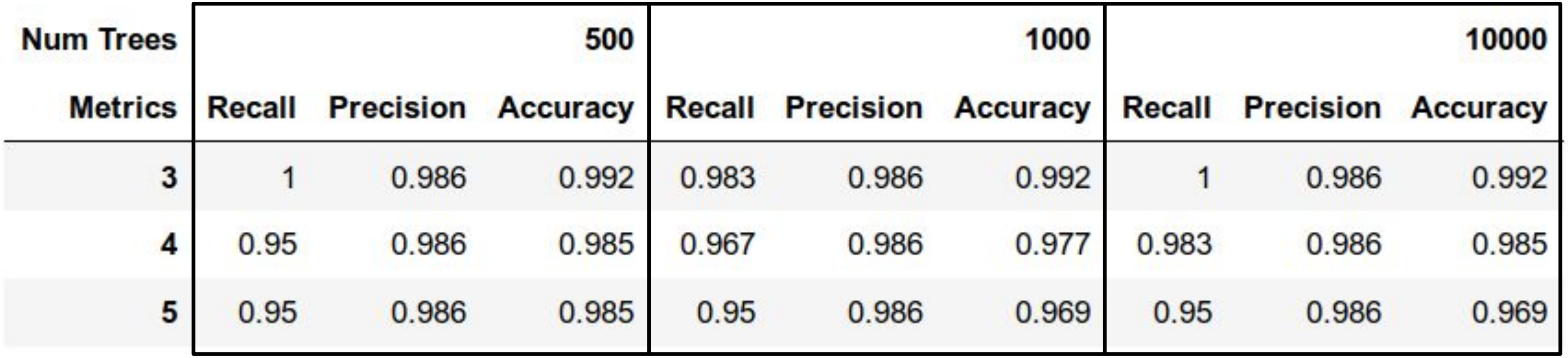
Relief									
Neurons	5						10		
Layers	2			3			2		
Metrics	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy
Learning rate									
0,1	0,93	0,975	0,954	0,933	0,817	0,864	0,7	0,884	0,809
0,05	0,93	0,975	0,954	0,933	0,817	0,864	0,7	0,884	0,809
0,01	0,93	0,975	0,954	0,933	0,817	0,864	0,7	0,884	0,809

Redes Bayesianas - Naive Bayes classifier



	Recall	Precision	Accuracy
Todas carac	0.967	0.971	0.97
PCA-10	0.933	0.983	0.962
SelectKbest5	0.967	0.986	0.977
Relief	0.967	0.983	0.978

Árvores de decisão - Random Forest



Num Trees	500			1000			10000		
Metrics	**Recall**	**Precision**	**Accuracy**	**Recall**	**Precision**	**Accuracy**	**Recall**	**Precision**	**Accuracy**
3	1	0.986	0.992	0.983	0.986	0.992	1	0.986	0.992
4	0.95	0.986	0.985	0.967	0.986	0.977	0.983	0.986	0.985
5	0.95	0.986	0.985	0.95	0.986	0.969	0.95	0.986	0.969

Conclusão

SVM foi o método que conseguiu uma performance de 100% em todas as métricas utilizadas, com validação cruzada (k=10), utilizando kernels com transformações não lineares (polinomial e RBF).

Random Forest também teve uma performance perto de 100%, acredito que por poder capturar separações não lineares.

Naive Bayes teve performance interessante, perto de 100%.

https://gitlab.com/bkemmer/ml-wine-analysis