

Universidade Federal do ABC



Relatório do Projeto - Mineração de Dados (Diurno) -  
Grupo 8

Discentes:

Diogo Eduardo Lima Alves - RA: 21012813

Marco Soares de Araujo - RA:11010112

Docente:

Prof. Dr. Thiago Ferreira Covões

Santo André  
Agosto - 2017

## 1. Introdução

Criada na década de 1990, essa base de dados mostra o resultado de uma análise química de vinhos que cresceram na mesma região da Itália, mas com três tipos diferentes de cultivo. Esta análise identifica treze constituintes em cada um dos três tipos de vinhos.

## 2. Análise Exploratória

Os atributos presentes na base de dados são:

- (V1) Class
- (V2) Alcohol
- (V3) Malic acid
- (V4) Ash
- (V5) Alcalinity of ash
- (V6) Magnesium
- (V7) Total phenols
- (V8) Flavanoids
- (V9) Nonflavanoid phenols
- (V10) Proanthocyanins
- (V11) Color intensity
- (V12) Hue
- (V13) OD280/OD315 of diluted wines
- (V14) Proline

Esta base de dados tem ao todo 178 instâncias, sem a presença de valores ausentes, apresenta característica multivariada, com características dos atributos podendo ser Reais ou Inteiros.

Os problemas possíveis de surgirem em um dataset deste tipo são a dificuldade em se particionar os objetos de acordo com suas classes, e encontrar combinação de atributos que possam facilitar este processo. Por conta disso, não foi possível descartar quaisquer atributos da base de dados inicialmente.

### 3. Metodologia

A linguagem utilizada no projeto será o R, que é uma linguagem de programação de código aberto e um ambiente para computação estatística amplamente utilizado para mineração de dados, tendo uma popularidade crescente nos últimos anos.

O primeiro procedimento a ser executado, após a importação dos dados, será obter as estatísticas descritivas da base de dados, com o objetivo de poder acessar rapidamente o quadro geral do que temos em mãos. Em seguida, serão gerados diversos gráficos com o intuito de observar o comportamento dos atributos presentes na base de dados.

Por último, será aplicado o classificador Naïve Bayes e realizado o algoritmo de validação cruzada sobre os resultados do classificador.

### 4. Resultados

```
> summary(wine)
```

V1		V2		V3		V4		V5		V6	
Min.	:1.000	Min.	:11.03	Min.	:0.740	Min.	:1.360	Min.	:10.60	Min.	: 70.00
1st Qu.	:1.000	1st Qu.	:12.36	1st Qu.	:1.603	1st Qu.	:2.210	1st Qu.	:17.20	1st Qu.	: 88.00
Median	:2.000	Median	:13.05	Median	:1.865	Median	:2.360	Median	:19.50	Median	: 98.00
Mean	:1.938	Mean	:13.00	Mean	:2.336	Mean	:2.367	Mean	:19.49	Mean	: 99.74
3rd Qu.	:3.000	3rd Qu.	:13.68	3rd Qu.	:3.083	3rd Qu.	:2.558	3rd Qu.	:21.50	3rd Qu.	:107.00
Max.	:3.000	Max.	:14.83	Max.	:5.800	Max.	:3.230	Max.	:30.00	Max.	:162.00

V7		V8		V9		V10		V11		V12	
Min.	:0.980	Min.	:0.340	Min.	:0.1300	Min.	:0.410	Min.	: 1.280	Min.	:0.4800
1st Qu.	:1.742	1st Qu.	:1.205	1st Qu.	:0.2700	1st Qu.	:1.250	1st Qu.	: 3.220	1st Qu.	:0.7825
Median	:2.355	Median	:2.135	Median	:0.3400	Median	:1.555	Median	: 4.690	Median	:0.9650
Mean	:2.295	Mean	:2.029	Mean	:0.3619	Mean	:1.591	Mean	: 5.058	Mean	:0.9574
3rd Qu.	:2.800	3rd Qu.	:2.875	3rd Qu.	:0.4375	3rd Qu.	:1.950	3rd Qu.	: 6.200	3rd Qu.	:1.1200
Max.	:3.880	Max.	:5.080	Max.	:0.6600	Max.	:3.580	Max.	:13.000	Max.	:1.7100

V13		V14	
Min.	:1.270	Min.	: 278.0
1st Qu.	:1.938	1st Qu.	: 500.5
Median	:2.780	Median	: 673.5
Mean	:2.612	Mean	: 746.9
3rd Qu.	:3.170	3rd Qu.	: 985.0
Max.	:4.000	Max.	:1680.0

Figura 1: Análise geral dos atributos pela função summary

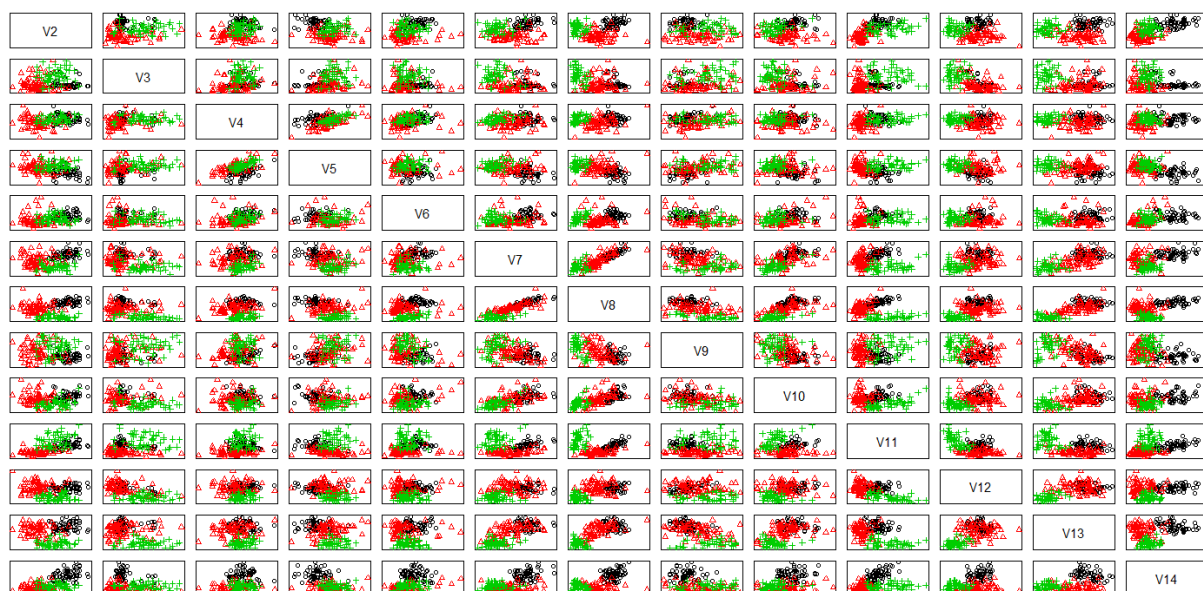


Figura 2: Gráfico de dispersão dos atributos

A Figura 1 nos mostra o sumário de todos os atributos, apresentando seus valores mínimos, máximo, médios, medianas e quadrantes. Através dos valores mínimos e máximos de todos os atributos é possível perceber que todos eles contam com uma grande dispersão e presença de outliers.

Observando o gráfico de dispersão dos atributos, representado pela Figura 2, a afirmação anterior se confirma, mostrando que os objetos das classes 1, 2 e 3 se sobrepõem de tal forma que fica muito difícil fazer a distinção entre os objetos das três classes de maneira que seja útil. No entanto, os atributos V13 e V14 se mostram os mais promissores entre todos os demais no que se refere a particionamento, e serão explorados mais adiante.

Através da análise de correlação mostrada na Figura 3, nota-se que as correlações entre os pares de atributos [V7, V8], [V7, V13], [V7, V14], [V8, V13], [V8, V14] e [V13, V14] se mostram especialmente azuis de acordo com a escala de cores do gráfico, indicando a força da relação linear entre eles.

É importante notar que, apesar da forte correlação entre os atributos, para inferir uma conexão entre eles de fato e obter mais informações foram necessários mais passos.

Portanto, para analisar ainda mais este grupo de atributos e sua relevância, foram realizadas as criações de uma árvore de decisões, apresentada na Figura 4, e um dendograma, apresentado na Figura 5.

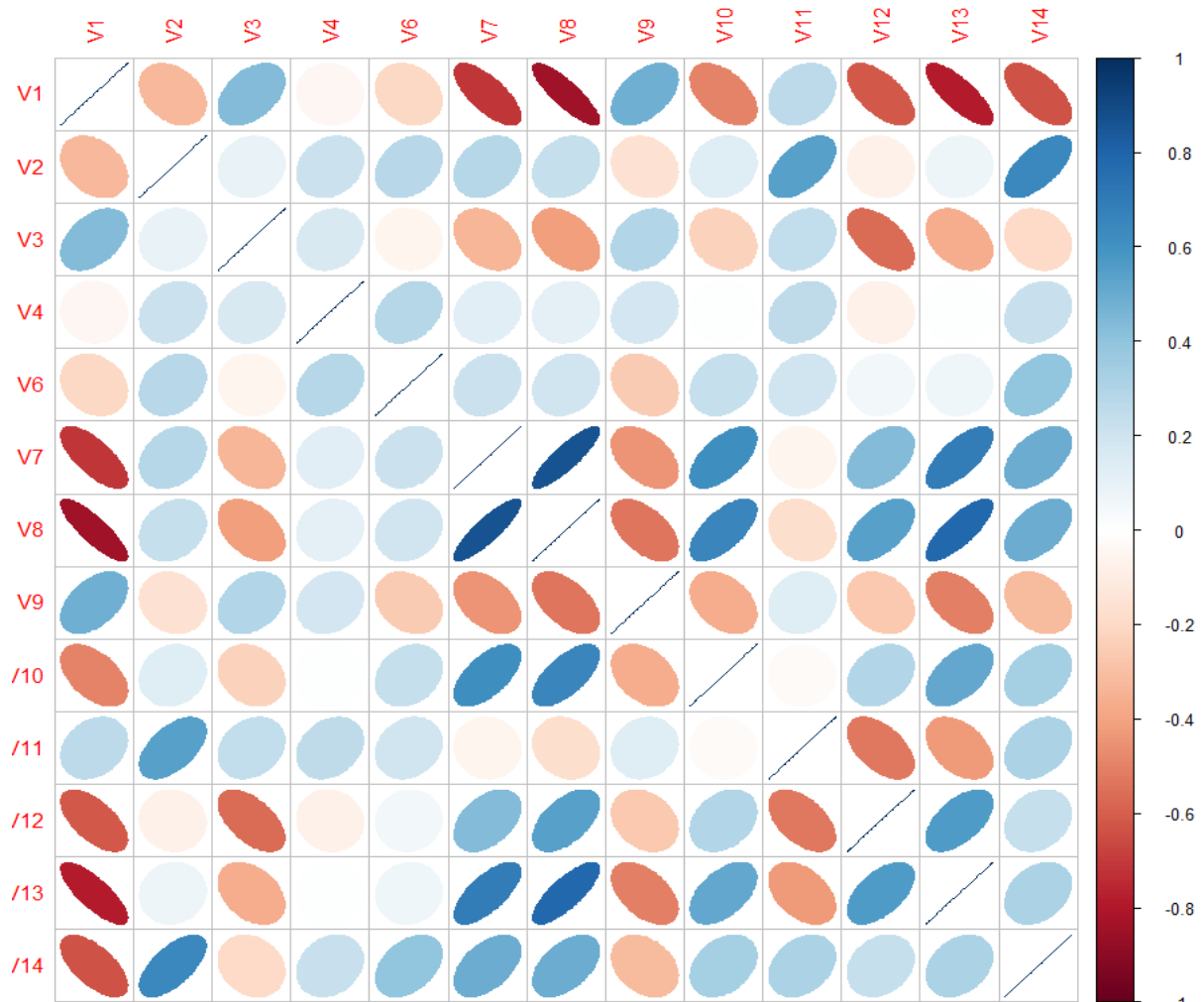


Figura 3: Análise de Correlação

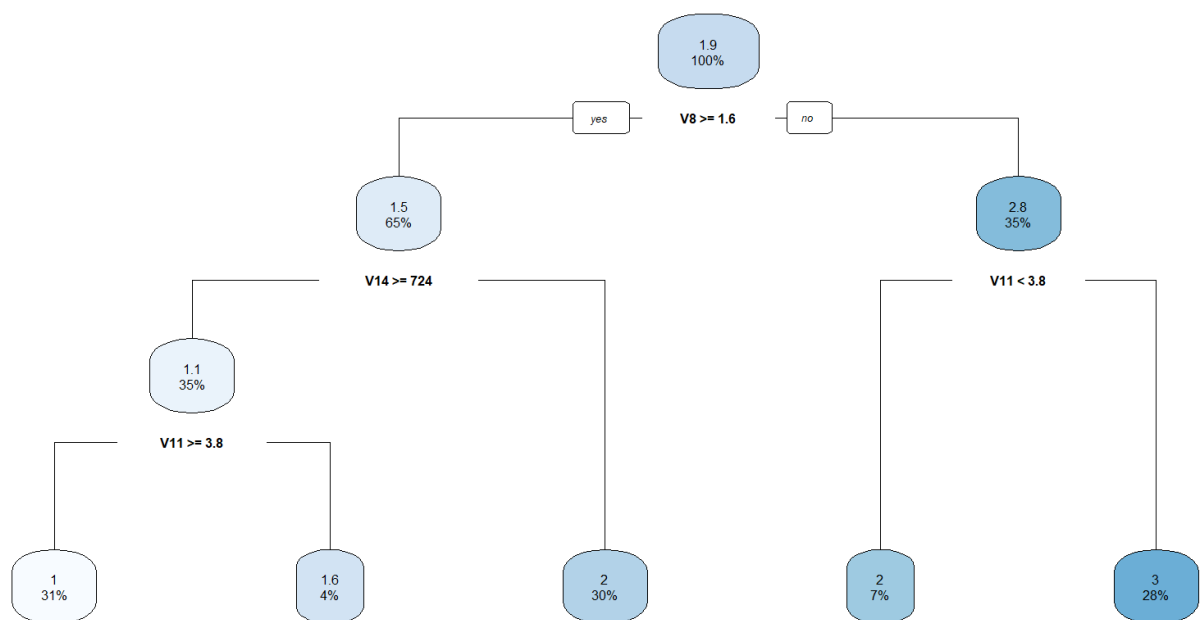


Figura 4: Árvore de Decisão



Figura 6: Gráfico de dispersão Atributos V14xV13

Dada essa relação de dispersão, foi de interesse testar a acurácia do algoritmo K-NN, que se saiu relativamente bem conforme dados abaixo, atingindo uma acurácia de 0,7347.

Para este teste foram usados apenas os atributos V1, V13 e V14 do banco de dados.

#### Overall Statistics

Accuracy : 0.7347  
95% CI : (0.5892, 0.8505)  
No Information Rate : 0.4694  
P-Value [Acc > NIR] : 0.0001481

Kappa : 0.6068  
McNemar's Test P-Value : 0.0601843

#### Statistics by Class:

	Class: 1	Class: 2	Class: 3
Sensitivity	0.9412	0.5652	0.7778
Specificity	0.9375	0.9615	0.7500
Pos Pred Value	0.8889	0.9286	0.4118
Neg Pred Value	0.9677	0.7143	0.9375
Prevalence	0.3469	0.4694	0.1837
Detection Rate	0.3265	0.2653	0.1429
Detection Prevalence	0.3673	0.2857	0.3469
Balanced Accuracy	0.9393	0.7634	0.7639

Figura 7: Estatísticas gerais

A eficácia do Naive Bayes foi muito acima do esperado, mesmo sendo utilizado com todos os 13 atributos ele chegou a atingir uma acurácia perfeita em uma das interações, gerando a tabela confusão representada na Figura 8.

```
> confusao = table(wineteste$V1, pred)
> confusao
      pred
      1  2  3
1  21  0  0
2   0 25  0
3   0  0 12
```

Figura 8: Tabela de confusão do Naive Bayes

## 5. Comentários Finais

Classificar e agrupar novas entradas no banco de dados, estudando correlações, aprendendo e extraíndo informação a partir dos dados é extremamente útil para tomar diversas decisões de negócio, por exemplo.

Saber a priori os atributos mais relevantes poderia ajudar a escolher modelos mais simples que resolvessem com eficácia os problemas propostos e também a excluir atributos irrelevantes, estas análises são feitas se apoiando em diversos fatores, como gráficos de dispersão, correlação e histogramas, por exemplo.

Existem muitos outros algoritmos que podem ser testados, porém como o Naive Bayes teve um desempenho impressionante em todas as vezes que foi usado (com uma média em torno de 0.98 de acurácia), é crível que seja a melhor solução, já que além de eficiente é um algoritmo bastante rápido e simples. Também foi usado um algoritmo mais simples, o K-NN. Uma vez que já havia sido encontrado um ótimo algoritmo para minerar os dados, foi de interesse selecionar um algoritmo mais simples que exigisse uma maior análise da base de dados, selecionando atributos, avaliando correlações e identificando as informações mais úteis, o que poderia ser útil caso a base de dados tivesse um número enorme de instâncias a serem computadas.