Ensaios em Técnicas de Pré-processamento Usando Dados Abertos

Artur Rodrigues Rocha Neto (431951) e Matheus Costa Mesquita Martins (371846)

Universidade Federal do Ceará - Departamento de Engenharia de Teleinformática
Campus do Pici, Acesso Público, Bloco 725, CEP: 60455-970 - Brasil

Resumo. O pré-processamento é uma etapa importante na análise de dados. Investigações preliminares valendo de descritores estatísticos, gráficos e tabelas auxiliam no planejamento de um modelo preditivo. Hoje, existe uma gama de dados abertos que podem ser usados para a prática do pré-processamento. Este trabalho traz ensaios de técnicas de pré-processamento usando um conjunto de dados de classificação de tipos de vidro. Uma investigação dos preditores, seus relacionamentos e capacidades descritivas, é apresentado e discutido. Por fim, uma normalização dos dados e uma análise usando componentes principais é mostrada.

1 Introdução

O processo de tomada de decisão é parte importante no comportamento dos seres humanos em suas relações com o ambiente onde vivem e com seus semelhantes. Decisões são tomadas com base em informações, que podem ser extraídas de dados e/ou experiências. Uma linha de pesquisa de grande importância no contexto tecnológico atual é a criação de mecanismos capazes de tomar decisões de forma automática. A modelagem preditiva é o processo de desenvolvimento de ferramentas matemáticas que gerão predições precisas sobre um certo fenônemo [1]. O tratamento dos dados é parte importante na confecção dos modelos preditivos. Técnicas de pré-processamento são utilizadas para garantir a consistência do conjunto de dados, além de minimizar problemas como ausências, ruídos (*outliers*) e diferença de ordem (escalonamento) [2]. Neste trabalho, algumas técnicas de pré-processamento e de análise exploratória dos dados são estudadas usando um conjunto de dados abertos.

2 Metodologia

Além dos dados, precisamos de medidas (descritores) que auxiliam na criação do modelo preditivo, seja em sua confecção ou na escolha dos melhores preditores. Essa seção apresenta as escolhas feitas nesse âmbito, bem como a extração dos descritores do conjunto de dados. Os experimentos, gráficos e resultados foram gerados com auxílio das linguagens R [3] e Python [4].

2.1 Conjunto de dados

O conjunto de dados escolhido para os ensaios apresenta amostras que caracterizão tipos de vidros com base atributos químicos (índice de refração e teores óxidos). A classificação de tipos de vidro foi motivada por investigações criminais, já que restos de vidro encontrados em cenas de crime pode ser usados como evidência [5].

O conjunto possui 214 amostras. Para cada amostra, temos: um índice númerico, valor do índice de refração da amostra, 8 percentagens de elementos estruturais (Na, Mg, Al, Si, K, Ca, Ba e Fe) e tiqueta de classe. O arquivo de dados original recebeu as seguintes modificações: adição de cabeçalho com nome das colunas e renomeação das etiquetas de classe (valor númerico para valor textual).

A Figura 1 mostra que as classes 1 e 2 possuem uma quantidade significativamente maior de amostras que as demais, revelando que o conjunto de dados é desbalanceado. Isso pode refletir em um viês do classificador a essas classes. Não há amostras da classe 4, o que invalidaria o modelo preditivo para amostras desse tipo.

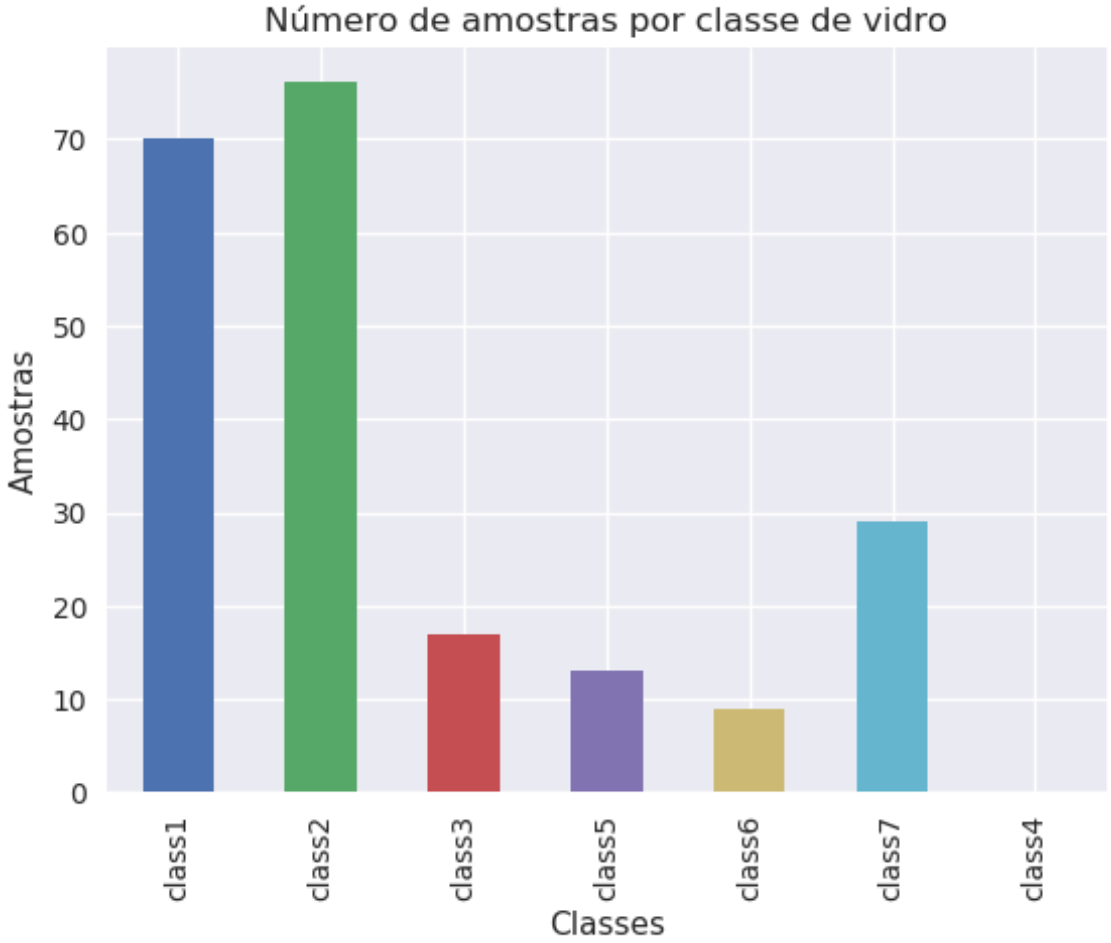

Figura 1: Número de amostras por classe

2.2 Análise monovariada

Os seguintes descritores estatísticos foram extraídos da distribuição dos dados de vidro: média, desvio padrão, variância e assimetria. Adicionamos a variância por acreditarmos ser uma media de dispersão conceitualmente mais fácil de ser interpretada. Os dois cenários de extração foram:

1. **Análise monovariada incondicional:** os analisadores são calculados para cada preditor englobando todo o conjunto de dados; e
2. **Análise monovarida por classe:** os dados são separados por classe antes das medições estatísticas.

Preditor	Média	Desvio Padrão	Variância	Assimetria
RI	1.5184	0.003	0.0	1.6254
Na	13.4079	0.8166	0.6668	0.4542
Mg	2.6845	1.4424	2.0805	-1.1526
Al	1.4449	0.4993	0.2493	0.9073
Si	72.6509	0.7745	0.5999	-0.7304
K	0.4971	0.6522	0.4254	6.5516
Ca	8.957	1.4232	2.0254	2.0471
Ba	0.175	0.4972	0.2472	3.4164
Fe	0.057	0.0974	0.0095	1.7543

Tabela 1: Análise monovariada incondicional

A Tabela 1 mostra que os preditores Na e Si estão em ordem de grandeza bem acima dos demais preditores. Essa diferença de escala indica a necessidade de uma normalização do conjunto de dados para melhor separação entre as classes. O preditor Ri possui uma baixíssima variância (arrendoda para zero com 4 casas decimais), o mesmo para o preditor Fe. Por não variarem significativamente entre as classes, isso indica que eles não possuem grande poder descritivo. Mg e Ca possuem as maiores variâncias globais, indicando grande poder descritivo para esses dois preditores. Os preditores K e Ba possuem uma forte assimetria positiva (6.5516 e 3.4164, respectivamente). Assimetrias indicam forte viês para um certo conjunto de valores, o que pode implicar em uma classificação viciada ao conjunto de dados de treinamento (Figura 2).

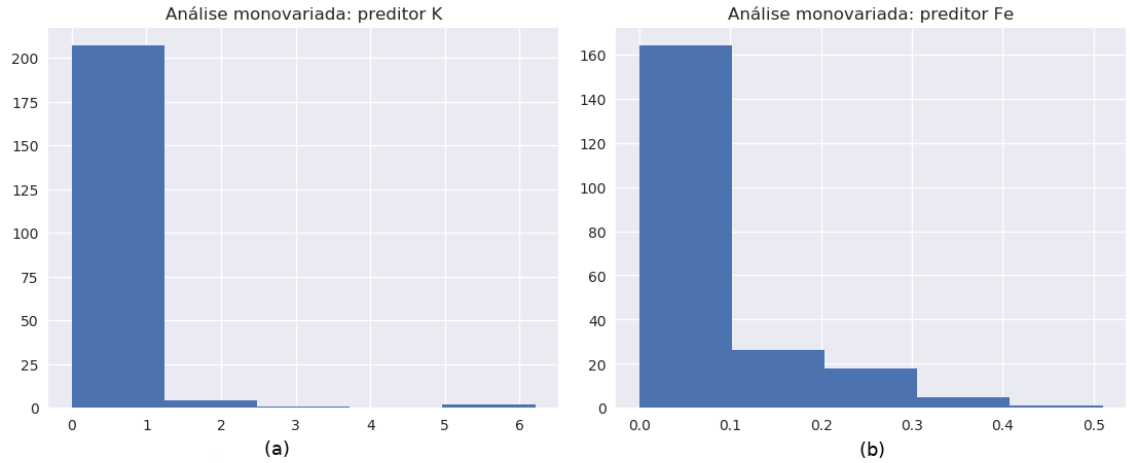

Figura 2: Histograma incondicional do preditores Ka e Fe. Note a forte assimetria a direita, revelando viês de certos valores.

A Figura 3 traz a dispersão geral dos preditores por classe. A lista completa dos descritores encontrasse no Apêndice A.

2.3 Análise bivariada

É de interesse na análise dos dadods investigar os relacionamentos entre pares de preditores. Podemos visualizar a correlção dois-a-dois com a ajuda de um gráfica de dispersão apresentado na

Figura 6 do Apêndice B. Podemos indicar que existe um relacionamento linear forte entre nos pares (Ri, Ca) e (Ri, Si).

Esses relacionamentos ficam mais evidentes quando calculamos a matriz de correlação, que resume em valores o relacionamento par-a-par entre os preditores. A correlação varia no intervalo [-1, 1]: valores positivo indicam relação direta, enquanto que valores negativos, inversa, e valore perto de zero significam baixa correlação. A Figura 4 traz essa matriz em formato de mapa de calor, onde as correlação positivas mais fortes ficam marcadas com cores escuras e as correlações negativas, com cores mais claras.

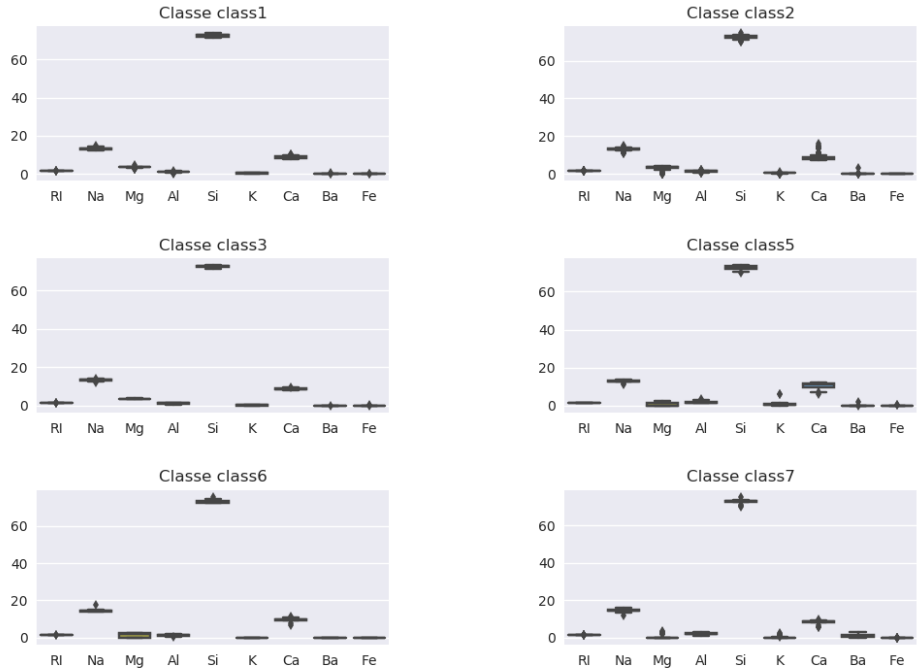

Figura 3: Análise monovariada por classe

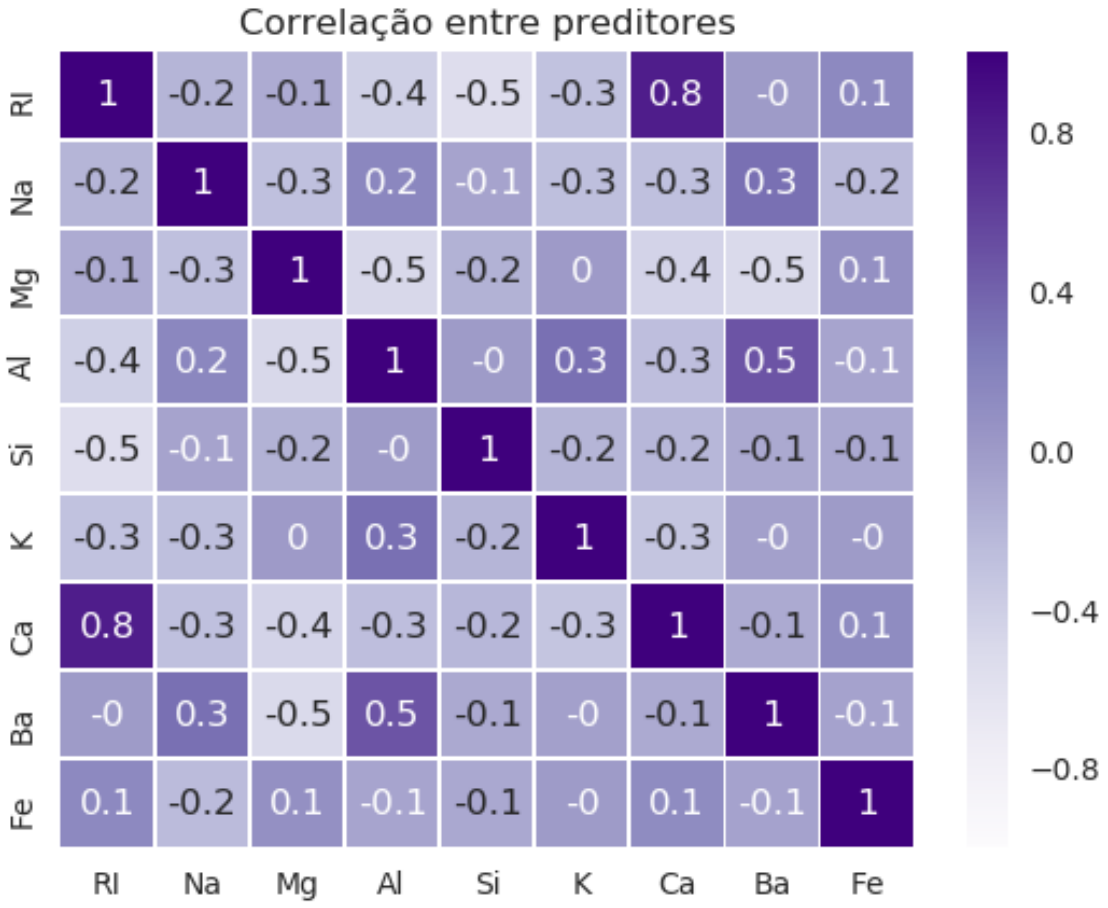

Figura 4: Análise bivariada: matriz de correlção estilo mapa de calor

3 Resultados

Finalizada a investigação preliminar, passamos para o pré-processamento: uma normalização dos preditores e uma análise usando componentes principais.

A normalização visa transformar os valores dos preditores de forma a se assemelharem com uma distribuição normal. Para cada preditor, os valores foram subtráidos da média daquele preditor (centralização) e depois dividio pelo desvio padrão (escalonamento).

A análise de componentes principais (PCA) é um método que age nos dados de forma a eliminar sobreposições e a melhor separar os conjuntos [6]. A PCA é uma transformação linear. Pode ser usada para redução de dimensionalidade ou para geração de novos preditores. A PCA retorna uma base que revela a maior dispersão dos dados [7]. A Figura 5 mostra o conjunto transformado.

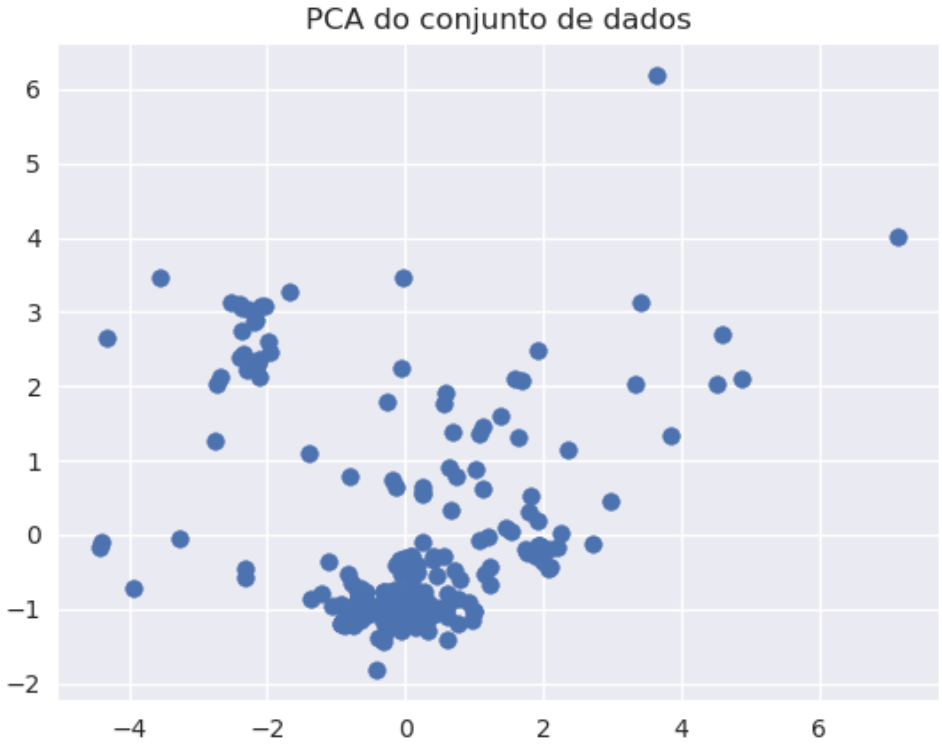

Figura 5: PCA

Referências

[1] Max Kuhn and Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.

[2] Ryszard S Michalski, Jaime G Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[4] Jake VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc., 1st edition, 2016.

[5] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

[6] Análise de componentes principais (pca). http://www.de.ufpb.br/ luiz/AED/Aula9.pdf. Online; acessado em 06-Setembro-2018.

[7] Análise de componentes principais (pca). http://www2.ic.uff.br/ aconci/PCA-ACP.pdf. Online; acessado em 06-Setembro-2018.

A Análise monovariada por classe: descritores estatísticos

Classe	Preditor	Média	Desvio Padrão	Variância	Assimetria
class1	RI	1.5187	0.0023	0.0	0.7767
class1	Na	13.2423	0.4993	0.2493	0.7872
class1	Mg	3.5524	0.247	0.061	-0.7067
class1	Al	1.1639	0.2732	0.0746	-1.1279
class1	Si	72.6191	0.5695	0.3243	-0.5788
class1	K	0.4474	0.2149	0.0462	-0.9397
class1	Ca	8.7973	0.5748	0.3304	0.7168
class1	Ba	0.0127	0.0838	0.007	7.8972
class1	Fe	0.057	0.0891	0.0079	1.3619
class2	RI	1.5186	0.0038	0.0	2.1414
class2	Na	13.1117	0.6642	0.4411	-1.0923
class2	Mg	3.0021	1.2157	1.4778	-1.8458
class2	Al	1.4082	0.3183	0.1013	-0.3867
class2	Si	72.598	0.7246	0.525	-1.4319
class2	K	0.5211	0.2137	0.0457	-1.0095
class2	Ca	9.0737	1.9216	3.6927	2.1664
class2	Ba	0.0503	0.3623	0.1313	8.574
class2	Fe	0.0797	0.1064	0.0113	0.9876
class3	RI	1.518	0.0019	0.0	1.169
class3	Na	13.4371	0.5069	0.2569	-0.5563
class3	Mg	3.5435	0.1628	0.0265	0.7251
class3	Al	1.2012	0.3475	0.1207	-0.4004
class3	Si	72.4047	0.5123	0.2624	-0.8326
class3	K	0.4065	0.2299	0.0528	-0.7679
class3	Ca	8.7829	0.3801	0.1445	0.9457
class3	Ba	0.0088	0.0364	0.0013	4.1231
class3	Fe	0.0571	0.1079	0.0116	2.0374
class5	RI	1.5189	0.0033	0.0	-0.7288
class5	Na	12.8277	0.777	0.6038	-1.1969
class5	Mg	0.7738	0.9991	0.9983	0.7503
class5	Al	2.0338	0.6939	0.4815	1.2781
class5	Si	72.3662	1.2823	1.6443	-0.8244
class5	K	1.47	2.1387	4.574	2.0382
class5	Ca	10.1238	2.1838	4.7689	-1.0112
class5	Ba	0.1877	0.6083	0.37	3.5344
class5	Fe	0.0608	0.1556	0.0242	2.5815
class6	RI	1.5175	0.0031	0.0	-1.6359
class6	Na	14.6467	1.084	1.1751	2.4281
class6	Mg	1.3056	1.0971	1.2037	-0.3175
class6	Al	1.3667	0.5719	0.327	-0.9695
class6	Si	73.2067	1.0795	1.1652	1.4428
class6	K	0.0	0.0	0.0	0.0
class6	Ca	9.3567	1.4499	2.1024	-0.7935
class6	Ba	0.0	0.0	0.0	0.0
class6	Fe	0.0	0.0	0.0	0.0
class7	RI	1.5171	0.0025	0.0	1.0863
class7	Na	14.4421	7 0.6864	0.4711	-1.61
class7	Mg	0.5383	1.1177	1.2492	1.8156
class7	Al	2.1228	0.4427	0.196	-0.3258
class7	Si	72.9659	0.9402	0.884	-1.3477
class7	K	0.3252	0.6685	0.4469	2.3801
class7	Ca	8.4914	0.9735	0.9477	-2.1532
class7	Ba	1.04	0.6653	0.4427	0.5012
class7	Fe	0.0134	0.0298	0.0009	1.9808

Tabela 2: Análise monovariada por classe: descritores estatísticos

B Análise bivariada: gráfico de dispersão

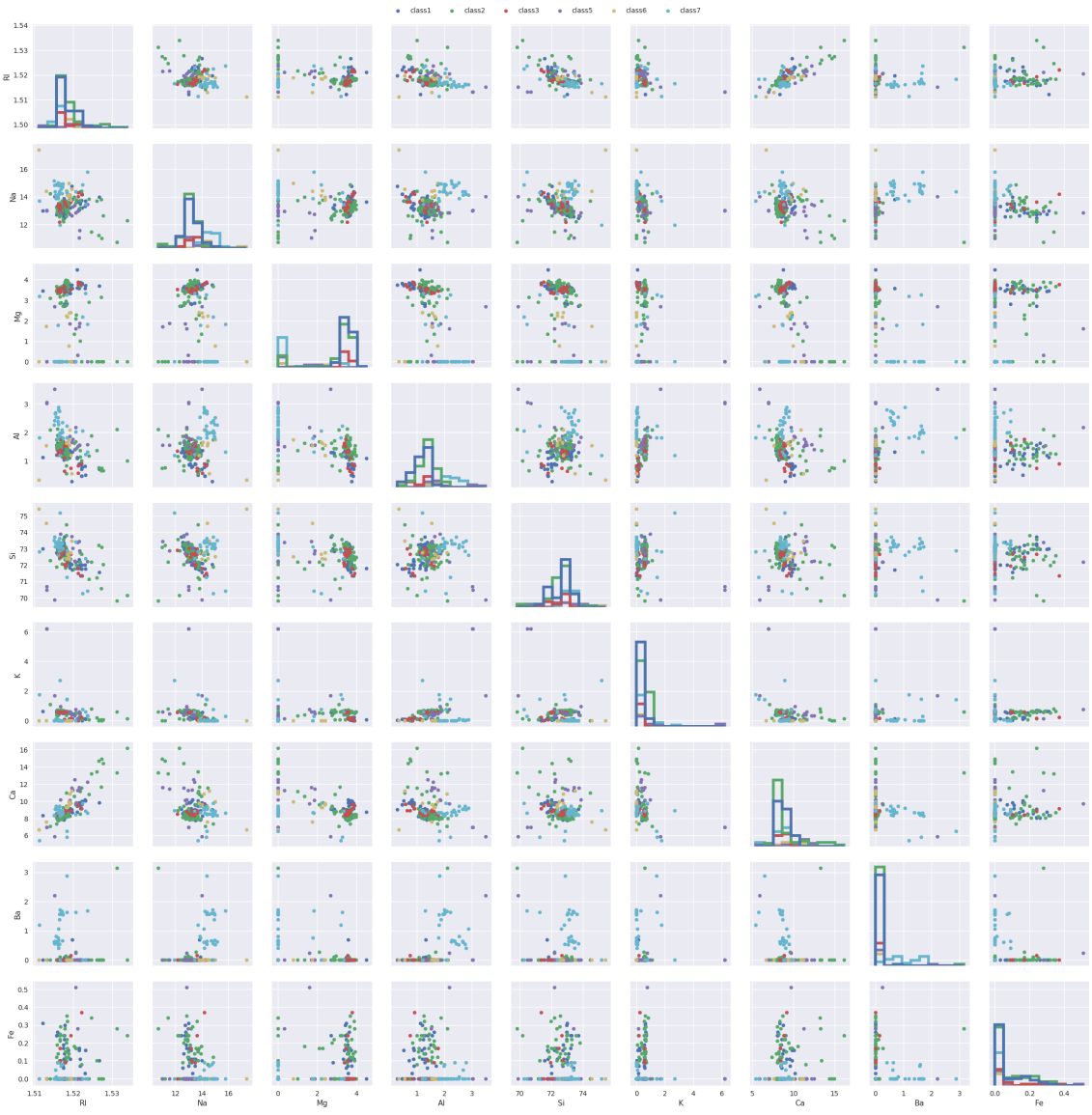

Figura 6: Análise bivariada: gráfico de dispersão