Atividade 2 - Visualização de informação

Leonardo Cesar Silva dos Santos Fernando Augusto Cardoso Candalaft

1 Explorando o conjunto de dados

Nesta atividade exploramos o conjunto de dados *Periodic Tables of Elements*, composto por diversas informações relacionadas à tabela periódica, como informações dos próprios elementos, nome dos pesquisadores que descobriam o elemento, ano de descoberta, etc.

Exploramos os dados inicialmente utilizando a ferramenta *Orange* (Figura 1). Observamos

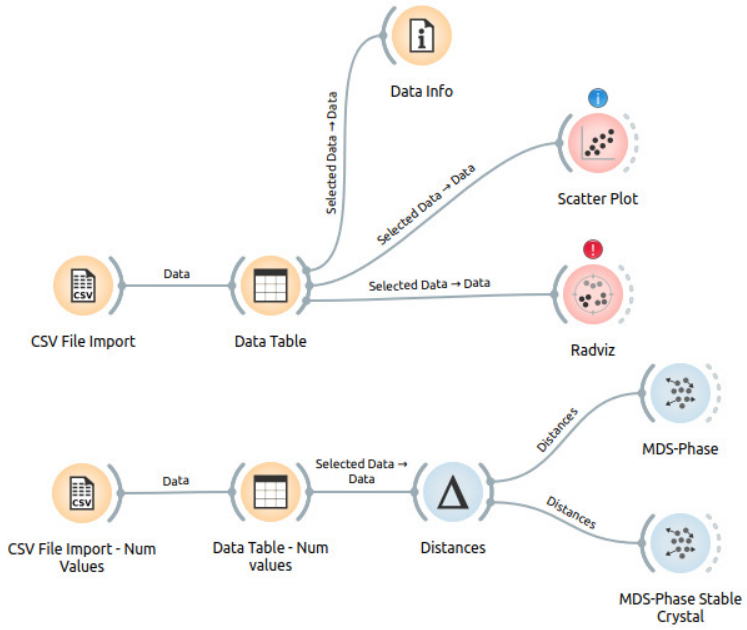

Figura 1: Passos utilizados durante a exploração de dados com o *Orange*.

inicialmente que nossa tabela era composta por 118 linhas correspondendo às informações de cada elemento e 23 colunas, sendo 16 (*features*) inicialmente com valores numéricos (a *feature group* por exemplo era numérica inicialmente, mas sua informação era essencialmente categórica), 3 categóricas e 4 relacionadas a metadados.

2 Analisando os dados visualmente

Durante nossa análise exploratória dos dados buscamos responder às seguintes perguntas:

- Existem elementos que podem ser *outliers* nos dados analisados?
- Qual o relacionamento existente entre as variáveis numéricas do problema, isto é, qual a correlação entre elas?
- Existe algum tipo de separação clara entre os dados, isto é, existe algum tipo de agrupamento no qual possamos identificar características de agrupamento?
- Existe algum relacionamento entre variáveis fornecidas? Por exemplo o ano de descobrimento de um grupo de elementos está relacionado a alguma variável numérica?

2.1 Respondendo às perguntas via análise visual da informação

Buscamos responder às perguntas acima analisando as Figuras 2 a 6.

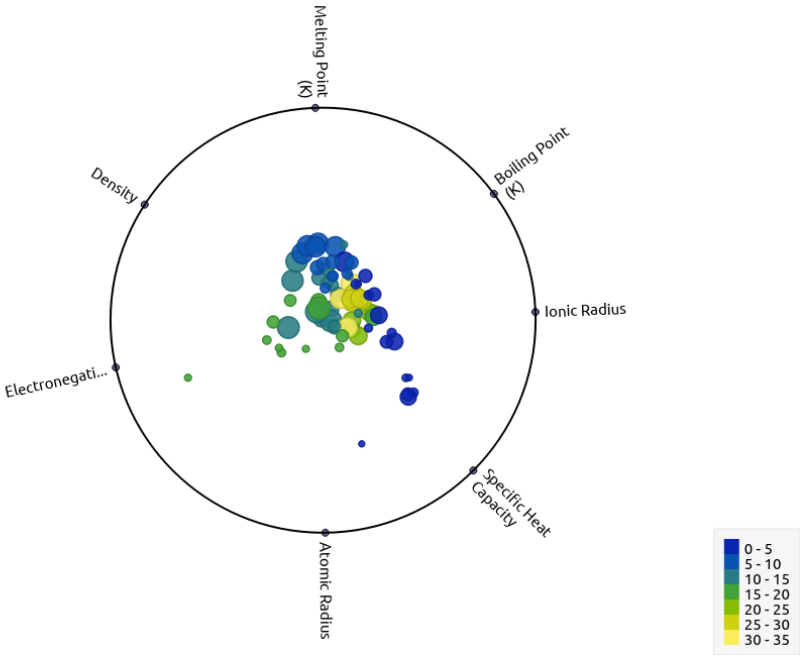

Figura 2: Relacionamento radial entre os elementos da tabela periódica.

Na Figura 2 buscamos identificar com o auxílio do gráfico ***RadViz*** se há um equilíbrio entre os elementos da tabela periódica com relação às variáveis *Density, Eletronegativity, Melting Point, Boiling Point, Ionic Radius, Specific Heat Capacity* e *Atomic Radius* (eixo radial), agrupadas por faixas de valores da coluna *group* (atributo *color*) e a relação com o *atomic weight* (atributo *size*).

Na Figura 3 analisamos o relacionamento entre as variáveis numéricas da tabela periódica por meio de um gráfico de **matriz de dispersão**. Nos eixos x e y temos os valores de cada variável e definimos o parâmetro *color* como sendo os valores encontrados na coluna *Phase*. (Note que renomeamos as colunas para melhorar a visualização do gráfico. Fizemos o seguinte mapeamento: Atomic Weight: AWeight, Ionic Radius: IRadius, Atomic Radius: ARadius, Electronegativity: ENeg, First Ionization Potential: FIP, Density: Dens, Melting Point (K): MP(K), Boiling Point (K): BP(K), Specific Heat Capacity: SHC.)

Na Figura 4 utilizamos o auxílio do gráfico gerado pelo método ***MDS*** para analisar o relacionamento dos elementos da tabela periódica e a possível separação dos mesmos por meio de alguma coluna categórica. Nos eixos x e y temos as variáveis retornadas pelo método *MDS*, exibimos os valores da coluna *Phase* por meio do atributo *color* e os respectivos *atomic weight* pelo atributo *size*.

Na Figura 5 temos a mesma análise realizada via Figura 4 , porém aqui também analisamos os valores da coluna *Most Stable Crystal* via atributo *label*.

Na Figura 6 analisamos o relacionamento das colunas *Atomic Weight, Atomic Radius, Ionic Radius, Eletronegativity, Density, Year of Discovery* e *Specific Heat Capacity* por meio do gráfico de **coordenadas paralelas**, onde cada coordenada é dada pelas colunas analisadas.

Analisamos também o resultado sobre o mesmo gráfico mas agora com foco em valores baixos de *Atomic Weight* (7).

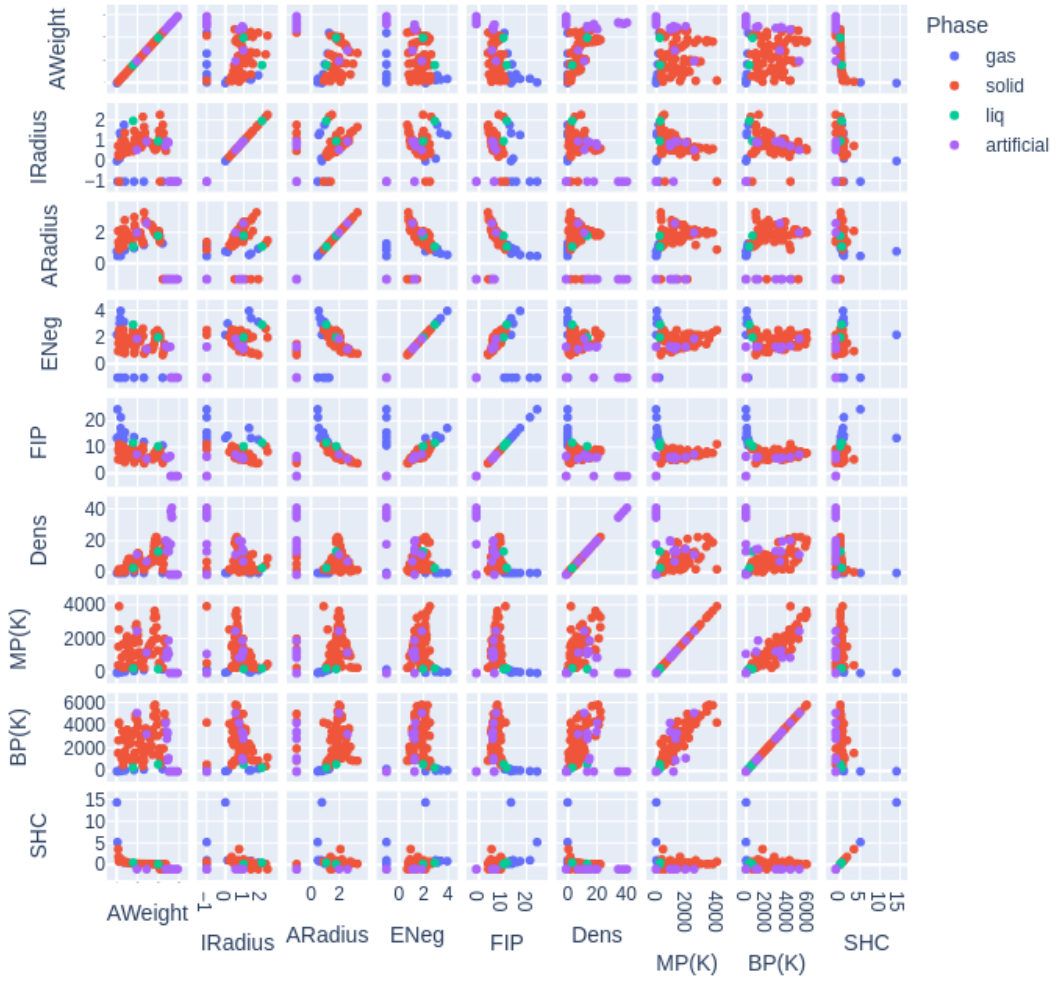

Figura 3: Correlação entre os valores numéricos fornecidos.

2.2 Resultados obtidos

- Analisando o gráfico da Figura 2 é possível notar que não há uma variável que se sobressaia sobre outra de modo extremo, de maneira que todos os elementos ficam concentrados na região mais ao centro, exceto dois valores que necessitam uma investigação mais profunda, dado que se caracterizam como pontos isolados: *Nitrogen* e *Hydrogen*. Também notamos que a coluna *group* não traz uma distinção clara entre os elementos.

- Na Figura 3 podemos notar alguns relacionamentos interessantes entre as variáveis numéricas do problema. Por exemplo, em termos de distinguir os dados por meio das classes presentes na coluna *Phase* podemos utilizar a relação entre as *features Atomic Weight (AWeight), Melting Point (MP)* e *Boiling Point (BP)*. Podemos notar também que a correlação da variável *Atomic Radius (ARadius)* é considerável quando em conjunto com as colunas *First Ionization Potential (FIP)* e *Electronegativity (ENeg)*.

- A Figura 4 exemplifica bem a separação dos dados por meio das classes em *Phase*. Também é possível notar que o *Atomic Weight* dos elementos da classe *artificial* são bem próximos uns dos outros. A Figura 5 assegura que a melhor separação dos dados é dada pelas classes em *Phase*, dado que os elementos ficam muito dispersos quando classificados

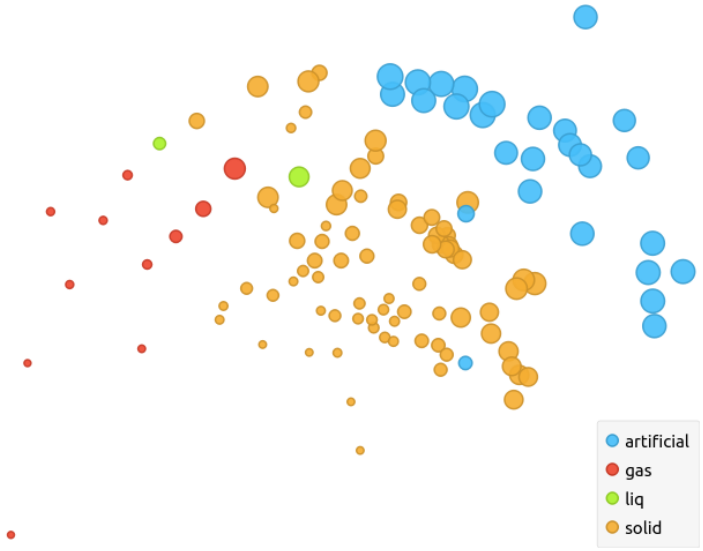

Figura 4: Análise do relacionamento entre os elementos via redução de dimensionalidade.

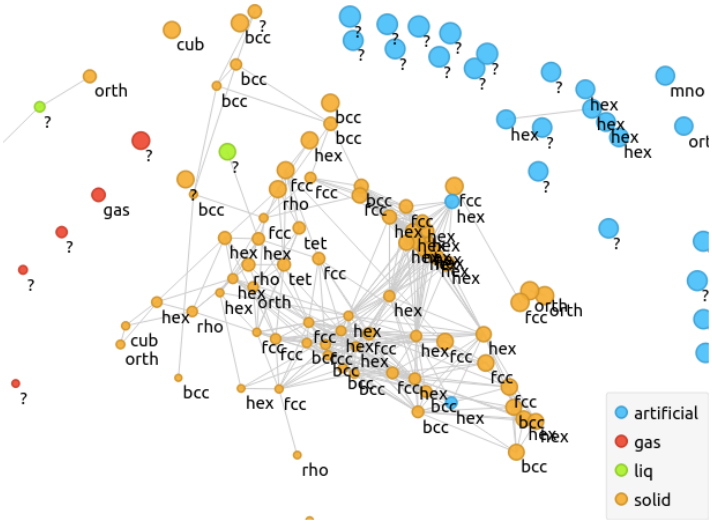

Figura 5: Análise do relacionamento entre os elementos via redução de dimensionalidade.

de outra maneira. Ainda analisando a Figura 5 é possível notar que possuímos muitos valores de *Most Stable Crystal* não mapeados - algo esperado, mas que vale atenção.

- Por fim, nas Figuras 6 e 7 exibimos o comportamento das variáveis via coordenadas paralelas. É possível notar pela Figura 7 que valores baixos de *Atomic Weight* possuem uma *Eletronegativity* relativamente alta. Também é possível notar que elementos com valores baixos de *Atomic Weight* foram descobertos logo nos primeiros anos de pesquisa.

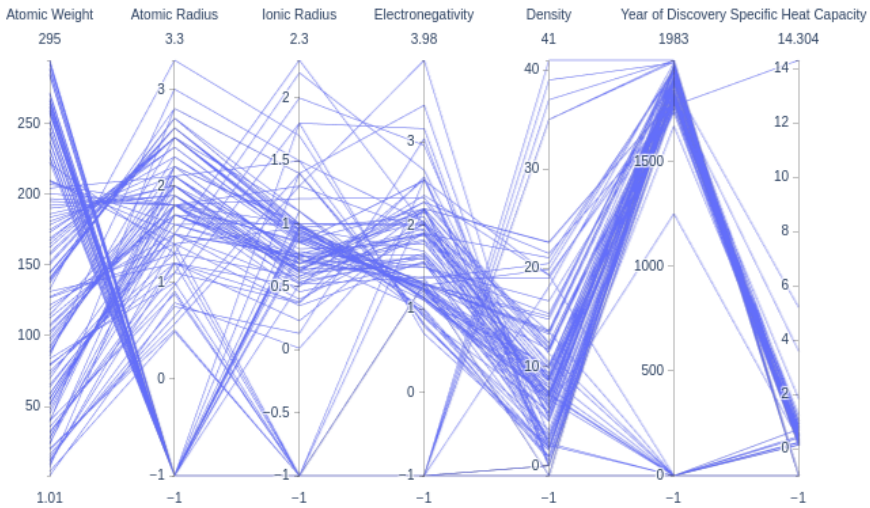

Figura 6: Relacionamento das variáveis via coordenadas paralelas.

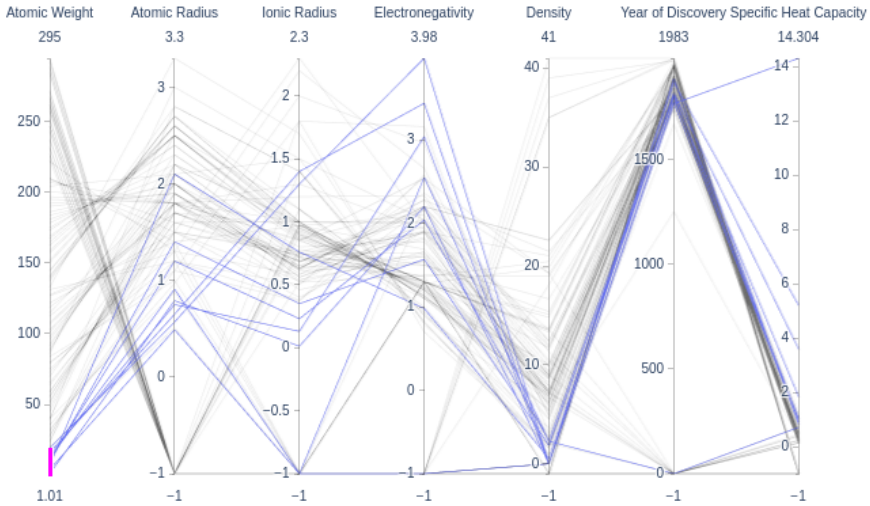

Figura 7: Relacionamento das variáveis via coordenadas paralelas.