Trabalho 2 - INF-0614

Kaleb Roncatti de Souza e Nelson Gomes Brasil Júnior

10 de outubro de 2022

1 Introdução

Seguindo a mesma linha do trabalho anterior, continuaremos usufruindo de dados para gerarmos visões a partir do conjunto bruto, abordando todo o aspecto de Visualização da Informação.

No Trabalho vigente, utilizaremos um `dataset` relacionado ao ramo da Química/Física, através dos elementos da tabela periódica e suas características respectivas, realizando-se uma análise gráfica e apresentando:

- Projeções Multidimensionais
- Coordenadas Paralelas
- Gráficos de Dispersão

Logo a seguir, analisaremos os gráficos obtidos, trazendo informações importantes que podem ser extraídas das respectivas visões.

Para a realização de tal projeto, utilizamos o software `Orange Data Mining` e `Python` aliado à bibliotecas tais como `pandas, matplotlib, seaborn` para a criação das visões.

2 Tarefas

2.1 Exploração inicial

Realizamos a exploração inicial do conjunto e notamos que se trata de uma base de dados dos elementos químicos, contendo 118 elementos entradas e 23 atributos. Além disso, esta base parece ser um conjunto desatualizado dos elementos, dado que os últimos 6 elementos ainda estão com os nomes provisórios `Ununtrium`, `Ununquadium`, `Ununpentium`, `Ununhexium`, `Ununseptium`, `Ununoctium` que já não são mais utilizados pela IUPAC[1]. Fizemos ainda uma análise exploratória sobre o número de elementos nulos por coluna para avaliar em quais análises futuras poderíamos excluir os elementos nulos. Das 23 colunas, 14 apresentavam elementos nulos e mostramos abaixo a quantidade destes elementos por coluna:

[1] `https://iupac.org/what-we-do/periodic-table-of-elements/`

- **Most Stable Crystal** 33
- **Type** 3
- **Ionic Radius** 28
- **Atomic Radius** 32
- **Electronegativity** 22
- **First Ionization Potential** 16
- **Density** 13
- **Melting Point (K)** 20
- **Boiling Point (K)** 20
- **Isotopes** 15
- **Discoverer** 9
- **Year of Discovery** 20
- **Specific Heat Capacity** 33
- **Electron Configuration** 23

A partir da análise dos dados usando o `pandas`, por exemplo conseguimos separar os atributos numéricos dos categóricos e podemos ter uma visão de quais deles podem ser utilizados para tentarmos entender um agrupamentos dos elementos químicos conhecidos.

2.2 Gráficos e análise

2.2.1 Projeções Multidimensionais

Para a geração de projeções multidimensionais, utilizamos o software `Orange`, gerando uma visão com as configurações mostradas na Figura 1.

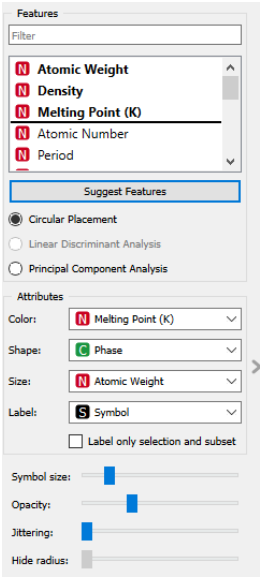

Figura 1: Eixos configurados para a geração da projeção multidimensional através do software `Orange`

.

Apresentamos, logo a seguir, o resultado do gráfico multidimensional gerado pela configuração anterior na Figura 2.

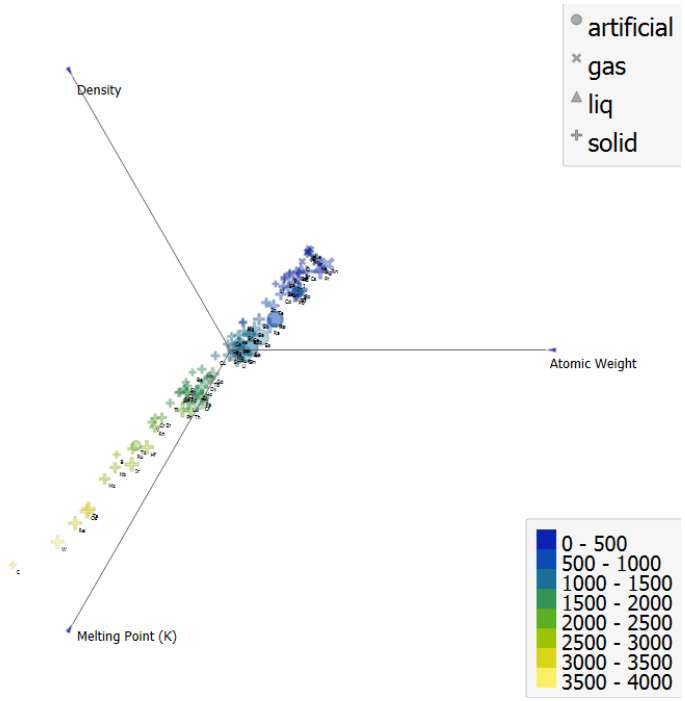

Figura 2: Visão multidimensional de `Atom Weight`, `Density`, `Melting Point (K)` com informações de `phase`.

.

Observe que, num simples gráfico de três eixos, conseguimos extrair informações de 4 diferentes `features` apresentadas no conjunto de dados. Conseguimos observar uma tendência que, olhando da visão em duas dimensões, parece até linear. Observe que, elementos cujo estado é sólido possuem `Melting Point (K)` mais elevados, e elementos cujo estado é gasoso, possuem `Melting Point (K)` mais reduzido.

Com relação à densidade e massa atômica, visualmente falando, não conseguimos extrair muitas informações assim como para o `Melting Point (K)`.

2.2.2 Coordenadas paralelas

Para a geração dos gráficos de coordenadas paralelas, consideramos o conjunto de dados removendo os elementos `NaN` e normalizando as medidas para que a visualização nos gráficos ficasse mais adequada. Para a realização de tais etapas, utilizamos `Python` e, as bibliotecas `pandas, sklearn e matplotlib`. A primeira biblioteca foi utilizada para a realização de um tratamento inicial dos dados, aliado à uma normalização do tipo `MinMaxScaling`. Um primeiro resultado é demonstrado na Figura 3.

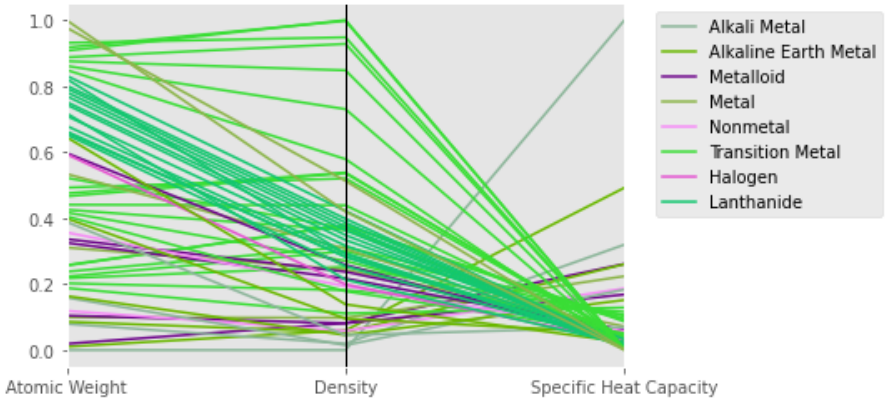

Figura 3: Gráfico de coordenadas paralelas explicitando relações entre `Atomic Weight`, `Density` e `Specific Heat Capacity` em função do `type` de cada um dos elementos químicos.

Através da visão acima, conseguimos extrair múltiplas informações, principalmente quando isolamos cada tipo de elemento da tabela periódica. Observe que, dado que removemos os elementos com `NaN`, praticamente deixamos de lado alguns conjuntos de elementos, tais como os gases nobres. De qualquer forma, para uma visualização fidedigna através do tipo de gráfico escolhido, optamos por deixar tais dados de fora da análise vigente.

- Observamos que, os elementos com **maior** peso atômico (`Atomic Weight`) existentes, pelo menos após a limpeza da base de dados, são os elementos do tipo Metal e Metal de Transição, seguidos dos lantanídeos, os quais também podem ser considerados como metais.
- A seguir, conseguimos observar também que, isoladamente, os elementos com **maior** densidade (`density`) são os metais de transição, também seguido por alguns metais.
- Observando-se o calor específico (`Specific Heat Capacity`), percebemos que, os metais de transição se agrupam nos elementos com **menor** calor específico, em acordo com a característica físico-química de tais elementos, dado que, o significado do calor específico se mostra como a quantidade de calor que um dado elemento precisa absorver para que a temperatura altere. Sabemos que para os metais, pouca aplicação de calor gera mudança de temperatura significativa, dado que tais elementos são conhecidos pela sua característica de alta condução de calor.

Observe que, utilizamos as medidas normalizadas, então obtivemos apenas algumas noções de ordem de grandeza para nossas análises. Numa análise mais detalhada poderíamos utilizar escalas diferentes para cada um dos eixos verticais, ajustando os valores em torno dos valores recorrentes de cada uma das medidas.

2.2.3 Gráficos de Dispersão

Para a análise dos gráficos de dispersão vamos considerar o conjunto de dados inteiro, sem desconsiderar os elementos `NaN` e também não normalizamos os dados. Fizemos isto pois notamos que para as características que vamos analisar não existem muitos elementos nulos e,

ao remover de todos os elementos da base poderíamos não analisar de maneira correta os dados. Utilizamos o pacote `seaborn` e a função `pairplot`[2] para criar a visualização que se encontra na Figura 4.

Escolhemos analisar os seguintes atributos da base dos elementos químicos: *Atomic Radius*, *Period*, *Group* e *Electronegativity*. Decidimos também exibir apenas a diagonal inferior do gráfico de dispersão pois este é um tipo de gráfico simétrico, mas em alguns casos a análise usando o gráfico completo pode ser mais interessante. Como a coloração do gráfico está seguindo o tipo do elemento químico (e são 12 categorias), decidimos manter os gráficos da diagonal na forma `kde - kernel density plot`, que é análogo a um histograma e representa o gráfico usando uma curva de densidade de probabilidade[3]. Algumas análises interessantes que conseguimos extrair à partir destes gráficos.

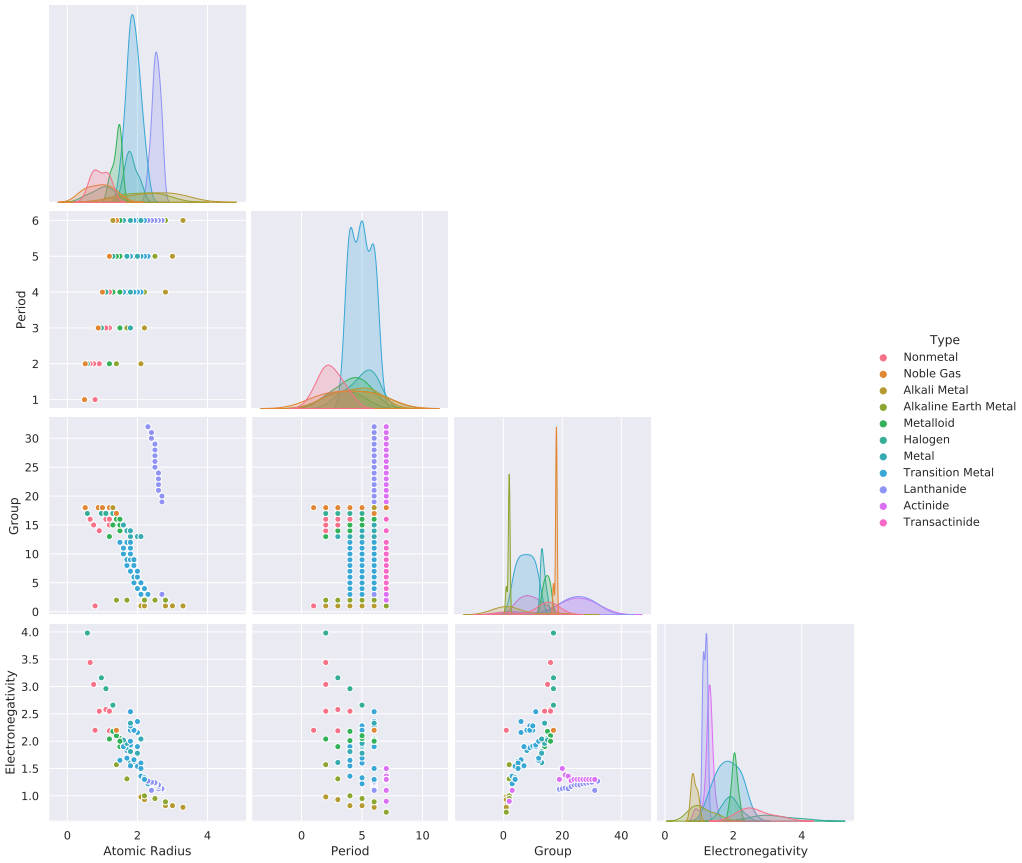

Figura 4: Gráfico Matriz de Dispersão dos elementos químicos tabelados

- Com o gráfico central podemos perceber a distribuição dos elementos químicos com relação ao atributo considerado. Por exemplo, tome o atributo "grupo". Na Tabela Periódica, cada coluna representa um grupo de elementos, por exemplo a coluna 1 é a coluna dos Metais Alcalinos, a 2 os Alcalinos Terrosos, e um dos mais importantes que é o grupo 18 que é o grupo dos Gases Nobres. Pelo gráfico da distribuição dos grupos, percebemos

[2] `https://seaborn.pydata.org/generated/seaborn.pairplot.html`
[3] `https://seaborn.pydata.org/generated/seaborn.kdeplot.html`

que os elementos do tipo `Alkali Metal` se concentram agrupados, como esperado, assim como os `Noble Gas`. Notamos porém que esperávamos um pico único para estes dois grupos, dado o que mencionamos anteriormente sobre os grupos, porém voltaremos a falar sobre isto logo mais. Notamos também que ps elementos do tipo `Transition Metal` estão concentrados nos grupos e nos períodos mais centrais. Ressaltamos também que os `Actinide` `Lanthanide` aparentam ter a mesma distribuição nos grupos;

- O caráter periódico da tabela fica explicitamente evidente quando olhamos para o gráfico *período versus grupo* que mostra exatamente a configuração da tabela periódica que conhecemos das aulas de Química, inclusive com a coloração dos tipos de elemento;

- Um outro aspecto interessante que percebemos ao analisar os dados foi a separação em clusters no gráfico *Atomic Radius versus Group* onde notamos que o tipo `Lanthanide` está bem separado dos outros elementos. O mesmo pode ser dito sobre os tipos `Lanthanide` e `Actinide` que formam um cluster separado dos outros elementos no gráfico *Electronegativity versus Group*;

- Por último, no mesmo gráfico *Electronegativity versus Group* notamos que existe um ponto pertencente à classe dos gases nobres que está próximo a um agrupamento do grupo `Metalloid`. Porém, dado que o atributo `Electronegativity` deveria ser nulo para gases nobres, fomos investigar. Voltamos ao gráfico *Period versus Group* e notamos que o período 6 conteria dois elementos do tipo `Alkali Metal` e que um deles está na coluna referente aos gases nobres. Voltando à base de dados original, olhamos para todos os elementos do período 6 e filtramos os Metais Alcalinos, chegando a duas possibilidades: `Cesium (55)` e `Radon (86)`. Como o elemento `Radon` possui o maior número atômico, pelo gráfico notamos que ele deveria ser o elemento do tipo `Alkali Metal`, o que não procede pois se trata de um gás nobre[4]. Entendemos que possa ser algum erro na base ou na leitura dos dados.

[4]https://en.wikipedia.org/wiki/Radon