ISCTE – Instituto Universitário de Lisboa

Mestrado em Engenharia de Telecomunicações e Informática

NOAA Global Historical Climatology Network Daily (GHCN-D)

Estudo das Diferentes Regiões Climáticas no Hemisfério Norte (com base na Temperatura)

Teresa Maria Garrido Felício – 87276

Projeto Final para a UC de Algoritmos para Big Data

Professor: João Pedro Oliveira

Maio de 2022

Índice

Índice de Figuras

Formulação do Problema e Contextualização

Na sequência da escolha do *dataset* "NOAA Global Historical Climatology Network Daily" foram formulados diversos problemas, em que o principal é o seguinte: quantas regiões climáticas diferentes consegue-se identificar apenas no hemisfério norte (com base na temperatura)? Para responder a esta questão foi utilizado o Modelo de Previsão Clustering. À priori desta seleção de problema, tinha sido testada a possibilidade de previsão do aquecimento global, através da utilização do Modelo de Regressão Linear. No entanto, devido à imprecisão dos resultados obtidos, concluiu-se que não era uma boa abordagem.

Para além da utilização deste tipo de modelos, foram estudadas outras características dos dados através de gráficos que mostram a variação da temperatura máxima, média e mínima em Portugal (ao longo do ano de 2019), a evolução da temperatura média em Portugal (num período de 50 anos), a evolução da temperatura média nos diferentes continentes (ao longo de 21 anos), a correlação entre as variáveis temperatura mínima e precipitação (para os anos de 1959 e 2019) e dois *heatmaps*, que mostram, através de um espetro de cores (quentes/frias), a variação da temperatura máxima ao longo dos meses do ano e consoante um intervalo de anos predefinido.

De forma a encontrar soluções para os problemas acima descritos, foi essencial começar por tratar os dados recolhidos dos diferentes *datasets* disponíveis. No decorrer deste trabalho terão sido utilizados os *datasets*: "ghcnd-stations" e "ghcnd-countries", assim como, o *dataset* que continha todos os dados em estudo, no intervalo de anos [1763-atualidade].

O dataset utilizado neste projeto é bastante extenso, o que não facilita a leitura de dados através da plataforma *cloud* AWS Academy. Esta plataforma decreta um limite de processamento de dados, o que implica, por vezes, a existência de erros no que diz respeito à execução do *notebook*.

Dados Utilizados

A base de dados do nosso projeto é constituída por diversos ficheiros de dados, sendo que cada um corresponde a um certo ano de observações meteorológicas. A primeira coluna de dados diz respeito ao número identificador da estação meteorológica que regista a observação e é composto por 11 caracteres, sendo os dois primeiros relativos ao código FIPS (código universal identificador dos países) do concernente país. De seguida, é apresentada a data referente à observação e o tipo de elemento meteorológico que foi recolhido. Embora existam inúmeros elementos, ao longo do nosso projeto apenas vamos trabalhar com as variáveis relativas à temperatura máxima, mínima, média e nível de precipitação. Na coluna seguinte são apresentados os valores para os diferentes tipos de elementos, sendo que os mesmos são apresentados num

formato de 10 x °C. Por fim, são apresentadas três colunas de *flags* e uma coluna com as horas e os minutos de observação que são irrelevantes para o nosso projeto e, por isso, são desprezadas.

O ficheiro de dados "ghcnd-stations" possui a informação descritiva de cada estação meteorológica, onde conseguimos identificar nove diferentes tipos de variáveis. A primeira coluna corresponde ao ID da estação meteorológica. De seguida são apresentadas as colunas que contêm os valores da latitude, longitude e elevação. Caso a estação esteja localizada nos Estados Unidos ou no Canadá, a mesma irá conter um código postal relativo ao estado onde se encontra. É também apresentado o nome da estação meteorológica e, por fim, três colunas que simbolizam *flags* referentes a redes meteorológicas, mas que serão desprezadas, visto não serem relevantes para o nosso projeto.

Inicialmente, o ficheiro de dados "ghcnd-countries" continha apenas duas variáveis: o código FIPS e o nome do país correspondente. Porém, devido ao objetivo de realizar um estudo da variação de temperatura média entre os diferentes continentes, adicionámos uma nova coluna que associa os países aos seus respetivos continentes.

Modelos de Previsão

Neste relatório vão ser abordados os modelos de previsão: Regressão Linear e Clustering. O primeiro será abordado mais brevemente, devido à ineficácia do seu modelo no conjunto de dados considerado, enquanto o segundo terá uma relevância muito superior, daí ter sido o modelo escolhido para solucionar o problema reformulado em estudo.

- ### Regressão Linear Simples

  A Regressão Linear Simples consiste na previsão do valor de uma variável (dependente), tendo por base uma outra variável (independente), pelo que é utilizada para modelar os dados em função de duas variáveis continuas. No nosso caso, o propósito seria a construção de um modelo em que se prevê a temperatura média, no continente europeu, para o intervalo de tempo [2000-2010] – em anos –, de forma a fazer o estudo da relação das variáveis temperatura média e ano. Dessa forma, a variável dependente utilizada foi a temperatura média, ao passo que a variável independente escolhida foi o ano. Através da sua observação e do valor baixíssimo medido através de uma função de eficiência, concluiu-se que não seria um bom modelo, tendo sido reformulado um novo problema, assim como uma nova técnica de previsão para o resolver.

- ### Clustering (K-Means)

  O Clustering do tipo K-Means – algoritmo de aprendizagem não-supervisionada – consiste em encontrar e agrupar conjuntos de dados com características semelhantes, consoante um determinado valor k, que irá corresponder ao número de clusters diferenciados.

No nosso caso, como pretendíamos realizar o estudo das diferentes regiões climáticas, no hemisfério norte, é de esperar diferentes grupos de clusters agrupados consoante o seu valor de temperatura média.

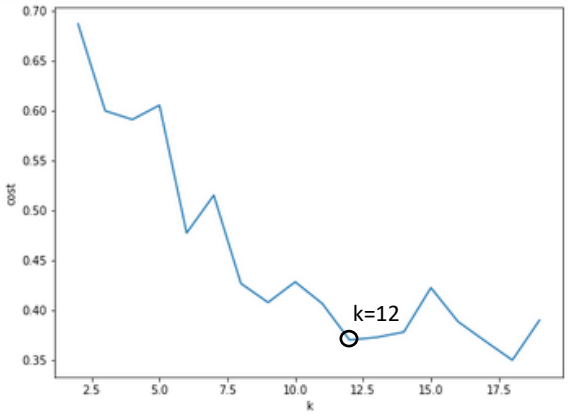

Figura 1 - Gráfico do k Ótimo

Na Figura 1 é possível visualizar o gráfico que foi analisado, de forma a extrair o valor de k – número de clusters. Esse valor foi escolhido de modo a obtermos o menor custo possível, com o menor número de agregações possíveis.

Após ter sido devidamente identificado o valor de k, os dados de cada estação meteorológica, do hemisfério norte, foram agrupados em k=12 clusters, com base na temperatura média registada ao longo do ano de 2021. Os resultados obtidos podem ser observados no mapa da Figura 2.

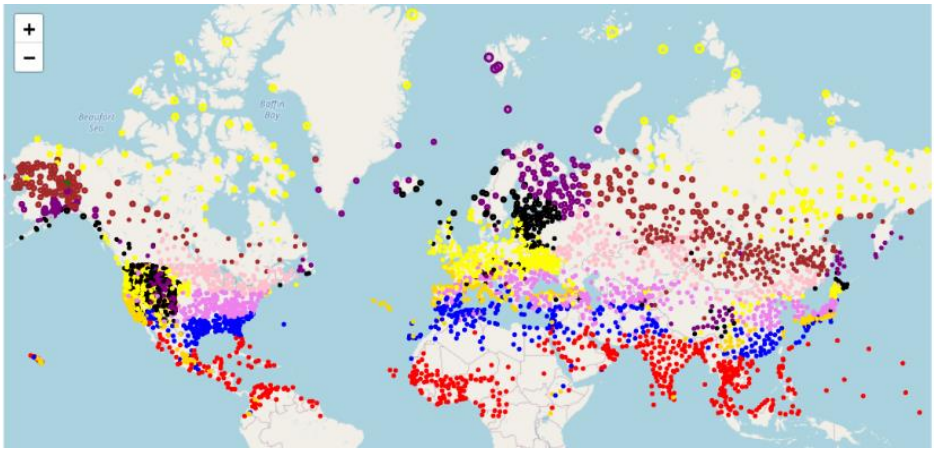

Figura 2 – Mapa de Clustering [Retirada do Notebook "World_map"]

A solução obtida, assim como a sua respetiva análise, mostra a zona do hemisfério norte dividida em 12 regiões – é possível visualizá-lo através do leque diverso de cores atribuído a cada cluster. Para além disso, também é possível observarmos que na horizontal (eixo da longitude) são visíveis linhas de clusters com a mesma cor – isto é, com o mesmo tipo de clima, baseado somente na temperatura média –, enquanto na vertical (eixo da latitude) os clusters têm sempre cores diferentes, mostrando, portanto, que estão em climas diferentes. A análise deste padrão está de acordo com o que é conhecido da história geográfica do planeta Terra, uma vez que os trópicos foram definidos consoante as zonas climáticas, que há semelhança do que foi descrito anteriormente, variavam na vertical e eram constantes na horizontal.

Análise Complementar

De forma a ser explorado e colocado em prática os conhecimentos adquiridos nesta UC, foram realizadas análises complementares aos dados contidos no *dataset* escolhido.

- ### Variação da Temperatura em Portugal

Na Figura 3 é possível a visualização das diferentes estações do ano através da interpretação da variação na curva no gráfico, ou seja, através dos máximos e mínimos da curva conseguimos identificar, sensivelmente, o intervalo de meses [Novembro-Março] a que correspondem estações como o Outono/Inverno devido às baixas temperaturas, enquanto os restantes meses dizem respeito a estações como a Primavera/Verão devido às altas temperaturas. Este gráfico compreende três curvas, avaliando a variação da temperatura máxima, média e mínima, respetivamente, ao longo do ano de 2019.

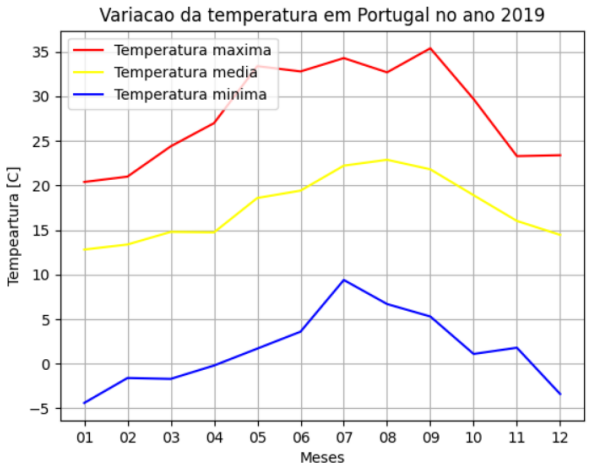

Figura 3 - Retirada do Notebook "Seasons"

- ### Evolução da Temperatura Média em Portugal

A evolução da temperatura média em Portugal é observada na Figura 4, dizendo respeito ao intervalo de anos [1970-2020]. No seguimento dessa análise conseguimos concluir que a temperatura em Portugal é, em média, aproximadamente, 17.068°C (Figura 5).

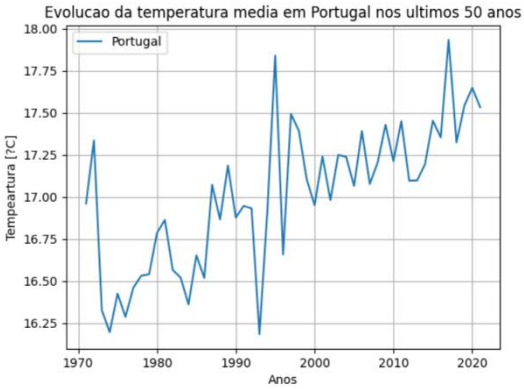

Figura 5 - Retirada do Notebook "Temperatura_media"

PAIS	avg(DATA VALUE)
Portugal	17.06844705666268

Figura 4 - Média da Temperatura em Portugal

- ## Evolução da Temperatura Média por Continente

Na Figura 6 está ilustrada a evolução da temperatura média, por continente, entre os anos [2000-2021]. Através da análise do gráfico, conseguimos concluir que, nos últimos 21 anos, o continente que registava uma maior temperatura, em média, seria África, ao passo que, de forma inversa, a Antárctica seria aquele com uma menor temperatura, em média. Adicionalmente, foi calculada a temperatura média no planeta Terra que corresponde a, aproximadamente, 9.6166°C (Figura 7).

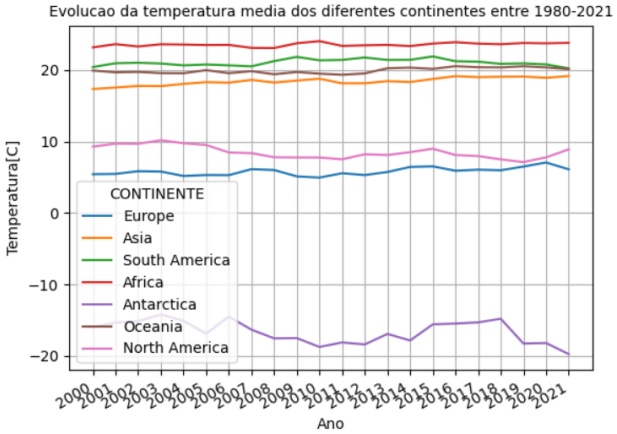

Figura 7 - Retirada do Notebook "Previsao_temperatura"

```
+------------------+
|  avg(DATA VALUE)|
+------------------+
|9.616642640843063|
+------------------+
```

Figura 6 – Média da temperatura no Planeta Terra

- ## Correlação entre Temperatura/Precipitação

A correlação entre as variáveis temperatura mínima e nível de precipitação consegue ser observada na Figura 8, para os anos de 1959 e 2019, respetivamente, em Portugal. Através da análise dos gráficos verificámos que no ano de 1959 o nível de precipitação era mais elevado, comparativamente ao ano de 2019. Para além disso, também foi possível apurar que antigamente (1959) era predominante chover quando as temperaturas mínimas registavam valores mais baixos [5-15]°C, do que em anos mais recentes (2019) [10-18]°C, algo que nos permite confirmar os fenómenos da seca e do aquecimento global.

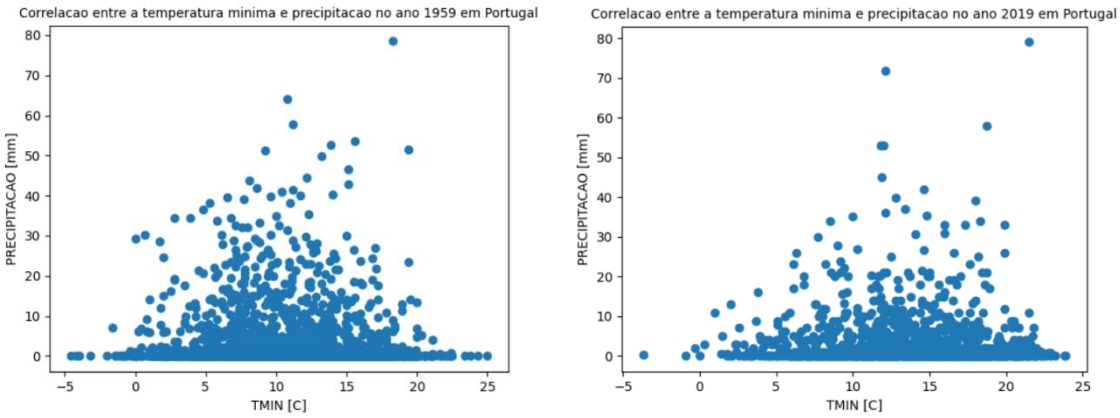

Figura 8 - Retirada do Notebook "Correlacao"

- ## HeatMap

Os *heatmaps* representam uma técnica de visualização de dados que se adapta perfeitamente aos dados do problema formulado, uma vez que a observação de padrões permite comparar os valores de temperatura em diferentes cenários. A Figura 9 transmite a variação da temperatura máxima, para cada mês, entre os anos [1980-2021], em Portugal.

Podemos observar, através da escala de temperatura, que existe um considerável aumento de temperatura máxima, principalmente, nos meses de Julho e Agosto, alusivos aos últimos 5 anos, e um aumento de temperatura relativamente aos meses mais frios, ao longo dos últimos 41 anos. Também é possível visualizar, mais uma vez, as diferentes estações do ano através dos diferentes grupos de cores. Por exemplo: os tons de laranja dizem respeito a estações do ano mais quentes, como o Verão, os tons de azul normalmente dizem respeito a estações mais frias, como o Inverno, enquanto todas as outras (tons de amarelo e verde) costumam estar associadas a estações como o Outono e a Primavera.

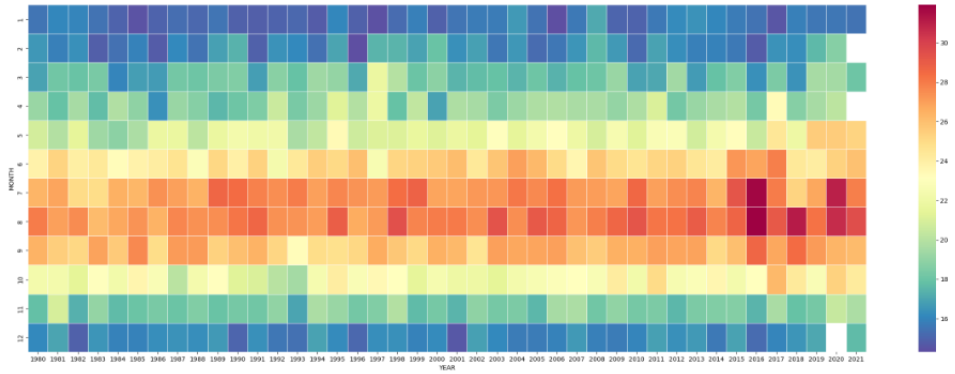

Figura 9 – Heatmap da Temperatura Máxima em função dos Meses e dos Anos, em Portugal (Retirada do Notebook "Heatmap")

Na Figura 10 é avaliada a variação da temperatura máxima, entre os anos [2000-2020], na Alemanha – país escolhido devido à elevada gama de valores de temperatura, sendo possível agrupar esses valores por cores, em diferentes categorias – *freezing cold*, *very cold*, *cold*, *normal*, *warm*, *hot* e *very hot*. De facto, para cada ano, é verificado o desenho de uma parábola invertida, ilustrando a subida crescente da temperatura de [Janeiro-Junho] e respetivo decréscimo da mesma de [Julho-Dezembro].

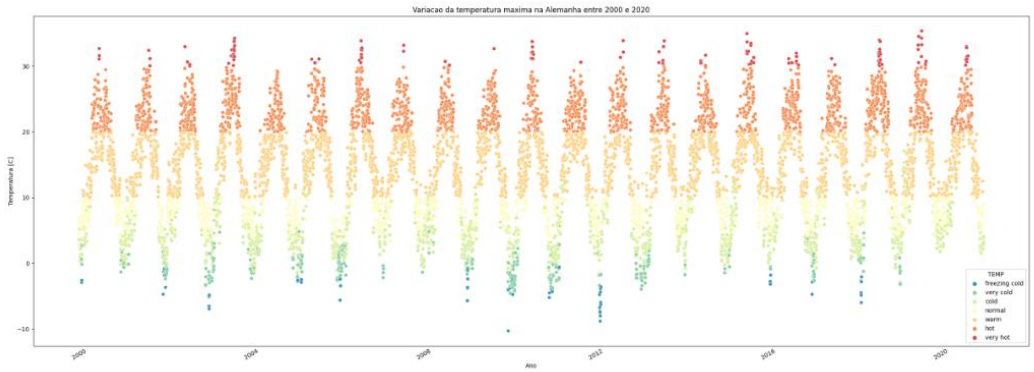

Figura 10 - Retirada do Notebook "Heatmap"

Referências

[1] https://www.relataly.com/pyspark-distributed-computing-tutorial-analyzing-zurich-weather-data-with-python/2739/ - “PySpark Weather Analytics”

[2] https://www.kaggle.com/code/liviucristianterebes/climate-clustering-based-on-temperatures - “Climate clustering based on temperatures”

[3] https://www.kaggle.com/code/hafsaezzahraouy/classification-clustering-with-pyspark/notebook - “Classification & Clustering with pyspark”

[4] https://risk-engineering.org/notebook/data-analysis-weather.html - “Analyzing weather data”

[5] https://github.com/dimajix/weather-analysis/blob/master/Weather%20Analysis.ipynb

- “Following the Climate Change”