Projeto Aplicado a Ciência de Dados I

# El Niño

Grupo 11

2021/2022

# Introdução

O presente trabalho desenvolvido no âmbito da UC Projeto Aplicado a Ciência de Dados I, cujos docentes são Diana Elisabeta Aldeia Mendes e Sérgio Miguel Carneiro Moro, tem como objetivo realizar um estudo sobre o tema escolhido pelos docentes, com base na metodologia CRISP-DM.

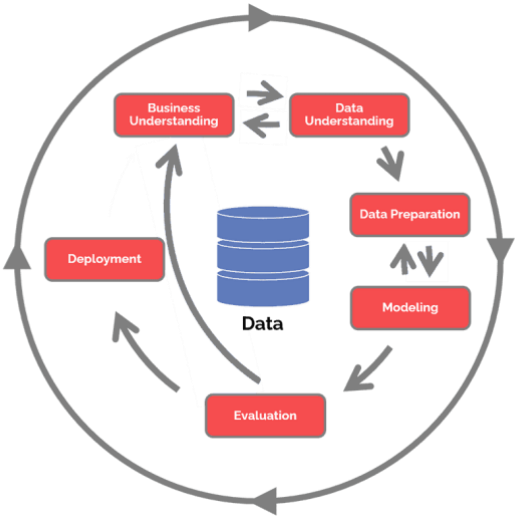



ILUSTRAÇÃO 1 - METODOLOGIA CRISP DM

O estudo desenvolvido ao longo do relatório visa aferir sobre as condições atmosféricas que conduzem ao fenómeno "*El Niño*", com base em informação registada por 70 boias distribuídas ao longo do Pacífico Equatorial, com o intuito de fornecer recomendações à sociedade.

Numa fase inicial, procedeu-se ao *Business Understanding*, mais especificamente, realizou-se alguma pesquisa com o intuito de compreender o fenómeno, razões para o seu acontecimento e consequências do mesmo. Seguidamente, procedeu-se em simultâneo ao *Data Understanding* e *ao Data Cleaning* realizadas através de informação obtida do repositório do conjunto de dados, da visualização gráfico e através de informação obtida na fase anterior.

Relativamente à fase de *Modeling*, optou-se por proceder tanto a métodos supervisionados, bem como a métodos não supervisionados. Primeiramente, realizou-se os métodos não supervisionados, isto é, o *Clustering*, pois o conjunto de dados não apresenta uma variável caracterizadora do *El Niño*. Consequentemente, realizou-se o agrupamento das condições que apresentem semelhança, todavia este método não permitiu uma interpretação clara sobre o objetivo de estudo. Em sequência, uma vez que o principal fenómeno caracterizador do *El Niño* é a subida da temperatura da superfície do mar, efetuou-se um modelo que prevê a subida da temperatura da superfície do mar através de informações geográficas e oceânicas.

Por fim, através dos conhecimentos obtidos no decorrer do trabalho sucedeu-se uma lista de recomendações feitas à população na possível presença do El Niño.

# Índice

# Business Understanding

## O que é o El Niño?

O El Nino é um termo usado para descrever o período em que a temperatura da água na superfície é acima do normal no largo da costa sul-americana ao longo do Pacífico Equatorial.

Durante um período normal, os ventos alísios da costa oeste da América do Sul provocam a deslocação das águas mais frias e profundas para a superfície. Estas águas trazem alimentos ricos em nutrientes para os peixes, mantendo-os a um nível onde os pescadores os conseguem apanhar. Contrariamente, durante períodos onde este fenómeno atmosférico-oceânico se sucede, a intensidade dos ventos alísios da costa oeste da América do Sul diminui, resultando na não subida das águas frias e ricas em nutrientes, provocando o aumento da temperatura da água na superfície e a descida dos peixes para zonas mais profundas onde existem alimentos ricos em nutrientes.

O aumento da temperatura da superfície do mar afeta o clima de diversas regiões do mundo, mais especificamente, as águas muito quentes no oceano pacífico equatorial "bombeiam" mais humidade para o ar, causando o aumento de chuvas, trovoadas e tempestades tropicais numa área maior.

Em países como a Austrália, Indonésia, Brasil, Índia e diversos países africanos pode ser possível experienciar condições de seca porque tempestades ricas em humidade são afastadas dessas áreas. Por outro lado, em países como a Argentina, a zona sul da China e Japão podem receber um aumento de tempestades húmidas, causando períodos de cheias e chuvas fortes. Adicionalmente, há uma diminuição de tempestades tropicais (furacões) no golfo do México e no oceano atlântico ocidental e um aumento de tempestades tropicais no pacifico.

# Data Understanding & Cleaning

Antes de se proceder à limpeza dos dados tornou-se necessário verificar a informação disponível pelo repositório "*UCI Machine Learning Repository* ", a partir do qual se realizou o *download* do *dataset*.

A informação disponibilizada permitiu verificar que os dados recolhidos consistem na informação recolhida por 70 boias distribuídas ao longo da zona do Pacífico Equatorial. Os registos recolhidos medem dados oceânicos e geográficos, com o intuito de detetar variações que possam conduzir aos ciclos do El Niño. Cada boia mede a temperatura do ar, a humidade relativa, os ventos e a temperatura da superfície do mar.

***NOTA: Os registos são efetuados sempre na mesma altura do dia.***

Após efetuar-se o *download* do *dataset* a partir do repositório "UCI Machine Learning Repository ", através da informação dada pelo repositório e também pelos comandos describe, info, head e tail obteve-se estatísticas descritivas, o tipo de dados e alguns exemplos de registos.

## Conhecimento sobre as variáveis

Através da análise conclui-se a existência de 12 variáveis e 178080 registos, em que cada variável corresponde às características captadas pelas boias e cada linha corresponde ao registo efetuado no momento.

1. "obs" - Identificação da observação;
2. "year" – O ano em que a observação foi obtida;
3. "month" - O mês em que a observação foi obtida;
4. "day" – O dia em que a observação foi obtida;
5. "date" – Junção das 3 variáveis anteriores, fornecendo a data da observação;
6. "latitude" - a posição latitudinal das boias;
7. "longitude"- Posição longitudinal das boias;

8. "zon_winds" - Velocidade(m/s) dos ventos que circulam na mesma latitude, paralelamente à linha do equador;

Valores compreendidos entre -10 e 10, nos quais os valores negativos apresentam a direção Oeste-Este e os valores positivos a apresentam a direção Este-Oeste

9. "mer_winds"- Velocidades(m/s) dos ventos que circulam paralelamente ao meridiano de greenwhich;

   Valores compreendidos entre -10 e 10, nos quais os valores negativos apresentam a direção Sul – Norte e os valores positivos a apresentam a direção Norte - Sul

10. "humidity" - Valores da humidade relativa do ar (%);
11. "air_temp" - Valores da temperatura do ar (C°);
12. "ss_temp" - Valores da água na superfície do oceano (C°).

Destas variáveis 5 apresentavam ser variáveis numéricas inteiras (obs, year, month, day, date), duas variáveis numéricas float (latitude e longitude) e 5 variáveis categóricas (zon_winds, mer_winds, humidity, air_temp, ss_temp).

## Valores omissos

Foi também notório através da observação dos registos que a notação "." representava os valores omissos no *dataset*. Substituímos então estes valores por *NA's*. Adicionalmente, verificou-se que como consequência da denominação "." para os valores omissos, as 5 variáveis categóricas na realidade tratavam—se de variáveis do tipo float, em seguimento, as mesmas foram transformadas em float.

| | |
|---|---|
| Zon_Winds | 14,13% |
| Mer_Winds | 14,12% |
| Humidity | 36,92 % |
| Air_temp | 10,24% |
| SS_temp | 9,55% |

Dada a existência de valores omissos, procedeu-se à sua análise, representada pela tabela à esquerda. Constatou-se que em termos percentuais, a maioria das colunas com valores omissos apenas apresentava 9%-15% de valores omissos, todavia a variável humidade apresentava cerca 37% de valores omissos, o que simboliza uma quantidade significativa do total de registos.

## Variável Humidade

Frente a este problema procedeu-se a uma pesquisa sobre a influência da variável humidade no El Niño. Verificou-se que apesar de não ser uma variável diretamente relacionada com a ocorrência do El Niño, uma vez que a zona equatorial em si já apresenta valores altos de húmida, considerou-se esta variável intrinsecamente relacionada com a temperatura do ar.

Ao verificar o heatmap apresentado na ilustração 2 verifica-se que existem valores que estão muito concentrados. Desse modo, verificou-se que o desvio padrão era bastante baixo, assim substituiu-se os valores omissos pela mediana.

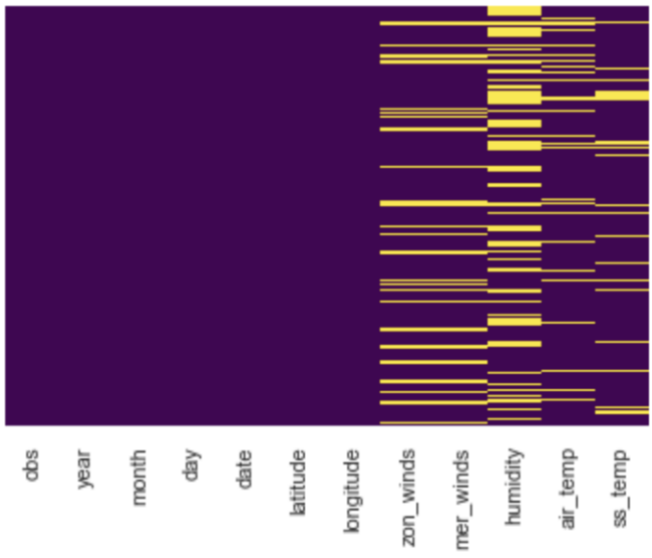



ILUSTRAÇÃO 2 - HEATMAP PARA VISUALIZAÇÃO DA DISTRIBUIÇÃO DE VALORES OMISSOS

### Variáveis com valores omissos

Como os valores omissos se tratam se registos meteorológicos, e também através da verificação do Heatmap apresentado na ilustração 2, optou-se por proceder à imputação dos valores omissos com os registos anteriores, pois verificou-se que maioritariamente, quando se trata de valores meteorológicos e as alterações de um dia para outro não são significativas, tanto se mantém o valor anterior registado ou regista-se em branco o valor.

***Nota: Esta decisão apenas foi possível pois se verificou através do heatmap que a dispersão dos valores omissos era aleatória, caso contrário, a substituição implicaria elevado enviesamento.***

## Outliers

Ao verificar as estatísticas descritivas verificou-se valores extremos e valores que levantaram alguma questão pois outra informação fornecida pelo repositório foi os intervalos de valores que as variáveis tomavam, tal como já foi referido acima.

### Zonal & Meridional Winds

Primeiramente observou-se que através da tabela devolvida pelo comando *describe* os valores referentes às variáveis 'zon_winds' e 'mer_winds' continham extremos fora do intervalo [-10,10]. Através do *boxplot*, verifica-se registos com velocidades acima e abaixo de 10m/s.

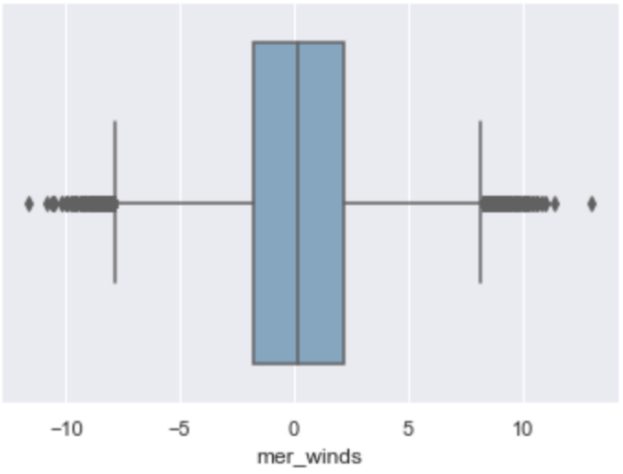



ILUSTRAÇÃO 3 - BOXPLOT MER_WINDS

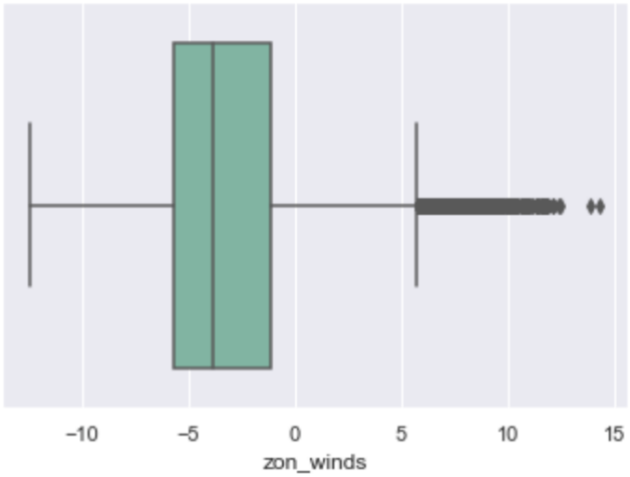



ILUSTRAÇÃO 4 - BOXPLOT ZON_WINDS

Verificou-se um nº total de 179 registos com velocidade acima e abaixo do intervalo, devido ao número elevado de registos do conjunto de dados, removeu-se esses registos.

## Humidity

De seguida foi analisada a variável humidade e verificou-se também valores significativamente baixos, sendo que no repositório encontra-se indicado que os níveis da humidade na zona tropical do Pacífico estão tipicamente entre os 70% e os 90%.

Através do boxplot abaixo, verifica-se que os valores se encontram maioritariamente entre 80%, sendo os valores 70% e 90% considerado extremos, ou seja, 60% claramente consiste num outlier, logo foram eliminados estes 11 registos extremos.

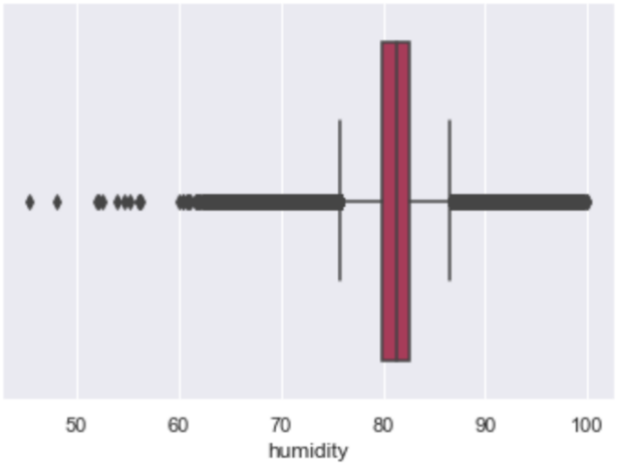



ILUSTRAÇÃO 5 - BOXPLOT HUMIDITY

# Feature Engineering

## Definição do Hemisfério e da Estação do ano

Sabe-se que as bóias que recolhem a informação estão localizadas nas extremidades da linha do Equador e que o El Niño decorre maioritariamente em Dezembro. Uma vez que as estações do ano diferem consoante a posição no Hemisfério Norte ou Sul, optou-se por definir a estação do ano consoante o Hemisfério em que se encontrava a boia.

Condições realizadas:

Hemisfério Norte
- Verão
  - Mês ∈ [7,9]
- Primavera
  - Mês ∈ [4,6]
- Inverno
  - Mês ∈ [1,3]
- Outono

Hemisfério Sul
- Verão
  - Mês ∈ [1,3]
- Primavera
  - Mês ∈ [10,12]
- Inverno
  - Mês ∈ [7,9]
- Outono
  - Mês ∈ [4,6]

## Definição da Direção dos Ventos

Outro fator que, após pesquisa online para melhor compreender o fenómeno, se considerou importante foi a velocidade dos ventos para o aumento da temperatura da superfície do mar, pois ficou entendido que um dos fatores causadores do El Niño era a diminuição da velocidade dos ventos vindos de este para oeste. Adicionalmente, tal como foi dito anteriormente, a

velocidade dos ventos zonais e meridionais apresentavam valores negativos para representar a direção dos ventos, assim com o intuito de obter uma interpretação mais clara e correta, a definição das duas novas variáveis que determinam a direção dos ventos zonais, ou seja, Oeste e Este e os ventos meridionais, ou seja, Norte e Sul, possibilitou a alteração dos valores negativos para valores positivos através do módulo do valor.

## Visualização das Variáveis e Dados

### Dispersão

Após a remoção dos outliers foram feitos histogramas de diversas variáveis de maneira a verificar a dispersão dos mesmos.

Após observação dos resultados concluiu-se que há uma concentração de valores no intervalo [80-85] na variável humidade e tanto a variável air temp (ilustração 7), como a ss_temp apresentam maioritariamente valores entre [26-28], A variável mer_winds (ilustração 6) também apresenta valores maioritariamente baixos, enquanto a variável zon_winds apresenta uma

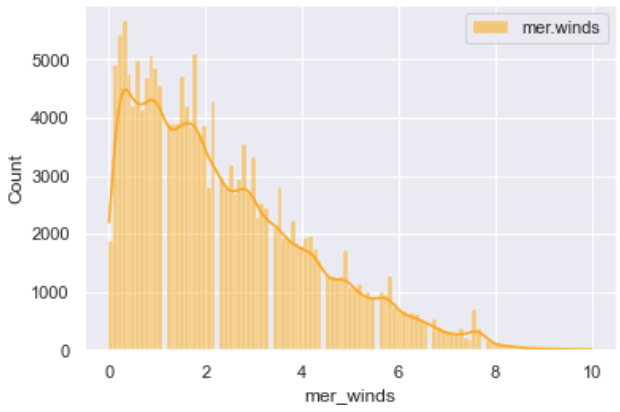



ILUSTRAÇÃO 6 - HISTOGRAMA MER_WINDS

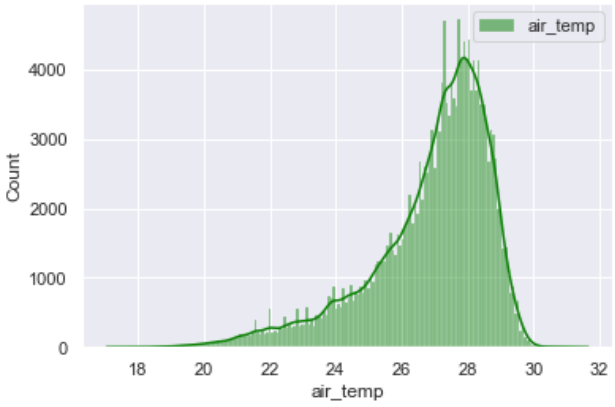



ILUSTRAÇÃO 7 - HISTOGRAMA AIR_TEMP

## Correlação

Com o objetivo de compreender a relação entre as variáveis procedeu-se à visualização da matriz (ilustração 8) de correlação, pois além de ser um bom indicador da relação entre duas variáveis permite também verificar o comportamento perante variação de uma das variáveis, no entanto este indicador não implica a causalidades.

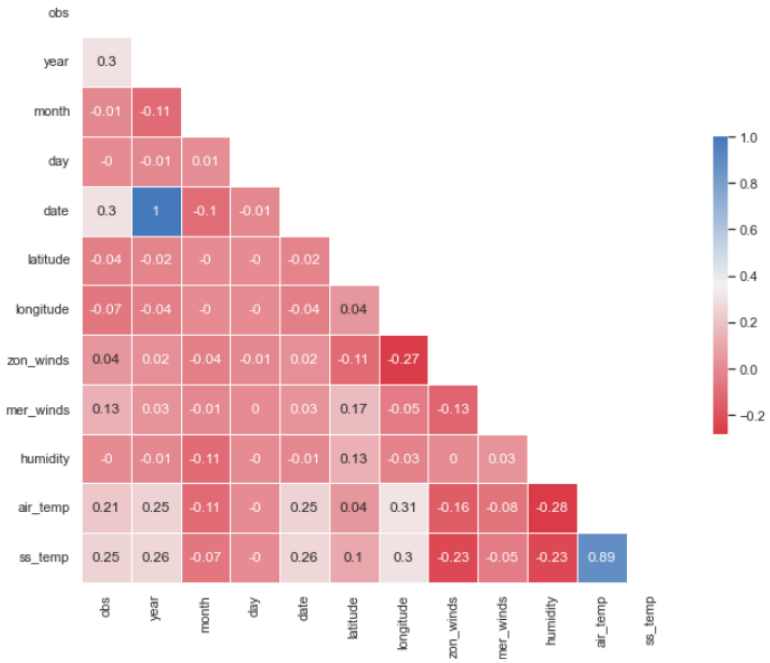



ILUSTRAÇÃO 8 - MATRIZ DE CORRELAÇÃO

Verifica-se que o valor em destaque é a correlação entre a variável ss_temp e air_temp (0.89), do ponto de vista lógico faz sentido pois sabe-se que a temperatura atmosférica tem grande influência na temperatura do mar. De seguida a longitude mostra também alguma correlação, apesar de fraca, com as variáveis ss_temp e air_temp (0.3). Seria de esperar que a latitude fosse a variável com maior correlação pois a proximidade ao equador é um fator relevante na temperatura, o posicionamento relativamente ao hemisfério não tem influência na temperatura.

## Relação

Através de um pairplot, veio-se a confirmar a conclusão realizada pela matriz de correlação, ou seja, as variáveis não apresentação relação entre si, exceto a relação forte e linear entre a temperatura do ar e temperatura da superfície mar.

# Modeling

## Aprendizagem Não Supervisionada

Inicialmente o objetivo do estudo incidiu no agrupamento das condições atmosféricas semelhantes de modo a proceder a recomendações à população na presença do fenómeno El Niño, isto porque não existe um variável que determine se o El Niño está de facto a acontecer.

Consequentemente, procedeu-se à aprendizagem não supervisionada, em que numa primeira fase será feita a redução dimensional, se aplicável, e seguidamente, através do algoritmo K-Means, serão determinados e caracterizados os clusters.

### Adequabilidade da redução dimensional

Primeiramente verificou-se se existia necessidade de efetuar redução dimensional, através do PCA ou do Factor Analysis, porém o PCA era o método ao qual o grupo apresentava maior conhecimento.

Para verificar a necessidade de aplicação do PCA ao conjunto de dados é necessário verificar:

1. **Correlação entre variáveis**

Tal como já foi referido as variáveis do conjunto de dados apresentam maioritariamente correlações baixas, exceto entre a temperatura do ar e a temperatura da superfície do mar, como se pode verificar na ilustração 8.

Temeu-se a não correlação entre as variáveis, apresentando um entrave à aplicação do PCA, pois este tem como principal objetivo a condensação da informação contida em várias variáveis em um conjunto menor, perdendo o mínimo de informação possível, para isso torna-se necessário que as variáveis apresentem aspetos comuns.

### *2.* Teste de Bartett

O presente teste consiste num teste de hipótese na qual a hipótese nula traduz a não correlação das variáveis, ou seja, a matriz de correlação é a matriz identidade. O output dado rejeita a hipótese nula, isto porque p-value < 0.05. Assim, as variáveis do conjunto de dados em estudo são correlacionadas. O teste veio a divergir da observação feita acima.

### *3.* KMO

Este teste permite comparar a magnitude do coeficiente de correlação com o coeficiente parcial. A sua interpretação é dada quanto mais perto o índice de encontra de 1, melhor é a significância da variável, valor abaixo de 0.50 não são tidos em conta. O output dado pelo teste, confirmou que todas as variáveis eram significativas.

## Aplicação do PCA

Dado os resultados dos testes acima descritos, verificou-se utilidade na aplicação da redução dimensional. Assim, procedeu-se à implementação do PCA nas variáveis, para isso estandardizou-se os dados, pois as variáveis apresentam escalas diferentes como graus, velocidade.

**Nota: Apenas se utilizou variáveis numéricas, pois as variáveis categóricas serviram para caracterizar os clusters.**

### 1º Implementação PCA

Numa primeira fase, definiu-se o número de componentes igual ao número de variáveis em estudo, ou seja, 7 componentes, pois consiste no máximo de componentes possíveis para as variáveis definidas. Seguidamente, com o intuito de verificar o nº adequado de componentes, procedeu-se primeiro à realização do critério de Kaiser que segundo o mesmo apenas componentes com valor próprio superior 1 são mantidas, pois explicam mais a variância do que a componente média.

**Critério de Kaiser** = 2 componentes

```
eigenvalues = pca.explained_variance_
eigenvalues
```

```
array([2.26909321, 1.32831176, 0.99129041, 0.91566955, 0.70826217,
```

ILUSTRAÇÃO 9 - CRITÉRIO DE KAISER

Em segundo lugar visualizou-se o gráfico cotovelo que procede à comparação entre o valor próprio e o nº de componentes, representando assim o total da variância associada a cada componente.

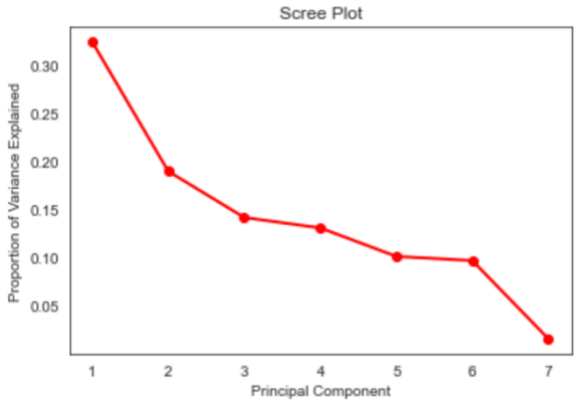



ILUSTRAÇÃO 10 - GRÁFICO COTOVELO

Ambos os testes definiram 2 PCAS, no entanto ao verificar a variância explicativa, optou-se por definir 3 componentes, pois apesar de se perder um grau de liberdade, permite uma melhor explicação das variáveis, cerca de 65% da variância.

```
Proportion of Variance Explained :  [0.32428292 0.18975776 0.14161341 0.1308092  0.10117974 0.09683954
 0.01551743]
Cumulative Prop. Variance Explained:  [0.32428292 0.51404068 0.65565408 0.78646328 0.88764302 0.98448257
 1.        ]
```

ILUSTRAÇÃO 11 - VARIÂNCIA EXPLICATIVA

## K- Means

Definidas as componentes realizou-se um gráfico cotovelo (Ilustração 12) que permite verificar qual o melhor nº de cluster a atribuir para o algoritmo KMeans, ao qual se concluiu que os melhores valores seriam 4 clusters.

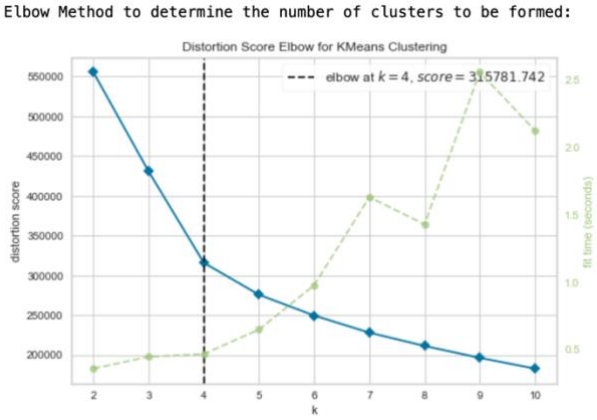



ILUSTRAÇÃO 12 - GRÁFICO COTOVELO

Através da ilustração abaixo que representa os clusters definidos, pode-se concluir que os dados se encontram muito próximos e consequentemente, os clusters também, levando a que estejam também sobrepostos, fazendo com que a sua interpretação não seja clara.

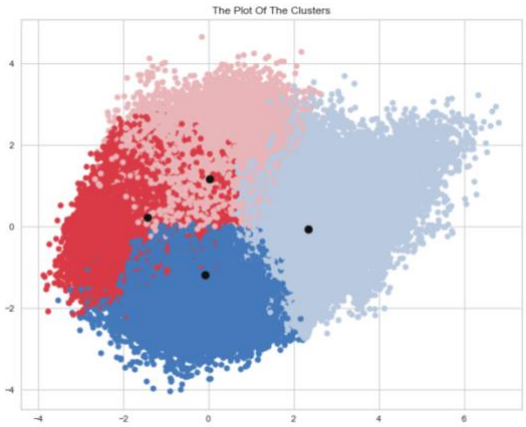



ILUSTRAÇÃO 13 - CLUSTERS FORMADOS PELO ALGORITMO

**Caracterização dos clusters**

Após definição dos clusters, caracterizou-se os mesmos, sendo necessário efetuar agrupamentos e também proceder à visualização de gráficos, dos quais se aferiu as seguintes conclusões:

| | |
|---|---|
| 1º Cluster | Predomínio de temperaturas extremas (28-30) |
| | Zona equatorial Norte |
| | Ventos (Este-Oeste) fracos |
| 2º Cluster | Predomínio de temperaturas elevadas (26-30) |
| | Zona equatorial Oeste Norte |
| | Ventos (Oeste - Este) fracos |
| | Ventos (Norte-Sul) fortes |
| 3º Cluster | Predomínio de temperaturas elevadas (26-30) |
| | Zona equatorial Sul Oeste |
| | Ventos (Oeste - Este) fortes |
| 4º Cluster | Predomínio de temperaturas amenas (26-30) |
| | Zona equatorial Oeste Sul |
| | Ventos (Oeste - Este) fortes |

Apesar de ser possível caracterizar os clusters, a sua interpretação é dificultada pois os valores apresentam-se muito próximos.

# Modeling

## Aprendizagem Supervisionada

Dada a difícil interpretação dos clusters, optou-se por prever o El Niño com base no aumento da temperatura da superfície do mar, dado que é a principal característica do mesmo e indicação do seu começo. Neste caso, a variável target consiste na ss_temp, consequentemente como se trata de uma variável métrica enfrenta-se um problema de regressão. As features correspondem a todas as variáveis presentes no conjunto de dados "Previous.csv", exceto o índice da observação, o dia, o mês, o ano do registo e a data.

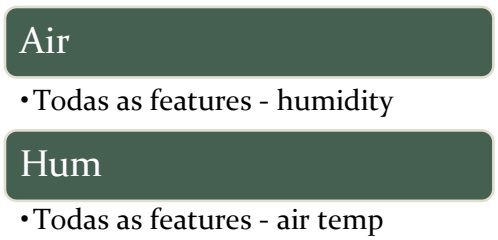
### Data Preparation


**Multicolinariedade**

Antes de proceder à modelação verificou-se a possível existência de multicolinariedade entre as variáveis features, dado que consiste num entrave para a modelação pois não permite uma interpretação clara e real do modelo, devido à dependência entre variáveis.

Verificou-se que as variáveis humidade e temperatura do ar apresentavam uma forte dependência entre si. Deste modo, uma vez que consideramos ambas as variáveis importantes na previsão da temperatura da superfície do mar, pois encontram-se intrinsecamente relacionados devido à condensação. A dependência destas duas variáveis é explicada pelo facto de as boias estarem distribuídas pelo equador, logo trata-se de um o clima essencialmente tropical, dando origem a um tempo quente e húmido.

Optou-se por efetuar duas bases de dados, em que a primeira contém variável temperatura do ar e não a humidade e a seguinte vice-versa.




**Variáveis Dummies**

Uma vez que foram criadas variáveis categóricas a partir das variáveis numéricas, decidiu-se utilizá-las para a modelação, consequentemente, procedeu-se ao processo de criação de variáveis dummy. Neste caso, tanto variável hemisfério como as variáveis ventos longitudinais e ventos meridionais têm 2 dummies, enquanto a variável estação tem 4 dummies.

**Conjuntos de treino e de teste**

Foram definidos dois conjuntos de treino para as features, pois como foi refiro anteriormente estamos presentes variáveis que apresentam multicolinariedade e consequentemente foram criadas duas bases de dados.

Foi definida uma divisão de 70% - 30%, pois existem bastantes registos e por isso optou-se por definir um conjunto de teste maior.

**Nota: Apesar de se estar presente variáveis com escalas diferentes, não se procedeu à estandartização ou normalização dos dados, pois os modelos implementados são a árvore de regressão e a random forest.**

## Regression Tree

**Melhor parâmetro**

Primeiramente utilizou-se a biblioteca GridSearchCV, cujo objetivo é devolver qual o melhor parâmetro para o algoritmo a desenvolver. Neste caso, o parâmetro que se pretendeu verificar foi a profundidade da árvore, com

base na comparação das performances, cujo resultado foi profundidade = 15, tanto para a base de dados air como a hum.

### 1º Modelo – Base de dados Air

Feature Importance

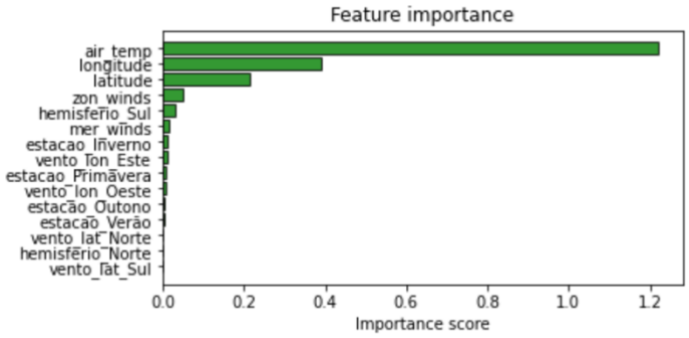



ILUSTRAÇÃO 14 - IMPORTÂNCIA VARIÁVEIS PARA MODELO 1

No primeiro modelo utilizou-se as features definidas na base de dados air e a profundidade recomendada acima. Verificou-se as importâncias de cada variável para a árvore e constatou-se segundo o gráfico acima (ilustração 14) que como esperado a temperatura do ar consiste na variável mais impactante na previsão da temperatura da superfície do mar.

Seguidamente verifica-se um impacto maior da longitude do que a latitude para a previsão da temperatura superfície do mar, observação questionável. Porém ao verificar a dispersão das boias e também da representação gráfica das boias sobre o globo verificou-se que há uma maior concentração de boias Este do pacífico do que a Oeste.

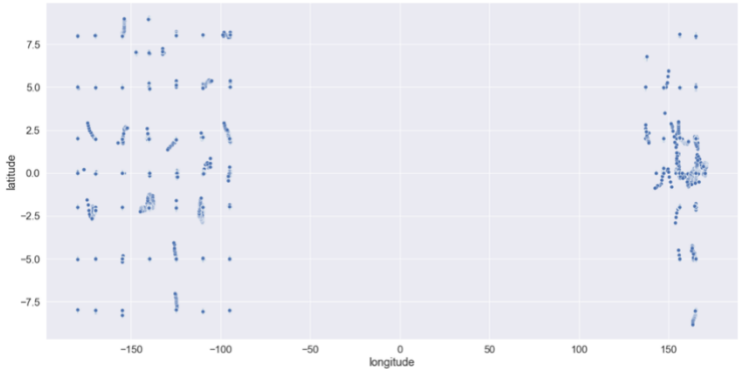



ILUSTRAÇÃO 15 - DISPERSÃO DAS BOIAS

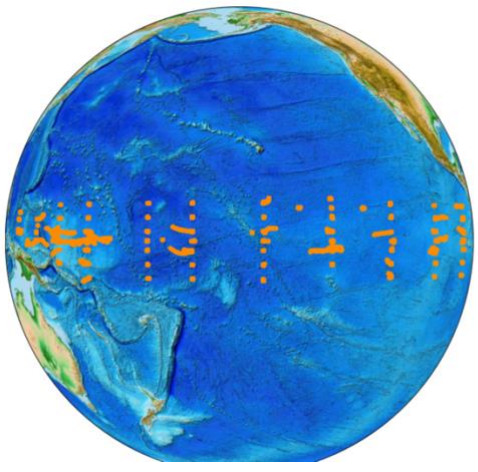


ILUSTRAÇÃO 16 - DISPERSÃO DAS BOIAS SOBRE O GLOBO

É possível verificar pela ilustração 15, na qual a longitude negativa representa as boias localizadas a este do pacífico e a longitude positiva a oeste do pacífico, verifica-se uma maior dispersão de valores a oeste e uma concentração de valores junto ao equador a oeste do pacífico.

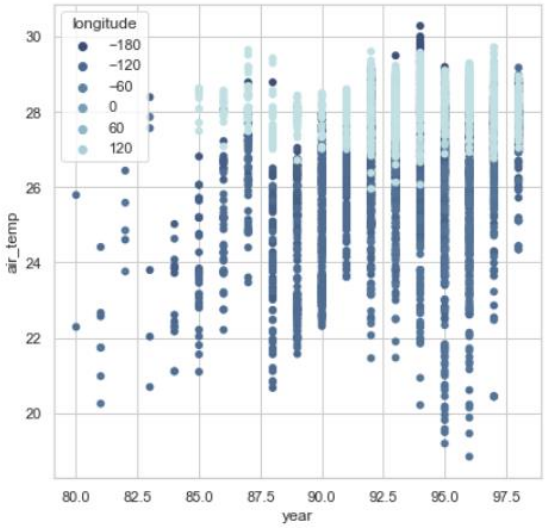



ILUSTRAÇÃO 18 - DISPERSÃO DOS VALORES DA TEMPERATURA DO AR FACE À LONGITUDE AO LONGO DOS ANO

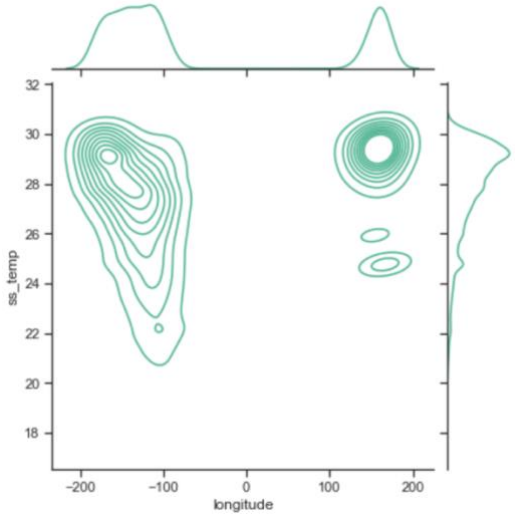



ILUSTRAÇÃO 17 - DISPERSÃO DOS VALORES DA TEMPERATURA DA SUPERFÍCIE DO MAR AO LONGO DA LONGITUDE

Através dos gráficos acima demonstram como varia a temperatura do ar (ilustração 18) e a temperatura da superfície do mar (ilustração 17) em função da longitude. Verifica-se que a Este do Oceano Pacífico há uma maior dispersão de valores, flutuando entre os 20-30 graus. Em contraste a Oeste do Oceano Pacífico verifica-se registo essencialmente mais elevados, entre os 26 e 32 graus.

Apesar de teoricamente a latitude influenciar a temperatura do ar e superfície do mar, mais concretamente, quanto mais próximo do equador tendencialmente a temperatura também é superior. Deve-se ter em conta a distribuição das variáveis descrita pelas ilustrações 13 e 14, pois tal como já foi referido anteriormente há uma maior concentração de boias junto ao equador a Oeste do Oceano Pacífico.

Verifica-se também importância da velocidade dos ventos zonais e da localização do hemisfério sul. Tal como já referido acima, o aumento da temperatura da superfície inicia-se com a redução da velocidade dos ventos alísios, ou seja, os ventos zonais (Este – Oeste), além a temperatura da água no Sul é superior à do norte do equador.

### Performance do Modelo

| Mape = 1,2% | MSE = 0,24 | MAE = 0,33 | $R^2$ = 0,94 |
|---|---|---|---|

Verifica-se acima que 94% da variância da temperatura da superfície do mar é explicada pelos features. Adicionalmente, é possível verificar que o MAPE indica que em média a previsão está incorreta em 1,2% e o MAE indica que a previsão pode estar errada em média 0,33 graus face ao valor correto. Por fim, o MSE indica um valor baixo. Assim conclui-se que o modelo apresenta boa capacidade preditiva.

### Validação

Com o intuito de validar o modelo, procedeu-se à validação cruzada com 10 folds, ao qual se aferiu que o modelo apresenta boa capacidade preditiva, como pode ser constatado pelo output abaixo o $R^2$ mantém-se constantemente elevado.

```
scores for k=10-fold validation: [0.92621859 0.92252423 0.92355577
0.92006385 0.9296921 0.92908547 0.92520451 0.92886684 0.92557643
0.91901895]
Score: 0.92 (+/- 0.01)
```

### Representação Gráfica

Por fim, realizou-se a representação gráfica entre os valores originais e os valores previstos pelo modelo, tal como se verificou pelas métricas acima descritas, não existe muita discrepância.

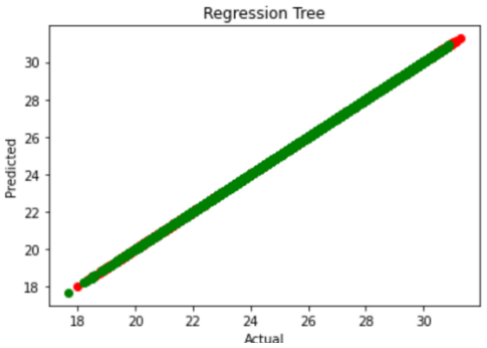



ILUSTRAÇÃO 19 - REPRESENTAÇÃO GRÁFICA VALORES ORIGINAIS E PREDITOS DO MODELO 1

**2º Modelo – Base de dados hum**

Feature Importance

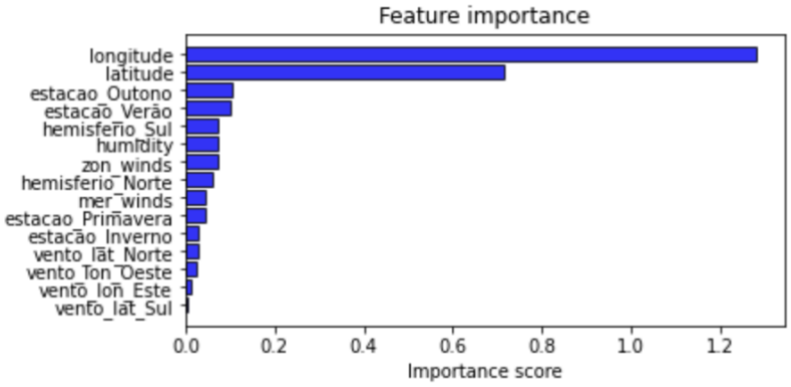



ILUSTRAÇÃO 20 – IMPORTÂNCIA FEATURES PARA MODELO 2

Ao verificar as variáveis importantes para o modelo, verificou-se novamente a importância da longitude e a latitude, justificadas pelas razões acima descritas. Porém contrariamente, ao modelo acima verificou-se a importância das estações do ano para a previsão da subida da temperatura da superfície do mar, justificadas pela alteração do tempo que indiretamente são influenciadoras da subida e descida da temperatura do ar.

Adicionalmente, verifica-se que variável hemisfério Sul e humidade sobrepõem-se à velocidade dos ventos zonais, isto porque se trata de um clima tropical. A humidade de forma indireta está relacionada com a localização geográfica e também com a estação do ano, pois estas representam a subida da temperatura do ar, variável com a qual demonstrou dependência.

Performance

| Mape = 1,8% | MSE = 0,60 | MAE = 0,49 | $R^2$ = 0,86 |
|---|---|---|---|

Através da tabela acima pode-se constatar que 86% da variância da temperatura da superfície do mar é explicada pelos features. Adicionalmente,

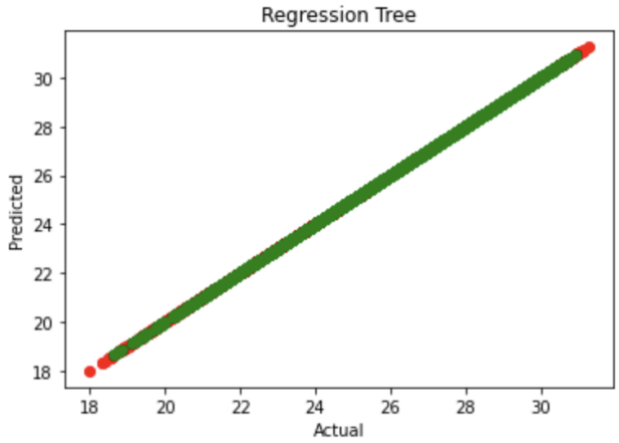
quanto às métricas de erro é possível verificar que o MAPE indica que em média a previsão está incorreta em 1,8% e o MAE indica que a previsão pode estar errada em média 0,60 graus face ao valor correto. Por fim, o MSE indica um valor baixo. Assim conclui-se que o modelo apresenta boa capacidade preditiva, apesar de o modelo anterior apresentar melhor performance, porém, é de salientar que esse fenómeno se deve à influência da temperatura do ar.

### Validação

Com o intuito de validar o modelo, procedeu-se à validação cruzada com 10 folds, ao qual se aferiu que o modelo apresenta boa capacidade preditiva, tal como no modelo anterior.

```
scores for k=10-fold validation: [0.83919099 0.83964986 0.84069141
0.81605249 0.82381519 0.84242932 0.83996997 0.82823661 0.83591649
0 .825284 ]
Score: 0.83 (+/- 0.02)
```

### Visualização




Verifica-se menor capacidade preditiva principalmente face a valores mais baixos, em comparação com a ilustração 19.

ILUSTRAÇÃO 21 - REPRESENTAÇÃO GRÁFICA VALORES ORIGINAIS E PREDITOS DO MODELO 2

## Random Forest

### Melhor Parâmetro

Tal como para a Regression Tree, procedeu-se à biblioteca GridSearchCV com o intuito de verificar o melhor parâmetro para a profundidade para ambas as bases de dados, ao qual se constatou um valor de 25 para ambas.

### 1 º Modelo – Base de Dados Air

Feature Importance

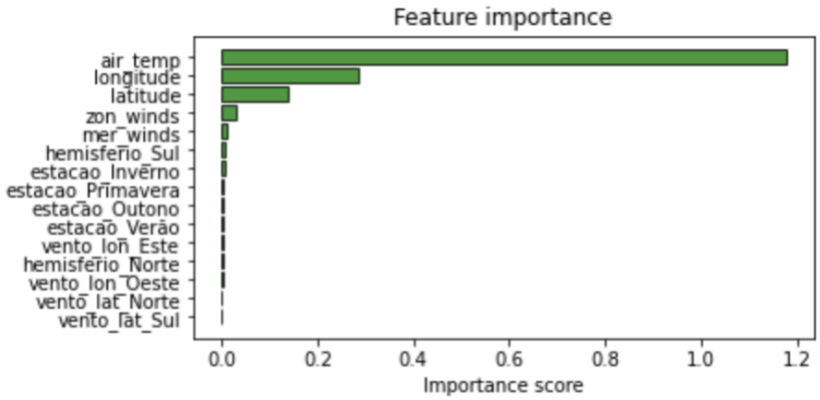



Ilustração 22 - Feature Importance modelo 1

Verificou-se qual a importância das variáveis e tal como as o 1º modelo da regression tree, as variáveis air temp, longitude e latitude apresentam maior importância na previsão da temperatura da superfície do mar. No entanto, constatou-se que a velocidade dos ventos meridionais ultrapassou a importância do hemisfério em que se localiza a boia.

Performance

| Mape = 1% | MSE = 0,27 | MAE = 0,27 | $R^2$ = 0,96 |
|---|---|---|---|

Verifica-se que 96% da variância da temperatura da superfície do mar é explicada pelos features. Além disso, relativamente às métricas de erro, verificam-se valores baixos, mais especificamente, o MAPE determina que em média a previsão está incorreta em 1%, o MAE indica que a previsão pode

estar errada em média 0,27 graus face ao valor correto. Por último, o MSE devolve também um valor reduzido.

### Validação

Contrariamente aos algoritmos anteriores não é necessário proceder à validação cruzada, isto porque o algoritmo em si aleatoriamente seleciona as features em cada ramo, de modo que não haja overfit. No entanto, é possível utilizar o parâmetro "oob_score = True" que estima em amostras aleatórias o $R^2$, quanto mais próximo de 1 melhor. Verificou-se o valor de 0,962 logo o mostra ter uma boa performance.

## 2º Modelo – Base de dados Hum - 20 Profundidade

### Feature Importance

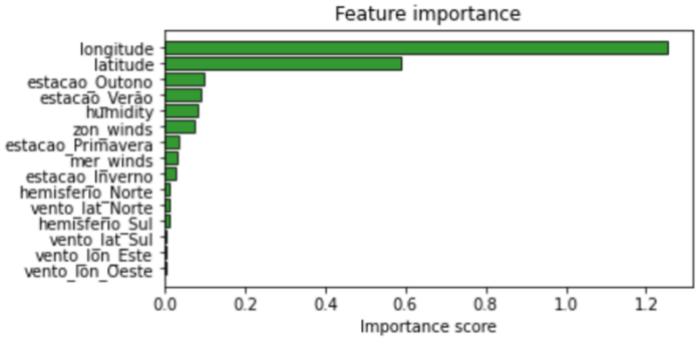



Ilustração 23 - Feature Importante modelo 2

Verifica-se o mesmo padrão de significância do modelo 2 da Tree Regression.

### Performance

| Mape = 1,4% | MSE = 0,36 | MAE = 0,39 | $R^2$ = 0,91 |
|---|---|---|---|

Tal como o segundo modelo desenvolvido através do algoritmo Tree Regression verifica-se uma perda de performance ao remover a variável temperatura do ar, porém a random forest apresentou melhor resultado em comparação com a Tree Regression. O MAPE determina que em média a previsão está incorreta em 1,4%, o MAE indica que a previsão pode estar

errada em média 0,36 graus face ao valor correto. Por último, o MSE devolve também um valor reduzido.

### Validação

O output dado pelo comando obb_score, consistiu no valor 0,91, ou seja, apresenta boa performance.

# Conclusões

Ao longo do processo de trabalho verificou-se que fatores como a velocidades dos ventos e a temperatura do ar são importantes para a determinar a temperatura da superfície do mar.

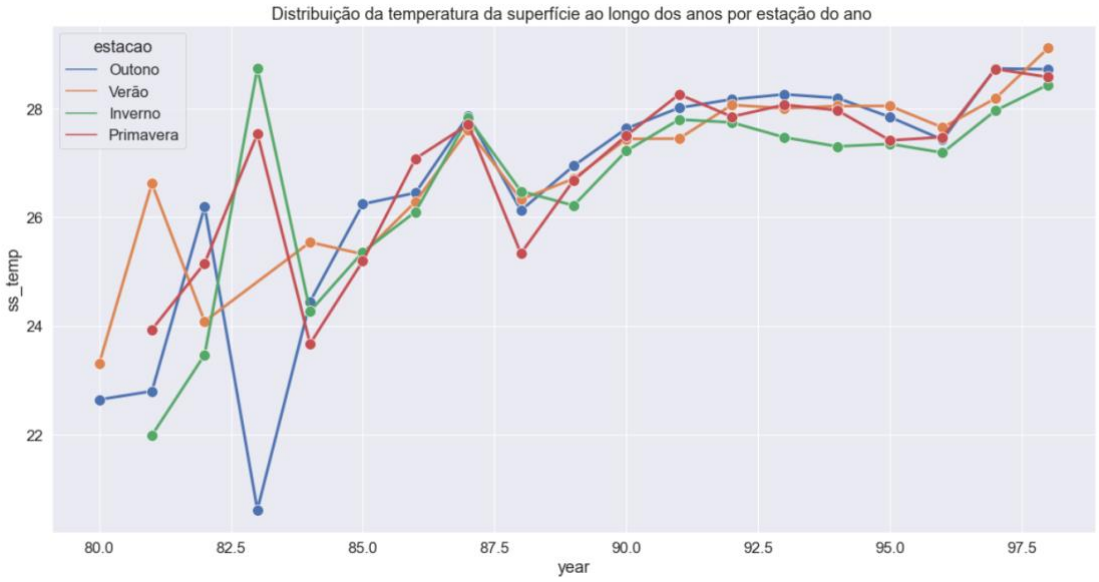



ILUSTRAÇÃO 24 - EVOLUÇÃO TEMPERATURA DA SUPERFÍCIE DO MAR AO LONGO DOS ANO CONSOANTE A ESTAÇÃO DO ANO

O gráfico (ilustração 24) acima permite verificar evolução da temperatura do ar ao longo dos anos consoante a estação do ano. Destaca-se a grande amplitude da variação da temperatura em 1982-1983, período no qual ocorreu o fenómeno do El Nino, seguido do fenómeno El Niña, ambos de grande intensidade. É possível verificar o aumento gradual da temperatura do mar no Inverno e Primavera (época seca no equador) e seguidamente a redução da temperatura no Verão e Outono (época fria no equador).

Adicionalmente, em 1986, observa-se uma subida generalizada da temperatura do mar, no qual se verificou uma nova ocorrência do fenómeno El Niño.

Novamente, em 1990, verifica-se uma tendência crescente da temperatura da superfície do mar, resultando em 1996 numa nova ocorrência do fenómeno com forte intensidade. O gráfico permite verificar o que acontecimento acontece principalmente na época seca.

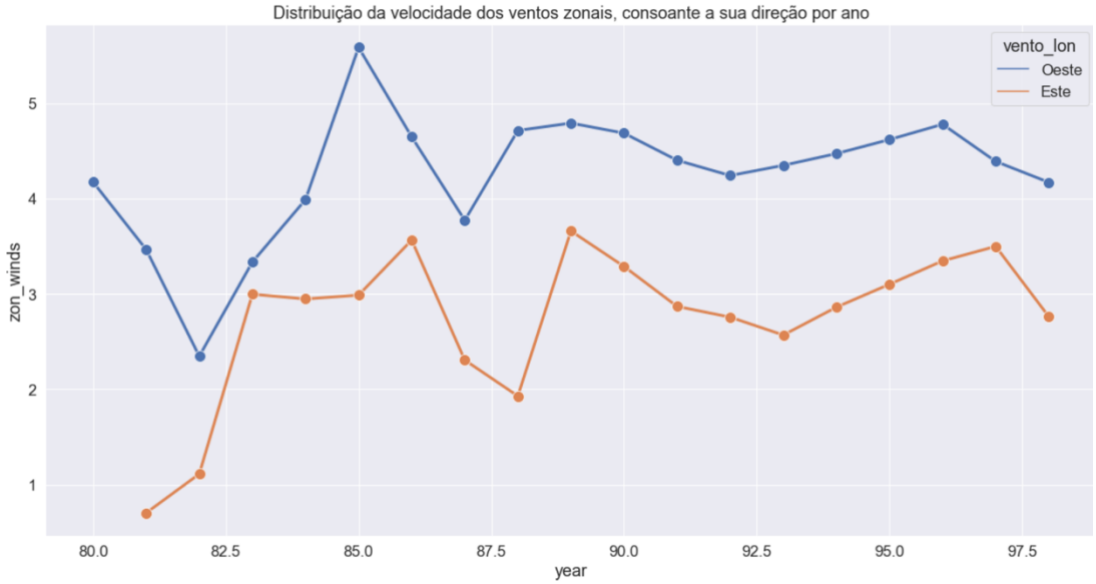



ILUSTRAÇÃO 25 - VARIAÇÃO DOS VENTOS ZONAIS AO LONDO DOS ANOS

O gráfico (ilustração 25) acima permite verificar a influência dos ventos zonais no fenómeno El Niño. Sabe-se que o El Niño ocorreu em 1982-1983, 1986-1987 e 1996-1997, sendo possível verificar uma descida dos ventos zonais em 82 e uma subida em 83, resultando na presença do fenómeno La Niña. Em segundo lugar, novamente em 86 verifica-se a descida da velocidade dos ventos. Por fim, verifica-se novamente uma redução da intensidade dos ventos.

O presente gráfico permite confirmar que a diminuição da intensidade dos ventos é um fenómeno associado ao El Niño.

Por fim, ao longo da modelação e da visualização de gráficos como os acima representados, tornou-se possível constatar que o conjunto de dados apresenta boa capacidade de previsão para a temperatura da superfície do mar, porém esta deve-se principalmente ao aumento da temperatura do ar, pois como se pôde verificar, apesar da variável air temp não ter sido utilizada, o modelo atribuía significância a variáveis que de forma indireta

apresentavam o aumento das temperaturas, como por exemplo o hemisfério Sul, estações do ano quentes, a humidade que consiste numa consequência da presença de um clima tropical e por fim a redução dos ventos que influencia temperaturas mais quentes.

Concluiu-se que a variável temperatura do ar é determinante do aumento da temperatura da superfície do mar.

# Recomendações

O fenómeno El Niño acontece com a subida da temperatura do ar, ou seja, em meses secos. Assim, dá-se a seguinte lista de recomendações aos diferentes setores da sociedade:

Na presença da época quente deve-se:

1.Sociedade no Geral

- Preparar os kits de emergência, visto que são uteis para qualquer tipo de emergência (Lanternas, baterias, dinheiro e suprimentos de primeiros socorros);
- Ter umas dezenas de sacos de areia para, caso haja uma tempestade, seja possível melhor proteger a sua casa;
- Cortar ramos de arvores ou até mesmo as arvores na sua propriedade que estejam em risco de cair. Envie solicitações de cortes de arvore para arvores que não pertençam a sua propriedade, mas que, frente a excesso de vento/tempestades possam danificar estruturas ou linhas de energia;
- Hidratação, principalmente crianças e idosos
- Não sair em hora de maior calor
- Não fazer queimadas

2. Agricultores

- Melhorar/fortificar infraestruturas relacionadas com a agricultura;
- Introdução a variedade de sementes tolerantes a secas, como tomate, abóboras e melancia;

3.Governo

- Limpar as sarjetas para caso haja chuvas torrenciais não as entupir;
- Investir em sistemas de alerta;
- Criar instituições de apoio aos cidadãos afetados pelo fenómeno;

# Bibliografia

[1] Sutton, W. & Srivastava, J. & Koo, J. & Vasileiou, I. & Pradesh, A. (2019, 2 abril). Striking a Balance: Managing El Niño and La Niña in Cambodia's Agriculture. Acedido em https://reliefweb.int/report/cambodia/striking-balance-managing-el-ni-o-and-la-ni-cambodias-agriculture

[2] The World Bank. (2019, 2 dezembro). Striking a Balance: Managing El Niño and La Niña in the East Asia and Pacific Region's Agriculture. Acedido em https://www.worldbank.org/en/topic/agriculture/publication/striking-a-balance-managing-el-nino-and-la-nina-in-the-east-asia-and-pacific-regions-agriculture

[3] NOAA. What is eutrophication? National Ocean Service website, https://oceanservice.noaa.gov/facts/eutrophication.html

[4] Navlani, A. (2019, 12 abril). Introduction to Factor Analysis in Python. Acedido em https://www.datacamp.com/tutorial/introduction-factor-analysis

NOAA. Reports to the Nation On Our Changing Planet: El Niño and Climate Prediction. Acedido em https://www.pmel.noaa.gov/elnino/sites/default/files/atoms/files/el_nino_report.pdf

[5] Hiregoudar, S. (2020, 4 agosto). Ways to Evaluate Regression Models. Acedido em https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70

[6] Bhor, Y. (2021, 31 maio). Yellowbrick : Visualization for model predictions. Acedido em https://www.analyticsvidhya.com/blog/2021/05/yellowbrick-visualization-for-model-predictions/

[7] ENSO Prediction Analysis. Acedido em https://github.com/dvsun/Final_Project

[8] Ujhelyi, T. (2022, 1 fevereiro). Regression Tree in Python Using Scikit-learn (Code Your Decision Tree Part #1). Acedido em https://data36.com/regression-tree-python-scikit-learn/

[9] Chakrabarti, S. (2021, 8 julho). AutoML using Pycaret with a Regression Use-Case. Acedido em https://www.analyticsvidhya.com/blog/2021/07/automl-using-pycaret-with-a-regression-use-case/

[10] Korstanje , J. The k-Nearest Neighbors (kNN) Algorithm in Python. Acedido em https://realpython.com/knn-python/

[11] Dutton Institute. (2015, 9 fevereiro). El Nino: 1997-1998. Acedido em https://www.youtube.com/watch?v=7SPRrCHC1RI

[12] Singh, A. (2018, 22 agosto). A Practical Introduction to K-Nearest Neighbors Algorithm for Regression (with Python code). Acedido em https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

[13] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. Acedido em https://scikit-learn.org/stable/about.html#citing-scikit-learn

[14] Salahat, F. (2022, maio). Credit Card Clustring. Acedido em https://www.kaggle.com/code/fayyadsalahat/credit-card-clustring

[15] Slater, S. (2022, fevereiro). Customer Clustering - K means. Acedido em https://www.kaggle.com/code/stevenslater/customer-clustering-k-means

[16] Naghshin, V. (2021, 2 julho). PCA and How to Interpret it — with Python. Acedido em https://medium.com/analytics-vidhya/pca-and-how-to-interpret-it-with-python-8aa664f7a69a

[17] Zuccarelli, E. (2021, 31 janeiro). Performance Metrics in Machine Learning — Part 3: Clustering. Acedido em

https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-d69550662dc6

[18] codebasics. (2019, 4 fevereiro). Machine Learning Tutorial Python - 13: K Means Clustering Algorithm. Acedido em https://www.youtube.com/watch?v=EItlUEPCIzM