

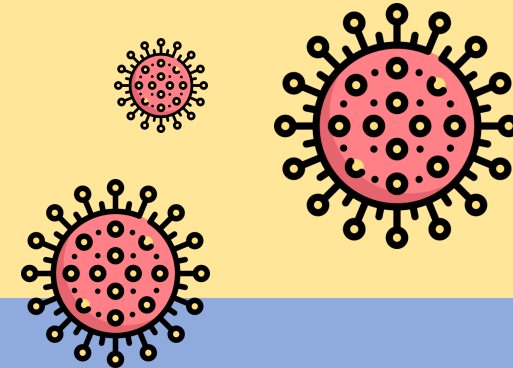


# COVID-19 Community Mobility Reports

Unidade Curricular: Inteligência Artificial

Maria Barros (up201608444) e Miguel Ferreira (up201606158)

# Definição do problema



Antes da definição do problema, é necessário primeiro conhecer o dataset. Para este trabalho, o dataset com que vamos trabalhar possui informação sobre **19 diferentes países**, com informações diárias durante **43 dias**, sobre diferentes informações:



Tendências de mobilidade a locais como supermercados, farmácias...



Tendências de mobilidade a locais como restaurantes, shoppings...



Tendências de mobilidade a locais de trabalho



Tendências de mobilidade a locais como parques, jardins, marinas...



Tendências de mobilidade a transportes públicos



Tendências de mobilidade a locais de residência



Número total de casos

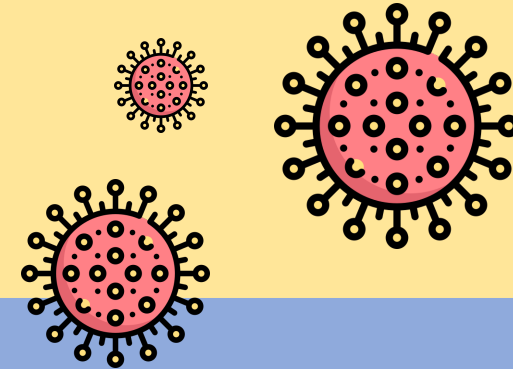


Número total de mortes

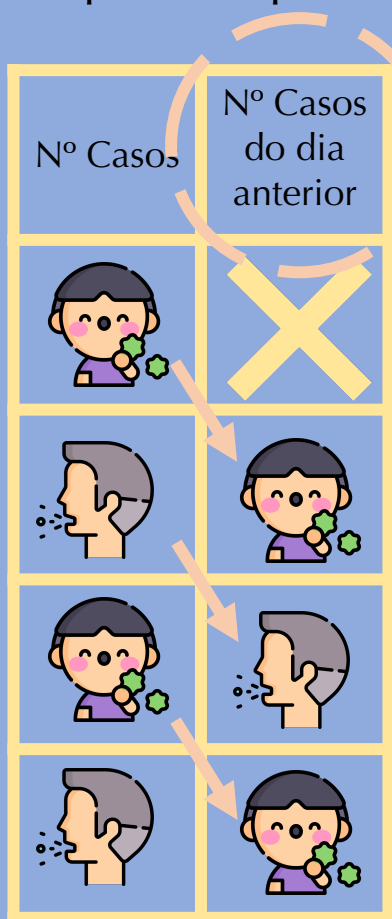


Como primeiro problema, decidimos então focar-nos na **previsão do número total de casos**, utilizando como **features todas as tendências de mobilidade**, e ainda o **número de casos do dia anterior**

# Ferramentas a utilizar



A primeira parte do trabalho consistiu então numa primeira avaliação do dataset

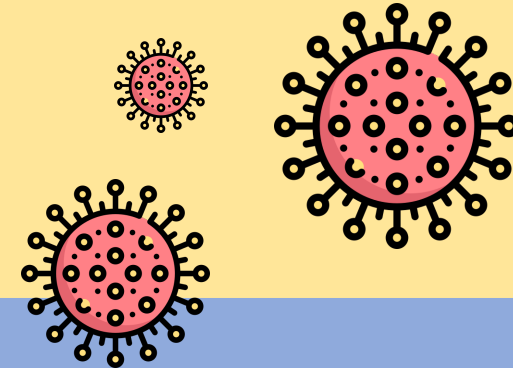


O primeiro passos consistiu então na criação de uma nova coluna, em que cada entrada corresponde ao número de **casos no dia anterior**. Desta vez, para prever o número de casos para cada dia, este valor pode ser utilizado como característica, e ajudar nessa previsão

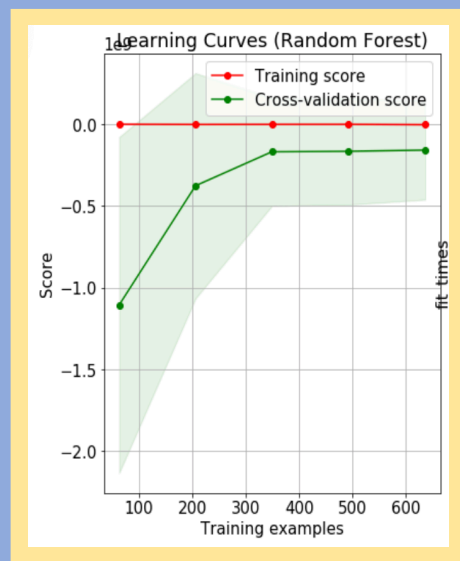


De forma a dificultar a previsão, considerámos, noutra abordagem, não o **número de casos do dia anterior** (dado que facilita bastante a tarefa da previsão), mas sim o número de casos conhecidos **na semana anterior**

# Ferramentas a utilizar



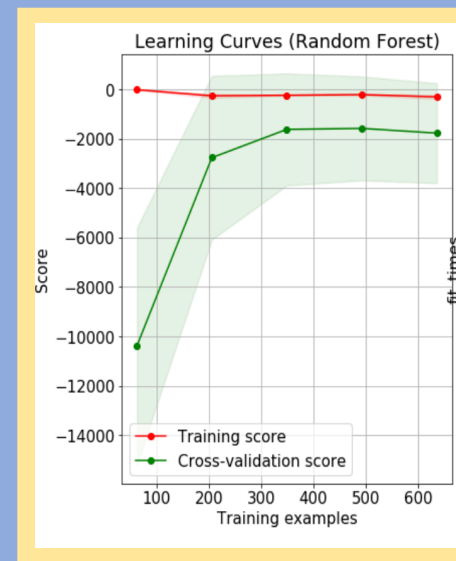
De seguida, e de forma a conseguir averiguar quais os algoritmos que produziriam melhores resultados para o dataset, traçaram-se as **curvas de aprendizagem** para diferentes algoritmos. Os algoritmos testados foram **Regressão Linear, KNN, Naïve Bayes, SGD, SVC, MLP, Decicion Tree e Random Forest**, estando os resultados da linha de aprendizagem apresentados a seguir. Várias métricas foram utilizadas para avaliar os resultados



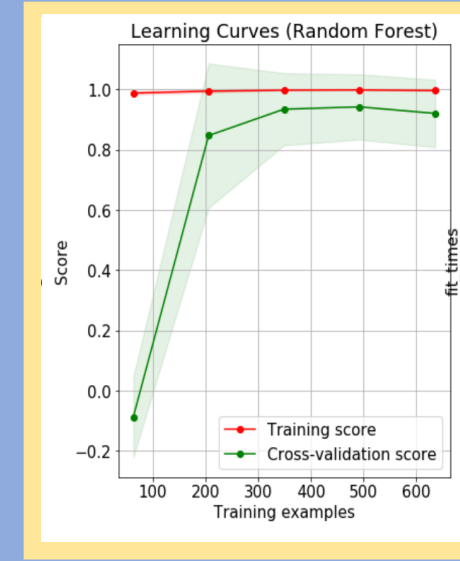
**Negative Mean Squared Error**



**Max Error**

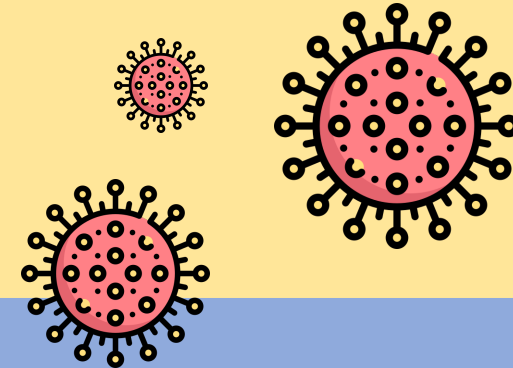


**Negative Mean Absolute Error**



**R<sup>2</sup>**

# Resultados



R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
0,998	0,994	2,1E+06	6E-05	2,6E+04	6,7E-02	2,7E+02	2,7E-03

TREINO

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
0,935	0,903	1,6E+08	2,4E-03	5,4E+04	1,9E-01	1,7E+03	1,2E-02

TESTE

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
0,993	0,983	8,6E+06	2,2E-04	3,6E+04	1,2E-01	8,2E+02	5,3E-03

TREINO

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
-5,274	0,524	4,5E+08	2,1E-02	7,8E+04	4,1E-01	7,2E+03	7,5E-02

TESTE

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
0,948	0,802	6,9E+07	2,4E-03	1,1E+05	3,0E-01	2,0E+03	2,0E-02

TREINO

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
-4,775	0,077	3,9E+08	3,5E-02	8,3E+04	6,3E-01	6,2E+03	9,7E-02

TESTE

PREVISÃO DO NÚMERO DE CASOS A PARTIR DO NÚMERO DE CASOS DO DIA ANTERIOR

Resultados muito satisfatórios tanto no **set de treino** e no **set de teste**

PREVISÃO DO NÚMERO DE CASOS A PARTIR DO NÚMERO DE CASOS DA SEMANA ANTERIOR

Resultados muito satisfatórios tanto no **set de treino** mas muito mais no **set de teste**

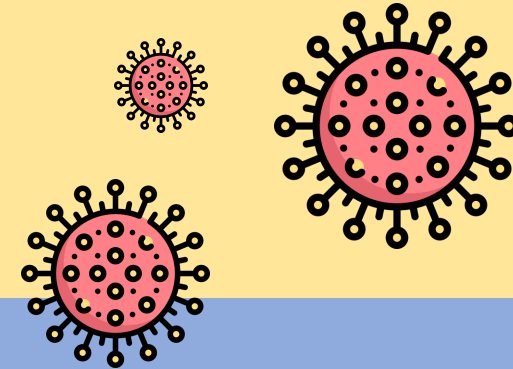


Indica **overfitting**

TUNNING DOS HYPERPARÂMETROS DO MODELO ANTERIOR

Melhoramento dos resultados, mas ainda existe bastante **overfitting**

# Introdução de novas features



Foram adicionadas algumas *features* novas de modo a tentar complementar e facilitar a previsão do número de casos



Contabilização do número de **milhões de habitantes** de cada país (de forma a normalizar número total de casos e de fatalidades)



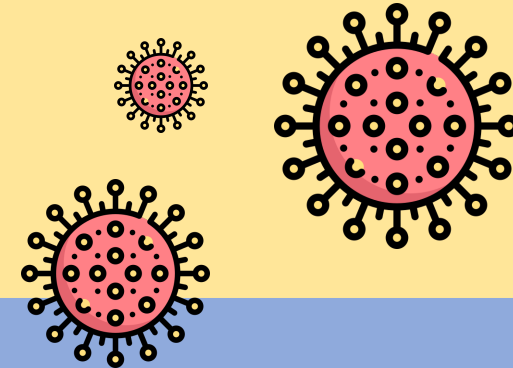
Valor do **Produto Interno Bruto** de cada país



Ranking da **eficiência do sistema de saúde** de cada país

Com o aumento da dimensão do espaço de *features*, foi também utilizado **PCA** e técnicas para determinar a **Feature Importance** de cada *feature*

# Resultados



R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
0,995	0,992	6,7E+02	1,4E-04	1,9E+02	1,0E-01	1,1E+01	4,7E-03

TREINO

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
-121,522	0,303	3,2E+04	2,8E-02	5,1E+02	5,4E-01	9,9E+01	9,3E-02

TESTE

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
0,997	0,995	3,6E+02	8,0E-05	1,7E+02	8,6E-02	6,7E+00	3,5E-03

TREINO

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
-0,552	0,844	9,7E+03	6,4E-03	3,6E+02	2,8E-01	4,2E+01	4,0E-02

TESTE

ADIÇÃO DE NOVAS FEATURES E  
REDUÇÃO DA DIMENSÃO COM PCA

Resultados muito bons para **set de treino**,  
mas piora com o **set de teste**

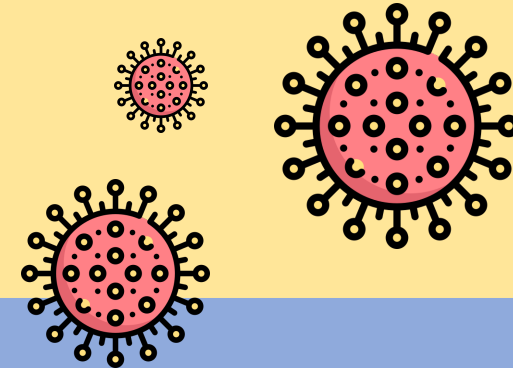
ADIÇÃO DE NOVAS FEATURES E  
ELIMINAÇÃO DAS MENOS IMPORTANTES







Apresenta melhor resultados no **set de teste**  
comparativamente aos resultados com PCA

Tuning de hiperparâmetros  
implementado em todos os splits

Estes resultados mostram que **remover *features* pouco importantes melhora significativamente** os resultados

# Adição de mais features



Nº Casos	Nº Casos 8 dias antes	Nº Casos 9 dias antes	Nº Casos 10 dias antes	(...)
				
				
				
GAP 7 DIAS				
				

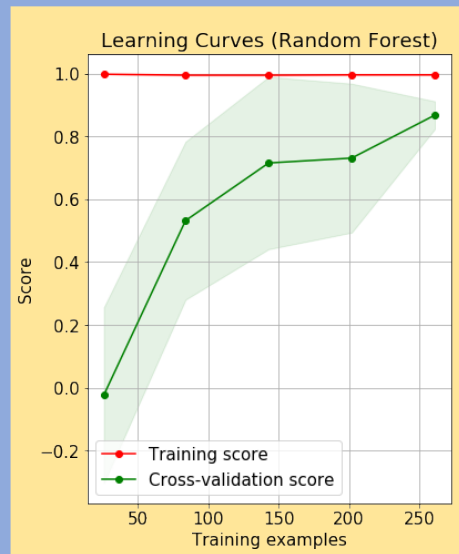
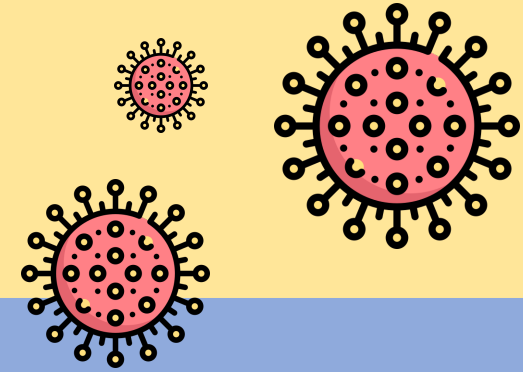
Foram adicionadas várias colunas relativas ao **número de casos na semana anterior**, e não apenas o dia de casos relativo ao dia correspondente a 1 semana antes da previsão



Ao adicionar todas estas *features*, ficamos com a informação não só do número de casos 1 semana antes, mas também **da tendência da evolução do número de casos durante uma semana**, na expectativa de ajudar na previsão



# Resultados



$R^2$

Para este problema, foram novamente traçadas **as curvas de aprendizagem** para os mesmos algoritmos mencionados anteriormente, e o que apresentou a melhor *performance* foi o **Random Forest**

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
0,993	0,974	2,0E+03	9,5E-04	2,1E+02	9,8E-02	1,5E+01	1,1E-02

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
0,831	0,900	3,2E+04	5,2E-03	5,5E+02	2,7E-01	9,2E+01	4,3E-02

TREINO

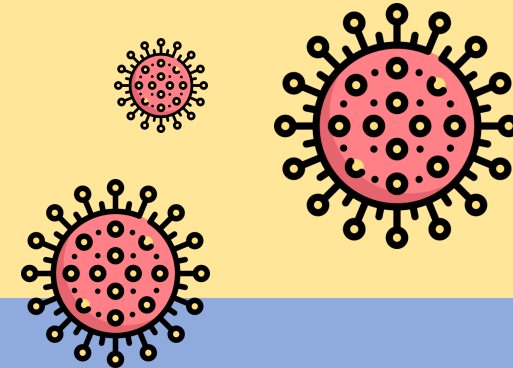
TESTE








RESULTADOS APÓS ADIÇÃO DOS DADOS RELATIVOS À SEMANA ANTERUOR

Podemos verificar que houve um **melhoramento** quando comparado com o modelo anterior

Tuning de hiperparâmetros implementado em todos os splits

# Novo Problema: Previsão das Fatalidades



Nº Fatalidades	Nº Fatalidades 8 dias antes	Nº Fatalidades 9 dias antes	Nº Fatalidades 10 dias antes
			
			
			
GAP 7 DIAS			
			

(...)

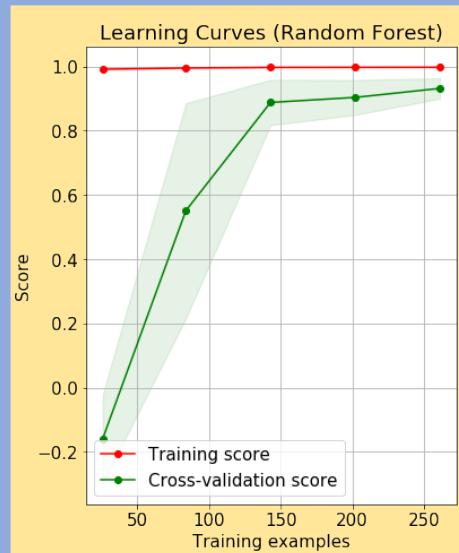
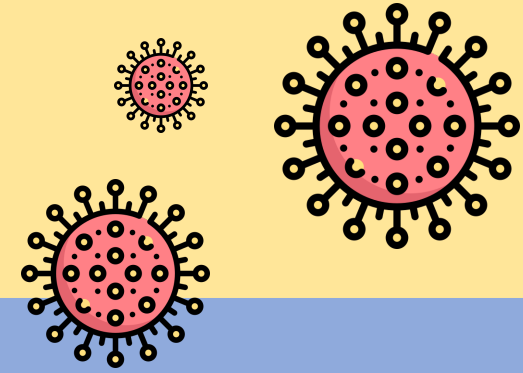
Outro problema escolhido foi a previsão do **número de fatalidades** para todos os países

Para isso, e de forma a complementar aos dados disponíveis, foram adicionados, em similaridade para a previsão do número total de casos, o **número de fatalidades na semana anterior**



Da mesma forma que no problema anterior, este número de fatalidades durante a semana anterior permite obter a **tendência de evolução do número de fatalidades** de forma a ajudar na previsão das mesmas

# Resultados



$R^2$

Para este problema, foram novamente traçadas **as curvas de aprendizagem** para os mesmos algoritmos mencionados anteriormente, e o que apresentou a melhor *performance* foi o **Random Forest**

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
0,998	0,997	5,2E+00	1,0E-04	1,3E+01	5,4E-02	9,5E-01	4,3E-03

R2	R2_s	MSE	MSE_s	ME	ME_s	MAE	MAE_s
0,936	0,962	1,2E+02	1,9E-03	3,1E+01	1,6E-01	5,1E+00	2,3E-02

TREINO

TESTE

RESULTADOS DA PREVISÃO DO NÚMERO DE FATALIDADES

Os resultados indicam que o modelo **conseguiu prever com alguma eficácia** o número de fatalidades

Tuning de hiperparâmetros  
implementado em todos os splits

