IART Projeto 2 - Classification of WHO Situation Report Data

Hugo Fernandes - up201909576
João Monteiro - up201705580
Ricardo Pinto - up201909580

Dataset inicial

Variáveis presentes no dataset inicial:

- Province/State: Especifica a província ou estado do país a que a linha de dados se refere.
- Country/Region: País ou região a que a linha de dados se refere.
- Lat: Latitude do país ou região.
- Long: Longitude do país ou região.
- Date: Data do dia do relatório da situação desta entrada.
- Confirmed: Número total de casos confirmados de COVID-19 até ao dia, inclusive.
- Deaths: Número total de mortes confirmadas de COVID-19 até o dia, inclusive.
- Recovered: Número total de pacientes que tiveram o vírus mas ficaram curados até ao dia, inclusive.

Dataset inicial

Province/State	Country/Region	Lat	Long	Date	Confirmed	Deaths	Recovered
NaN	Afghanistan	33.000000	65.000000	1/22/20	0	0	0
NaN	Albania	41.153300	20.168300	1/22/20	0	0	0
NaN	Algeria	28.033900	1.659600	1/22/20	0	0	0
NaN	Andorra	42.506300	1.521800	1/22/20	0	0	0
NaN	Angola	-11.202700	17.873900	1/22/20	0	0	0
...	...	...	...	...	...	...	...
NaN	Sao Tome and Principe	0.186360	6.613081	5/21/20	251	8	4
NaN	Yemen	15.552727	48.516388	5/21/20	197	33	0
NaN	Comoros	-11.645500	43.333300	5/21/20	34	1	8
NaN	Tajikistan	38.861034	71.276093	5/21/20	2350	44	0
NaN	Lesotho	-29.609988	28.233608	5/21/20	1	0	0

Alterações Realizadas

Alterações realizadas ao dataset inicial:

1. Date: Mudamos a data para um contador que começa em 0 no primeiro dia registado 22-01-2020.
2. Province/State: Removemos a coluna, juntando todos os dados das províncias/estados de um país ou região numa só entrada por dia.
3. *Active_Cases*: Criamos a variável que corresponde ao número de confirmados total - (número de mortos total + número de curados total).
4. *Yesterdays_Confirmed_Cases*: Criamos a variável que corresponde ao numero total de casos existentes no dia anterior.
5. *Increase_in_Cases*: Criamos a variável que corresponde ao aumento de casos que ocorreu em cada dia.
6. Will_Infection_Ratio_Increase: Variável dependente do nosso projeto, esta variável diz se o aumento de casos (*Increase_in_Cases*) do dia seguinte vai ser maior ou menor ao aumento do dia atual.
7. Removemos entradas de países quando eles não têm casos ou seja a primeira entrada de cada país será no dia em que se deteta o primeiro caso.

Dataset Alterado

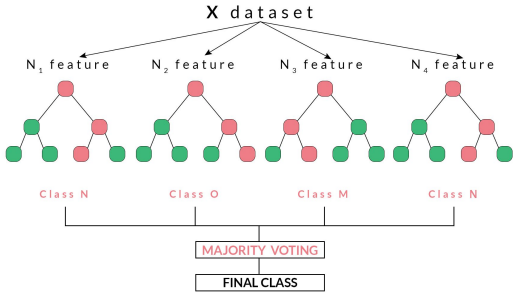
Country/Region	Day	Lat	Long	Confirmed	Deaths	Recovered	Active_Cases	Yesterdays_Confirmed_Cases	Increase_in_Cases	Will_Infection_Ratio_Increase
Afghanistan	33	33.0	65.0	1	0	0	1	0.0	1.0	False
Afghanistan	34	33.0	65.0	1	0	0	1	1.0	0.0	False
Afghanistan	35	33.0	65.0	1	0	0	1	1.0	0.0	False
Afghanistan	36	33.0	65.0	1	0	0	1	1.0	0.0	False
Afghanistan	37	33.0	65.0	1	0	0	1	1.0	0.0	False
...	...	...	...	...	...	...	...	...	...	...
Zimbabwe	116	-20.0	30.0	44	4	17	23	42.0	2.0	False
Zimbabwe	117	-20.0	30.0	46	4	18	24	44.0	2.0	False
Zimbabwe	118	-20.0	30.0	46	4	18	24	46.0	0.0	True
Zimbabwe	119	-20.0	30.0	48	4	18	26	46.0	2.0	True
Zimbabwe	120	-20.0	30.0	51	4	18	29	48.0	3.0	False

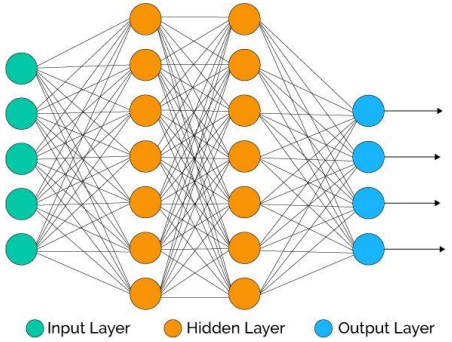
Pré-processamento

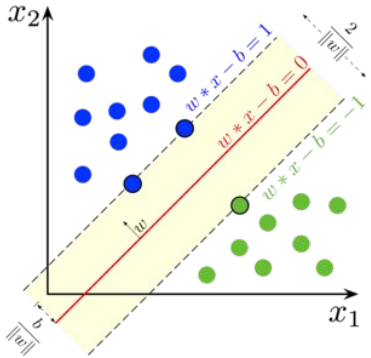
Ações realizadas aos dados na fase de pré-processamento:

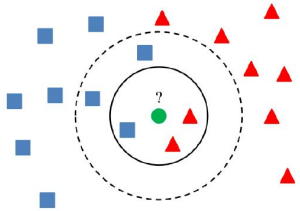
1. Transformar a variável Country/Region numa variável numérica.
2. Ordenar a tabela por dia e, de seguida, país/região.
3. Definir subconjunto X e subconjunto y para a modelagem.
4. Dividir X e y em conjuntos de treino e teste.
5. Escalar os dados.

Algoritmos Abordados

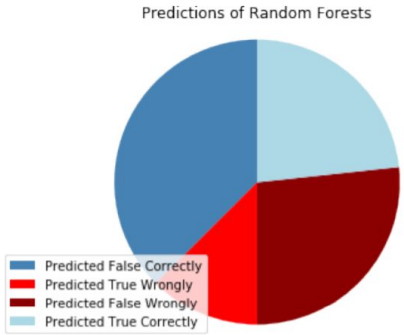


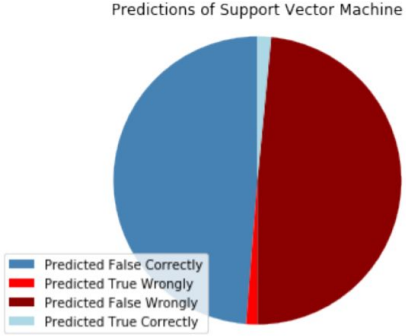


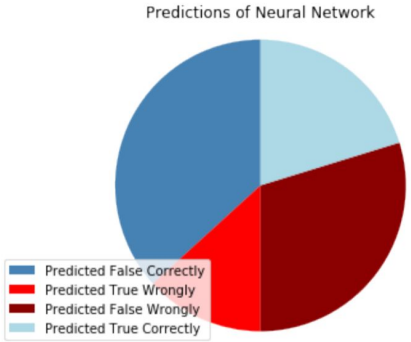


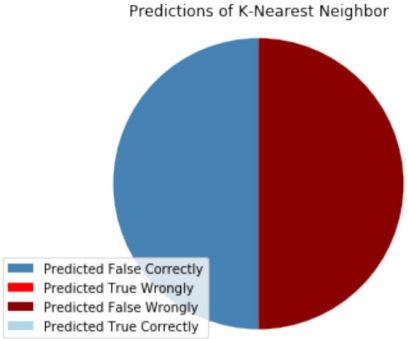


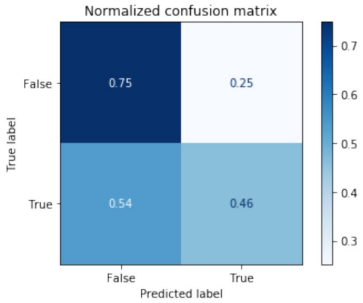
Matrizes de Confusão

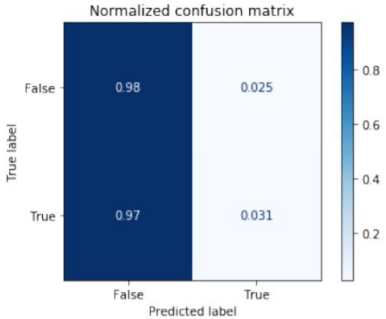


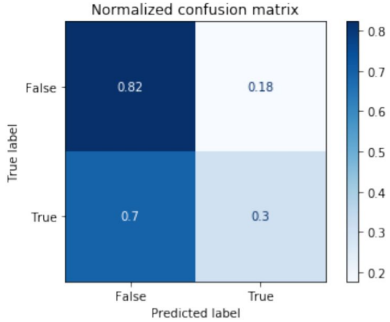


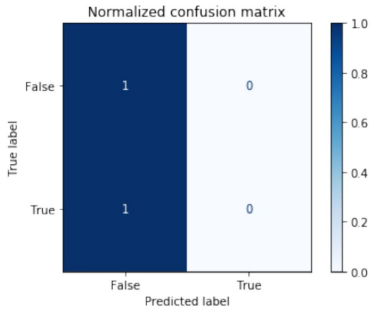






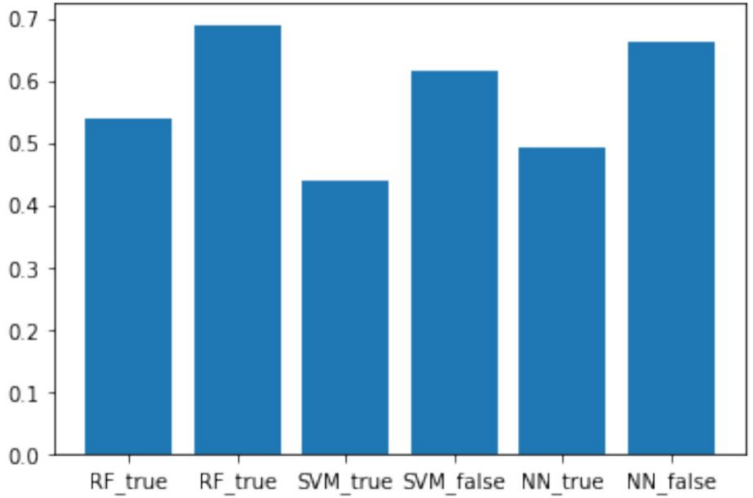






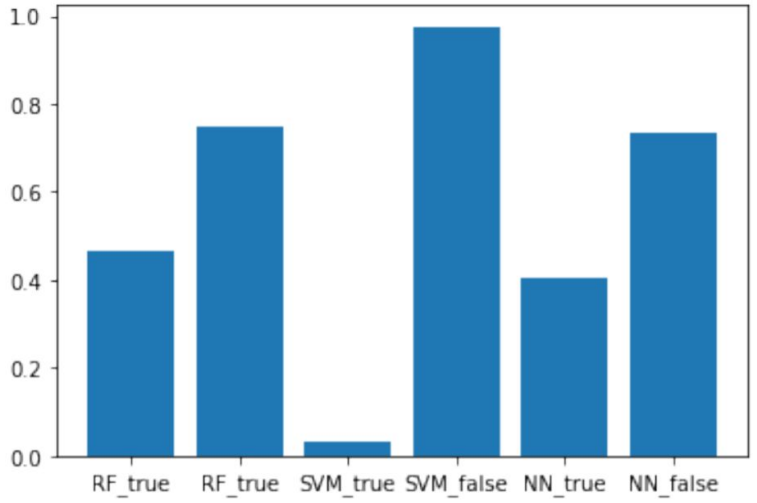
Comparação de métricas de performance - *Precision*

	Algorithm	Label	Precision
0	RF	True	0.540265
1	RF	False	0.689479
2	SVM	True	0.437500
3	SVM	False	0.614201
4	NN	True	0.490948
5	NN	False	0.661500



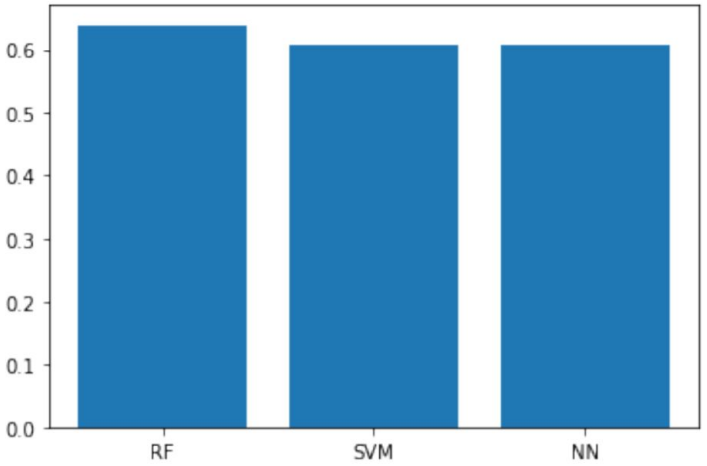
Comparação de métricas de performance - *Recall*

	Algorithm	Label	Recall
0	RF	True	0.465729
1	RF	False	0.749584
2	SVM	True	0.030756
3	SVM	False	0.975014
4	NN	True	0.405097
5	NN	False	0.734592



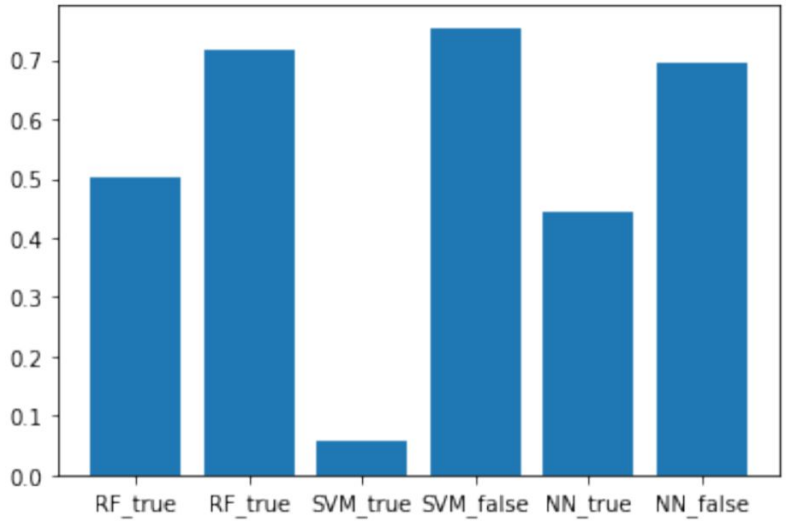
Comparação de métricas de performance - *Accuracy*

	Algorithm	Accuracy
0	RF	0.639673
1	SVM	0.609391
2	NN	0.607009



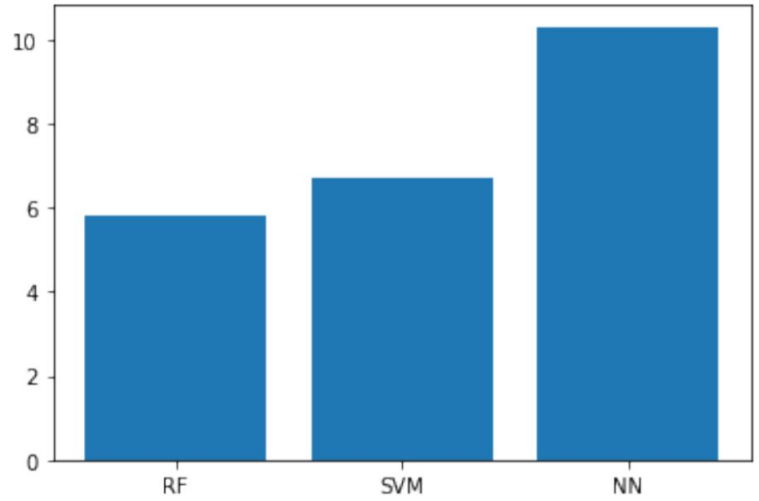
Comparação de métricas de performance - *f-score*

	Algorithm	Label	F1-Score
0	RF	True	0.500236
1	RF	False	0.718276
2	SVM	True	0.057471
3	SVM	False	0.753648
4	NN	True	0.443909
5	NN	False	0.696133



Comparação de métricas de performance - *Runtime*

	Algorithm	Runtime
0	RF	5.795721
1	SVM	6.714968
2	NN	10.312418



Conclusões

1. Comparação entre algoritmos
    a. Realisticamente apenas *Random Forest* e *Neural Networks* podem ser comparados.
    b. Diferença em *runtime* entre os dois algoritmos.
    c. Ligeiras diferenças entre outras métricas de performance
    d. Melhor algoritmo na nossa opinião.
2. Piores algoritmos
    a. KNN previu que a variável nunca seria *True*.
    b. SVM previu *True* apenas 80 vezes em aproximadamente 3000 entradas, e apenas 35 das 80 foram previsões corretas.
3. Resultados gerais da classificação
    a. Performance baixa até nos melhores dois casos.
    b. Tendencia a prever *False.*
    c. Baixa correlação das colunas com a variável dependente que conseguimos definir.