Data Science 4 Covid

Sistemas Inteligentes de Bioinformática – janeiro 2021

G1: Maria Adília Monteiro pg40961 | Sofia de Beir pg38263 | Rita Lopes Conde pg40974 | José Alexandre Carvalho pg38932

1. MOTIVAÇÃO

No início de dezembro de 2019, um surto de coronavírus (COVID-19), causado por uma nova síndrome respiratória aguda grave coronavírus 2 (SARS-CoV-2), surge na cidade de Wuhan, província de Hubei, China. Menos de um mês depois, a 30 de janeiro de 2020, a Organização Mundial da Saúde (OMS) declara o surto como uma Emergência de Saúde Pública de Interesse Internacional. A 11 de março de 2020, a OMS declara oficialmente o estado de pandemia [1]. Um ano depois, a 20 de janeiro de 2021, o COVID-19 resultou em mais de 96 milhões de casos confirmados e mais de 2 milhões de óbitos a nível mundial [2].

Os coronavírus são os principais patógenos em doenças respiratórias emergentes, constituindo uma grande família de vírus RNA de cadeia simples (ssRNA) que podem ser isolados em diferentes espécies animais. Estes vírus podem atravessar barreiras inter-espécies e causar, em humanos, doenças que vão desde constipações comuns a doenças mais graves, como *Middle East respiratory syndrome* (MERS) e SARS [3].

O espectro clínico de COVID-19 varia desde formas assintomáticas a condições clínicas caracterizadas por insuficiência respiratória que requer ventilação mecânica, manifestações multiorgânicas e sistémicas, choque séptico e síndromes de disfunção de múltiplos órgãos, febre, mal-estar, tosse seca e dispneia [3].

A incidência de infeção por SARS-CoV-2 é observada mais frequentemente em pacientes adultos do sexo masculino, com a idade mediana dos pacientes entre 34 e 59 anos. O SARS-CoV-2 tem, também, maior probabilidade de infetar pessoas com comorbidades crónicas, como doenças cardiovasculares e diabetes. A maior proporção de casos graves ocorre em adultos ≥ 60 anos de idade e naqueles com certas condições subjacentes [1].

1.1. OBJETIVOS

Com este trabalho, pretende-se estudar a temática em que vivemos atualmente, isto é, da pandemia do COVID-19. Deste modo, iremos incluir métodos de *data science*, análise estatística, *machine learning* e *deep learning* de forma a melhor compreender as possíveis relações entre os dados socioeconómicos e meteorológicos na evolução de casos diários e, consequentemente, número de mortes. Em adição, analisar-se-á a distribuição de mortalidade por género/faixa etária em Portugal.

2. MÉTODOS

2.1. DADOS

Os dados de COVID-19 foram extraídos da Direção Geral de Saúde, sendo atualizados diariamente no Github[1]. Por outro lado, os dados meteorológicos foram obtidos diretamente do Instituto Português do Mar e da Atmosfera (IPMA)[2], usando o *notebook* "DL IPMA". Finalmente, os dados socioeconómicos foram descarregados do site da PORDATA[3] e reunidos num único ficheiro usando o *notebook* "*more_pordata process*".

Data: *dataset* diário a nível nacional, contendo fatores tais como:

- casos confirmados e mortes acumulados por grupo etário
- novos casos confirmados diários
- pacientes sob cuidado hospitalar, casos recuperados e ativos
- percentagem de sintomas reportada pelos casos confirmados

Conc_data: casos confirmados acumulados por concelho, que passam de diários a semanais nas últimas entradas, e com término do *dataset* em outubro. Entradas posteriores foram colocadas num 2º ficheiro, ainda que com outra escala de dados e alguns valores negativos, pelo que não foram usados.

Conc_get: população total recenseada em 2019 e distribuição por concelho, para efeitos de normalização.

Temp_precip: temperatura mínima e máxima e precipitação diárias, por concelho. *Dataset* com início em setembro.

More_pordata: variáveis socioeconómicas por concelho:

- rácio entre médicos e farmacêuticos e população do concelho
- número total de hospitais, escolas com vários ciclos de ensino, universidades e hóspedes em hotéis
- salário médio mensal, no geral, e por género
- rácio de poder de compra
- taxa de abstenção de voto

Conc_data_new: taxa de risco de infeção e incidência de casos por concelho

2.2. PRÉ-PROCESSAMENTO

De modo a tratar os dados adquiridos e obter os resultados pretendidos de determinadas análises, foi necessário realizar um pré-processamento para remover dados em falta e outras inconsistências. Assim, esse pré-processamento passou pelas seguintes etapas:

- Converter os nomes dos concelhos em maiúsculas sem acentos e outros carateres ("ç")
- Conversão dos grupos etários para os grupos utilizados no *dataset* principal de dados COVID-19 (conc_get)

[1] https://github.com/dssg-pt/covid19pt-data

[2] http://api.ipma.pt/#services

[3] https://www.pordata.pt/

- Uma coluna de dias da semana foi acrescentada ao *dataframe* data
- Filtrar concelhos com número de Na's > quantil 75 (conc_data)
- Filtrar as primeiras semanas dos dados, com alta concentração de Nas (conc_data)
- Assegurar que *dataframes* têm os mesmos concelhos
- Separação dos dados por concelho diários e semanais (conc_data)
- *Backfill* – substituir entradas em falta pelas do(s) dia(s) seguinte(s)
- Depois das operações anteriores, substituir os restantes Nas por 0s
- Normalização dos dados por concelho, por 100 000 habitantes (conc_data)

2.3. ANÁLISE ESTATÍSTICA

Inicialmente, os dados foram tratados como dados sem natureza sequencial usando testes de normalidade de *Shapiro-Wilk* e testes (não) paramétricos. Desta forma, foi analisada a hipótese do número de confirmados e de mortes diferir não só por grupo etário, mas também por género. Os resultados do teste contradizem a literatura, não rejeitando a hipótese nula, a qual assumia que não havia diferença significativa entre confirmados e mortes em homens e mulheres.

Outra análise errónea da mesma vertente foi a comparação do número de novos confirmados por dia nas semanas antes e depois do Natal. Na altura da análise, os dados disponíveis terminavam uma semana após o Natal, pelo que os 2 grupos de dados a comparar também se encontravam desequilibrados. Estas análises foram removidas do trabalho final.

A correlação de *Spearman* (não paramétrica) entre o número de confirmados acumulados por concelho (conc_data) num único dia e as variáveis socioeconómicas foi realizada tanto com os dados não normalizados e normalizados por 100 000 habitantes (por concelho). Desta forma, os dados perdem a sua natureza sequencial.

Devido à falta de dados diários por concelho após a 1ª vaga, e ao começo do *dataset* de dados meteorológicos em setembro, a relação entre a temperatura e o número de casos acabou por não ser explorada.

2.4. ANÁLISE NÃO SUPERVISIONADA

Um *clustering* hierárquico dos casos confirmados acumulados por concelho, num dado dia, foi realizado com os métodos de *single* e *average linkage*. A diferente região geral do país (Norte, Centro, Lisboa, Alentejo e Algarve) foi usada para colorir as legendas dos concelhos correspondentes. Devido a diferenças de nomes, concelhos sem área correspondente foram colocados a preto.

Foram utilizados 2 métodos de redução de dimensionalidade, *Principal Component Analysis* (PCA) e *t-distributed stochastic neighbor embedding* (t-SNE), nos mesmos dados utilizados no *clustering*.

2.5. MACHINE LEARNING

A eficácia de diferentes modelos de regressão – *Random Forest*, *K-nearest neighbours*, *Neural Network* e *Linear Regression* – foi testada usando dados não normalizados, validação cruzada para 3 *k-folds* e a métrica R^2. Os dois problemas de regressão testados foram prever o número de novos confirmados por dia tendo como *input* os do dia anterior, e prever o número de óbitos acumulados usando o número de confirmados no grupo etário com mais de 70 anos. Dada a natureza dos dados

time series, não foi realizada permutação dos dados, e o processo de validação cruzada recorreu à classe *TimeSeriesSplit* do package *sklearn*.

A eficácia de diferentes métodos de normalização – divisão pelo desvio padrão, substração pela média, *z-score* e *fitted normalization* (classe *StandardScaler* do *sklearn*) - também foi testada com o tipo de modelo com melhor desempenho na análise anterior.

Depois destas análises preliminares, modelos mais complexos foram gerados usando a melhor combinação de tipo de modelo de regressão e método de normalização, usando como *input* os últimos 7 ou 14 dias de novos confirmados por dia para prever o número de casos no dia seguinte.

Além dos métodos de *machine learning* tradicionais, foram gerados vários modelos de *deep learning* (*output* único) – *Linear*, *Dense*, LSTM – com base num tutorial do *tensorflow* específico a dados time series[4]. Após a criação das classes utilizadas, os mesmos dados dos 14 dias anteriores de confirmados para prever os do dia seguinte, normalizados, foram divididos em *datasets* de treino, validação e teste. Dado o tamanho reduzido do *dataset* de treino, foram implementados métodos de redução de *overfitting*, nomeadamente Early Stopping e Dropout (simultaneamente).

A janela de previsão representa o número de time steps do *input* (14), o número de variáveis de saída (1) e o desfasamento entre eles (1). Foi utilizado como modelo *baseline* uma janela de previsão de 1 variável de input para 1 variável de saída com 1 dia de diferença. O desempenho dos modelos foi avaliado pela métrica de erro médio absoluto.

3. RESULTADOS E DISCUSSÃO

Após o pré-processamento dos dados, foram calculadas algumas métricas epidemiológicas[5], nomeadamente a população exposta, ou percentagem de casos confirmados acumulados em relação à população total portuguesa (3,2%), *crude mortality rate*, ou percentagem de mortes acumuladas em relação à população (0,1%) e *case fatality rate*, ou percentagem de óbitos em relação ao número de confirmados. Esta última métrica varia de forma menos consistente ao longo do tempo, tendo atingido um máximo de 4,2% no mês de maio.

Nenhuma destas métricas representa o risco de morte ou de infeção por COVID-19, uma vez que são demasiado simplistas e não consideram, por exemplo, os diferentes fatores de risco, como idade avançada e condições pré-existentes. Por outro lado, a melhor estimativa destas métricas incluiria o maior número de dados possível, a nível mundial, de forma a aproximar-se o mais possível dos valores reais.

Verificou-se também que o número de novos casos por dia varia com o dia da semana, tendo, em média, um menor número de casos às Segunda e Terça-feiras. Além disso, apresentámos o número de casos ativos, internados e novos óbitos por dia máximos mensais. Em novembro de 2020, o máximo de casos ativos atingiu quase os 90 000 casos, com mais de 3000 internados e 91 mortos por dia.

[4] https://www.tensorflow.org/tutorials/structured_data/time_series?hl=en

[5] https://ourworldindata.org/covid-mortality-risk

Apresentámos também os 5 concelhos com maior número de casos por concelho, normalizados por 100 000 habitantes, na data mais recente disponível (2020-07-05). Com base no *dataframe* da incidência, apresentámos os 5 concelhos com maior número de casos no início de janeiro de 2021.

A secção da análise exploratória contem 7 figuras. Na Figura 1, apresentamos o número de casos confirmados acumulados, casos ativos, pacientes hospitalares e óbitos acumulados até dezembro de 2020. Com base nesta figura, considerámos que a 2ª vaga da pandemia em Portugal aparenta ter começado em setembro.

A Figura 2 contrasta a variação do número de casos a nível nacional e por concelho, nos concelhos de Braga, Porto e Lisboa. Dos 3 concelhos, apenas Lisboa aparenta ter um crescimento do número de casos por 100 000 habitantes aproximadamente linear, até julho de 2020. A Figura 3 compara o número de casos relativos por concelho. O Norte de Portugal aparenta ter a maior concentração de concelhos de alta incidência na data analisada (fim da 1ª vaga, 2020-07-05).

A Figura 4 apresenta o número de sintomas reportados pelos casos confirmados até agosto de 2020. Os sintomas disponíveis no *dataset* incluem tosse, febre, dificuldade respiratória, cefaleia, dores musculares e fraqueza generalizada. A tosse é o sintoma mais comum após a estabilização das percentagens em abril. Infelizmente, os dados perdem consistência depois de agosto, limitando o número de dados disponíveis para uma possível abordagem de *machine learning*.

A Figura 5 explora a distribuição do número de casos confirmados e óbitos normalizados por grupo etário e género. O grupo etário entre os 20 e 29 anos apresenta o maior número de casos, seguido do grupo com mais de 70 anos. Relativamente aos óbitos, estes afetam predominantemente o grupo com mais de 70 anos, apesar de existirem alguns óbitos nos grupos etários com menos de 40 anos.

A Figura 6 continua a análise dos casos e óbitos por idade, e a relação entre eles. Dado que o grupo com mais de 70 anos tem o maior número de óbitos, a relação entre casos e óbitos é que mais se aproxima da linear. A Figura 7 apresenta a temperatura mínima e máxima, tal como a precipitação, em Braga entre setembro e dezembro de 2020. Um gráfico mais eficaz compararia o número de casos em Braga diretamente com a temperatura.

Relativamente à secção da análise estatística, as variáveis socioeconómicas testadas não tinham uma correlação significativa (valores aproximadamente nulos) com o número de casos por concelho após normalização, pelo que estes dados não foram mais usados.

Na secção de análise multivariada, a Figura 10 mostra que os concelhos do Norte do país, a verde, são os que melhor se organizam em clusters distintos (*average linkage*). O *cluster* que mais se distingue dos outros, à direita, inclui 2 dos concelhos com maior número de casos por 100 000 habitantes no fim da 1ª vaga, Vila-Nova-de-Foz-Coa e Ovar, apesar de não estarem próximos geograficamente. A Figura 11 apresenta as análises PCA e tSNE utilizando os mesmos dados, mas nenhuma consegue discriminar bem os concelhos em *clusters* distintos.

A secção principal do trabalho inclui os métodos tradicionais de *machine learning*, que apresentaram melhor desempenho que os métodos *deep learning*. A Figura 12 pretende exemplificar a aparente relação linear entre casos confirmados com um dia de diferença.

Os modelos de regressão para prever o número de mortes tendo como input número de confirmados num dado grupo etário tem com base alguns pressupostos errados. Em primeiro lugar,

restringe apenas ao grupo etário com o maior número de mortes, mas não responsável pelo número total de mortes. Esta decisão foi tomada com base na relação quase linear entre casos e óbitos, no mesmo dia, neste grupo etário (Figura 6).

Por outro lado, não entra em consideração com o desfasamento entre número de infeções e mortes. Uma melhor opção teria em conta o número de infetados de cada grupo etário, possivelmente relacionado com o número de mortes uma semana depois ou mais tarde. Podíamos, por exemplo, ter procurado o valor médio de dias entre infeção e morte, apesar de este provavelmente ser altamente variável conforme as caraterísticas de cada paciente.

Dos modelos de regressão testados, os de *Linear Regression* demonstraram melhor desempenho na regressão do número de casos, atingindo um R^2 de 0.5 (Figura 15). O método de normalização que resultou no melhor desempenho dos modelos de *Linear Regression* para a regressão do número de casos confirmados foi a *fitted*. Repetimos o processo utilizando a melhor combinação e mais dados de *input*, nomeadamente os últimos 7 e 14 dias de casos confirmados. Os modelos com mais dados tiveram melhor desempenho, com um R^2 de aproximadamente 0.7.

Finalmente, os modelos *deep learning* gerados obtiveram um erro médio absoluto bastante maior que os modelos tradicionais (Figura 17), apresentando *overfitting* mesmo com a implementação dos métodos de *Early Stopping* e *Dropout* em simultâneo (Figura 18). Este facto deve-se principalmente ao número muito reduzido de dados. Mesmo acrescentado mais variáveis aos modelos, como por exemplo o número de novos confirmados por dia por grupo etário e o número de casos ativos, é provável que este problema continue a afetar o desempenho dos modelos *deep learning*.

Apesar do foco desta análise ser a nível nacional, o trabalho poderia ter beneficiado de comparações simples entre Portugal e outros países, principalmente países membros da União Europeia. Por exemplo, seria pertinente apresentar comparações entre o número de casos por milhão de habitantes em diferentes países – aquando da escrita deste relatório, em janeiro de 2021, Portugal aparece nos primeiros lugares do ranking de países com valores mais elevados do mundo [4].

REFERÊNCIAS

[1] Harapan, H., Itoh, N., Yufika, A., Winardi, W., Keam, S., Te, H., ... & Mudatsir, M. (2020). Coronavirus disease 2019 (COVID-19): A literature review. Journal of Infection and Public Health.

[2] COVID-19 Dashboard. https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 Last Accessed: 21/01/2021

[3] Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2020). Features, evaluation and treatment coronavirus (COVID-19). In Statpearls [internet]. StatPearls Publishing.

[4] Statista. Rate of COVID-19 cases in the most impacted countries worldwide as of January 18, 2021 (per million population). https://www.statista.com/statistics/1174594/covid19-case-rate-select-countries-worldwide/ Last Accessed: 22/01/2021