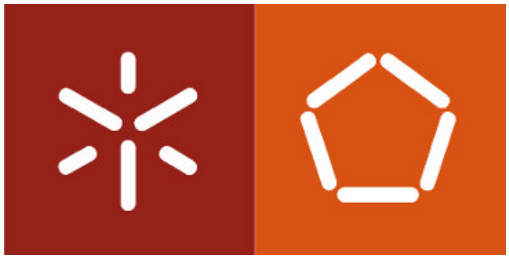

Universidade do Minho

Mestrado em Engenharia Informática

Previsão do número de infetados/mortes de doenças pulmonares

Diogo Tavares, pg42826

Hugo Nogueira, A81898

Luís Abreu, A82888

Aprendizagem Automática 2

4º Ano, 2º Semestre

Departamento de Informática

Junho 2021

Índice

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Identificação do Projeto e Objetivos

No âmbito da Unidade Curricular de Aprendizagem Automática 2 foi-nos proposta a realização de modelos de *Deep Learning* capaz de trabalhar séries temporais.

O principal objetivo passa pela criação de modelos de previsão do número de óbitos causados por uma determinada doença, e assim auxiliar uma melhor resposta no combate à mesma. Atualmente a pandemia COVID tem sobrecarregado os sistemas de saúde consumindo a maioria dos seus recursos. Neste trabalho, pretende-se o desenvolvimento de uma ferramenta, com recurso a métodos de deep learning e à linguagem Python, que seja capaz de prever o número de infectados e mortos provocados por doenças pulmonares. A previsão será feita com base nos dados fornecidos por diversas entidades de saúde sobre o número de pessoas com doenças pulmonares ao longo de um período de tempo, bem como por outras entidades dos mais diferenciados campos de atividades, como por exemplo meteorologia, aviação entre outros.

A previsão do número de mortes será realizada utilizando Séries Temporais com diferentes registos de periodos temporais, nomeadamente:

- registos diários;
- registos semanais;
- registos mensal.

1.2 Etapas de Trabalho

Conforme planeado com os orientadores do projeto, este trabalho foi dividido em três partes distintas. Uma vez que não existia um dataset criado para poder ser feita uma previsão do número de óbitos por uma doença pulmonar, uma dessas partes, a mais importante e a mais longa e onde ocorreram vários processos iterativos para melhoria desse dataset, consistiu na sua criação. Após a criação do dataset segui-se a fase de exploração e tratamento de dados, onde foram retirados features

de pouca importancia para uma previsão do número de óbitos bem como tratados os dados para melhoria do desempenho do modelo, em específico o tratamento de *missing values* Após esta fase procedeu-se à escolha de atributos relevantes para a previsão objetivo. Por fim, seguiu-se a implementação de modelos de previsão utilizando Redes Neuronais Recorrentes e Redes neuronais Convolucionais, e aqui foram experimentadas optimizações diferentes por forma a fazer um benchmark de qual o melhor melhor modelo a utilizar para este dataset.

1.3 Estrutura do relatório

Este relatório encontra-se dividido em 7 capítulos.

No capítulo 1.**Introdução** é feita uma breve abordagem ao que se pretende com a realização deste projeto, bem como quais as diferentes etapas do trabalho.

Seguidamente, no capítulo de 2.**Dataset** faz-se uma explicação detalhada dos diferentes *datasets* utilizados para a criação do *dataset* utilizado nas diferentes previsões do número de óbitos. Esta preparação ocupou uma boa parte da realização do trabalho.

No capítulo 3.**Exploração de Dados, Tratamento de dados e *Feature Selection*** , uma vez que o *dataset* final criado para a alimentação dos modelos de previsão de Séries Temporais contêm inúmeras features, e sabendo que quanto mais simples o *dataset* menos complexos são os modelos de previsão são apresentadas as várias observações realizadas para determinar quais as *features* mais pertinentes de se manter ou quais as menos interessantes e assim eliminá-las do *dataset* e obter um modelo mais simples mas sem nunca descurar a sua performance. Encontram-se também explicadas ao pormenor os vários tratamentos de dados aplicados ao dataset, em particular no seu tratamento de *missing values* e na sua transformação de *dataset* de fequência diária para frequencia semanal e mensal. Aqui é possível também encontrar a exploração de dados que nos permitiu conhecer melhor os dados do datset bem como nos auxiliou na escolha da melhor forma para realizar a *feature selection*.

No capítulo 4.**Modelos de previsão** são enumerados e explicados ao por-

menor quais os modelos e suas layers constituintes utilizados neste trabalho para a previsão do número de óbitos por COVID-19 em Portugal e o número de óbitos por Pneuminia nos EUA

No capitulo 5.**Otimização** é exposto os vários hiper-parâmetros que foram testados no intuíto de encontrar e melhorar os diferentes modelos de previsão.

No capitulo 6.**Resultados** fazemos a reflexão e análise dos resultados obtido com a aplicação dos modelos de previsão.

Por fim, em 7.**Conclusões** é feita uma breve conclusão do trabalho realizado, onde são retiradas ilações sobre o mesmo e se o grupo concluiu os objetivos iniciais a que se propôs.

2 Dataset

O *dataset* utilizado neste projeto para a previsão do número de ótbitos em Portugal foi criado de raíz pelo grupo.

Este processo foi algo moroso devido aos requisitos que os dados do *dataset* deveriam obedecer. Como neste projeto utilizamos *datasets* com frequência de registos diários, semanais e mensais tornou a procura de dados que obedecessem a estes requisitos algo complicado, assim o grupo optou por procurar dados cuja sua frequência de registos fosse diária, para mais tarde serem transformados de forma fácil, através da soma dos mesmos, em registos semanais e mensais.

Como o objetivo deste trabalho é a previsão do número de mortes por doença respiratória em Portugal, e devido aos recentes acontecimentos da atualidade no que à pandemia COVID-19, optamos por nos focar em encontrar datasets referentes ao COVID-19, bem como outros datasets que fossem pertinentes e que contribuíssem para um incremento na qualidade dos nossos dados. No entanto apesar de partir-mos do principio que seria fácil encontrar encontrar estes dados, isso não se veio a revelar, pois apesar de haverem muitos datasets sobre a doença COVID-19, o facto desta doença ser recente faz com que outros *datsets* que no nosso entender poderiam ser relevantes sejam difíceis de encontrar.

Como base para a construção do *dataset* recorremos ao repositório no *GitHub* criado pela *Data Science for Social Good Portugal*, onde nele é possível encontrar dados referentes a Portugal, nomeadamente o número de casos confirmados de COVID-19, número de casos cofirmados por regiões do país, número de casos por faixa etária e género, o número de óbitos e número de ótitos por regiões, nível de sintomas apresentados e número de recuperados e respetivas regiões. Ainda neste repositório da *Data Science for Social Good Portugal* encontramos dados referentes ao número de testes realizados pelos laboratórios em Portugal e ao número dos diferentes tipos de teste existentes, PCR e Antigénio, ralizados.

No entanto para incrementar valor ao *dataset* que servirá para realizar previsões acrescentamos ainda dados relativos à meteorologia em Portugal. utilizando a plataforma *Visualcrossing* foi-nos possível obter dados atmosféricos em Portugal,

onde obtivemos as temperaturas minimas, máximas, médias para Portugal bem como o nível de precipitação, humidade relativa, visibilidade, estado do céu (limpo, parcialmente nublado, nublado).

Foi ainda acrescentado dados relativos à evolução diária dos acionamentos de meios de emergência médica em Portugal. Esses dados foram obtidos junto do SNS - Serviço Nacional de Saúde. Nele constavam o número de acionamento de helicóptero, viaturas, ambulâncias, motociclos de emergência médica.

Por forma a enriquecer ainda mais o *dataset* foram adicionados dados relativos a todos os países do mundo no que toca ao número de casos registados, número de novos casos, numero total de mortos nesses países, número de testes realizados, número de vacinados parcialmente e completamente, número de paciente internados em unidade de cuidados intensivos, numeros de pacientes internados nos hospitais por COVID-19, rendimentos per capita do respetivo país, número de mortes cardiovasculares, número de pessoas com diabetes, dados relativos a condições sanitária entre outros. Todos estes dados foram obtidos pela plataforma *our World Data* através do seu *GitHub*.

Com todos estes acrescentos ao *dataset* de base obtido na plataforma *Data Science for Social Good Portugal*, oiginou um *dataset* com 422 registos e 4146 features. Como é possível observar, obtivemos um *dataset* algo extenso e para fazer uma redução de *features* por forma a que os nossos modelos sejam mais simples, iremos nos capítulos seguintes fazer uma exploração dos seus valores e *features*.

Nos modelos de previsão semanal tivemos a oportunidade de testar e prever o número de mortes por Pneumonia nos Estados unidos da América, sendo que a construção deste *dataset* se baseou emm dados recolhidos no portal do Center for Disease Control and Prevention dos EUA.

Os *datasets* utilizados para a criação do *dataset* que será posteriormente tratado e explorado com o intuíto de ser utilizado na previsão do número de óbitos em Portugal, encontra-se em **Datasets auxiliares**, e os processos que foram aplicados para a sua criação encontram-se na pasta **tratamento data** .

Outros *datasets* foram também pesquisados, mas devido à dificuldade em encontrar dados foram postos de parte, e o grupo optou por se focar unica e exclusivamente no *dataset* com dados sobre a pandemia COVID-19 em Portugal e dados sobre Pneumonia nos EUA. Estes outros *datasets* encontram-se em **Datasets ignorados**.

3 Exploração de Dados, Tratamento de dados e *Feature Selection*

A exploração de dados e o seu tratamento, é das fases mais importantes em todo o processo de criação de um modelo de *machine learning*, pois permite-nos aferir a qualidade dos dados que estamos a utilizar bem como os seus pontos fortes e fracos. Nesta fase é essencial entender os tipos dos dados, definir quais as variáveis independentes e qual a classe objetivo, e é também importante retirar dos dados medidas ou características das diferentes entidades que compõem o *dataset*.

Para entender melhor os dados foram realizada uma série de visualizações gráficas para identificar-mos possíveis padrões nos dados e a possível existência de valores nulos no *datase*. Toda a exploração efetuada ao *dataset* encontra-se na pasta **exploracao data**. Nesta pasta encontram-se o ficheiro *Jupyter Notebook* **Final exploration** que demonstra passo a passo as diferentes explorações realizadas sobre o *dataset* dataframe tratado.

Para esta exploração de dados ser possível de ser realizada, foi necessário tratar os dados, eliminado os valores nulos e possíveis atributos desnecessário. Este processo foi sendo realizado de forma iterativa e á medida que mais dados de diferentes *datasets* foram sendo adicionados. Daí a pasta tratamento data conter dois ficheiros *Jupyter notebook*, o ficheiro process data que contem o tratamento de dados relativos ao *dataset* base daily_covid_19_portugal.csv e tratamento relativo às condições meteorológicas que se encontram nos *datasets* temp_portugal_XXX.csv

Este tratamento inicial originou o *dataset* daily_covid.csv, a quem posteriormente foram adicionados mais dados. Dados estes que foram tratados no ficheiro tratamento.ipynb.

O tratamento dos dados foram focados especialmente na resolução dos valores nulos no *dataset*, que foram o principal problema encontrado nos dados. Para resolver esse problema, em alguns casos demos-lhes valor 0, pois os registos encontravam-se ainda muito no inicio do periodo de pandemia, por exemplo não haviam vacinas ou ainda não tinham ocorrido óbitos. Para a falta de valores pelo

meio do dataset utilizamos o valor médio ou valor dia anterior.

Findada toda esta etapa de tratamento de dados, foi criado o *dataset* dataframe_tratado.csv, que foi então utilizado para aprofundar a nossa exploração de dados e efetuar ***feature selection*** por forma a simplificar o modelo de aprendizagem e melhorar a sua qualidade.

Esta ***feature selection*** encontra-se no ficheiro Final_exploration.ipymb, e inicialmente algumas *features* foram eliminadas baseadas na intuição, mas sempre orientada ao objetivo do problema, após isso a seleção de atributos consistiu na utilização de matriz de correlações de Spearman, pois talm como visto na exploração de dados, estes são não paramétricos. Consistiu ainda na aplicação do algoritmo SelectKBest da biblioteca *Scikitlearn*, e ta,bém dessa mesma biblioteca a utilização do algoritmo *RandomForestRegressor* aplicado ás *Lag Variables*.

Após a aplicação deste métodos decidimos manter *features* cuja sua presença fosse comum em ambos os algoritmos e que tivessem uma correlação superior a 0.5, e mantivemos também as *features* que ambos os algoritmos sugeriram mas que não se encontravam em simultânio nos mesmos. Esse *dataset* designa-se por dataset_final.csv, que contem 422 registos e 77 atributos.

É importante e necessário referir que ao *dataset* utilizado para a previsão semanal foram adicionadas *features* que não se encontram nem nos *dataset* diário nem no *dataset* mensal, e todo o processamento desses novos dados encontram-se no ficheiro tratamento_usa_dataset.ipynb, e deu origem ao *dataset_final_semanal.csv* que é constituído por 422 registos e 87 atributos. Também com este *dataset* se fez a previsão de mortes por Pneumonia para os EUA

4 Modelos de previsão

O objetivo deste trabalho centrava-se na previsão do número de óbitos em Portugal. Para alcançar tal objetivo utilizamos modelos de *deep learning* de redes neuronais recorrentes **LSTM - Long Short Term Memory** e **CNN - Concolutional Neural Network**. Estes modelos de previsão são os mais utilizados na atualidade na previsão utilizando Séries Temporais. Na figura 1 observamos a constituição dos modelos que utilizamos no nosso trabalho.

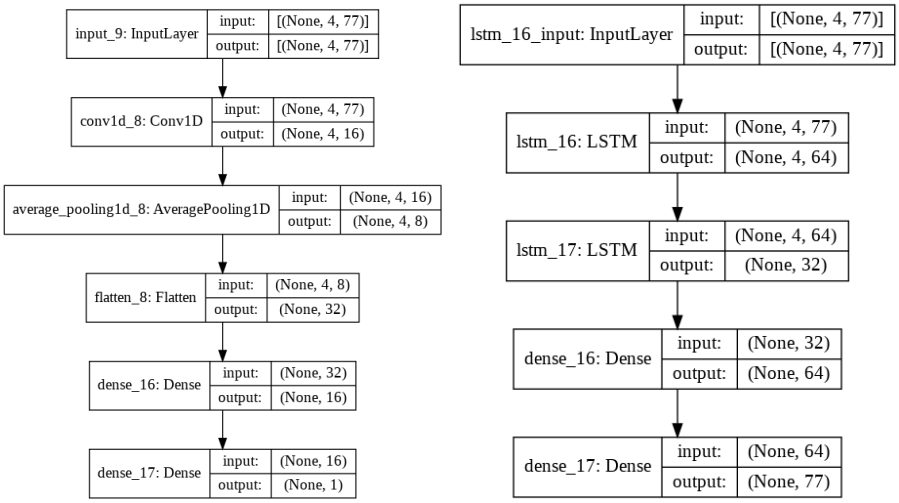

Figura 1: Modelos de *Deep learning* CNN (esquerda) e LSTM (direita)

A construção e treino destes diferentes modelos, uma vez que utilizamos *datasets* de frequencia diária, semanal e mensal encontram-se nas pastas **Daily Model**, **Weekly Model** e **Monthly Model** respetivamente.

5 Hiperparâmetros de otimização

Como forma de obtermos e encontrar-mos o melhor modelo possível, e de forma a poder-mos comparar a *performance* dos diferentes modelos, optamos por aplicar a ambos os modelos de *machine learning* para a previsão do número de óbitos diários e semanais em Portugal os mesmos parâmetros de otimização. Na previsão do número de mortes mensais em Portugal por COVID-19, devido ao seu registo temporal mensal que tornam o *dataset* pequeno o que por sua vez afetaria a qualidade da previsão, outros parâmetros foram utilizados, como é possível constatar na tabela 2. Os hiperparâmetros que optamos por experimentar foram o número de épocas de treino, taxa de aprendizagem e o *batch size*.

	Número de épocas	Timesteps	Taxa Aprendizagem	*Batch size*
1 Otimização	25	1	0.001	1
2 Otimização	50	1	0.0001	2
3 Otimização	100	1	0.00001	3
4 Otimização	60	2	0.001	1
5 Otimização	90	2	0.0001	2
6 Otimização	120	2	0.00001	3
7 Otimização	50	4	0.001	1
8 Otimização	90	4	0.0001	2
9 Otimização	120	4	0.00001	3

Tabela 1: Valores utlizados nas optimizações nos modelos Diários e Semanais

	Número de épocas	Timesteps	Taxa Aprendizagem	*Batch size*
1 Otimização	25	30	0.001	1
2 Otimização	50	30	0.0001	2
3 Otimização	100	30	0.00001	3
4 Otimização	60	30	0.001	1
5 Otimização	90	30	0.0001	2
6 Otimização	120	30	0.00001	3
7 Otimização	50	30	0.001	1
8 Otimização	90	30	0.0001	2
9 Otimização	120	30	0.00001	3

Tabela 2: Valores utlizados nas optimizações dos modelos Mensais

Estas otimizações são possíveis de serem observadas nos *notebooks* do repositório onde são feitas a construção e treino dos diferentes modelos, nas pastas **Daily Model**, **Weekly Model** e **Monthly Model**.

6 Resultados

Todos os resultados obtidos no treino dos diferentes modelos com as diferentes optimizações encontram-se, dependendo da frequência de registo do *dataset* nas pastas **Daily Model**, **Weekly Model** e **Monthly Model** no repositório do *GitHub*.

6.1 *Dataset* Diário

6.1.1 CNN

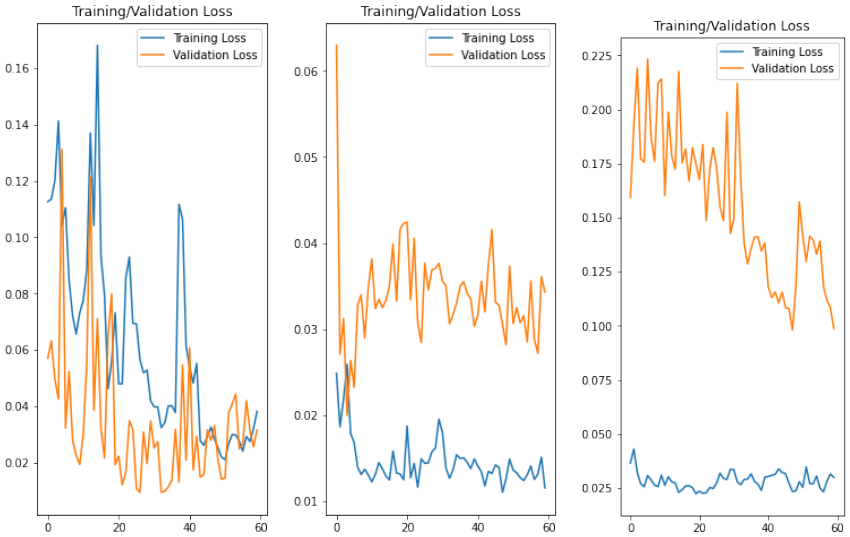

Figura 2: Evolução da Loss e Validation Loss observadas no melhor modelo CNN encontrado

	Loss	Val Loss	RMSE
1 Tuning	0.038020	0.078915	0.038020
2 Tuning	0.055468	0.123096	0.055541
3 Tuning	0.101778	0.185751	0.101730
4 Tuning	0.034948	0.074552	0.034948
5 Tuning	0.037923	0.136972	0.037973
6 Tuning	0.106314	0.173387	0.106971
7 Tuning	0.045820	0.092129	0.045820
8 Tuning	0.037118	0.131635	0.037103
9 Tuning	0.174079	0.277094	0.174086

Tabela 3: Valores médios de cada uma das optimizações testadas no modelo CNN

Através dos resultados obtidos, utilizando como medida de desempenho RMSE, o modelo com a optimização em que são utilizadas 60 épocas de treino, um *timestep* de 2, *batch size* de 1 e uma taxa de aprendizagem 0.001 foi aquele que obteve melhor resultado. Com esta otimização o modelo previu que ocorreriam 23 mortes no dia 23-04-2021, no entanto o valor real foi de 1 óbito nesse dia. Curiosamente a otimização 1 apesar de obter valores de RMSE ligeiramente piores previu um total de 9 óbitos para 23-04-2021.

Intuindo através da análise dos gráficos da evolução das *losses* de treino e validação, é-nos sugerido que o modelo poderá estar em *underfitting*, como tal seria ideal aumentar o número de épocas para aprendizagem ou a taxa de aprendizagem, para este convergir mais rapidamente para um *good fitting*.

6.1.2 LSTM

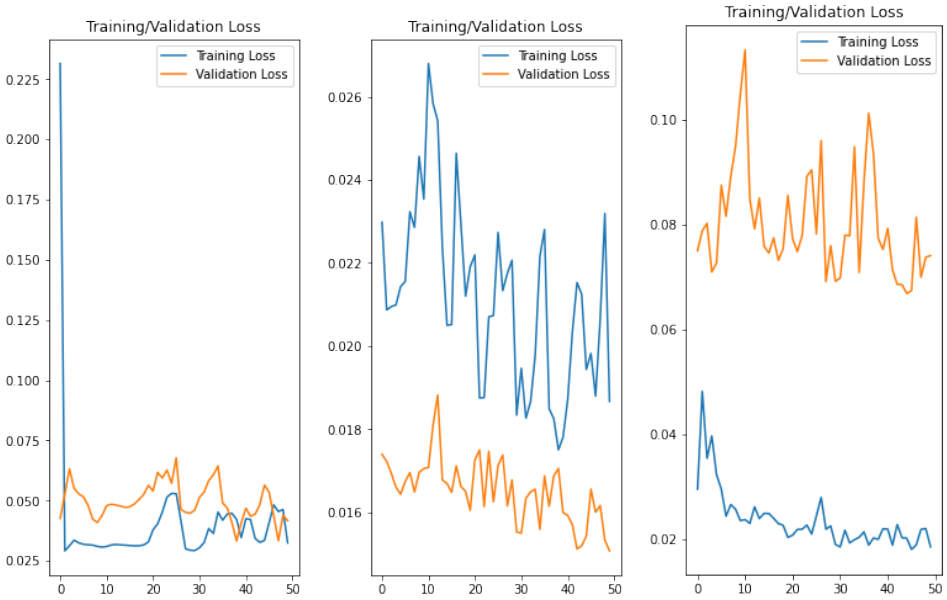

Figura 3: Evolução da Loss e Validation Loss observadas no melhor modelo LSTM encontrado

	Loss	Val Loss	RMSE
1 Tuning	0.031370	0.042616	0.031370
2 Tuning	0.044053	0.054807	0.044027
3 Tuning	0.154328	0.187572	0.154469
4 Tuning	0.028392	0.066768	0.028392
5 Tuning	0.034011	0.058998	0.033977
6 Tuning	0.122296	0.159483	0.122296
7 Tuning	0.028437	0.048779	0.028437
8 Tuning	0.037118	0.052955	0.034664
9 Tuning	0.174079	0.139032	0.101069

Tabela 4: Valores médios de cada uma das optimizações testadas no modelo LSTM

Após observar os resultados obtidos, constatamos que o melhor hiperparâmetro utilizando a métrica RMSE como decisor é a otimização 4, no entanto observando o valor médios obtidos para a Loss de validação com hiperparâmetro de otimização 7, verificamos que é consideravelmente melhor que a otimização 6, e tanto a Loss e a RMSE com a otimização 7 é muito semelhante aos resultados da otimização 4, como tal o modelo com a otimização 7 foi por nós escolhido como sendo o melhor modelo. Com este modelo o número de óbitos para o dia 23-04-2021 foi de 24. Os hiperparâmetros da otimização 7 consistia na utilização de 50 épocas, 2 *timesteps*, taxa de aprendizagem de 0.001 e um *batchsize* de 1.

Analisando os valores das *losses* à medida que as épocas de treino vão avançando observamos que há uma possibilidade de o modelo estar perto de *underfitting*, necessitando este de ser mais treinado aumentando o número de épocas e aumentar a sua *learning rate.*

6.2 *Dataset* Semanal

6.2.1 LSTM

6.2.1.1 Covid

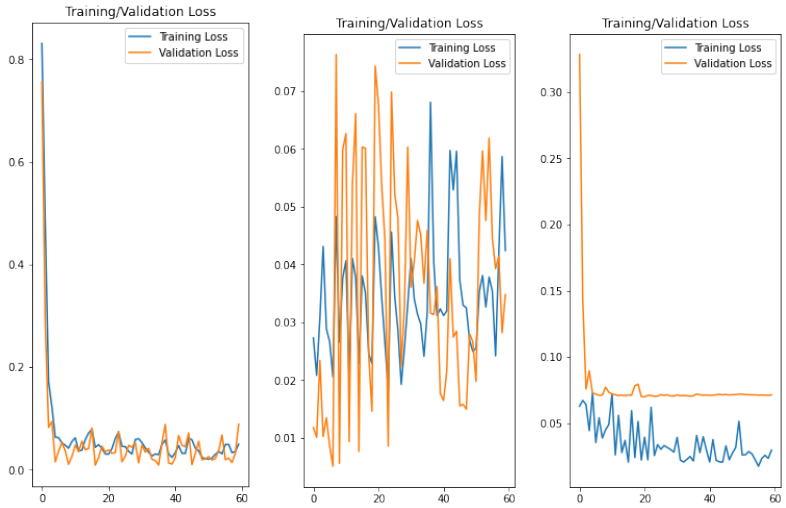

Figura 4: Evolução da Loss e Validation Loss observadas no melhor modelo LSTM encontrado (Covid)

	Loss	Val Loss	RMSE
1 Tuning	0.056273	0.060026	0.056273
2 Tuning	0.143953	0.224563	0.144107
3 Tuning	0.611841	0.700849	0.614166
4 Tuning	0.045703	0.055969	0.045703
5 Tuning	0.079359	0.132047	0.079487
6 Tuning	0.452286	0.594321	0.453417
7 Tuning	0.046808	0.063597	0.046808
8 Tuning	0.088008	0.178477	0.087965
9 Tuning	0.497985	0.626104	0.498655

Tabela 5: Valores médios de cada optimização testada - LSTM (Covid)

Através dos resultados obtidos, utilizando como medida de desempenho o RMSE, o modelo com a otimização em que são utilizadas 60 épocas de treino, um *timestep* de 2, *batch size* de 1 e uma taxa de aprendizagem 0.001 foi aquele que obteve melhor resultado. Com esta otimização o modelo previu que ocorreriam 106 mortes por Covid-19 em Portugal na semana de 26-04-2021 a 2-05-2021. Através dos *plots* da evolução do *training loss* e *validation loss* podemos verificar, principalmente através dos *plots* das iterações 1 e 3 do *time series cross validation*, que a convergência está a ser atingida logo nas primeiras épocas. Pondo isto, poderia-se reduzir um pouco o número de épocas e/ou reduzir também um pouco o *learning rate*.

6.2.1.2 Pneumonia

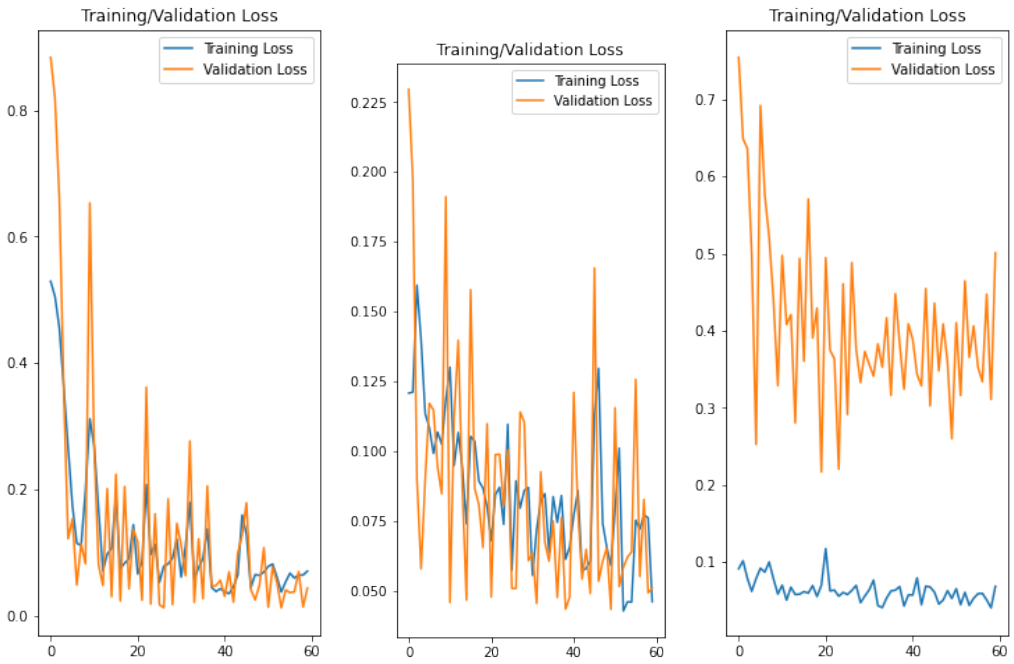

Figura 5: Evolução da Loss e Validation Loss observadas no melhor modelo LSTM encontrado (Pneumonia)

	Loss	Val Loss	RMSE
1 Tuning	0.113006	0.275551	0.113006
2 Tuning	0.167794	0.505305	0.171513
3 Tuning	0.483828	0.845477	0.486233
4 Tuning	0.091368	0.210401	0.091368
5 Tuning	0.109159	0.407833	0.111038
6 Tuning	0.412543	0.790084	0.420210
7 Tuning	0.099023	0.209202	0.099023
8 Tuning	0.094518	0.334615	0.095226
9 Tuning	0.385115	0.790497	0.393135

Tabela 6: Valores médios de cada uma das optimizações testadas - (Pneumonia)

Através dos resultados obtidos, utilizando como medida de desempenho o RMSE, o modelo com a otimização em que são utilizadas 60 épocas de treino, um *timestep* de 2, *batch size* de 1 e uma taxa de aprendizagem 0.001 foi aquele que obteve melhor resultado. Com esta otimização o modelo previu que ocorreriam 5720 mortes por Pneumonia nos Estados Unidos na semana de 26-04-2021 a 2-05-2021. Através dos *plots* da evolução do *training loss* e *validation loss* podemos verificar que existe uma enorme volatilidade principalmente por parte da *validation loss*. Para contrariar isto é necessário reduzir um pouco os valores relativos ao *learning rate*. Esta redução do *learning rate* foi aplicada noutras otimizações, mas como podemos observar na tabela a cima em nada melhorou o modelo. Como a redução foi de 0.001 para 0.0001 provavelmente uma redução ligeiramente menor poderia ajudar a verificar ainda melhor resultados.

6.2.2 CNN

6.2.2.1 Covid

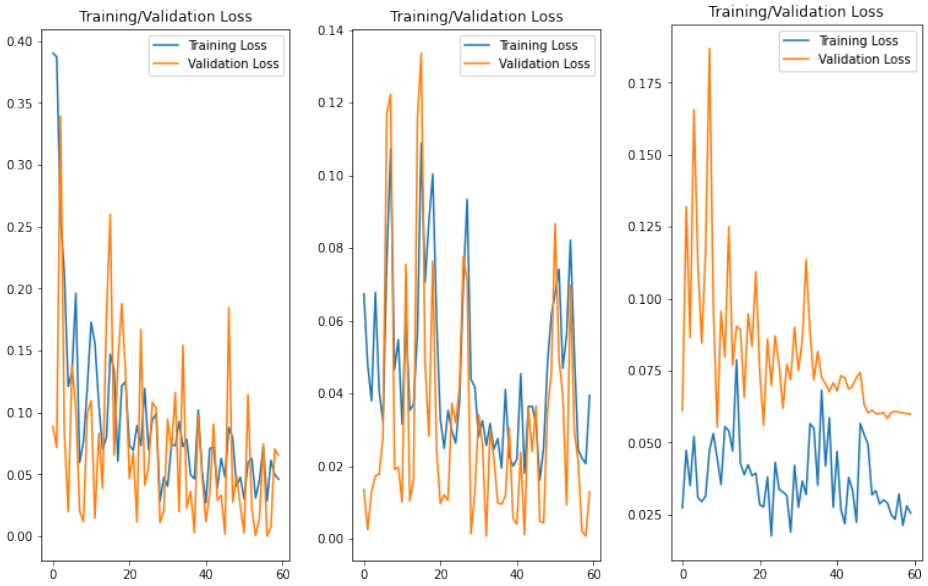

Figura 6: Evolução da Loss e Validation Loss observadas no melhor modelo CNN encontrado (Covid)

	Loss	Val Loss	RMSE
1 Tuning	0.082534	0.100594	0.082534
2 Tuning	0.100126	0.487570	0.103117
3 Tuning	0.206230	0.206967	0.206933
4 Tuning	0.059373	0.061783	0.059373
5 Tuning	0.072030	0.250014	0.072409
6 Tuning	0.361532	0.176111	0.350621
7 Tuning	0.063488	0.067193	0.063488
8 Tuning	0.069832	0.109693	0.069794
9 Tuning	0.683829	0.851989	0.685041

Tabela 7: Valores médios de cada uma das optimizações testadas - CNN (Covid)

Através dos resultados obtidos, utilizando como medida de desempenho o RMSE, o modelo com a otimização em que são utilizadas 60 épocas de treino, um *timestep* de 2, *batch size* de 1 e uma taxa de aprendizagem 0.001 foi aquele que obteve melhor resultado. Com esta otimização o modelo previu que ocorreriam 390 mortes por Covid-19 em Portugal na semana de 26-04-2021 a 2-05-2021. Através dos *plots* da evolução do *training loss* e *validation loss* podemos verificar, uma grande volatilidade nas curvas de *training loss* e *validation loss*. Esta volatilidade é muito provavelmente devida a um *learning rate* demasiado alto e/ou valores de *beta_1* e *beta_2* não óptimos. Foram testados modelos com *learning rate* inferior ao modelo que obteve o melhor resultado, mas estes obtiveram um pior RMSE do que o modelo escolhido. Como a descida do *learning rate* de modelo para modelo foi de 0.001 para 0.0001, um possível futuro passo a seguir é verificar da existência de um valor intermédio entre esse intervalo que obtenha melhores resultados.

6.2.2.2 Pneumonia

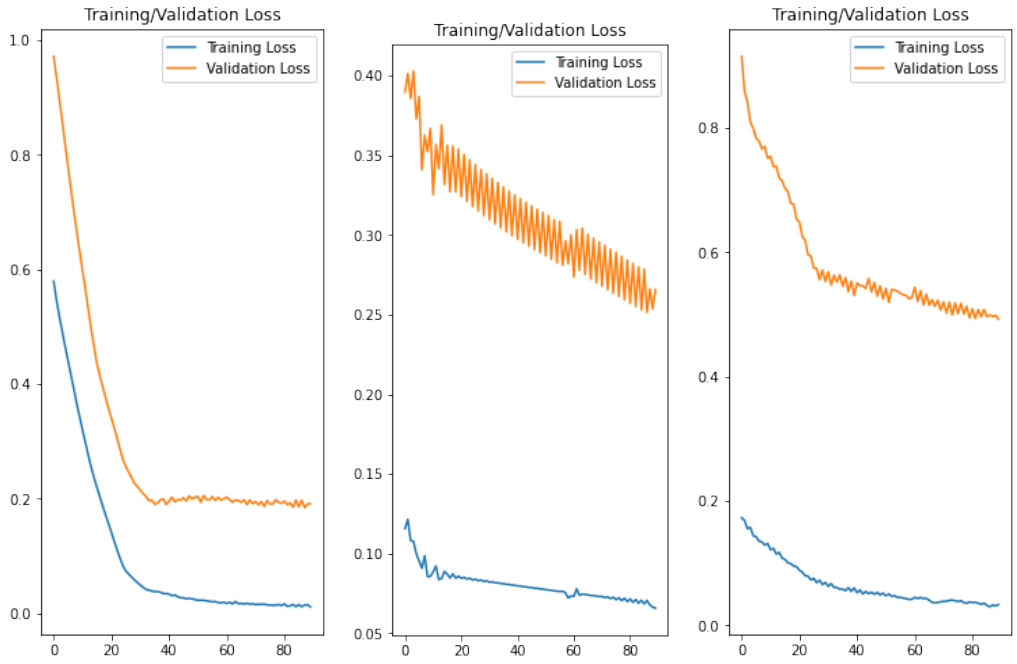

Figura 7: Evolução da Loss e Validation Loss observadas no melhor modelo CNN encontrado (Pneumonia)

	Loss	Val Loss	RMSE
1 Tuning	0.111530	0.307866	0.111530
2 Tuning	0.172333	0.474916	0.176522
3 Tuning	0.312084	0.787107	0.320815
4 Tuning	0.108374	0.353663	0.108374
5 Tuning	0.109602	0.203892	0.109839
6 Tuning	0.379726	0.556863	0.386966
7 Tuning	0.099632	0.190068	0.099632
8 Tuning	0.082921	0.399261	0.086426
9 Tuning	0.541940	0.908098	0.552806

Tabela 8: Valores médios de cada uma das optimizações testadas - CNN (Pneumonia)

Através dos resultados obtidos, utilizando como medida de desempenho o RMSE, o modelo com a otimização em que são utilizadas 90 épocas de treino, um *timestep* de 4, *batch size* de 2 e uma taxa de aprendizagem 0.0001 foi aquele que obteve melhor resultado. Com esta otimização o modelo previu que ocorreriam 12336 mortes por Pneumonia nos Estados Unidos na semana de 26-04-2021 a 2-05-2021. Através dos *plots* da evolução do *training loss* e *validation loss* podemos verificar, que ao contrário do que se passava nos modelos apresentados anteriormente, já não se verifica um enorme volatilidade nas curvas de *training loss* e *validation loss*. Isto deve-se ao facto de a melhor otimização obtida, contrariamente às melhores otimizações obtidas para os outros modelo, ter um *learning rate* mais baixo de 0.0001. Apesar disso é possível verificar através do *plot* de *loss* que o modelo se encontra em claro *underfitting* pois a *validation loss* apesar de ainda se encontrar em decréscimo é bastante superior à *training loss*. Para contrariar isto é necessário correr o mesmo modelo mas com mais épocas ou então aumentar um pouco o *learning rate* usado. Na tabela acima é possível verificar também que a otimização 7, apesar de apresentar um RMSE ligeiramente superior à 8, a diferença entre os valores de *loss* e *validation loss* é bastante inferior o que

parece suportar a sugestão de aumentando um pouco o *learning rate*, o modelo melhorará.

6.3 *Dataset* Mensal

6.3.1 CNN

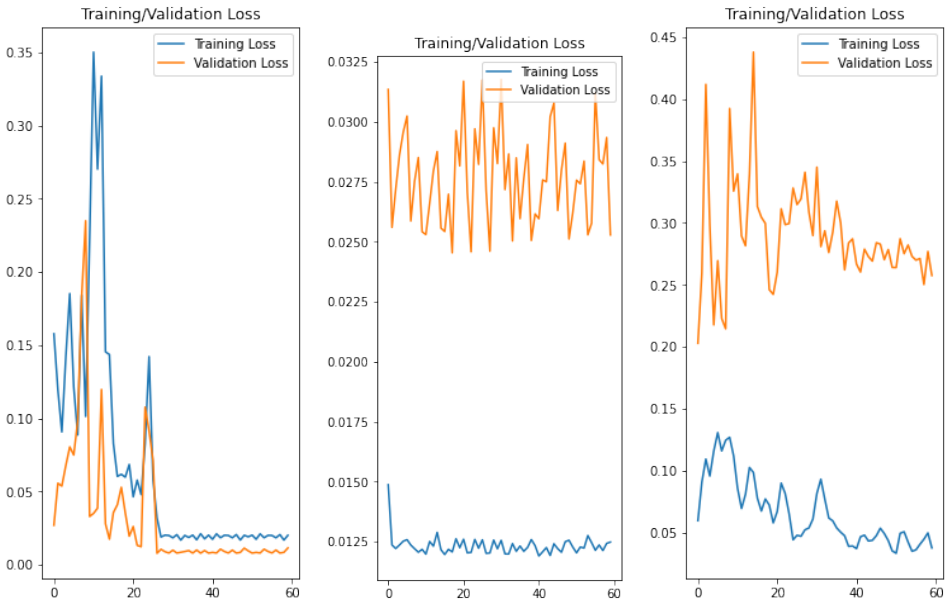

Figura 8: Evolução da Loss e Validation Loss observadas no melhor modelo CNN encontrado

	Loss	Val Loss	RMSE
1 Tuning	0.061841	0.098640	0.061841
2 Tuning	0.083244	0.200880	0.083191
3 Tuning	0.068085	0.203084	0.068085
4 Tuning	0.048834	0.116525	0.048834
5 Tuning	0.075497	0.143922	0.075439
6 Tuning	0.108152	0.202707	0.108152
7 Tuning	0.047317	0.148868	0.047317
8 Tuning	0.079366	0.162857	0.079357
9 Tuning	0.070490	0.190862	0.070490

Tabela 9: Valores médios das optimizações testadas no modelo CNN

Através dos resultados obtidos, com a medida de desempenho RMSE, o modelo que obteve melhor resultado, foi o modelo 4 com a otimização que utiliza 60 épocas, 30 *timesteps*, *batch size* de 1 e taxa de aprendizagem 0.001. Dada esta otimização, obtivemos uma previsão do modelo, para o mês 5 de 2021, de 52 mortes por COVID-19. Através dos gráficos da evolução do *training loss* e do *validation loss*, podemos verificar uma grande volatilidade nas curvas de *training loss* e *validation loss*. Esta volatilidade é muito relativa, provavelmente a um *learning rate* demasiado alto e/ou valores de *beta_1* e *beta_2* não ótimos. Foram feitos testes com modelos com *learning rate* inferior ao modelo que obteve o melhor resultado, mas estes obtiveram um pior RMSE do que o modelo escolhido. Como a descida do *learning rate* de modelo para modelo foi de 0.001 para 0.0001, um possível futuro passo a seguir é verificar da existência de um valor intermédio entre esse intervalo que obtenha melhores resultados.

6.3.2 LSTM

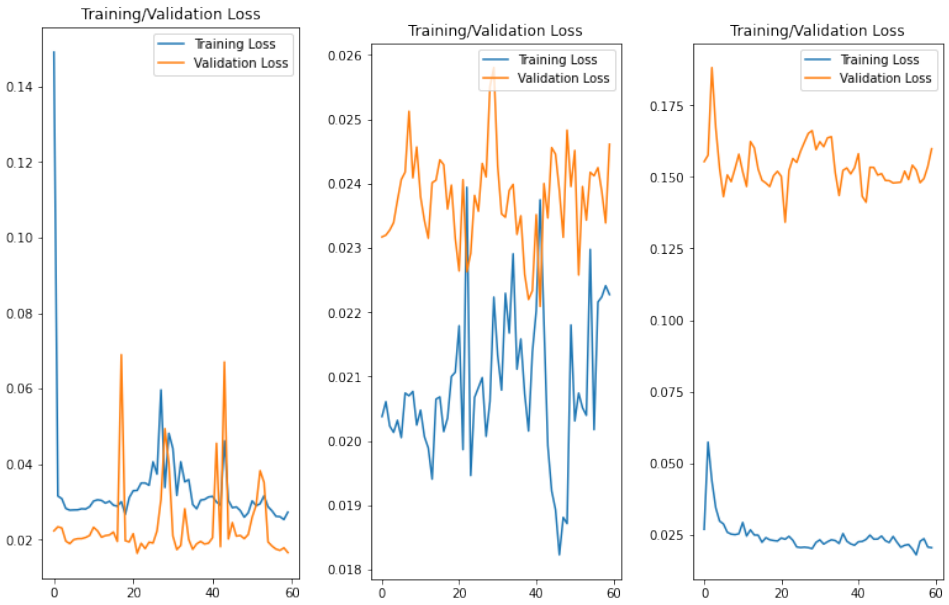

Figura 9: Evolução da Loss e Validation Loss observadas no melhor modelo LSTM encontrado

	Loss	Val Loss	RMSE
1 Tuning	0.029256	0.062958	0.029256
2 Tuning	0.042193	0.078947	0.042163
3 Tuning	0.129381	0.200988	0.129381
4 Tuning	0.026260	0.067147	0.026260
5 Tuning	0.035609	0.080651	0.035583
6 Tuning	0.104296	0.157153	0.104296
7 Tuning	0.026683	0.065414	0.026683
8 Tuning	0.035191	0.080621	0.035169
9 Tuning	0.114897	0.172908	0.114897

Tabela 10: Valores médios de cada uma das optimizações testadas no modelo LSTM

Através dos resultados obtidos, com a medida de desempenho RMSE, o modelo que obteve melhor resultado, foi o modelo 4 com a otimização que utiliza 60 épocas, 30 *timesteps*, *batch size* de 1 e taxa de aprendizagem 0.001. Dada esta otimização, obtivemos uma previsão do modelo, para o mês 5 de 2021, de 50 mortes por COVID-19. Através dos gráficos da evolução do *training loss* e do *validation loss*, podemos verificar uma grande volatilidade nas curvas de *training loss* e *validation loss*. Esta volatilidade é muito relativa, provavelmente a um *learning rate* demasiado alto e/ou valores de *beta_1* e *beta_2* não ótimos. Foram feitos testes com modelos com *learning rate* inferior ao modelo que obteve o melhor resultado, mas estes obtiveram um pior RMSE do que o modelo escolhido. Como a descida do *learning rate* de modelo para modelo foi de 0.001 para 0.0001, um possível futuro passo a seguir é verificar da existência de um valor intermédio entre esse intervalo que obtenha melhores resultados.

7 Conclusão

Com o objetivo de prever o número de óbitos associados a doenças respiratórias foi necessário construir *datasets* que contivessem *features* relevantes,e para isso foi necessária uma vasta e extensa pesquisa. Aliás, foi neste passo onde foi despendido mais tempo, uma vez que como a doença COVID-19 é uma doença recente, não existem muitos dados relativos a outras variáveis não relacionadas com esta doença e que podiam ser pertinentes para o auxilio do calculo do número de mortes.

Findada esta etapa, procedeu-se ao tratamento e exploração dos dados, e esta nova etapa foi muito importante para tomarmos decisões acerca das *features* a escolher e quais os tipos de modelos *deep learning LSTM* e *CNN* a implementar.

Estes modelos foram otimizados utilizando várias configurações de valores, como número de épocas, *timesteps*, taxa de aprendizagem e *batch size*, de modo a conseguirmos perceber qual a configuração que traz melhores valores de previsão para o modelo.

Nos modelos criados, obtivemos previsões de 23 e 24 mortes para o dia 23-04-2021, nos modelos diários CNN e LSTM, respetivamente, 106 e 390 mortes na semana de 26-04-2021 a 02-05-2021, nos modelos semanais, para o dataset do COVID-19, *LSTM* e *CNN*, respetivamente, e 52 e 50 mortes no mês 05-2021, nos modelos mensais *CNN* e *LSTM*, respetivamente. Relativamente ao *dataset* da pneumonia, obteve-se a previsão de 5720 e 12336 mortes por pneumonia nos Estados Unidos na semana de 26-04-2021 a 02-05-2021, para os modelos *LSTM* e *CNN*, respetivamente.

De uma forma geral é possível verificar que para os diversos modelos para as diferentes séries temporais, os modelos *LSTM* obtiveram melhor performance que os modelos *CNN*, como podemos verificar através dos seus valores de RMSE.

8 Referências

Comprehensive data exploration with Python https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

Training Loss and Validation in Deep Learning https://stackoverflow.com/questions/48226086/training-loss-and-validation-loss-in-deep-learning

How To Prevent Overfitting https://programming-review.com/machine-learning/overfitting