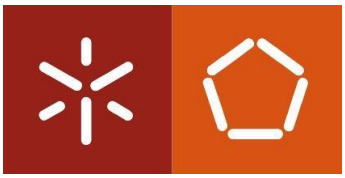

Universidade do Minho

Departamento de Informática

TP2 - Inteligência Ambiente e Sensorização

Bruno Martins (a80410) Catarina Machado (a81047)
Filipe Monteiro (a80229) Jéssica Lemos (a82061)

19 de Abril de 2020

Conteúdo

Resumo

O COVID-19 tem sido um problema crescente no mundo inteiro e o ritmo a que se dispersa pelo mundo aumenta a cada dia que passa. Tendo em conta todos os dados de cada país, da situação deste vírus, a nossa plataforma irá mostrá-la de forma simples ao utilizador, assim como tentar prever o futuro destes países. Nesta primeira fase apenas explicámos o que iremos fazer, como o pretendemos fazer, que API's utilizaremos e o que faremos com as informações recolhidas.

1 Introdução

Neste projeto foi nos pedido a criação de uma plataforma que, através de sensores físicos e virtuais para obtenção de dados, realize análises nestes dados e obtenha resultados, respondendo a algumas possíveis questões. No nosso caso, este projeto incidirá sobre a situação que vivemos actualmente da pandemia COVID-19. Esta plataforma irá utilizar dados provenientes de API's (sensores virtuais) sobre o estado de cada país em termos de mortes, número de infetados, histórico do estado do país tentar criar uma previsão futura do estado dos países.

2 Arquitetura

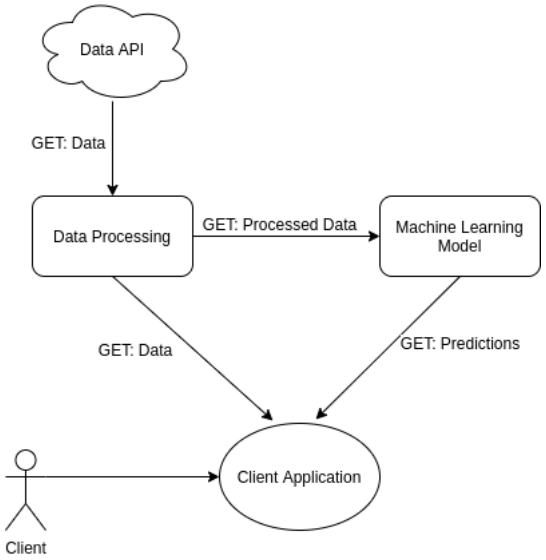

Figura 1: Arquitetura base do projeto.

3 Sensores Virtuais

3.1 *The virus tracker*

Para a obtenção dos dados relativos ao vírus COVID-19 será utilizada a API providenciada pelo site *thevirustracker.com*. Iremos utilizar três *endpoints*:

- Dados totais mundiais;
- Dados totais de país;
- Dados históricos de país.

Nas amostras de código seguintes podemos ver um exemplo de um pedido e a respostas em formato *json*.

```
OkHttpClient client = new OkHttpClient().newBuilder()
  .build();
Request request = new Request.Builder()
  .url("https://thevirustracker.com/free-api?countryTimeline=PT")
  .method("GET", null)
  .build();
```

```
Response response = client.newCall(request).execute();
```

Listing 1: Pedido para histórico da doença em Portugal


```
"countrytimelinedata": [
                {
                        "info": {
                                "ourid": 127,
                                "title": "Portugal",
                                "code": "PT",
                                "source": "https://thevirustracker.com"
                        }
                }
        ],
        "timelineitems": [
                          "3/2/2020": {
                                "new_daily_cases": 2,
                                "new_daily_deaths": 0,
                                "total_cases": 2,
                                "total_recoveries": 0,
                                "total_deaths": 0
                        },
                        "3/3/2020": {
                                "new_daily_cases": 0,
                                "new_daily_deaths": 0,
                                "total_cases": 2,
                                "total_recoveries": 0,
                                "total_deaths": 0
                        },
                        "3/4/2020": {
                                "new_daily_cases": 3,
                                "new_daily_deaths": 0,
                                "total_cases": 5,
                                "total_recoveries": 0,
                                "total_deaths": 0
                        },
                        "3/5/2020": {
                                "new_daily_cases": 3,
                                "new_daily_deaths": 0,
                                "total_cases": 8,
                                "total_recoveries": 0,
                                "total_deaths": 0
                        },
                        "3/6/2020": {
                                "new_daily_cases": 5,
                                "new_daily_deaths": 0,
                                "total_cases": 13,
                                "total_recoveries": 0,
                                "total_deaths": 0
                        }
        ]
```


Listing 2: Resposta ao pedido anterior (reduzido)

Uma informação importante de referir é que esta API é aberta ao público não necessitando de nenhuma chave para a utilização

4 Tratamento de dados

Como foi referido na secção anterior os dados são provenientes de uma API externa. Este facto faz com que nos pedidos feitos venha muita informação acessória não necessária, tanto para o modelo de aprendizagem como para a apresentação dos dados.

Para o tratamento dos dados foi desenvolvida uma API em *Go* que consiste em fazer pedidos à API *thevirustracker.com*, processar as respostas e disponibilizar *endpoints* para que os outros serviços da aplicação possam obter esses dados processados. Os *endpoints* criados foram os seguintes:

- */overallData* - Retorna os dados mundiais
- */countryData* - Retorna os dados de apenas um país, tem como argumento a abreviatura do país
- */countryHistory* - Retorna os dados históricos de um país, tem como argumento a abreviatura do país
- */countries* - Retorna todos os países disponíveis para fazer *querys* subsequentes

4.1 *JSON Schema*

A resposta aos pedidos efetuados é feita tendo em conta os seguintes esquemas:

4.1.1 Dados totais

```
{
  "$schema": "http://json-schema.org/draft-04/schema#",
  "type": "object",
  "properties": {
    "total": {
      "type": "integer"
    },
    "newToday": {
      "type": "integer"
    },
    "cured": {
      "type": "integer"
    },
    "deaths": {
      "type": "integer"
    }
  },
  "required": [
    "total",
    "newToday",
    "cured",
    "deaths"
  ]
}
```

Listing 3: Resposta ao pedido /overallData

4.1.2 Dados de país

```
{
  "$schema": "http://json-schema.org/draft-04/schema#",
  "type": "object",
  "properties": {
    "total": {
      "type": "integer"
    },
    "newToday": {
      "type": "integer"
    },
    "cured": {
      "type": "integer"
    },
    "deaths": {
      "type": "integer"
    },
    "critical": {
      "type": "integer"
    },
    "active": {
      "type": "integer"
    },
    "date": {
      "type": "string"
    }
  },
  "required": [
    "total",
    "newToday",
    "cured",
    "deaths",
    "critical",
    "active",
    "date"
  ]
}
```

Listing 4: Resposta ao pedido /countryData

4.1.3 Dados do histórico de um país

```
{
  "$schema": "http://json-schema.org/draft-04/schema#",
  "type": "array",
  "items": [
    {
      "type": "object",
      "properties": {
        "date": {
          "type": "object",
          "properties": {
            "new_daily_cases": {
              "type": "integer"
            },
            "new_daily_deaths": {
              "type": "integer"
            },
            "total_cases": {
              "type": "integer"
            },
            "total_deaths": {
              "type": "integer"
            },
            "total_recoveries": {
              "type": "integer"
            }
          },
          "required": [
            "new_daily_cases",
            "new_daily_deaths",
            "total_cases",
            "total_deaths",
            "total_recoveries"
          ]
        }
      },
      "required": [
        "date"
      ]
    }
  ]
}
```

Listing 5: Resposta ao pedido /countryHistory

5 Análise dos Dados

Como visto anteriormente, a plataforma irá tentar realizar previsões para mostrar aos utilizadores. Para isto, haverá num micro-serviço, uma pequena aplicação responsável por analisar os dados e criar previsões para a plataforma utilizar. Neste, iremos utilizar um conjunto pequeno de algoritmos *machine-learning* para realizar as previsões, desde simples regressões lineares a pequenas redes neuronais. Claro que iremos utilizar a que melhor desempenho tiver, ou uma combinação destes.

5.1 Prophet

Criado pelo Facebook como uma ferramenta de previsão automática para *python* e *R*, **Prophet** foi utilizado para fazer então as mesmas previsões que os restantes modelos, para analisar e comparar sobre que modelo teria a melhor *accuracy*. Esta foi uma boa escolha para este *dataset* porque este não apresenta qualquer tipo de "moda", sendo que o algoritmo foi desenhado para prever casos assim definidos. É uma ferramente simples de utilizar, num estilo *sklearn*, onde apenas temos de passar o *dataset* formatado em duas colunas - **date** e **y** (valor a prever) - e depois de treinado, pedir uma previsão do tamanho que queremos. Esta ferramenta é muito versátil, e percebemos que a sua capacidade de previsão, é muito elevada comparada com os outros, apresentando margens de confiança que fazem querer estimar confortavelmente os nossos valores.

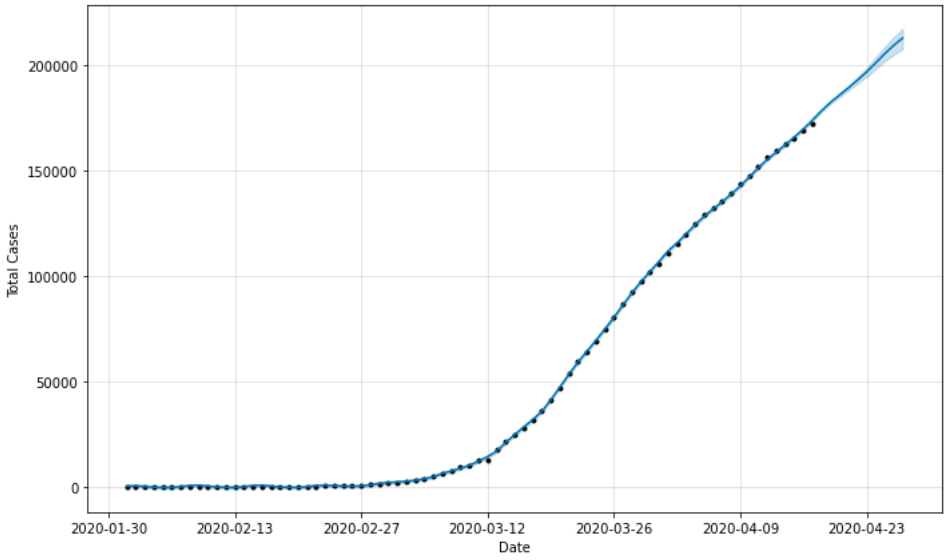

Figura 2: Previsão usando **Prophet**, para Itália, em 10 dias.

5.2 ARIMA

ARIMA é um tipo de modelo bastante utilizado no mundo da estatística, para previsão também de valores com teor temporal. Esta necessita de 3 campos para representar o modelo:

- p: número de dias utilizados para prever o próximo (*lag*);
- d: grau de diferenciação dos valores. Este serve para tentar estacionar o *dataset*, ou seja, os valores da *target* estabilizam e variam sempre da mesma forma, evitando assim a introdução de erros;
- q: valor do tamanho da *moving average*, ou seja, quantidade de pontos a usar para calcular esta métrica.

Este algoritmo é muito utilizado em séries temporais devido à sua facilidade de utilização e na sua capacidade de evitar *overfitting*. Após alguns testes, o conjunto de parâmetros que melhor resolveram o problema foram (5,1,0).

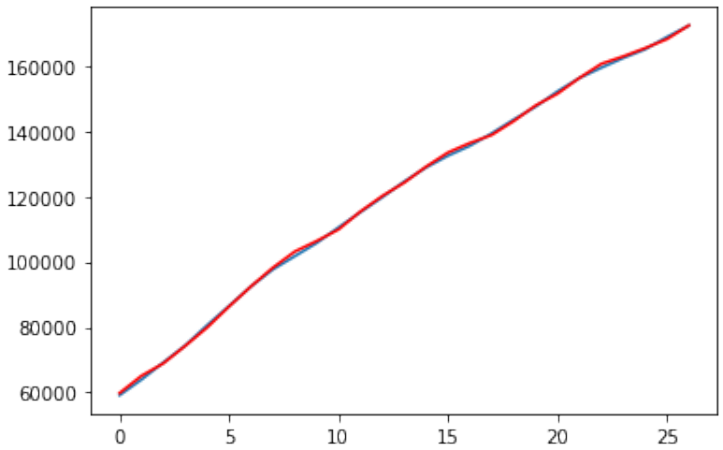

Figura 3: Treino do **ARIMA** para Itália. Verifica-se que a previsão acompanha a linha real de forma muito próxima.

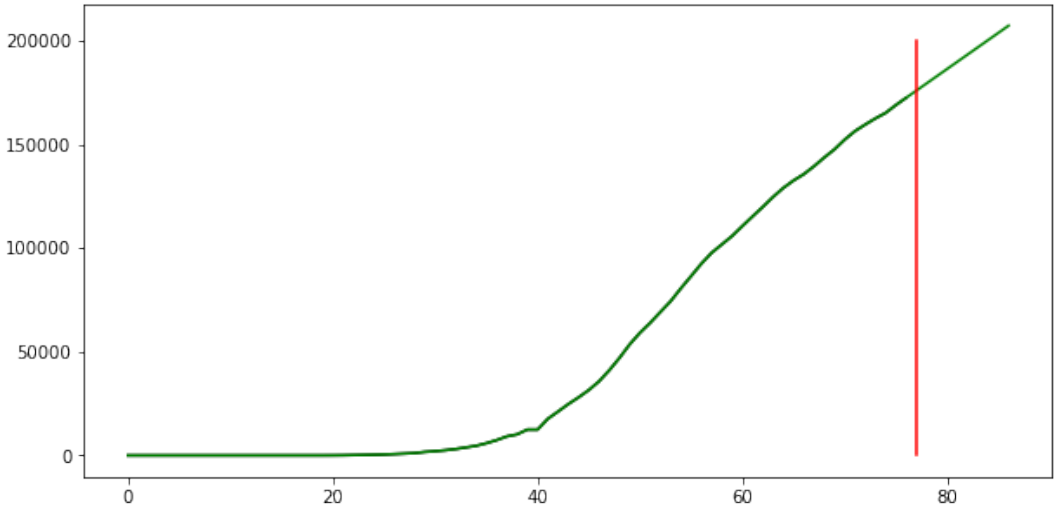

Figura 4: Previsão usando **ARIMA**, paraItália, em 10 dias. A previsão começa depois da linha vermelha.

5.3 Exponential Smoothening

Descoberto enquanto analisávamos a documentação de implementação do *ARIMA* acima referido, encontramos um outro tipo de modelo capaz de executar previsões: **Exponential Smoothening**. Este é um subtipo do anterior, e tem diversos tipos de submodelos, entre eles, a destacar, a simples *Exponential Smoothening* e a *Holt's Exponential Smoothening*. Entre elas, utilizámos a última pois foi a única que mostrou resultados reais, sendo esta uma versão mais restrita da primeira. Em relação ao *ARIMA*, este vai variando os pesos das variáveis dependentes para a previsão ao longo do tempo, mais concretamente diminuindo o peso de variáveis mais antigas. Isto permite que o algoritmo, apesar de olhar para o passado, prevê o próximos valores com mais atenção das variáveis mais recentes. Este ajuste dos pesos pode ser realizado por nós, mas, dado o problema, os resultados não variavam muito, sendo que ficou com os parâmetros *default*. Uma grande diferença deste modelo e o anterior referido, é que o anterior executa regressões automáticas nos dados, de forma a estacionar estes para previsões.

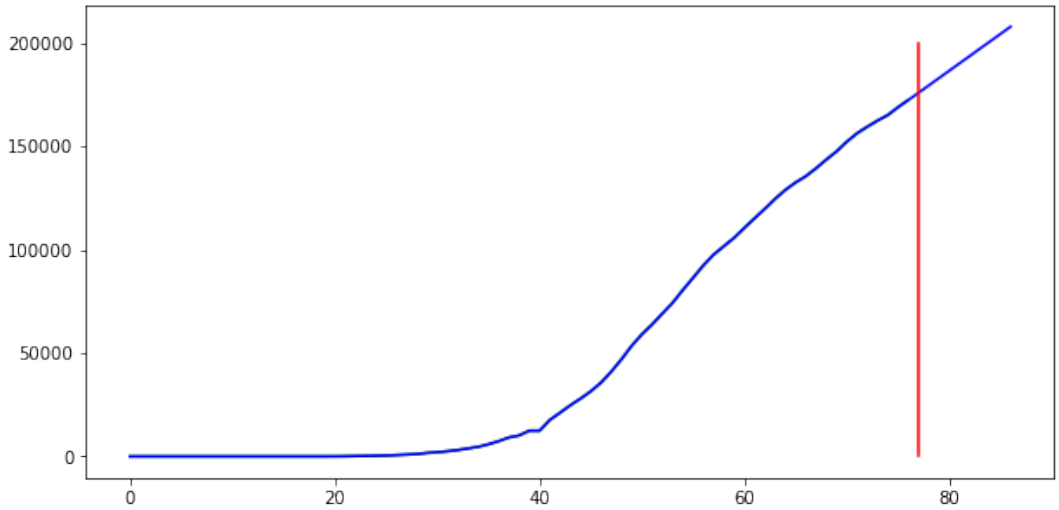

Figura 5: Previsão usando **Exponential Smoothening**, para Itália, em 10 dias. A previsão começa depois da linha vermelha.

Em termos de curva esta deu resultados muitos semelhantes à curva do *ARIMA*. Talvez com um maior quantidade de dados se verificasse uma diferença de maior valor.

5.4 LSTM

Dado que o nosso problema é de carácter temporal, **LSTM**'s são um outro tipo de modelo de *machine learning* muito utilizadas neste tipo de situação. A sua capacidade de encontrar padrões (redes neuronais) juntamente com a capacidade de guardar memória sobre os *inputs* (redes neuronais recorrentes) fazem desta uma forte candidata em dar ótimos resultados na análise do estado dos países no futuro. Inicialmente, planeávamos ter uma LSTM que, ajustada ao dados de um país, previa o futuro desse mesmo, contudo, verificámos que não era possível fazê-lo, pelo menos com resultados aceitáveis, pois os *datasets* entre eles eram diferentes e, o que funciona para um, não tem de necessariamente funcionar para outro. Por isso, optámos por criar dois tipos de **LSTM**'s:

- Modelo de previsão de um país, consoante o crescimento de outro;
- Modelo de previsão de alguns países mais sérios para o mundo.

No entando, não estava a ser fácil criar, por exemplo, um modelo para Itália. Com mais alguns testes e exploração da arquitetura da rede, descobrimos umas que já conseguia apresentar resultados aceitáveis para qualquer *dataset* utilizado: 3 camadas internas de 256 nodos *LSTM*. Aumentando mais o número de nodos/camadas demonstrou pior desempenho, e a diminuição caía no problema anterior falado, por isso decidimos ficar por esta. Nem com otimização do tamanho do *batch* ou a utilização de *LSTM*'s *stateful* o desempenho melhorou.

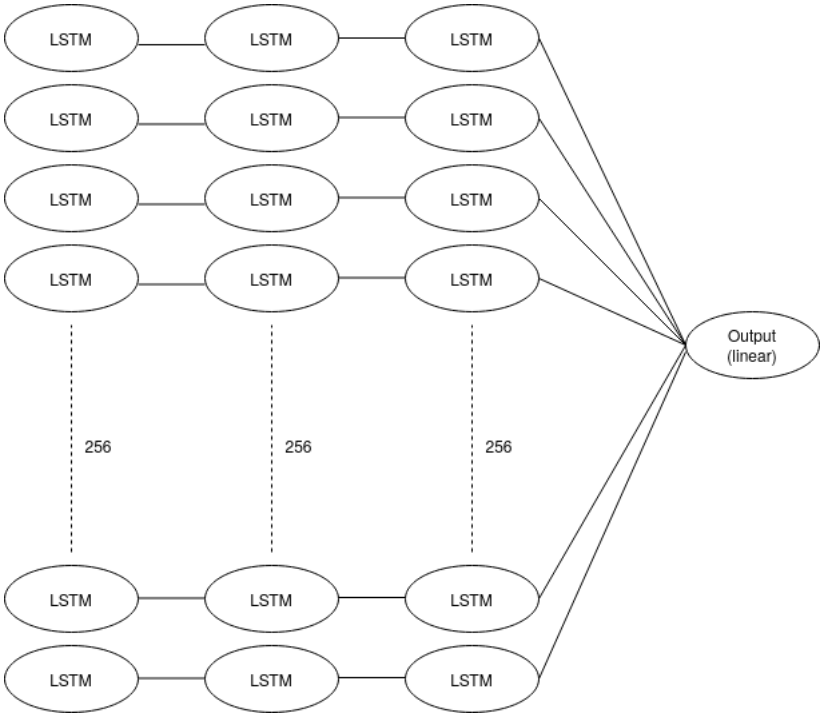

Figura 6: Arquitectura base da *LSTM*. Nas camadas internas utilizámos, respectivamente, um *dropout* de 0.2, 0.2 e 0.5 e um treino de 1000 *epochs*, com *early stopping* para evitar aumentos da *loss*.

Mas apesar dos resultados já mais correctos que apresentava, veremos mais à frente que estes estariam um pouco longe dos outros modelos - pela negativa.

Em relação à previsão de um país consoante o crescimento de outro, utilizando *LSTM*'s, o modelo gerado não era capaz de responder corretamente: por alguma razão, os valores estabilizavam sempre num determinado valor. É provável que seja possível criar este modelo, mas era necessário mais conhecimento e investigação nas *LSTM*'s.

5.5 Comparação entre Modelos

Apesar de mostrar-mos os 4 modelos na nossa plataforma, achámos por bem fazermos uma análise comparativa entre os resultados destas, percebendo qual aquele que de facto se aproxima mais a valores reais. Por isso, pegando no caso de Itália, truncámos os dados que tínhamos para ficar com 15 dias para previsão e executámos os 4 modelos a ver que soluções apresentavam. Estes foram os resultados obtidos:

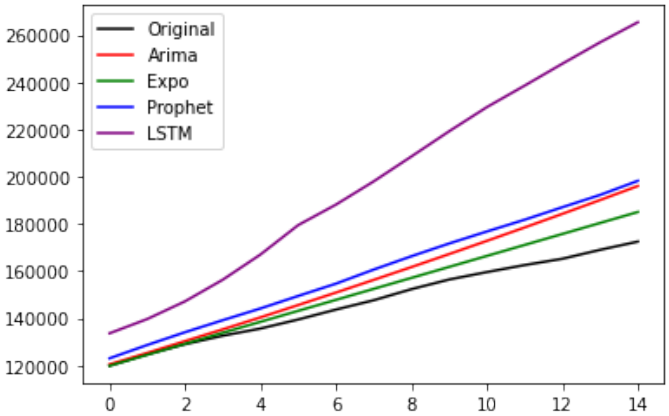

Figura 7: Comparação entre os diferentes modelos para uma previsão de 15 dias no futuro, no número total de casos.

Todos os modelos apresentam resultados que podem ser interpretados como "o pior dos casos", ou seja, se nada for feito durante este espaço de tempo. Mas claro que não é o que se verifica. Interessante foi reparar que as *LSTM's*, que inicialmente achava ser o melhor modelo para esta previsão, foi o que deu pior resultados, exagerando nos valores. Em contra partida, todos os outros saíram-se de forma semelhante, tendo sido o *Prophet* o único que apresentou um crescimento não linear. Outro modelo importante a analisar é o *Exponential Smoothening*, que apresentou os resultados mais próximos dos originais, ou seja, entre todos, foi o que "exagerou"menos.

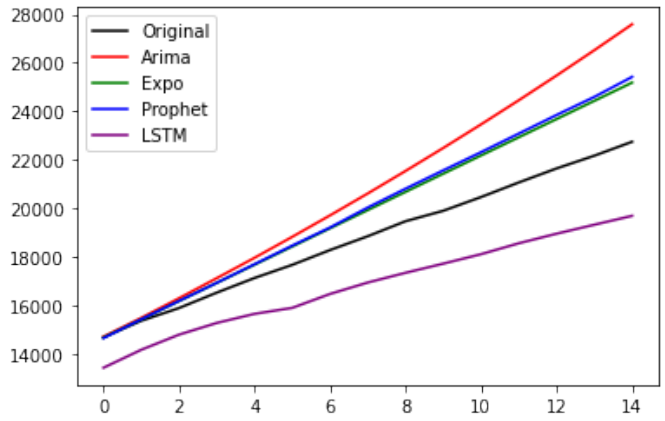

Figura 8: Comparação entre os diferentes modelos para uma previsão de 15 dias no futuro, no número total de mortes.

5.6 Deploy da *ML API*

Neste momento a API responsável pelas previsões é corrida localmente e, para cada chamada, os dados são recolhidos da *API base*, são treinados os modelos e só depois é entregue a resposta. Como tal as respostas aos pedidos demoram imenso tempo (especialmente com o uso da *LSTM*). Como tal, numa situação ideal, os dados iriam ser guardados numa base de dados e seria executado diariamente uma chamada à *API base* para atualizar os dados e, de igual forma, os modelos iriam ser treinados para todos os países diariamente, evitando assim o compasso de espera que temos atualmente. Isto não foi implementado por falta de tempo e acharmos mais importante focarmos-nos na implementação crua, deixando as otimizações para mais tarde.

6 Plataforma COVID-19 World Analyser

Esta plataforma tem como principal objetivo expor informações relevantes através de uma interface simples e intuitiva. Esta é constituída por quatro partes distintas, contudo complementares. O utilizador tem ao seu dispor uma listagem dos vários países que estão a combater esta epidemia e o número de infetados correspondentes. Selecionando um país terá ao seu dispor diversas informações através dos mais variados gráficos tais como:

- Número total de infetados por dia;
- Número total de mortos por dia;
- Número total de recuperados por dia;
- Número de novos casos por dia;
- Número de novas mortes por dia.

Nestes gráficos para além de ser possível observar os dados existentes até ao momento podemos ainda analisar quais as previsões obtidas através dos diferentes algoritmos de *machine learning* apresentados anteriormente para 7, 15 ou 30 dias.

O utilizador poderá ainda consultar os dados globais tais como o número total de infetados, o aumento verificado, o número de recuperados e o número de mortos. Para além disso, poderão ser selecionados vários países em simultâneo com o intuito de se comparar a situação nos mesmos. O utilizador tem ainda a possibilidade de visualizar o estado da epidemia no mundo consultando o mapa do mesmo que representa a incidência da epidemia em cada zona.

6.1 Desenvolvimento do Frontend

O interface gráfica deste serviço foi criada utilizando HTML, Bootstrap 4 e Javascript (jQuery) e pode ser consultada a qualquer momento através de https://covid-19-world-analyser.netlify.com. No entanto, como a API apenas corre localmente, não é possível visualizar a página completa através desse link (terá que ser localmente).

Desta forma, quando se acede à página, o resultado obtido encontra-se na figura seguinte. É de frisar que o gráfico apresenta os dados do país em que o utilizador que está a aceder ao mesmo se encontra (informação recolhida através do ipinfo.io) e que a tabela inferior apresenta o cumulativo dos dados de todos os países.

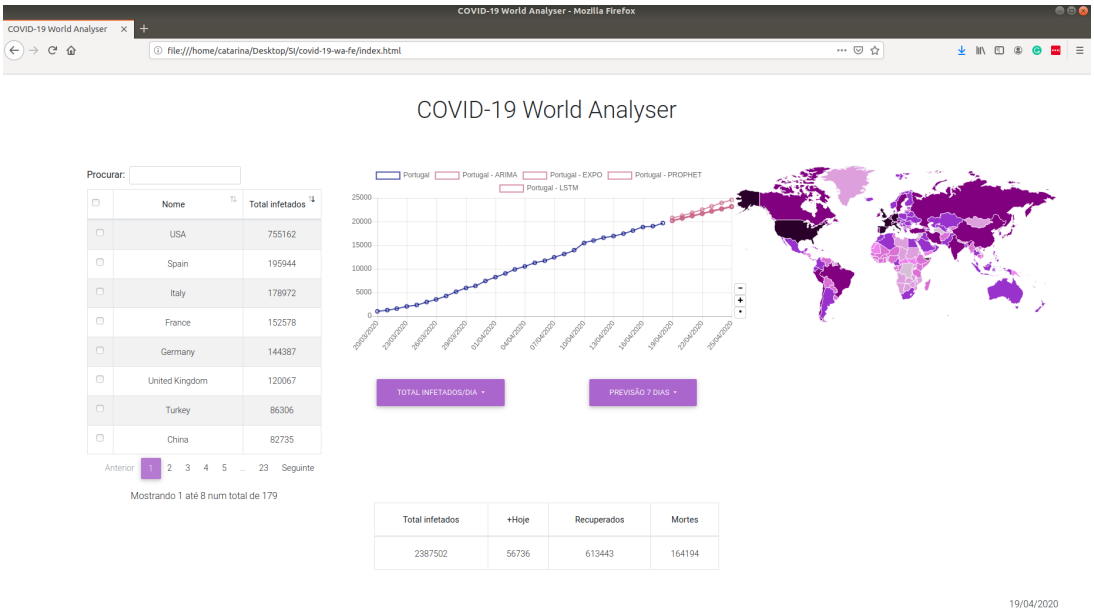

Figura 9: Interface gráfica.

Tal como já mencionado, a interface gráfica é constituída por quatro diferentes componentes e, em seguida, apresenta-se detalhadamente cada uma deles.

6.1.1 Tabela dos Países

A tabela, presente no lado esquerdo da interface gráfica, começa por estar ordenada decrescentemente pelo número de infetados dos países, sendo possível ordenar também crescentemente ou através do nome do país. São apresentados 8 países em cada página da tabela, sendo possível consultar outros avançando pelas páginas. É também possível pesquisar qualquer nome de um país através da caixa de texto superior (tal como se pode ver na Figura 10).

A *checklist* tem como objetivo permitir a seleção de vários países em simultâneo de forma a que os respetivos dados possam ser comparados. Selecionando a checklist, o utilizador pode consultar os dados desse determinado país (gráfico e tabela dos dados globais) (Figura 11) ou selecionando vários consulta o cumulativo de todos eles (Figura 12).

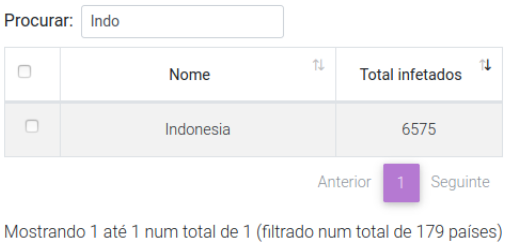

Figura 10: Procura de país.

6.1.2 Tabela dos dados globais

Na parte mais inferior da página, encontra-se uma tabela que apresenta o número total de infetados, o aumento de infetados diário, o número total de recuperados e o número total de mortes. Assim que se acede à página, quando não está nenhum país selecionado, a tabela apresenta os dados de todos os países (tal como se pode ver na Figura 9). Selecionando um país, apresenta somente os dados desse país (Figura 11) e, selecionado vários países, apresenta a soma desses dados (Figura 12).

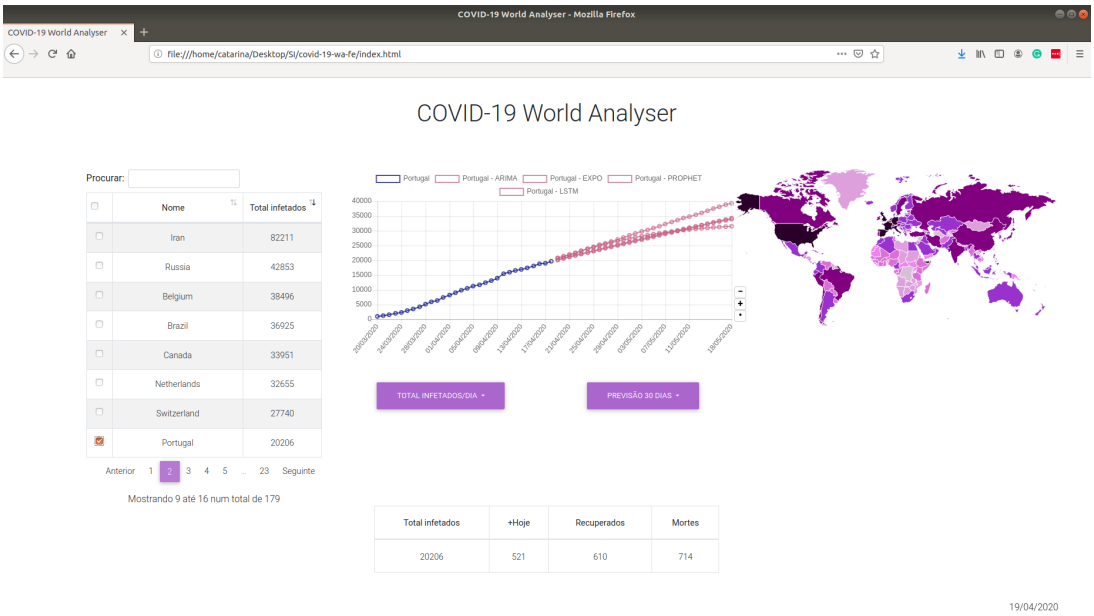

Figura 11: Portugal.

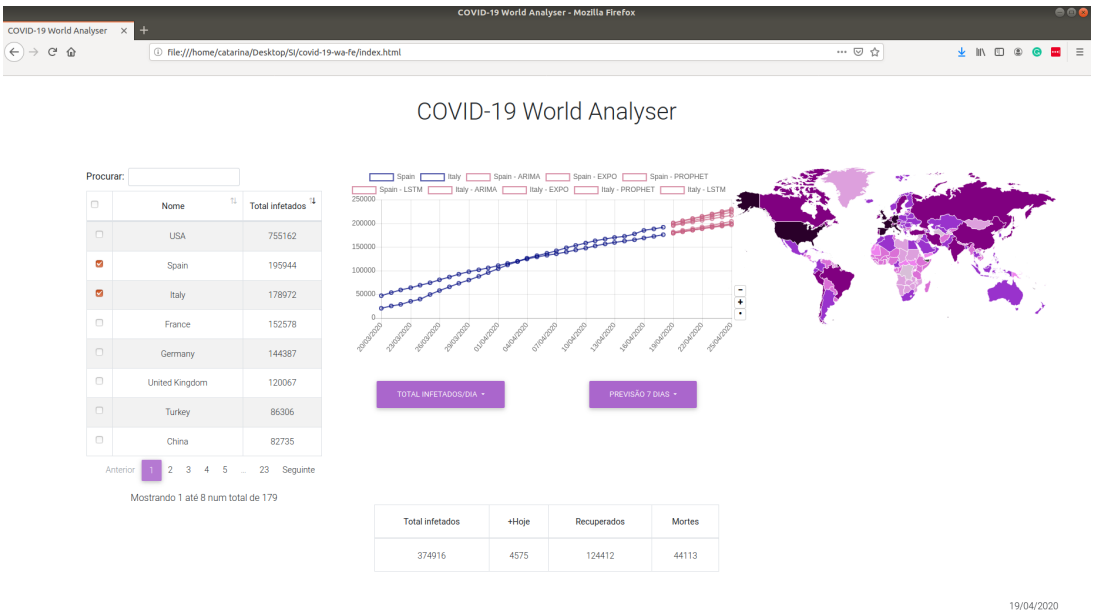

Figura 12: Itália e Espanha.

6.1.3 Gráficos

Tal como já explicado, quando se acede à página o gráfico apresenta os dados do país onde se encontra o utilizador que está a aceder à página (Figura 9) e podem ser selecionados um ou mais em simultâneo através da tabela do lado esquerdo.

Por omissão, o gráfico apresentado corresponde ao **Total infetados/dia** e a previsão é de **7 dias**. No entanto, o utilizador pode optar por selecionar outro tipo de gráfico e outro número de dias de previsão (explicado na Secção 6). Para tal, basta selecionar o botão correspondente ao tipo de gráfico (botão mais à esquerda) e ao número de dias (botão mais à direita), respetivamente, tal como se pode ver através da Figura 13 e 14.

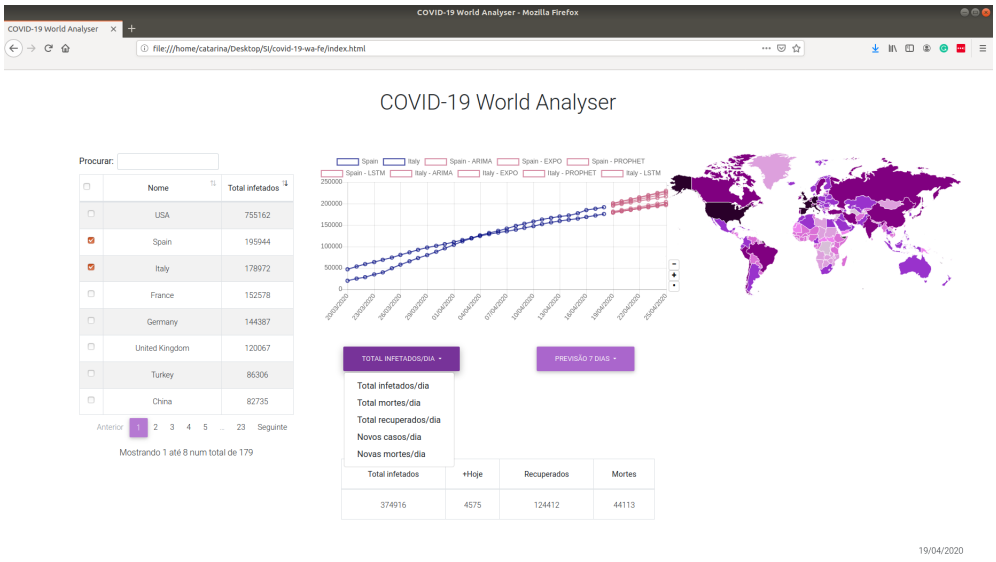

Figura 13: Botão de selecção de tipo de gráfico.

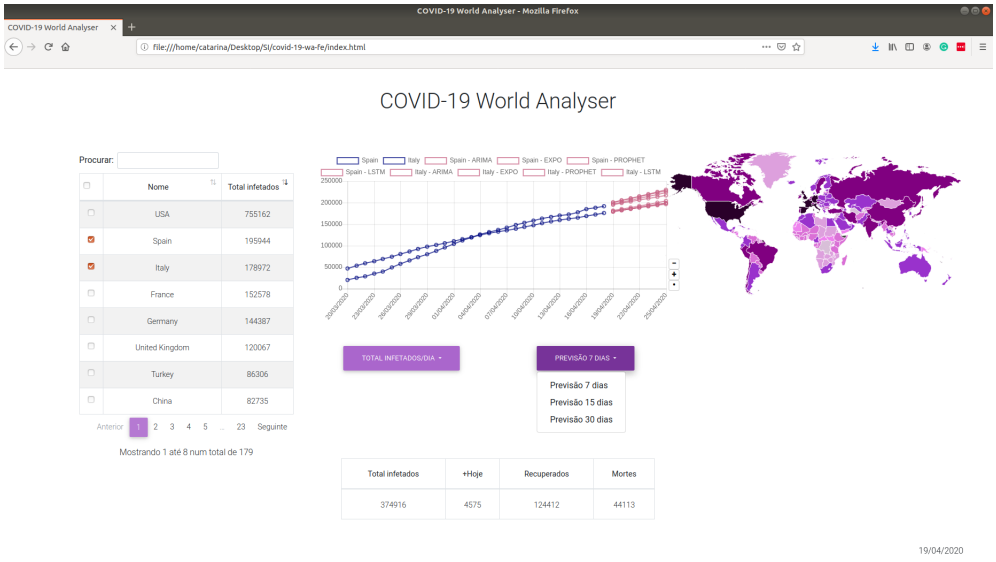

Figura 14: Botão de selecção de número de dias.

Para consultar com mais pormenor o número exato do eixo dos x e y (data e número total de infetados, respetivamente) e a informação do que a linha representa (nome do país e o nome do algoritmo aplicado para a previsão) basta passar o rato pelo círculo desejado na linha, tal como apresentado na figura seguinte.

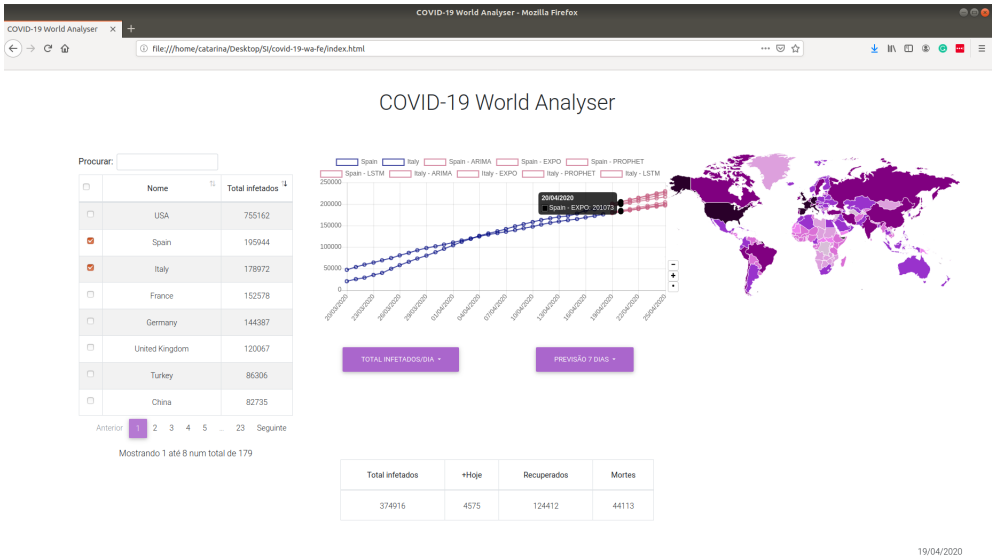

Figura 15: Consultar informação da linha.

Apenas a título demonstrativo, nas figuras seguintes encontram-se, respetivamente, os gráficos correspondentes a **30 dias de previsão do número total de infetados em Portugal** e o gráfico correspondente ao número de **7 dias de previsão do número de novos casos diários em Portugal**.

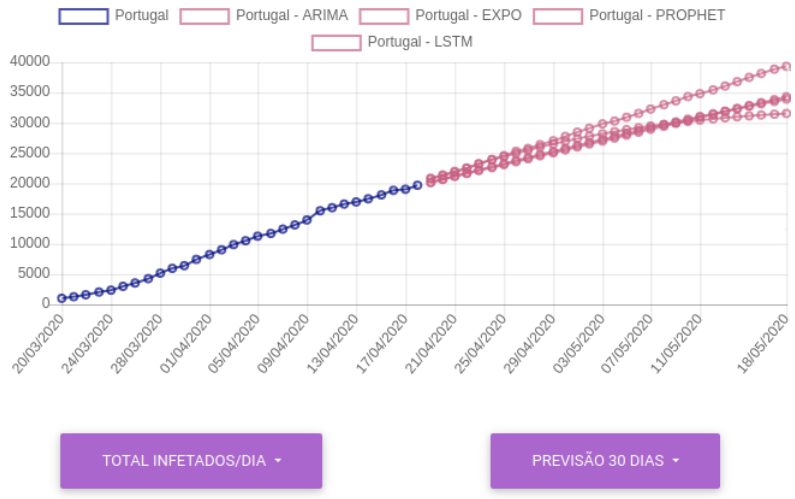

Figura 16: 30 dias de previsão do número total de infetados em Portugal.

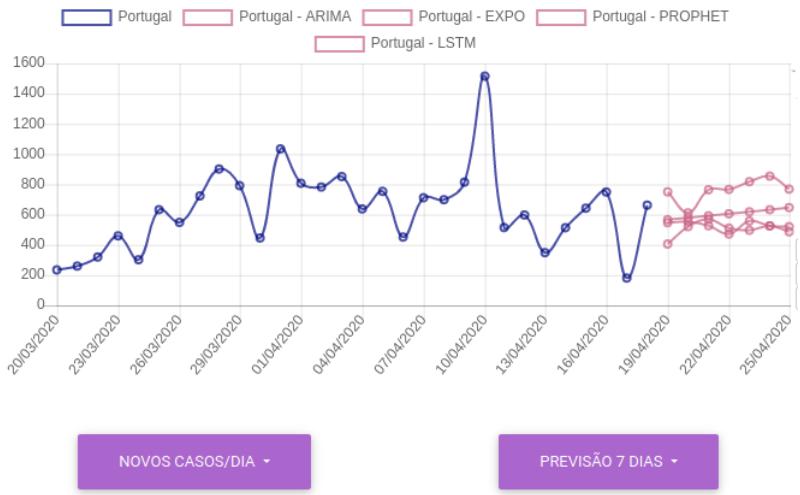

Figura 17: 7 dias de previsão do número de novos casos diários em Portugal.

6.1.4 Mapa Mundo

Com o intuito de se obter uma perspectiva mais geral do estado da pandemia optamos por apresentar um mapa mundo que permita de forma instantânea perceber a situação atual em cada um dos países.

Para a implementação do mapa recorremos ao *plugin jQuery Mapael* que permite exibir mapas vetoriais dinâmicos. Desta forma, exibimos o mapa mundo de uma forma bastante simples, sendo possível fazer zoom e explorar o mapa conforme o pretendido.

De modo a apresentar o estado da pandemia através de uma legenda que tem em conta o número de infetados por COVID-19 é possível visualizar o mapa com cores, ou seja, a gravidade em cada país. Assim sendo, quanto maior for o número de casos existentes mais escura será a cor do país apresentada no mapa. Para além disso, é possível consultar o número exato de infetados, casos recuperados e mortes em cada ponto do mundo tal como podemos verificar na Figura 18. Esta informação reflete-se no mapa usando a API já referida anteriormente.

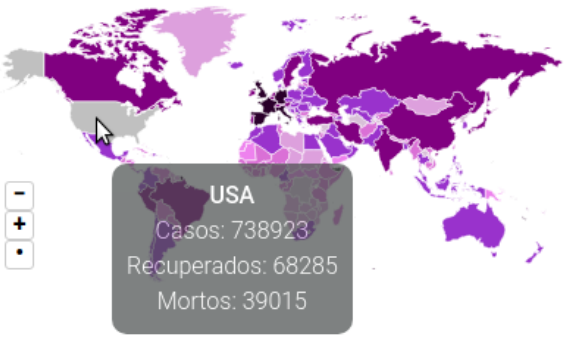

Figura 18: Mapa Mundo.

6.1.5 Data

A data do presente dia é apresentada automaticamente no canto inferior direito da página.

7 Conclusão

O tratamento de dados vindos de sensores físicos/virtuais é algo poderoso que pode ser usado em vários aspectos da nossa vida. Neste caso, usando uma API e a atual calamidade que o mundo atravessa, conseguimos criar uma plataforma que, não só mostra o estado atual de vários países, em vários aspectos, mas também tenta responder à pergunta "o que nos espera no futuro?".

Devido à quantidade de dados atuais, estas previsões são dadas num ponto de vista do pior que pode acontecer - apesar de se saber que os números são maiores e mais assustadores, apenas não se sabe quanto exatamente - , sendo que assim, se assim virmos, talvez estas previsões retratem de momento uma situação mais real do que se acha. De todos os modelos, alguns são mais corretos que outros, mas infelizmente a nossa *LSTM* não foi desenvolvida com conhecimento suficiente para produzir resultados convincentes. Talvez seja falha nossa, mas os dados também diríamos serem culpados. Mas esperemos que nunca cheguemos ao que as nossas previsões nos dizem.

Por fim, algumas das grandes dificuldades enfrentadas durante o desenvolvimento foram as falhas aleatórias da API usada. Sem sabermos, esta API estava a necessitar de apoios para se manter online, sendo que não sabemos quanto tempo irá permanecer *online* e/ou aberta ao público para uso. Para além disto, numa fase inicial, esta foi sofrendo mudanças nas respostas fornecidas, sendo que o uso do nosso lado foram feitos *updates* até ao final do desenvolvimento.