3.1 Contexto dos dados

Os dados utilizados na análise a que nos propusemos tratam-se de dados gerais relativos ao COVID-19, mais especificamente, o desenvolvimento do vírus no contexto nacional. São fornecidos pela Direção Geração de Saúde e colocados no repositório podendo ser encontrados no seguinte link ( <https://github.com/dssg-pt/covid19pt-data> ). Devemos também ter em consideração que a equipa por trás do repositório está empenhada em atualizar os dados que lhes é fornecido diariamente no mesmo. O ficheiro está estrutura de forma a facilitar a manipulação dos dados em Excel e Python devendo os utilizadores, ter em consideração que a possibilidade de surgir novas colunas está em aberto, algo que pode influenciar o estudo dos dados efetuados previamente.

Finalmente, queremos realçar que apesar da precisão dos dados fornecidos pela Direção Geral de Saúde estejam a melhorar com o tempo, estes não representam com total exatidão a situação nacional com que lidamos, considerando que alguns fatores para que isto se verifique estão completamente fora do controlo da DGS.

3.2 Recolha de dados

Os dados presentes no repositório são, como já referimos, fornecidos pela Direção Geral de Saúde sendo atualizados de acordo com a evolução diária da doença respiratória sendo que a recolha dos mesmos por parte da DGS está parcialmente limitada visto que muitos casos positivos da doença não são detetados dado que a população não é testada na sua totalidade e apenas em caso de suspeita prévia. Isto é algo prejudicial tanto para a recolha de dados como para a evolução do vírus porque não é possível ter uma perceção absoluta daquilo que realmente se passa e em vez disso, observamos números parciais dos ativos, óbitos e recuperados por exemplo estendendo-se isto a todos os outros fatores causados pelo vírus.

3.3 Preparação de dados

Visto que as complexidades dos dados fornecidos na versão original eram excessivas para a análise que pretendemos efetuar, decidimos filtrar os dados para obter apenas o que consideramos essencial para refletir sobre o objetivo do trabalho. Esta filtragem foi feita no ficheiro Jupyter, onde, através da data frames originais, reduzimos o número de colunas deixando apenas os essenciais ao trabalho, sendo estes os data frames que utilizámos para a realização do trabalho

3.4 Exploração de dados

Com os data frames preparados para análise, decidimos definir o que iriamos de facto analisar com a informação recolhida. Dado que o tema do nosso trabalho se foca na região norte do país, decidimos fazer uma análise da evolução do vírus no distrito do Porto, mas considerando que isto não seria suficiente, achámos benéfico fazer uma comparação entre o crescimento da doença entre o distrito e o panorama que observamos no contexto nacional. Para elaborar um pouco mais o nosso estudo, decidimos observar algumas taxas que podíamos observar, a distribuição do vírus relativamente ao número de confirmados e para finalizar, observar a tendência de crescimento da doença utilizando variáveis especificas.

3.5 Modelização de dados

Para iniciar a análise foi necessária uma prévia manipulação dos dados de modo a que o estudo dos mesmo seja adequado para chegar às conclusões a que nos propusemos.

Modelizámos os dados de forma a facilitar a sua análise e os modelos que achamos essenciais para o seguimento do trabalho foram um heatmap, um piechart, análise da regressão linear, análise de clusters e como um extra com menor, mas relevante importância, algumas taxas que consideramos importantes. Aqui está o resultado desta modelização:

3.6 Avaliação

Através da análise observámos uma amplitude de diversos resultados. Pudemos observar que a nível nacional a taxa de óbitos causados pela doença é de 1,51% dos confirmados e a taxa de recuperação é de 69,11% o que nos indica que apesar de se tratar de uma doença extremamente infeciosa, a grande maioria dos habitantes portugueses diagnosticados com a doença acabam por recuperar. Infelizmente, isto não se verifica em todo o mundo, e, apesar de Portugal não ser o país com a menor taxa de óbitos, está relativamente bem posicionado devido ao alto nível profissional no ramo da saúde e aos métodos de tratamento adotados pelos profissionais de saúde no combate à doença a nível nacional.

Ao analisar o Heatmap podemos observar a correlação entre diferentes variáveis e, podemos observar que as variáveis que mais se correlacionam são confirmados e recuperados e as que menos se correlacionam são suspeitos e ativos, o que faz sentido porque todos os recuperados tiveram de ser previamente confirmados com a posse do vírus e no caso oposto todos os ativos são indivíduos diagnosticados com o vírus, logo não há qualquer suspeita, dado que existe a confirmação de infeção. Consideramos o Heatmap algo interessante de analisar apesar de não ser possível tirar grandes conclusões através da sua utilização.

O nosso PieChart permite-nos observar claramente quais os concelhos mais afetados dentro do distrito do Porto, sendo os três concelhos mais afetados o do Porto verificando 15% dos ativos do distrito, seguido por Vila Nova de Gaia demonstrando 14% dos infetados e por fim Matosinhos com 10%. Os concelhos que verificam menor propagação do vírus são Marco de Canaveses e Amarante onde ambos verificam 2% e os menos afetados, Baião e Trofa com 1%.

Esta análise permite-nos entender que os concelhos mais afetados são aqueles que observam maior densidade populacional, oque faz todo sentido visto que se trata de um vírus com muita capacidade de propagação. [[-0.00987277 0.01640735] [-0.99014072 0.98359227]], a média de quadrados dos erros obtida foi 124094,39 e o score, ou seja, o coeficiente de determinação R^2 da previsão foi 0.95.

Com a análise de clusters que consideramos pouco relevante para o objetivo final do nosso trabalho, experimentamos diversos métodos, mas, aquele que consideramos principal é a utilização das duas variáveis que contém os concelhos mais afetados do distrito do Porto e agrupar os dados dado o seu grau de semelhança. Os resultados obtidos não foram os mais relevantes para a realização do relatório.

© Jorge Balula