**Universidade dos Açores**

**Licenciatura em Informática (9384)**

**Laboratório de Ciência de Dados**

**RELATÓRIO FINAL**

Pedro Sousa ([2019101451@uac.pt](mailto:2019101451@uac.pt))

Salif Faustino ([20172005@uac.pt](mailto:20172005@uac.pt))

**Resumo**

Este trabalho foi realizado no âmbito da disciplina de Laboratório e ciências de dados e teve como objetivo a analise e exploração de dados. Este projeto pretende descrever de forma sucinta os resultados obtidos no processo de analise e exploração de dados sobre o momento pandémico que atravessamos concretamente em Brasil.

**BREVE INTRODUÇÃO**

Na cadeira de Laboratório e ciências de dados, do 3º Ano da Licenciatura em Informática, foi-nos proposto a interpretação e implementação de algoritmos aprendidos durante o semestre, utilizando dados escolhidos por nós. Os dados escolhidos nesse trabalho correspondem aos registos acerca do vírus cov-19 no ano de 2020, entre os meses de março e junho. O objetivo desse trabalho é através dos algoritmos estudados durante o semestre recolher informação que possa ser útil e identificar padrões que possam ser uteis para o cliente (pessoas que não percebem nada sobre algoritmos). Com base nesse objetivo iremos explorar e interpretar todos os resultados obtidos.

A versão final do projeto encontra-se disponível no github, através do seguinte link:

[https://github.com/SalifNTC/ Projeto-analise-de-dados](https://github.com/SalifNTC/projeto-2-g10)

1. **SUMÁRIO EXECUTIVO**

Com a situação pandémica que o mundo atravessa, os alunos resolveram fazer uma investigação e através dela recolher dados sobre o número de casos no brasil no ano de 2020, entre os meses de março e junho.

Cada registo contém a informação do **ID, CONTADOR, DATA NOTIFICAÇÃO CLASSIFICAÇÃO, SEXO, IDADE, FAIXA ETÁRIA, MUNICIPIO\_RESIDENCIA\_COD, MUNICIPIO RESIDENCIA, COMORBIDADE, INTERNACAO, UTI, DATA\_ATUALIZACAO**. Com base nessa informação dos dados o aluno tem o intuito de aplicar os conhecimentos adquiridos ao longo do semestre, identificar padrões ou relações que possam ser uteis na discussão.

As adversidades que surgiram no trabalho foram são várias. Os dados apresentam uma estrutura complexa e os alunos entenderam que poderiam ter outros campos para permitir uma análise mais detalhada. Os dados também apresentam um défice de atributos com valores numéricos, esses fatores contribuíram para que houve uma maior dificuldade na identificação de padrões.

De um modo geral os objetivos os objetivos do trabalho foram parcialmente cumpridos, ficando por limar alguns gráficos e conclusões em relação a aplicabilidade dos algoritmos. Todavia, os conhecimentos adquiridos durante as aulas foram todos aplicados de forma adequada, onde também exploraram os conhecimentos para que houvesse um maior enriquecimento do trabalho.

1. **COMPREENSÃO DOS DADOS E PRÉ-PROCESSAMENTO** 
   1. **. Descrição dos dados**

Com a situação pandémica que o mundo atravessa, os alunos resolveram fazer uma investigação e através dela recolher dados sobre o número de casos no brasil no ano de 2020, entre os meses de março e junho.

**2.2. Descrição dos dados na tabela dos dados**

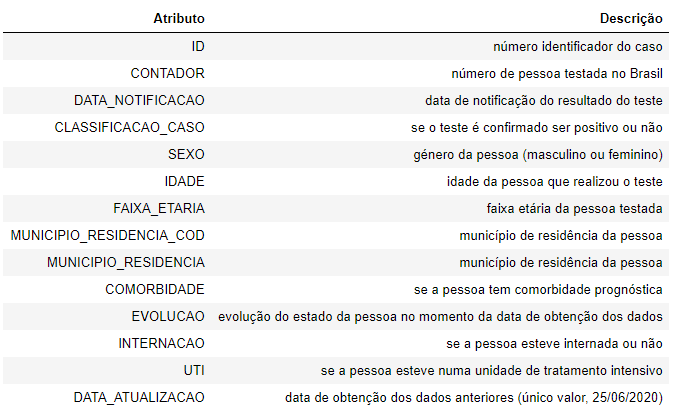
****

Figura 1- Descrição dos dados na tabela de dados

**2.3 Escalas de medida**

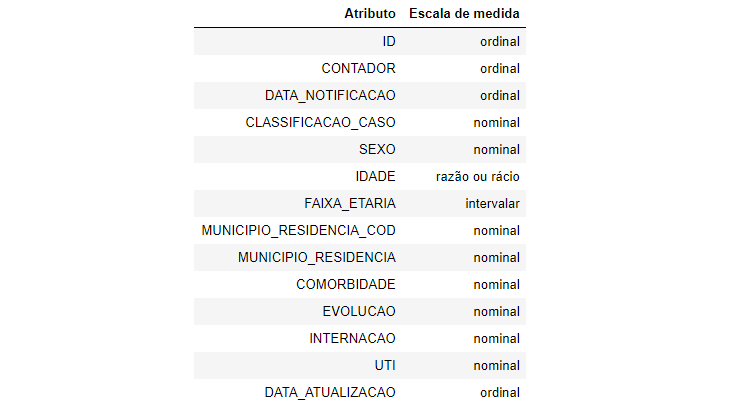
****

Figura 2- Escalas de medida dos dados da tabela

Conforme é possível perceber na **Figura 1** acima, os atributos apresentam uma variedade no que diz respeito ao tipo de atributos (numéricos, caracteres), embora haja um défice de atributos número que possam ser relevantes para aplicabilidade dos algoritmos na regressão linear.

Na **Figura 2** acima, é possível perceber que os atributos apresentam uma variedade do que diz respeito as escalas de media (escala nominal, escala ordinal e escala métrica), o que irá facilitar na escolha dos atributos para aplicar os algoritmos.

De um modo geral, os atributos apresentam uma boa variedade de dados, todavia alguns atributos irão precisar de algum tratamento para facilitar a aplicabilidade dos diversos algoritmos.

**2.5. ANÁLISE EXPLORATÓRIA DOS DADOS´**

O objetivo desse trabalho é através dos algoritmos estudados durante o semestre recolher informação que possa ser útil e identificar padrões que possam ser uteis para o cliente (pessoas que não percebem nada sobre algoritmos). Para este efeito, foram aplicados conhecimentos e técnicas de aprendizagem usadas na área da Ciência de Dados, que foram estudados ao longo do semestre.

Para que houvesse uma boa prática dos conhecimentos optamos por tratar alguns atributos, onde tivemos que ter em conta o tipo de aprendizagem e os algoritmos utilizados. Para algoritmos que apenas aceitam atributos numéricos com por exemplo o algoritmo de regressão linear foi escolhido o atributo *target* a **IDADE,** que representa a idade de cada individuo que realizou o teste.

Este atributo foi escolhido com objetivo que verificar que relação pode existir entre A IDADE e outros atributos, com por exemplo a possibilidade de um individuo com uma idade elevada que testou positivo a CVD-19 ser internada, obtido, fazer parte dos cuidados intensivos, etc...

Para algoritmos de aprendizagem que permitem atributos categóricos como target, como por exemplo as árvores de classificação e modelos de classificação, foi escolhido o atributo INTERNACAO\_SIM como o atributo *target,* que resultado de um tratamento (***dummys***) e representa se um determinado individuo foi/não, foi internado.

**2.5.1. DESCRIÇÃO DOS DADOS**

Para identificar melhor os atributos e obter-se uma informação mais detalhada dos mesmos forma aplicados algumas funções disponibilizadas, tais como o ***describe***. Essas Informações são uteis para saber tipo de atributo, variedade das informações dos atributos, etc…

**2.5.1.1. Descrição dos dados numéricos**

****

Figura 3- Discrição dos dados númericos

**2.5.1.2. Descrição dos dados não numéricos**

****

Figura 4- Discrição dos dados não númericos

A informação apresentada na **Figura 3 e 4** permite recolher uma serie de informações bastante uteis, como o número de registos, número de vezes que um determinado valor aparece em cada coluna e o tipo de atributo.

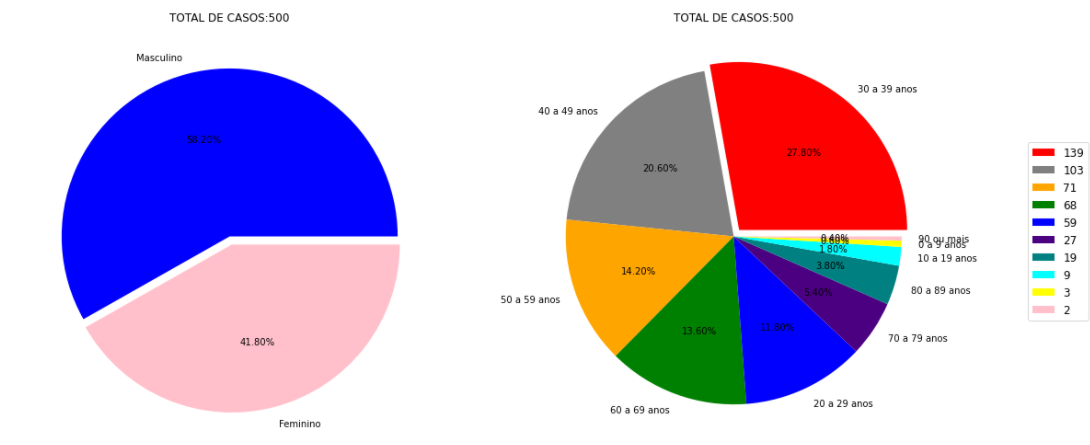


Figura 5- Taxa de infeção por sexo e por grupo etário

Na **Figura 5** acima, no primeiro Gráfico indicamos as percentagens das pessoas infetadas com o coronavírus, onde podemos observar que os indivíduos do sexo masculino apresentam maior percentagem de infeção em relação aos indivíduos do sexo feminino.

No segundo gráfico indicamos as percentagens de pessoas infetadas por grupo etário, onde podemos observar que o grupo compreendido entre as idades dos 30 aos 39 apresenta maior percentagem de infeção

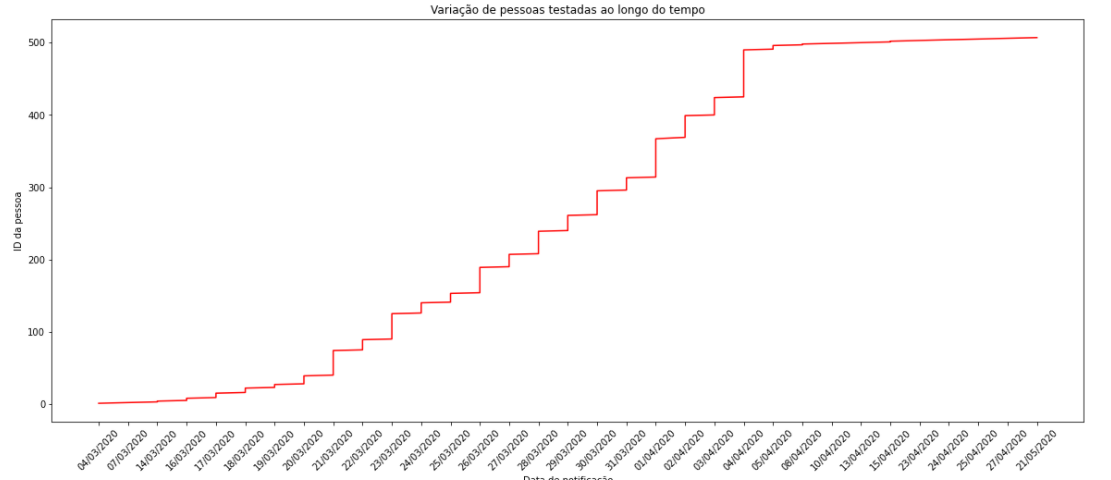


Figura 6 - Número de pessoas testadas ao logo do tempo.

Na **Figura 6** acima, gráfico mostra o número de pessoas testadas ao logo do tempo, onde podemos observar que o crescimento é exponencial, ou seja, a medida que o tempo passa mais testes são realizados devido ao aumento de pessoas infetadas, é importante realçar que chega uma dada altura que o crescimento dos casos deixa de ser exponencial, ficando constante e essa informação reflete-se na **Figura 6** acima conforme podemos ver.

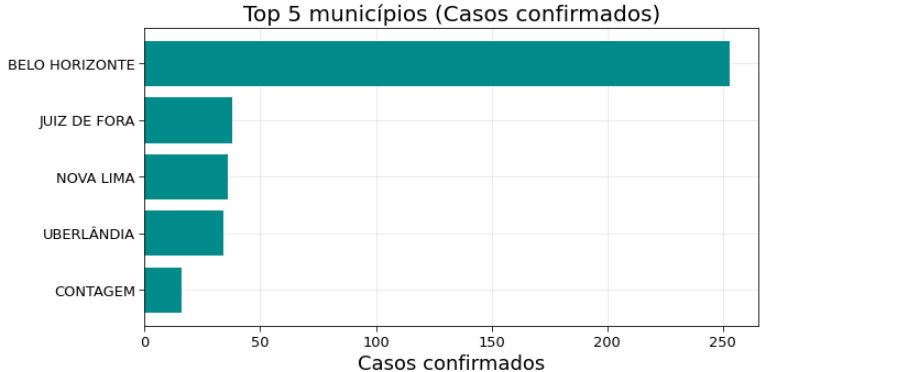


Figura 7- Top 5 Municípios (Casos Confirmados)

Na **Figura 7** acima, o gráfico mostra o top 5 das regiões com maior número de casos, onde o município de BELO HORIZONTE apresenta maior número de casos.

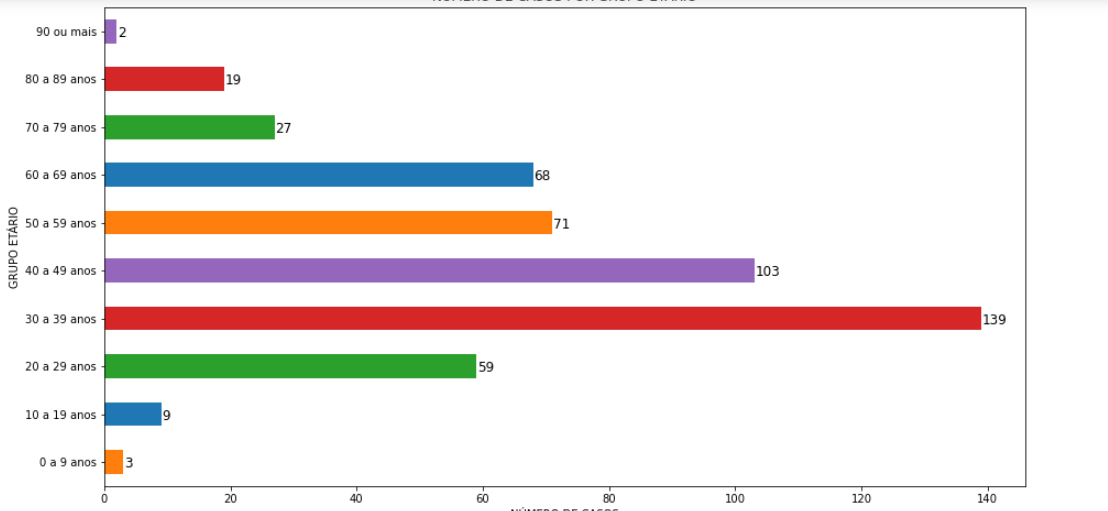


Figura 8 - Número de casos por grupo etário

Na **Figura 8** acima, gráfico mostra o número de casos registados por faixa etária, podemos observar que o grupo com as idades compreendidas entre os 30 e os 39 apresentam um maior número de casos em relação às demais faixas etárias e o grupo com as idades compreendidas entre os 0 e 9 apresenta um menor número de casos.

É importante salientar que apesar dos grupos etários com idades superiores a 60 tiverem o menor número de casos pertencem ao grupo de que tem maior possibilidade de serem internados.

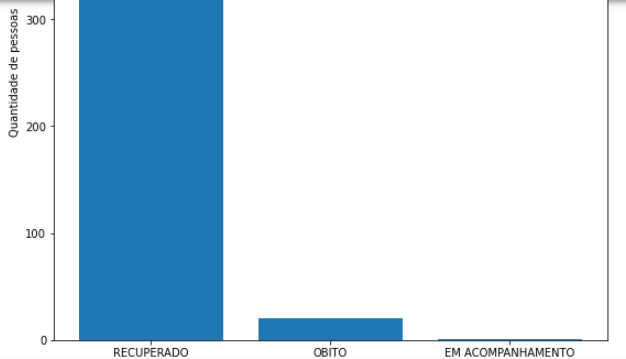
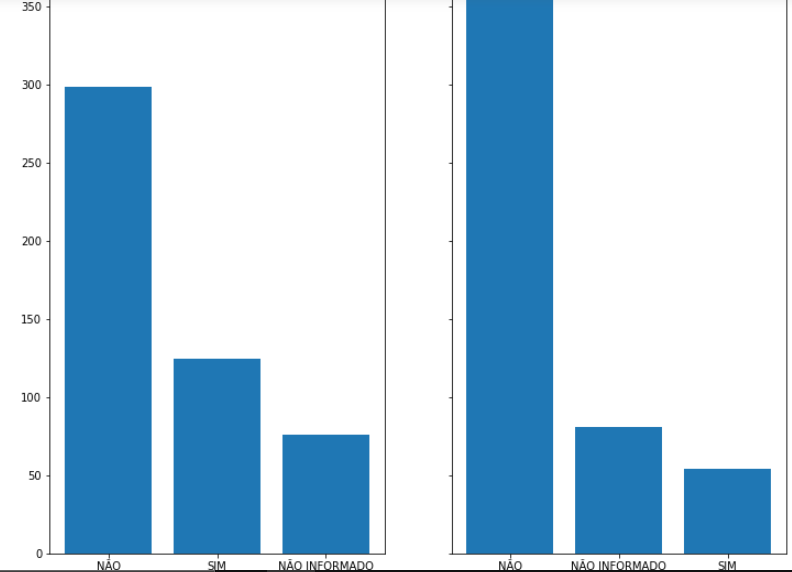


Figura 9- Evolução do vírus

Na **Figura 9** acima, o gráfico a direita compara a quantidade de pessoas das quais se sabe ou não (ou não foi informado) que estiveram internadas com as que estiveram nos cuidados intensivos e o gráfico a esquerda ilustra a evolução dos vírus, onde podemos observar que temos um número de recuperados bastante considerável quando comparado com os óbitos e casos em acompanhamento.

**2.6. Qualidade dos dados**

De um modo geral podemos considerar que os dados apresentam uma boa qualidade embora seja preciso fazer um pré-processamento conforme iremos ver nas seções …

**2.6.1 Identificação de outliers**

Os principais outliers foram detectados durante a aplicação da regressão multivariada. Como a

diferença consequente à remoção de alguns outliers se revelou pouco significante para a qualidade

do modelo de regressão, decidimos manter os dados originais. Como o modelo aprende sobre

20475 instâncias, a remoção de algumas baseadas no gráfico de influência terá pouco importância.

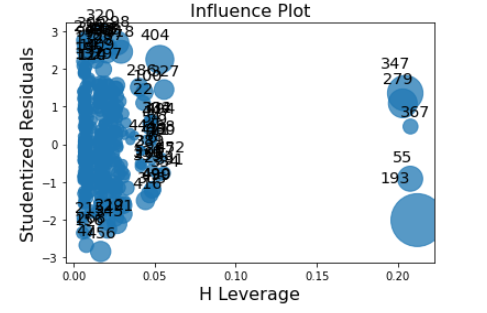


Figura 10 - Gráfico de influência

**2.6.2 Pré-processamento**

Os atributos por vezes apresentaram informações pouco relevantes para a aprendizagem na nossa perspetiva, onde optamos por excluir dos mesmos (exemplo DATA\_NOTIFICACAO, MUNICIPIO\_RESIDENCIA). Estes atributos foram excluídos por possuírem diversos valores distintos (data de notificação possui 33 valor distintos e o município de residência possui 58 valores destintos)

Nesta seção também foram criados dummys para os atributos com valores de um domínio inferior a 4, após criar os mesmos foram eliminados os respetivos atributos “originais”. A Figura **11** abaixo ilustra o resultado obtido.

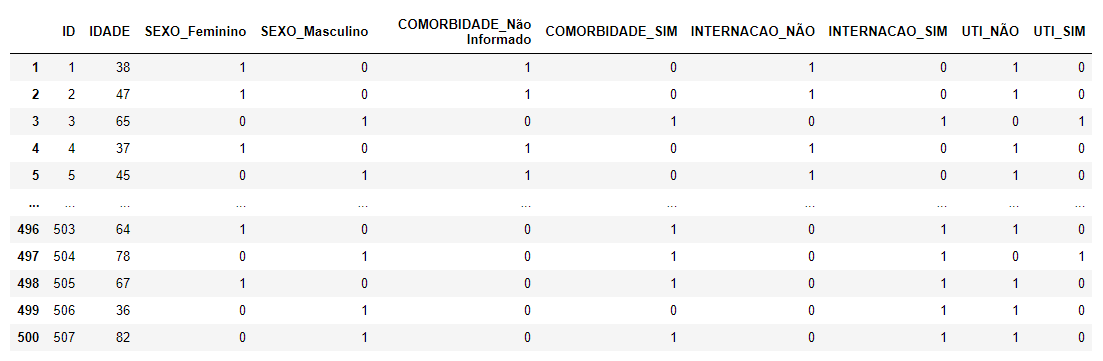


Figura - Criação de dummys