# Grupo 1 - Classificação e Associação

**Edilaine L. Pena[1], Guilherme S. Patricio[2], Igor Sant'Anna Siqueira José das Neves[3], João Vitor de Moura Rosa[4]**

Instituto de Ciências Exatas e Aplicadas – Universidade Federal de Ouro Preto (UFOP)
Cep 35931-008 – João Monlevade – MG – Brazil

## 1. Introdução

Neste trabalho prático implementamos scripts em Python para realizar as tarefas de classificação e associação, usando uma base de dados SRAG 2020 disponibilizada pelo governo. Para realização do mesmo foi feito um pré-processamento para remover dados que consideramos redundantes ou não necessários.

## 2. Algoritmo de classificação

Para que fosse realizada a classificação foi utilizado o algoritmo KNN visto em sala de aula, utilizando os 20 vizinhos mais próximos, além disso a base de dados foi dividida em um conjunto de testes com o tamanho igual a 33% do tamanho total, e um conjunto de treino com o tamanho de 66% do tamanho total.

Foram escolhidos dois Targets para o nosso projeto, a evolução do caso, que representa que fim teve um determinado paciente, e a classificação final do caso, que mostra o diagnóstico final do caso.

➢ Na primeira análise de dados, utilizamos alguns sintomas comuns do COVID-19 para serem classificados conforme a evolução do caso, sendo o valor 1 dado para cura e 2 dado para óbito. Utilizando as métricas FEBRE, TOSSE, DISPNEIA e DESC_RESP, que significam febre, tosse, dispnéia e desconforto respiratório respectivamente, obtivemos os seguintes resultados:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.64 | 0.93 | 0.76 | 320 |
| 2 | 0.62 | 0.17 | 0.27 | 202 |
| accuracy | | | 0.64 | 522 |
| macro avg | 0.63 | 0.55 | 0.52 | 522 |
| weighted avg | 0.64 | 0.64 | 0.57 | 522 |

**Figura 1. Resultado da classificação com Features igual a sintomas e Target igual a Evolução do caso.**

Precision: Exatidão das previsões positivas.

Recall: Fração de positivos que foram identificados corretamente.

F1-score: Média harmônica entre precision e recall

Support: Número de ocorrências reais da classe no conjunto de dados especificado.

A acurácia dessa classificação foi de 63%, o que significa que de acordo com a amostra dos dados presentes na nossa base de dados, não é possível classificar com precisão se o paciente irá viver ou morrer, apenas com a combinação dos sintomas.

Utilizando o mesmo conjunto de atributos que representam os sintomas, fizemos a classificação para o target classificação final do caso, onde o valor 4 é dado para Síndrome respiratória aguda grave não especificada, e o valor 5 é dado para Síndrome respiratória aguda grave por Covid-19, foram obtidos os seguintes resultados:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 4 | 0.29 | 0.02 | 0.04 | 164 |
| 5 | 0.69 | 0.97 | 0.80 | 358 |
| accuracy | | | 0.67 | 522 |
| macro avg | 0.49 | 0.50 | 0.42 | 522 |
| weighted avg | 0.56 | 0.67 | 0.57 | 522 |

**Figura 2. Resultado da classificação com Features igual a sintomas e Target classificação final do caso.**

A acurácia dessa classificação foi de 67%, o que significa que de acordo com a amostra dos dados presentes na nossa base de dados, não é possível classificar com precisão se o paciente terá Covid ou uma outra síndrome respiratória, apenas verificando a combinação dos sintomas.

➢ A segunda análise foi feita utilizando todos os fatores de risco presentes na base de dados, denominados pelas siglas: PUERPERA, CARDIOPATI, HEMATOLOGI, SIND_DOWN , HEPATICA, ASMA, DIABETES, NEUROLOGIC, PNEUMOPATI, IMUNODEPRE, RENAL e OBESIDADE, que significam respectivamente mulher que pariu recentemente, Doença Cardiovascular, Doença Hematológica Crônica, Síndrome de Down, Doença Hepática Crônica, Asma, Diabetes, Doença Neurológica, pneumopatia crônica, Imunodeficiência, Doença Renal Crônica e Obesidade.

Utilizando esses atributos como Features e a evolução do caso como Target os resultados obtidos foram:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.62 | 0.97 | 0.76 | 320 |
| 2 | 0.54 | 0.06 | 0.12 | 202 |
| accuracy | | | 0.62 | 522 |
| macro avg | 0.58 | 0.51 | 0.44 | 522 |
| weighted avg | 0.59 | 0.62 | 0.51 | 522 |

**Figura 3. Resultado da classificação com Features igual ao fator de risco e Target evolução do caso.**

Totalizando uma acurácia de apenas 61%, o que significa que de acordo com a amostra dos dados presentes na nossa base de dados, não é possível classificar com precisão se o paciente irá viver ou morrer, apenas com a combinação dos fatores de risco.

➢ A terceira análise foi realizada apenas com o atributo idade, e seus resultados foram:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.66 | 0.85 | 0.74 | 320 |
| 2 | 0.57 | 0.31 | 0.40 | 202 |
| accuracy | | | 0.64 | 522 |
| macro avg | 0.61 | 0.58 | 0.57 | 522 |
| weighted avg | 0.63 | 0.64 | 0.61 | 522 |

**Figura 4. Resultado da classificação com Features igual a idade e Target evolução do caso.**

Totalizando uma acurácia de apenas 64%, que significa que de acordo com a amostra dos dados presentes na nossa base de dados, não é possível classificar com precisão se o paciente irá viver ou morrer, levando em conta apenas a sua idade.

Utilizando o atributo idade para a classificação final do caso, foram obtidos os seguintes resultados:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 4 | 0.60 | 0.11 | 0.19 | 164 |
| 5 | 0.70 | 0.97 | 0.81 | 358 |
| accuracy | | | 0.70 | 522 |
| macro avg | 0.65 | 0.54 | 0.50 | 522 |
| weighted avg | 0.67 | 0.70 | 0.62 | 522 |

**Figura 5. Resultado da classificação com Features igual a idade e Target classificação final do caso.**

Totalizando uma acurácia de 70%, o que significa que de acordo com a amostra dos dados presentes na nossa base de dados, não é possível classificar com precisão se o paciente terá Covid ou uma outra síndrome respiratória, apenas pela sua idade.

➢ Os Features que retornaram uma melhor acurácia para o Target Classificador final do caso foram: RES_IGG, RES_IGM, PCR_RESUL, SUPORT_VEN, que significam respectivamente Resultado da Sorologia para SARS-CoV-2 hemoglobina IGG, e Resultado da Sorologia para SARS-CoV-2 hemoglobina IGM, testes moleculares (RT-PCR), feitos a partir da coleta de mucosa do nariz e da garganta e suporte ventilatório. Os resultados obtidos por eles foram:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 4 | 0.89 | 0.93 | 0.91 | 164 |
| 5 | 0.97 | 0.95 | 0.96 | 358 |
| accuracy | | | 0.94 | 522 |
| macro avg | 0.93 | 0.94 | 0.93 | 522 |
| weighted avg | 0.94 | 0.94 | 0.94 | 522 |

**Figura 6. Resultado da classificação com Features igual a RES_IGG, RES_IGM, PCR_RESUL, SUPORT_VEN e Target classificação final do caso.**

Totalizando uma acurácia de 94%, o que significa que com esses atributos é possível classificar com uma boa precisão se o paciente terá Covid ou se ele terá uma outra síndrome respiratória usando apenas com esses atributos.

➢ Não foram encontrados atributos que geram uma alta acurácia(acima de 90%) para classificar o Target evolução final do caso.

## ➢ 3. Algoritmo de associação

Para o segundo algoritmo de análise de dados, utilizamos o método Apriori visto durante a disciplina, como visto em sala o algoritmo funciona buscando associações ou itemsets. Para se formar estes itemsets primeiro foi feito uma removido da tabela os itens os campos e deixado apenas os sintomas e evolução. Para denominá los foi utilizado já os nomes do dicionário passado para guiar o uso, que no caso foram utilizados as métricas: EVOLUCAO, FEBRE, TOSSE, DISPNEIA, FADIGA, DESC_RESP, PERD_OLFT e PERD_PALA, como demonstrado no dicionário para os sintomas eram atribuído 1 para sim, 2 para nao e 9 para ignorado.

Parte do algoritmo necessita que tenhamos valores binários para verdadeiro e falso como média utilizamos os campos 1 e 9 para 1 e verdadeiro e 2 para 0 e falso.

| | EVOLUCAO | FEBRE | TOSSE | DISPNEIA | FADIGA | DESC_RESP | PERD_OLFT | PERD_PALA |
|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **1** | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| **2** | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| **3** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **4** | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **1576** | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| **1577** | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| **1578** | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| **1579** | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| **1580** | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |

**Figura 7. Resultado da remoção de dados que nao seriam utilizados.**

Após essa etapa os valores são convertidos para booleanos e define o threshold (hiperparâmetro) precisamos calcular o support de todas as combinações de itens e extrair um subconjunto de itens frequentes, por padrão definimos por 0,5 e removemos os itemsets com o valor de support menor que 0,5.

| | antecedents | consequents | support | confidence |
|---|---|---|---|---|
| 0 | (EVOLUCAO) | (TOSSE) | 0.470588 | 0.752275 |
| 1 | (TOSSE) | (EVOLUCAO) | 0.470588 | 0.672087 |
| 2 | (EVOLUCAO) | (DISPNEIA) | 0.445920 | 0.712841 |
| 3 | (DISPNEIA) | (EVOLUCAO) | 0.445920 | 0.597458 |
| 4 | (TOSSE) | (FEBRE) | 0.472486 | 0.674797 |
| 5 | (FEBRE) | (TOSSE) | 0.472486 | 0.798930 |
| 6 | (FEBRE) | (DISPNEIA) | 0.457306 | 0.773262 |
| 7 | (DISPNEIA) | (FEBRE) | 0.457306 | 0.612712 |
| 8 | (DESC_RESP) | (FEBRE) | 0.405440 | 0.620523 |
| 9 | (FEBRE) | (DESC_RESP) | 0.405440 | 0.685561 |
| 10 | (TOSSE) | (DISPNEIA) | 0.540164 | 0.771454 |
| 11 | (DISPNEIA) | (TOSSE) | 0.540164 | 0.723729 |
| 12 | (TOSSE) | (DESC_RESP) | 0.481973 | 0.688347 |
| 13 | (DESC_RESP) | (TOSSE) | 0.481973 | 0.737657 |
| 14 | (DESC_RESP) | (DISPNEIA) | 0.552815 | 0.846079 |
| 15 | (DISPNEIA) | (DESC_RESP) | 0.552815 | 0.740678 |
| 16 | (TOSSE, DESC_RESP) | (DISPNEIA) | 0.413030 | 0.856955 |
| 17 | (TOSSE, DISPNEIA) | (DESC_RESP) | 0.413030 | 0.764637 |
| 18 | (DESC_RESP, DISPNEIA) | (TOSSE) | 0.413030 | 0.747140 |
| 19 | (TOSSE) | (DESC_RESP, DISPNEIA) | 0.413030 | 0.589883 |
| 20 | (DESC_RESP) | (TOSSE, DISPNEIA) | 0.413030 | 0.632139 |
| 21 | (DISPNEIA) | (TOSSE, DESC_RESP) | 0.413030 | 0.553390 |

**Figura 4. Resultado da associação dos itemsets, supports maiores que 0,5 e o nível de confiança.**

Depois iremos filtrar os que tem o tamanho menor que 2 e valor de support menor que 0,5

| | support | itemsets | length |
|---|---|---|---|
| **10** | 0.540164 | (TOSSE, DISPNEIA) | 2 |
| **12** | 0.552815 | (DESC_RESP, DISPNEIA) | 2 |

**Figura 4. Resultado dos itemsets e supports maiores que 0,5 e tamanho igual a 2.**

Assim obtivemos um nível de confiança (TOSSE, DISPNEIA) e consequência (DESC_RESP) de 76% e (DESC_RESP, DISPNEIA) e consequência (TOSSE) de 74,7%.

## 4. Conclusão

Concluímos que para classificar a evolução final dos casos não foram encontrados atributos que geram uma alta acurácia, e para a classificação final dos casos foi encontrado um conjunto de atributos que resultaram em uma alta acurácia.

E associações sobre a combinação de sintomas também não obteve um nível satisfatório de confiança, sendo ambas combinações com níveis inferiores a 80%.

## 5. Organização do trabalho

| Participantes | Tarefas |
|---|---|
| Edilaine | Pré-processamento |
| Guilherme | Classificação |
| Igor | Associação |
| João Vitor | Associação |

## 6. Repositorio do GitHub

https://github.com/jvmr535/projeto-SAD-grupo-1