Disciplina de Visualização de Dados para a Área de Saúde

Projeto - Análise de condições clínicas respiratórias com base em anotações médicas e áudios de tosse

Brunna Raphaelly Amaral da Silva - 144566

Carolina Vieira Campos - 263081

Gabriel Bianchin de Oliveira - 230217

Taciana Alessandra Gomes Cruz - 107132

INTRODUÇÃO

Neste trabalho fizemos a classificação dos dados relacionados aos áudios de tosse, que foram anotados por especialistas. Para a classificação, utilizamos algoritmos de aprendizado de máquina supervisionado.

DATASET

INSPEÇÃO E PRÉ-PROCESSAMENTO DOS DADOS

O dataset possui 20.000 registros tanto de áudio quanto de metadados referentes aos áudios. Os áudios de tosse possuem durações diferentes entre eles e em alguns áudios é possível perceber a presença de ruídos ambientes tais como pessoas conversando e aparelho de televisão ligados, assim como áudios sem tosse ou ainda áudios apenas com silêncio.

O dataset possui a priori as classes COVID-19, healthy, symptomatic e NaN(classe ausente). Desses, filtramos os relevantes para a nossa pergunta, sendo: COVID-19, healthy, symptomatic. A Tabela abaixo mostra a quantidade de amostras em cada uma das categorias.

COVID-19	1.010
healthy	8.562
symptomatic	1.742

Inicialmente, as amostras que não possuíam tosse foram excluídas utilizando uma feature presente no dataset que classifica com um grau de certeza a presença de uma tosse ou não no áudio. Esta feature apresenta valores de **0.0** até **1.0**. Levando isso em consideração, excluímos das análises os áudios que possuem graus iguais ou menores de certeza de serem tosse igual a 0.5.

Outro ponto relevante de filtragem dos dados foi a identificação que alguns exemplos do dataset não possuem avaliação por médicos especialistas e outros possuem de 1 avaliação ao máximo de 3 avaliações. Dessa forma, removemos também amostras que não foram avaliadas por especialistas.

Cada um dos especialistas fez anotações em relação a diferentes características do áudio. A Tabela a seguir apresenta as características utilizadas na base de dados e os valores possíveis para cada uma das características.

Característica	Valores possíveis	Descrição da característica
quality	good, ok, poor, no_cough	Qualidade do áudio
cough_type	wet, dry, unkown	Tipo de tosse
dyspnea	True, False	Presença de falta de ar
wheezing	True, False	Presença de respiração ofegante
stridor	True, False	Presença de estridor
choking	True, False	Presença de som de asfixia
congestion	True, False	Presença de congestão
nothing	True, False	Falta de sons específicos

diagnosis	upper_infection, lower_infection, obstructive_disease, COVID-19, healthy_cough	Diagnóstico feito a partir do som
severity	pseudocough, mild, severe, unknown	Impressão do especialista para o áudio

Como cada amostra pode ser avaliada por mais de um especialista, fizemos a média da análise da amostra a partir dos diagnósticos de cada um dos especialistas. Para isto, atribuímos valores para as características, como “good” - 0, “ok” - 1, “poor” - 2 e “no_cough” - 3, e fizemos a média dos valores. Caso existisse uma discordância entre os especialistas e o valor final da média seja um número real, arredondamos para a classe mais próxima. Ao final, todos os dados que foram avaliados por especialistas possuíam apenas um único valor para cada uma das características acima.

Com os dados de treinamento gerados, utilizamos a categoria status como a categoria para ser predita. Ou seja, o objetivo é classificar em qual das seguintes categorias os dados pertencem: healthy, symptomatic e COVID-19.

O dataset foi dividido entre conjunto de treinamento, validação e teste. Separou-se inicialmente o conjunto de teste com 20% exemplos e dentro os exemplos restantes separamos 80% para treino e 20% para validação, sendo assim ficando com 20% para o teste, 16% para validação e 64% para treinamento. A mesma divisão adotada para os dados anotados de especialistas foi utilizada para a classificação de áudio.

Efetuamos em seguida o mapeamento das classes que até então eram categóricas, para os valores 0, 1 e 2. Logo, a classe healthy passou a ser 0, a classe symptomatic passou a ser 1 e a classe COVID-19 passou a ser 2.

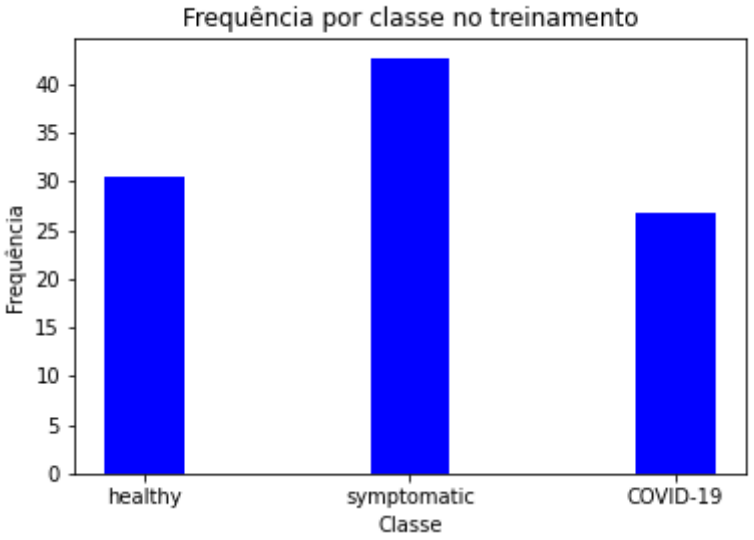
Como o problema de classificação é desbalanceado, adotamos pesos para cada uma das classes. Resolvemos utilizar os pesos para cada classe para que os modelos tivessem uma penalização maior para erros nas classes minoritárias. Calculamos os pesos a partir da divisão da classe com mais amostras pela classe em questão, de modo que o peso da classe majoritária fosse igual a 1 e as outras classes tivessem pesos maiores que 1. Os pesos foram calculados no conjunto de treinamento.

healthy - com o peso 1.3982300884955752

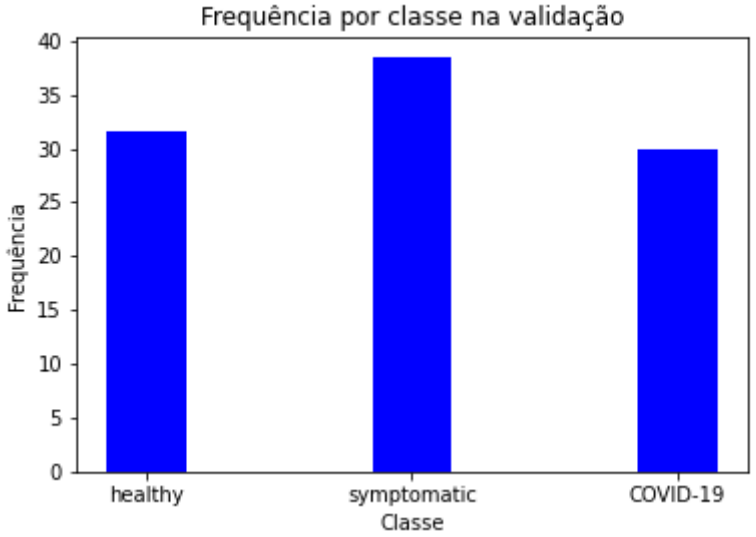
symptomatic - - com o peso 1.0

COVID-19 -- com o peso 1.5906040268456376

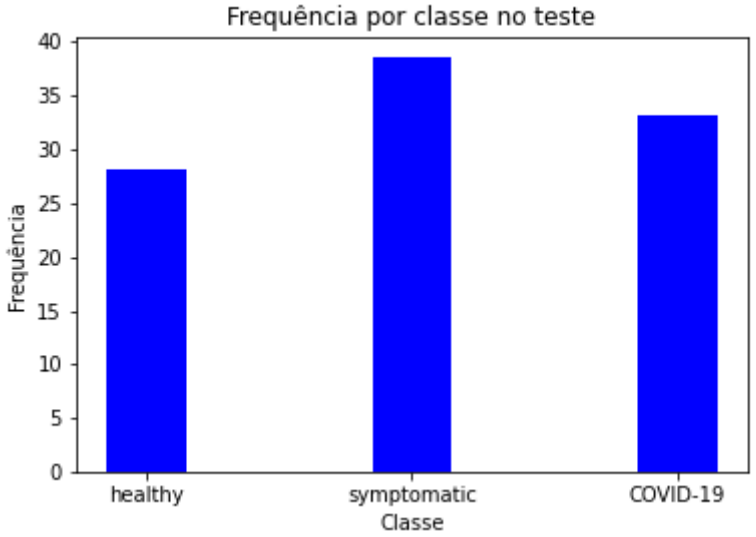
A distribuição dos dados no conjunto de treinamento é dada pela Figura abaixo. A classe mais comum é symptomatic, com mais de 40% dos dados, seguida por healthy, com 30% dos dados, e COVID-19, com 27% dos dados.



Já no conjunto de validação, a distribuição segue próxima ao conjunto de treinamento, conforme apresentado na Figura abaixo.



Já para o conjunto de teste, a distribuição segue com mais de 35% dos dados sendo da classe sympotimatic, mais de 30% de COVID-19 e cerca de 27% como healthy.



EXPERIMENTOS

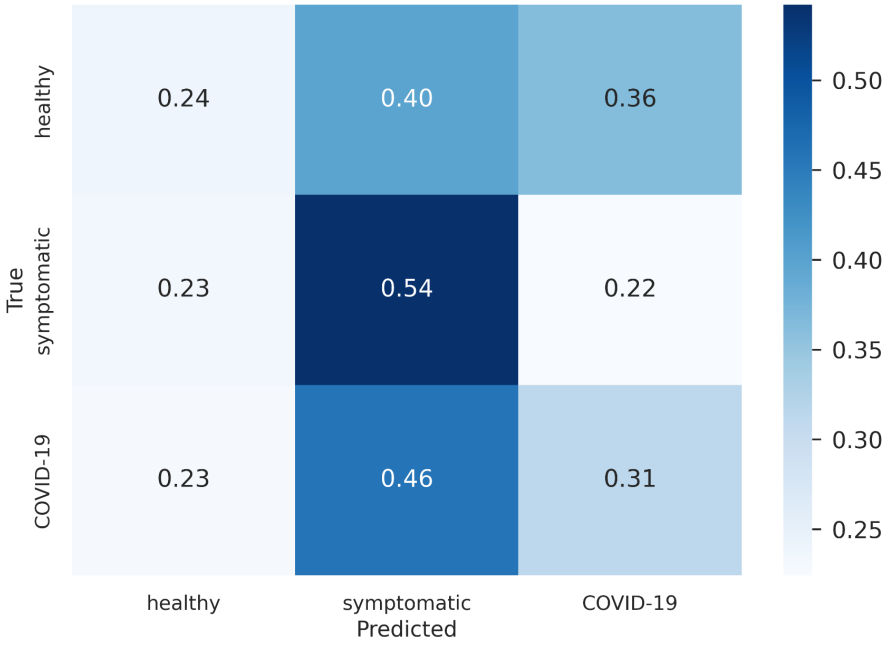
Para os experimentos, utilizamos três diferentes classificadores de aprendizado de máquina, sendo eles: árvore de decisão, floresta aleatória e máquinas de vetores de suporte. Para cada um dos classificadores, realizamos uma busca de parâmetros usando o método Grid-Search. A Tabela a seguir apresenta os parâmetros utilizados para cada classificador no Grid-Search.

Algoritmo	**Parâmetros**
Árvore de decisão	Tamanho máximo da árvore
Floresta aleatória	Tamanho máximo da árvore, quantidade de árvores
Máquina de vetores de suporte	Kernel, C, graus do polinômio, gamma

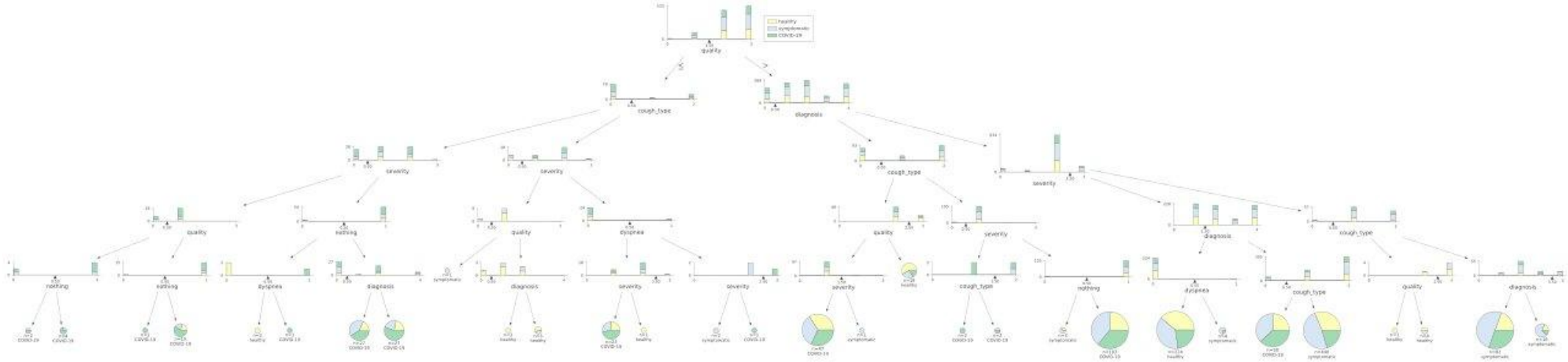
Com cada um dos classificadores otimizados, realizamos o treinamento e avaliamos os resultados no conjunto de validação. A Tabela a seguir apresenta os resultados de acurácia balanceada de cada um dos classificadores.

Algoritmo	**Acurácia Balanceada na Validação**
Árvore de decisão	36,46%
Floresta aleatória	35,60%
Máquina de vetores de suporte	36,44%

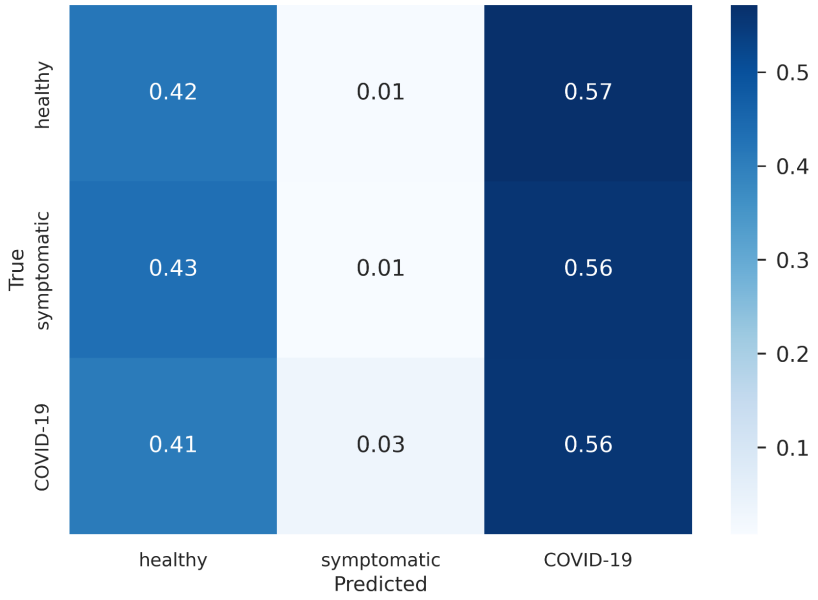
Dentre os classificadores, o melhor resultado foi obtido pela árvore de decisão, com a acurácia balanceada igual a 36,46% no conjunto de validação. A Figura abaixo apresenta a matriz de confusão no conjunto de validação. A matriz mostra muitos erros na classificação de dados da classe healthy e COVID-19, classificados erroneamente como symptomatic.



Além da matriz de confusão, geramos a visualização da árvore. A Figura abaixo mostra a árvore gerada.


healthy
symptomatic
COVID-19
quality
cough_type
diagnosis
severity
cough_type
severity
quality
dyspnea
quality
severity
diagnosis
cough_type
nothing
nothing
dyspnea
diagnosis
diagnosis
severity
severity
severity
cough_type
nothing
dyspnea
cough_type
quality
diagnosis

Após a avaliação do melhor classificador no conjunto de validação, aplicamos este preditor no conjunto de teste. O resultado obtido foi igual a 32,75% de acurácia balanceada. A Figura abaixo apresenta a matriz de confusão no conjunto de teste.



Pela matriz de confusão, podemos observar que o classificador não foi capaz de classificar corretamente os dados, principalmente para a classe symptomatic.

CONCLUSÃO

A classificação de dados das anotações de especialistas pode ajudar, juntamente com o áudio da tosse, a diagnosticar a presença de COVID-19. Nesta parte do trabalho, realizamos a classificação de dados de especialistas em healthy, symptomatic e COVID-19.

O melhor resultado na validação foi encontrado por uma árvore de decisão. Porém, quando avaliamos o classificador no conjunto de teste, o resultado obtido não foi satisfatório. Acreditamos que a pouca quantidade de dados pode ter impactado negativamente no aprendizado e na generalização do modelo.