UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Aplicações de Machine Learning para Diagnóstico de Covid-19: Análise de Imagens Tomográficas

Fabiane Renata de Santana Yassukawa

Trabalho de Conclusão de Curso - MBA em Ciência de Dados (CEMEAI)

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Aplicações de Machine Learning para Diagnóstico de Covid - 19

Fabiane Renata de Santana Yassukawa

[Nome do Aluno]

USP - São Carlos

2020

FABIANE RENATA DE SANTANA YASSUKAWA

Aplicações de Machine Learning para Diagnóstico de Covid - 19: Análise de Imagens Tomográficas

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciência de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Rafael Izbicki

USP - São Carlos

2020

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Y29a

Yassukawa, Fabiane Renata de Santana
Aplicacões de Machine Learning para Diagnóstico de COVID-19: Análise de Imagens Tomográficas / Fabiane Renata de Santana Yassukawa; orientador Rafael Izbicki. -- São Carlos, 2020.
p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2020.

1. Machine Learning. 2. Análise Classificatória. 3. Suport Vector Machine. 4. Regressão Logística do Tipo Lasso. 5. K Nearest Neighbor. I. Izbicki, Rafael, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

FOLHA DE AVALIAÇÃO OU APROVAÇÃO

DEDICATÓRIA

Dedico este trabalho aos meus pais Kendi e Luzia pelo apoio e amor incondicional, por toda dedicação à minha formação como pessoa e profissional.

Aos meus pais pelo amor incondicional.

AGRADECIMENTOS

Agradeço primeiramente a Deus por estar à frente de cada novo desafio na minha vida, ao meu esposo Eduardo pela parceria para que eu pudesse realizar este trabalho, ao meu filho Arthur por dividir suas horas de brincadeira com a mãe com a realização deste trabalho e ao meu orientador Prof. Dr. Rafael Izbicki por toda dedicação e paciência na condução deste trabalho.

EPÍGRAFE

"*A mente que se abre a uma nova ideia jamais voltara ao seu tamanho original*"

Albert Einstein

RESUMO

YASSUKAWA, S. R. F. **Aplicações de Machine Learning para Diagnóstico de Covid - 19**: Análise de Imagens Tomográficas. 2020. 57 f. Trabalho de conclusão de curso (MBA em Ciência de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2020.

Desde o final de 2019 até o momento atual o mundo tem vivenciado uma nova realidade. Desde a detecção do primeiro caso de infecção pelo Coronavírus até o momento atual a onda de incertezas e o número crescente de mortes tem motivado uma grande mobilidade das mais diversas áreas de pesquisa à procura de melhor entendimento da nova doença e dos impactos deste novo patógeno nas vidas das pessoas. Não está sendo diferente para os pesquisadores de machine learning, muitos foram os trabalhos publicados até o momento e com resultados promissores para identificação de novos infectados e projeção do número de mortes e de casos confirmados.

O uso de machine learning para os mais variados fins tem ganhado grande notoriedade nos últimos anos. O emprego das técnicas de aprendizado de máquina tem servido como grande aliada para auxiliar profissionais das mais diversas áreas na tomada de decisões.
Este trabalho apresenta a aplicações de algorítmos de machine learning para discriminação de pessoas contaminadas pelo Coronavíus através dos dados de imagens tomográficas do tórax.

Palavras-chave: Covid - 19. Machine Learning. Dados não estruturados.

ABSTRACT

From the end of 2019 until the present moment the world has experienced a new reality. Since the detection of the first case of Coronavirus infection to the present moment, the wave of uncertainties and the growing number of deaths has motivated a great mobility of the most diverse research areas in search of a better understanding for the new disease and the impacts of this new pathogen on people's lives. It is not being different for machine learning researchers, many papers have been published so far with very promising results in identifying new infected cases and projecting the number of confirmed deaths.

The use of machine learning for the most varied purposes has gained great notoriety in recent years. The use of machine collection techniques has served as a great ally to assist professionals from the most diverse areas in the decision making.

This work presents an application of machine learning algorithms for the discrimination of people infected by the Coronavirus through the tomographic images of the chest.

Keywords: Covid - 19. Machine Learning. Unstructured data.

SUMÁRIO

1 INTRODUÇÃO

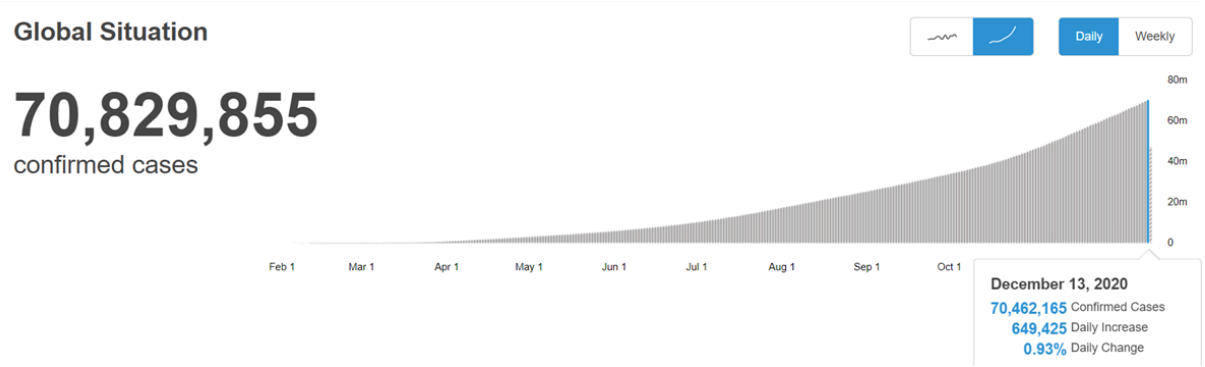
1.1 O que é Covid-19?

Um problema atual de grande relevância é a pandemia causada pelo novo vírus SARS-CoV-2. Um dos pontos críticos e motivador de grande mobilização mundial em várias áreas de pesquisa é a dimensão da transmissibilidade, sintomas respiratórios e taxa de mortalidade causada por este novo patógeno. Esta nova cepa de coronavírus e a doença infecciosa causada por ele foram nomeados respectivamente por SARS-CoV-2 e COVID-19 [1].

O surto teve sua origem em Wuhan na China em Dezembro de 2019 se espalhou rapidamente para todos os continentes com uma curva de crescimento exponencial de novos infectados. Em 30 de janeiro de 2020, o comitê internacional de emergência da OMS declarou oficialmente o surto uma "emergência de saúde pública de interesse internacional" e em 11 de março de 2020 a doença foi declarada pandemia.

Em 13 de Dezembro de 2020 o número de casos confirmados de infecção pelo SARS-Cov-2 somavam 70.829.855 com 1.605.091 mortes segundo dados publicados pela World Health Organization.

Figura 1 – Número Acumulado de Casos Confirmados Mundialmente



Fonte: https://covid19.who.int/

Até o momento a principal forma de transmissão do vírus é de pessoa para pessoa através de gotículas respiratórias de infectados e gotículas no ar e superfície de contato.

1.2 Sintomas

Há uma grande variação nos sintomas causados pelo Covid-19. Infectados pelo novo patógeno podem não apresentar qualquer sintoma como podem apresentar sintomas graves e culminar na morte do paciente. Os sintomas mais frequentemente relatados são febre (98%), tosse seca (76%) e fadiga (44%). Sintomas atípicos também observados incluem escarro (28%), dor de cabeça (8%), hemoptise (5%) e diarreia (3%) [2]. Segundo dados da OMS as estatísticas sugerem que 80% das infecções são leves ou assintomáticas, 15% são graves e requerem oxigênio e 5% são crítica e exigem ventilação.

1.3 Introdução e Motivação

A área de aprendizado estatístico ou aprendizado de máquina têm-se mobilizado frente ao desafio de auxiliar no diagnóstico rápido e eficiente de infecção pelo SARS-CoV-2 e inúmeros projetos de pesquisa já deram contribuições expressivas para auxiliar a área médica no diagnóstico correto.

Neste contexto, a possibilidade de fornecer rapidamente uma probabilidade de diagnóstico de infecção pelo vírus para que munir a área médica de informação na hora de avaliação de um paciente pode ser um bom recurso dado que o exame utilizado atualmente para este diagnóstico RT-PCR além de caro demora cerca de 3 dias úteis para obtenção do resultado.

1.4 Objetivos

Este trabalho tem o objetivo de investigar a aplicação do aprendizado de máquina na identificação de pessoas infeccionadas pelo Covid-19 e avaliar métricas estatísticas das técnicas estudadas na predição de novas observações. Utilizamos neste trabalho uma base de imagens de tomografias de pulmões de pessoas contaminadas pelo Covid-19 e de pessoas sadias, disponível em uma fonte de dados pública disponibilizada por estudiosos da universidade de San Diego [2].

1.5 Estrutura do Trabalho

A estrutura deste trabalho está organizada da seguinte maneira. No presente capítulo fazemos uma contextualização do objetivo do trabalho e uma breve visão geral do problema estudado.

O Capítulo 2 é composto por uma revisão bibliográfica do problema estudado com o objetivo de contextualizar o leitor sobre a doença estudada e os métodos de diagnóstico utilizados.

O Capítulo 3 é destinado a apresentar as metodologias utilizadas neste trabalho, suas particularidades, contextualizar o leitor nas terminologias utilizadas, notações, suposições feitas e apresentar técnicas de avaliação e seleção de modelos.

O Capítulo 4 é o ponto central do trabalho onde descreve as etapas do estudo desde o processamento das imagens, tratamento dos dados e estruturação das bases de modelagem e teste e os resultados obtidos em cada uma das técnicas estudadas.

Por fim o Capítulo 5 apresentamos as conclusões obtidas pelo presente estudo.

2 REVISÃO BIBLIOGRÁFICA

Esta seção está destinada a apresentar os principais métodos utilizados para diagnóstico de contaminação pelo SARS-CoV-2 suas vantagens e desvantagens.

2.1 Diagnóstico

As formas de diagnóstico para o COVID-19 atualmente utilizadas são o diagnóstico clínico que envolve a avaliação de sintomas e análise de exames de imagens como radiografia do tórax, tomografia computadorizada e o diagnóstico laboratorial que engloba a análise de material como a secreção da nasofaringe e traqueia, sangue e fezes do paciente.

2.2 Diagnóstico e Tratamento

2.2.1 Diagnóstico clínico

O diagnóstico clínico é composto pela avaliação dos sintomas, exames físicos e exames de imagem. Na avaliação dos sintomas o médico avalia se há sintomas que caracterizam a infecção do sistema respiratório superior como congestão nasal, coriza, dispneia, tosse seca, fadiga e ainda apresentação de febre.

O exame físico avalia sinais evidentes para pacientes no estado grave que frequentemente apresentam sons do sistema respiratório enfraquecidos e falta de ar SARS (síndrome respiratória aguda). Apesar de haver casos de infecção assintomáticos estes sintomas são as primeiras avaliações para o diagnóstico clínico [2].

O exame de imagem radiográfico do tórax e a tomografia computadorizada fazem parte do diagnóstico clínico e até o momento são utilizados para diagnosticar infectados pelo COVID-19. Alguns estudos apontam que a tomografia computadorizada supera a sensibilidade do exame RT-PCR e defendem o potencial da TC de tórax como ferramenta primária para a triagem de infectados pelo novo vírus [2]. Nos exames de tomografia computadorizada as lesões pulmonares são mais nitidamente identificadas que nos exames radiográficos. A lesões comumente encontradas nos acometidos pela contaminação pelo

SARS-CoV-2 são opacidade em vidro fosco, consolidação segmentar bilateral nos pulmões [1].

Um achado no estudo de pessoas acometidas pela infecção do COVID-19 mostrou que há diferenças nos achados tomográficos para pessoas que apresentaram sintomas e pessoas assintomáticas. O estudo mostrou que dentre as anormalidades encontradas para os casos assintomáticos há uma predominância do ground glass opacity (GGO), área na tomografia que apresenta opacidade e que dificulta a visualização das bordas de vasos pulmonares. E para os casos sintomáticos a característica com maior incidência foi a consolidação, área caracterizada por apresentar substituição do ar alveolar por líquido ou tecido conjuntivo [3].

2.2.2 Diagnóstico laboratorial

A necessidade do exame laboratorial se dá pela diferenciação das infecções do trato respiratório por outros vírus que também apresentam sintomas similares com os causados pelo SARS-CoV-2 como é o caso dos vírus da pneumonia e influenza por exemplo.

Entre os exames laboratoriais o RT-PCR (*Reverse transcription polymerase chain reaction*) é considerado como padrão ouro para método de diagnóstico para o Covid-19, este exame se baseia na detecção de sequências únicas de RNA viral na amostra coletada do paciente. Há estudos que mostram que a sensibilidade desse método em swabs da garganta no COVID-19 é de cerca de 59% [2]. Entenda-se por sensibilidade de um determinado teste a probabilidade de classificação de um verdadeiro positivo corretamente e especificidade a probabilidade de classificação de um verdadeiro negativo corretamente [4]. Em relação à sensibilidade do teste há a possibilidade de falsos negativos o que pode gerar a necessidade de testes adicionais para outras amostras [8].

Alguns outros exames laboratoriais como exame de sangue para contagem do número total de leucócitos, linfócitos e monócitos auxiliam no diagnóstico no estado inicial da infecção pelo vírus.

2.2.3 Estudos de Imagem

A pandemia causou uma grande escassez dos kits para teste do RT-PCR o que gerou a necessidade de hospitais utilizarem métodos alternativos para diagnosticar a doença. Entre os

métodos a tomografia computadorizada ganhou destaque. Segundo especialistas a Tomografia Computadorizada não serve para diagnosticar se uma pessoa foi infectada pelo SARS-CoV-2 ou por qualquer outro vírus, mas auxiliam no julgamento se um paciente está com uma infecção causada por pneumonia viral. Durante uma pandemia com alto crescimento de infectados é muito provável que uma pneumonia viral detectada em um exame de imagem seja causada pelo vírus SARS-CoV-2 [2].

O trabalho médico para detecção de uma infecção do trato respiratório por exames de imagens é uma tarefa manual e demorada especialmente no caso de uma pandemia em que o número de imagens para análise é elevado.

A partir dessa motivação vários trabalhos de machine learning para análise de imagens foram publicados e com resultados expressivos para auxiliar no diagnóstico assertivo do COVID-19. Porém as imagens utilizadas nestes estudos não podem ser compartilhadas com o público por questões de confidencialidade. Diante dessa necessidade Xie P. et all, criaram um banco de dados composto por 349 imagens tomográficas de 216 pacientes COVID-19 positivo e 397 imagens tomográficas de pacientes COVID-19 negativo para promover o estudo e desenvolvimento de modelos para diagnóstico do COVID-19 e com esse banco de imagens desenvolveram modelos de Deep Learning com resultados promissores [2].

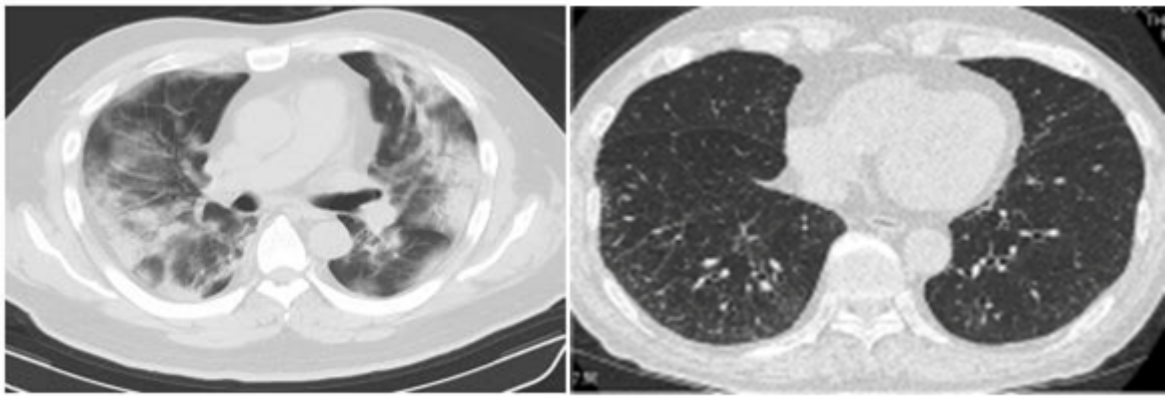
No estudo, os autores relatam o problema de se usar bases de dados pequenas para estudo de machine learning que podem levar ao overfitting. As duas seguintes abordagens foram utilizadas para contornar esse problema.

A primeira se trata da inclusão de dados adicionais na base de treinamento que se trata de segmentações de imagens de regiões pulmonares sadias e segmentações de regiões lesionadas. Os autores ressaltam que essa prática ajuda o modelo a aprender que deve prestar atenção em determinados padrões das imagens.

Uma segunda abordagem foi utilizar melhores representações visuais. Utilizou-se modelos pré-treinados como ImageNet nas imagens de treinamento e a utilização de aprendizagem auto supervisionada contrastiva que cria exemplos aumentados de tomografias e em seguida aprende a rede de representação visual. Para mais detalhes veja o artigo dos autores [2].

Abaixo na figura 2 apresentamos duas tomografias utilizadas no estudo de Xie P. et all sendo uma de caso de Covid positivo e outra de Covid negativo.

Figura 2 – Imagens Tomográficas de pulmão para Covid positivo e Covid negativo



Covid-19 positivo Covid-19 negativo

No estudo os autores descrevem que utilizaram um algoritmo de aprendizagem multitarefa e aprendizagem auto-supervisionada contrastiva. As seguintes métricas foram consideradas para avaliação do método proposto.

(1) Acurácia
(2) Score F1
(3) Área sob a cura Roc (AUC)

Os valores das estatísticas obtidas para a base de validação no estudo dos autores foram de respectivamente, Score F1 de 0.90, AUC de 0.98 e acurácia de 0.89.

Neste trabalho utilizaremos o banco de imagens criado pelos autores Xie P. et all.

Nosso objetivo é a avaliação do desempenho de diferentes modelos de machine learning na classificação de pacientes com Covid-19.

3 METODOLOGIA

3.1 Notação e Suposições Estatísticas

Esta seção e as seções subsequentes apresento os aspectos teóricos que envolvem o estudo estatístico e métodos de aprendizado de máquina.

Todo estudo estatístico requer algumas suposições. Suposições neste contexto é a ação de admitir uma hipótese não comprovada como verdadeira para dar seguimento à um estudo. No caso do estudo do Covid-19 não estamos isentos das suposições. Um exemplo de suposição forte na previsão do número de infectados pelo novo vírus é de que a base de infectados é representativa de toda população de interesse, ou seja, todas as pessoas infectadas pelo vírus. Porém essa suposição não se confirma pois ela representa apenas à parte deste público que é a parcela que realizou o teste para Covid-19. O público testado não representa o público realmente infectado pelo vírus e sim parte dele que é a parcela do público infectado que apresentou algum sintoma que motivou o teste para Covid-19. Outro exemplo de suposição é que para estudos de aprendizado de máquina temos a suposição de que os dados são independentes e identicamente distribuídos. A suposição de independência significa que saber o resultado de exame para Covid-19 em um paciente não altera a probabilidade de resultado para outro paciente. Já a suposição de identicamente distribuídos significa que os dados analisados têm a mesma distribuição de probabilidades para todos os elementos o que não é verdade para o caso de uma pandemia pois há uma grande heterogeneidade na população [4].

Seja uma Y uma característica que se tem interesse de prever em uma determinada população. Na nossa aplicação Y representa a variável aleatória binária ter ou não ter sido contaminado pelo Covid-19. Suponhamos também que para um grupo de pessoas estudadas temos algumas outras variáveis que ajudam a identificar se uma pessoa foi ou não contaminada pelo patógeno, no contexto deste estudo estas variáveis são os pixels das imagens.

A variável Y é chamada de variável resposta ou variável dependente. Para as n unidades amostrais denotaremos Y_i, $i = 1, ..., n$. As demais variáveis que acima exemplificamos como sintomas presentes ou ausentes no público estudado são chamadas de covariáveis, variáveis explicativas ou variáveis independentes e denotamos por x_i.

Denotamos por $x_{i,j}$ o valor da j-ésima covariável na i-ésima unidade amostral, ou seja, para a unidade amostral x_i temos o vetor das d covariáveis estudadas

$$x_i = \left(x_{i,1}, ..., x_{i,d}\right),\ i = i, 2, \dots, n.$$

O objetivo de um modelo de predição é criar uma relação matemática entre as covariáveis e a variável resposta $g(X) \to y$, possibilitando a predição da variável resposta para novas observações das covariáveis.

3.2 Validação Cruzada

Existem várias técnicas que podem ser utilizadas na construção de um modelo preditivo e é natural que escolhamos mais de uma técnica para modelar e então escolher o modelo que melhor se adequa aos dados. Uma importante etapa na modelagem é a avaliação de qual dentre os modelos candidatos tem melhor desempenho na predição da variável resposta. Para realizar essa avaliação particiona-se a base em partes mutuamente excludentes que denominamos base de treinamento e base de validação ou teste. Basicamente a base de treinamento utilizamos para estimar os parâmetros do modelo e a base de teste para validar o modelo estimado. Esse procedimento denomina-se validação cruzada e objetiva avaliar a capacidade de generalização, ou seja, o risco do modelo proposto, definido na seção 3.3.

Várias são as propostas de particionamento da base, dentre elas, o método k-fold que foi o escolhido para este trabalho. Este método propõe a divisão da base em k conjuntos

mutuamente exclusivos e de mesmo tamanho onde k-1 destes conjuntos são utilizados para treinamento e estimação dos parâmetros do modelo e um para validação do modelo. Essa estimação segue uma sistemática circular em que cada uma das k rodadas um conjunto diferente é utilizado para validação do modelo e os demais para estimação dos parâmetros. Ao final das k rodadas calcula-se as estatísticas sobre os erros de estimação encontrados.

3.3 Seleção de Modelos

Para selecionar um modelo dentre outros candidatos um critério importante de avaliação é feito através da probabilidade de o modelo errar. Essa probabilidade é calculada através da função de risco definida a seguir.

A função risco R(g) de um modelo de classificação genérico g, é dada pela esperança da função indicadora de erro de predição, que assume o valor 1 quando $Y \neq g(X)$ na base de validação e 0 caso contrário.

$$R(g) \;=\; E[I(Y \neq g(X))] = \; P(Y \neq g(X))$$

A função risco é estimada pela proporção de erros do modelo avaliado na base de validação como definido a seguir:

$$\hat{R}(g) = \frac{1}{n} \sum_{k=1}^{n} I\left(Y_k^{'} \neq g\left(X_k^{'}\right)\right),$$

onde $\left(X_1^{'}, Y_1^{'}\right)$, ..., $\left(X_n^{'}, Y_n^{'}\right)$ são as unidades amostrais da base de validação.

Podemos definir outras funções para medir o risco de erro de classificação, mas neste trabalho utilizaremos a função indicadora definida acima como como critério de seleção de modelos.

3.4 Medidas de Avaliação do Modelo

Retomando o exemplo do Covid-19, considere que necessitamos avaliar dentre dois modelos qual deles possui melhor assertividade na previsão da variável resposta *Y*. Como

vimos no tópico anterior, a função risco é uma importante estatística de avaliação que auxilia na escolha do modelo final. Algumas outras métricas têm o papel complementar nessa avaliação para seleção de um modelo e as apresentaremos a seguir. Essas estatísticas são importantes pois utilizando somente a função de risco podemos cair no seguinte problema de seleção. Um modelo pode apresentar o risco de classificar um doente como saudável aproximadamente igual a outro modelo e então as estatísticas complementares são úteis para auxiliar a tomada de decisão.

Na tabela 3.1 apresentamos as possíveis classificações feitas por um teste ou modelo versus a real categoria dos dados. Tomando como exemplo o teste RT-PCR para avaliação da pessoa ter ou não ter sido contaminada pelo Covid-19 o teste pode erroneamente classificar uma pessoa doente como sadia, que abaixo nomeamos de Falso Negativo (FN) como também pode classificar incorretamente o uma pessoa sadia como doente que nomeamos de Falso Positivo (FP).

Matriz de Confusão			
Teste		**Doença**	
		Positivo	**Negativo**
	Positivo	VP	FP
	Negativo	FN	VN

VP: Verdadeiro Positivo, **FN**: Falso Negativo, **FP**: Falso Positivo, **VN**: Verdadeiro Negativo

Tabela 3.1 – Matriz de Confusão

Quando comparamos os modelos candidatos utilizamos a proporção de observações nas classes da tabela acima para avaliar o desempenho do modelo em uma base de teste.

Nomeamos de Sensibilidade ou Recall de um teste a probabilidade de um indivíduo avaliado e doente ter seu teste classificado como positivo. Sensibilidade = VP/(VP+FN), ou seja, a proporção das pessoas doentes classificadas corretamente como positivo.

Nomeamos de Especificidade de um teste a probabilidade de um indivíduo avaliado e normal ter seu teste classificado como negativo. Especificidade = VN/(VN+FP), ou seja, a proporção das pessoas sadias classificadas corretamente como negativo.

Acurácia de um teste é a probabilidade de indivíduos positivo ou negativo serem classificados corretamente. Acurácia = (VP + VN)/(P + N), ou seja, a proporção de pessoas doentes ou não classificadas corretamente pelo modelo ou teste.

Precisão ou Precision de um teste é a probabilidade de os positivos classificados pelo teste serem assertivos. Precisão = VP/(VP + FP), ou seja, a proporção de verdadeiros positivos dentre todos positivos do teste.

Uma opção visual que utilizam as métricas explicadas acima é o gráfico sob a curva ROC. Os gráficos são opções interessantes pois trazem uma visão multidimensional do problema avaliado. O gráfico da curva ROC utiliza as duas estatísticas, Sensibilidade que é a proporção dentre as pessoas doentes serem classificadas corretamente como doentes e 1 – Especifidade, apresentadas anteriormente. Quanto maior a curva sob a curva melhor o desempenho do modelo. Um exemplo de gráfico é apresentado na figura 3 abaixo.

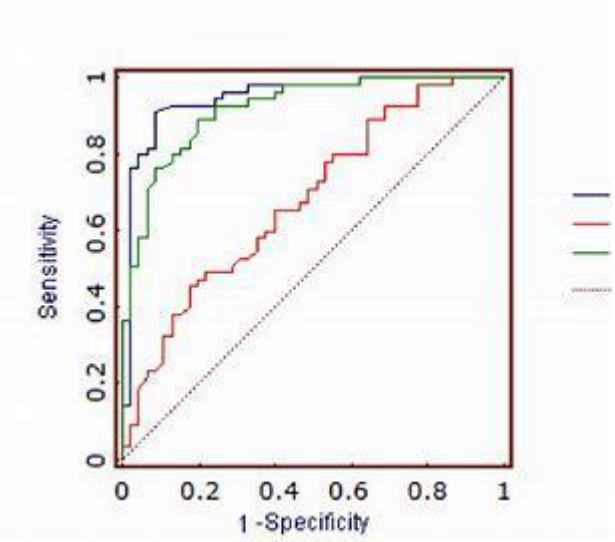


Figura 3: Curva ROC

A seguir faremos uma revisão dos métodos de predição utilizados neste trabalho. A partir deste ponto assumiremos que Y, variável resposta, pertence ao intervalo $Y \epsilon \{0, 1\}$ a menos que seja definido de outra forma.

3.5 K Vizinhos Mais Próximos

O método dos k vizinhos mais próximos (k-nearest neighbours, knn) é um dos métodos mais utilizados quando se deseja classificar padrões [5]. O algoritmo knn pode ser utilizado nos casos de classificação e de regressão. Entende-se por classificação a divisão dos dados em classes categóricas pré-definidas através de uma função matemática ou modelo (classificador). O método baseia-se na disposição espacial dos dados, entende-se que dados mais próximos espacialmente tendem a serem similares e dados distantes tendem a serem não similares. A medida de distância mais frequentemente utilizada é a distância Euclidiana, dada pela equação (3.1)

$$d\left(x_i, x_j\right) = \sqrt{\sum_{l=1}^{d} \left(x_{i,l} - x_{j,l}\right)^2} (3.1)$$

A finalidade é estimar uma função de regressão r(**x**) para um dado conjunto de variáveis explicativas ou covariáveis **x** baseado nos valores de da variável resposta Y dos k-vizinhos mais próximos a **x** ou seja, a função de regressão é estimada através da média da variável resposta dos k vizinhos mais próximos a **x** [5]. Exemplificando no contexto do Covid-19 uma aplicação seria utilizar o algoritmo para classificar imagens tomográficas para classificação como positivo ou negativo para contaminação do patógeno, onde as variáveis explicativas seriam os pixels das imagens e baseado na distância euclidiana das variáveis explicativas das k imagens mais próximas a imagem teste classifica-se como positivo ou negativo pelo cálculo da proporção de zeros e uns da variável da resposta das k imagens.

Um ponto que muitas vezes melhora muito a acurácia do método é a normalização das covariáveis. Outro ponto que merece destaque é a escolha do número de k. Valores de k muito altos levam a modelos muito simples e com baixa acurácia. E valores de k muito baixos geram modelos complexos com alta acurácia, porém com overfitting, ou seja, o modelo só consegue explicar os dados da base utilizada para construção do mesmo não sendo possível utilizá-lo para explicar novos dados [5]. Assim é importante fazer a escolha de k via validação cruzada.

3.6 Regressão Logística

Outro classificador comumente utilizado na divisão dos dados em duas ou mais classes categóricas como nosso exemplo de classes com ou sem Covid, é a regressão logística. A

regressão logística usa uma medida de similaridade calculada pelos valores das *d* covariáveis $X_i = \left(X_{i,1} \ldots X_{i,d}\right),\ i = 1,\ \ldots, n$ que caracterizam cada um *n* dos indivíduos da base de dados.

Utilizando a regressão logística estimamos a probabilidade *p* de um indivíduo *i* pertença ao grupo onde a varável resposta *Y = 1* (com Covid) e que denominamos como grupo evento e também a probabilidade de que pertença ao grupo *Y = 0* (sem Covid) que denominamos grupo não evento onde essa última é a probabilidade complementar *(1 – p)*.

Definindo um modelo geral de regressão logística temos a seguinte equação:

$$ln\left\{\frac{p}{(1-p)}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots\ + \beta_d x_d = z,$$

Matematicamente, elevando os dois lados da igualdade acima podemos expressar a probabilidade $p = P(Y = 1|\boldsymbol{x})$ por

$$P(Y = 1|x) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

O gráfico da probabilidade *p* versus os valores de *Z* apresentamos abaixo na figura (3.1).

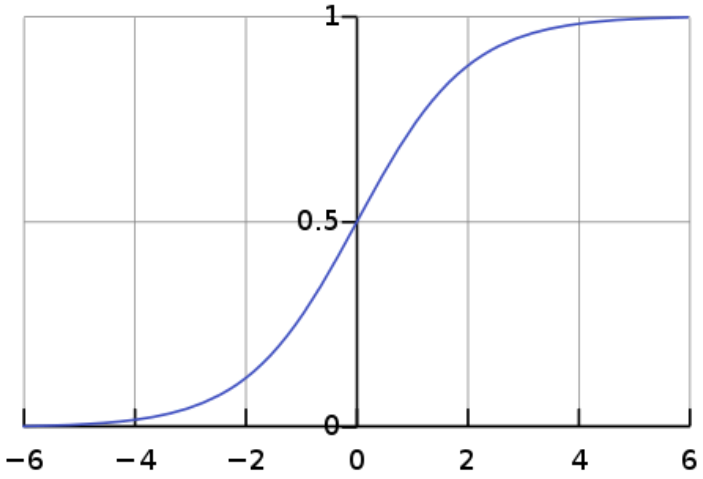


Figura 3.1: Curva logística

Note que a probabilidade p é uma função de Z e quanto maior o valor de Z maior a probabilidade de p.

A estimação dos parâmetros da regressão logística para uma amostra independente e identicamente distribuída $(X_1 Y_1, \ldots, X_n Y_n)$, ou seja os valores de $\beta = \left(\beta_0\, \beta_1 \ldots \beta_d\right)$ pode ser feita através da função de máxima verossimilhança dada por:

$$L(y; (x, \beta)) = \prod_{k=1}^{n} (P(Y_k = 1|x_k, \beta))^{y_k} (1 - P(Y_k = 1|x_k, \beta))^{1-y_k}$$

$$\prod_{k=1}^{n} \left(\frac{e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_{d,i}}}{1 - e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_{d,i}}} \right)^{y_k} \left(\frac{1}{1 - e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_{d,i}}} \right)^{1-y_k}$$

As estimativas dos coeficientes de β é obtida maximizando $L(y;(x,\beta))$ através de algoritmos numéricos.

Um valor pre-determinado k de p = P(Y = 1) será utilizado para classificar um indivíduo na categoria Y = 1 ou Y = 0 dessa forma descriminaremos os dois públicos de acordo com a probabilidade calculada através dos valores das covariáveis, ou seja se a probabilidade de um dado indivíduo for maior que o valor de corte k o classificamos com Y=1. Uma boa prática quando se tem uma base desbalanceada, ou seja quando o percentual de observações do grupo evento é muito baixo como é o caso quando estudamos uma doença rara em uma população é escolher um valor de k também baixo e compatível com a taxa de ocorrência na população.

Neste trabalho utilizamos a regressão logística juntamente com o Lasso (Least Absolute Selection and Shrinkage Operator) que permite que alguns coeficientes β sejam iguais a zero. A regressão logística do tipo Lasso é também chamada de método de regressão penalizado. O Lasso estabelece um limite superior para a soma dos valores absolutos dos parâmetros $\sum_{j=1}^{d} |\beta_j|$ do modelo. Essa restrição de limite faz que os valores de alguns dos coeficientes β das variáveis do modelo convirjam para zero enquanto outros tem seus valores aumentados essa calibração de pesos nada mais é de que uma forma de seleção de covariáveis que possibilita a obtenção de um subconjunto de covariáveis que minimiza o erro de predição [5].

3.7 Support Vector Machines

O método de máquinas de vetores de suporte (Support Vector Machines, SVM) tem sua base na teoria de aprendizado estatístico, desenvolvida por Cortes e Vapnik (1995). O método tem objetivo de estimar um classificador linear que maximiza a margem de separação das duas classes e minimiza o erro de treinamento. Neste trabalho descrevemos quando temos classes separadas linearmente mas uma generalização para margens não lineares de separação pode ser feita através das funções de kernel que medem a similaridade entre objetos [5].

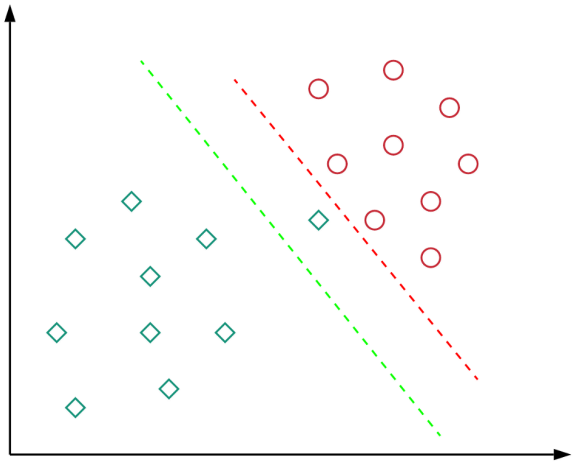


Figura 3.2 Margens de separação linear.

O algoritmo dessa técnica não se baseia em cálculo de probabilidades como visto nas técnicas anteriores. A única saída de um modelo de SVM são as classes estimadas de novas observações.

Suponha que nossa variável resposta Y assuma um dos seguintes valores do conjunto $C = \{-1, 1\}$. Considere a seguinte função linear das d covariáveis em estudo

$$f(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d.$$

O classificador g(x) dado pelo SVM tem a seguinte regra de classificação segundo os valores obtidos por f(x):

$$Se f(x) < 0, g(x) = -1$$
$$Se f(x) \geq 0, g(x) = 1$$

Para se chegar na expressão de f(x) o algoritmo se baseia na construção de hiperplanos de separação entre as duas classes estudadas e busca a função que define o hiperplano que tem maior margem de separação entre as duas classes e menor erro de classificação.

No caso de houver um hiperplano que separa perfeitamente os dados entre suas categorias o algoritmo SVM busca o hiperplano com coeficientes β que satisfazem

$$(\beta, M) = argmax_{\beta,M} M$$

Com as seguintes restrições:

(1) $\sum_{i=1}^{d} \beta_i^2 = 1$ e

(2) $para\ todo\ i\ =\ 1,\ \ldots,n\ y_i f_\beta(x_i) \geq M$

O somatório dos valores de beta ao quadrado garante a comparação de hiperplanos com base na sua norma.

O problema da definição acima é que ela exige a suposição de haver um hiperplano que separa perfeitamente bem os dados uma exigência muito difícil de ser encontrada na prática. Assim o SVM permite que pontos estejam do lado errado da margem de separação do hiperplano porém com o limitador de distância das margens de separação C.

Matematicamente temos a expressão acima reescrita da seguinte forma

$$(\beta, M)\ =\ argmax_{\beta,M} M$$

Com as seguintes restrições:

(1) $\sum_{i=1}^{d} \beta_i^2 = 1$ e

(2) $paratodoi\ =\ 1, \ldots, n y_i f_\beta(x_i) \geq M(1 - \epsilon_i)$, em que $\epsilon_i > 0$ e $\sum_{i=1}^{n} \epsilon_i \leq C$.

3.8 Boosting

O boosting utilizado para problemas de classificação busca a agregação de classificadores fracos para a variável resposta para compor classificadores mais poderosos na discriminação da variável reposta. Quanto maior a proporção de erro de um classificador menor será o peso dele na combinação final. Em outras palavras o algoritmo busca novos classificadores que visam corrigir os erros de classificação de classificadores usados anteriormente de forma sequencial, ou seja o b-ésimo classificador utilizado pelo algoritmo dependerá do b-1-ésimo classificador e assim sucessivamente. Abaixo descrevemos o algoritmo Adaboost.M1 (Freund e Schapire, 1995).

Considere que a variável resposta $y_i,\ i\ =\ 1,\ \ldots\ ,\ n$ assuma os seguintes valores {-1, 1}, no contexto deste trabalho -1 se refere ao paciente com Covid e 1 ao paciente sem Covid.

Primeiramente o algoritmo atribui pesos iguais a cada uma das observações $w_1 = w_n = \frac{1}{n}$. Para os B classificadores $g_b(x)$, com b = 1, ..., B temos os seguintes passos do algoritmo utilizando a base de treinamento:

(1) Ajuste de $g_b(x)$ utilizando os pesos iguais para as observações $w_1 = w_n = \frac{1}{n}$

(2) Cálculo do erro ponderado obtido pelo ajuste $er_b \frac{\sum_{i=1}^{n} w_i I\left(y_i \neq g_b\left(x_i\right)\right)}{\sum_{i=1}^{n} w_i}$

(3) Cálculo $\alpha_b = \log log\left(\frac{1-er_b}{er_b}\right)$

(4) Atualização dos pesos $w_i \leftarrow w_i exp\left\{\alpha_b I\left(y_i \neq g_b\left(x_i\right)\right)\right\},\ i = 1, \ldots, n$

Finalmente o modelo é retornado por $g(x) = sinal\left(\sum_{b=1}^{B} \alpha_b g_b(x)\right)$. Observe que o peso de cada classificador α_b é tanto menor quanto maior for a proporção do erro do classificador [5].

3.9 Redes Neurais

Redes Neurais (RNA) como o próprio nome diz uma técnica de modelagem inspirada na capacidade de se adaptar e aprender do sistema nervoso humano. Comparativamente ao cérebro humano que processa informação através de impulsos elétricos captados pelos dendritos de um neurônio uma rede neural adquire aprendizado através do processamento das informações das unidades de entrada que são as covariáveis ou variáveis explicativas do modelo.

A estrutura de uma rede neural apresenta uma estrutura como da figura (3.3).

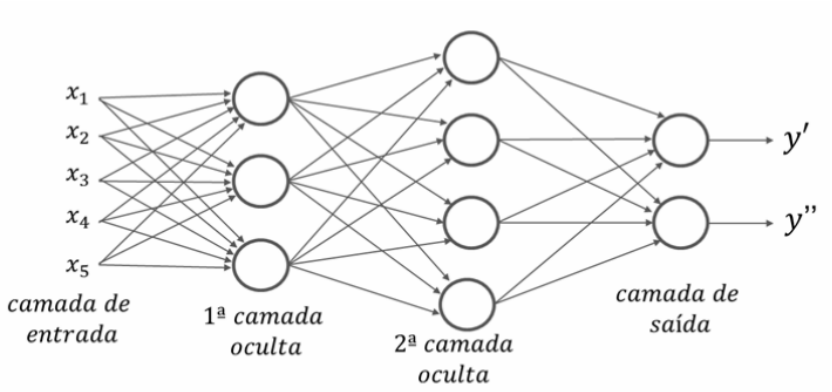


Figura 3.3 Estrutura de uma rede neural

Lendo a figura do lado esquerdo para o direito temos os nós iniciais que são as entradas da rede, ou seja, cada uma das variáveis explicativas do banco de dados (no exemplo da figura, temos 5 variáveis). Cada flecha que une os neurônios da camada de entrada e da

primeira camada oculta são pesos β, que ponderam os valores das covariáveis de entrada. Cada um dos nós da segunda camada representa uma transformação dos nós da camada anterior que nada mais é que a aplicação de uma função no valor da combinação linear das covariáveis ponderadas pelos pesos β. No contexto de redes neurais chamamos essa função de função ativação que é pré-definida anteriormente. A dinâmica dessa transformação das covariáveis da camada de entrada para a camada oculta i é ilustrada na figura (3.4)

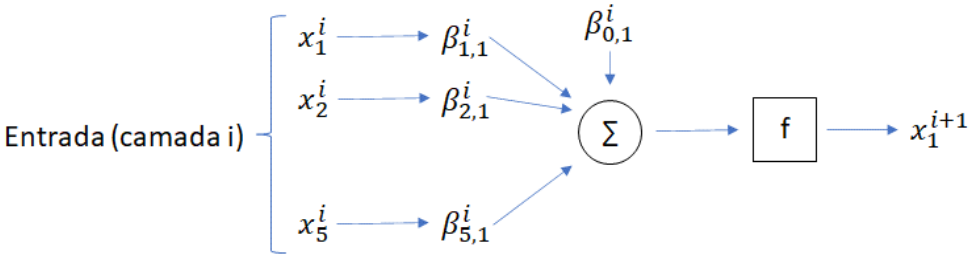


Figura 3.4 Processamento de cálculo da saída da camada i para i + 1

Genericamente uma rede neural pode ter mais de uma camada intermediária com diferentes números de neurônios. Além disso uma rede neural pode apresentar estruturas de retroalimentação onde o valor de nós intermediários pode voltar para entradas camadas anteriores da rede [5].

A estimação do vetor de parâmetros **β** deve considerar uma função a ser minimizada. Porém o cálculo do vetor **β** nem sempre é possível de ser feito de forma analítica ou quando é possível pode ser muito custoso computacionalmente, por exemplo quando n é muito grande.

O algoritmo gradiente descendente objetiva buscar de forma interativa, os valores dos parâmetros que minimizam uma função de interesse.

Suponhamos que o objetivo é ajustar uma regressão logística para classificar um paciente nas categorias Covid positivo ou Covid negativo, baseado nos sintomas que apresenta. No caso de uma classificação binária a função mais comumente utilizada como função custo é a entropia cruzada binária data por:

$$Custo\left(h_\theta(x), y\right) = -\ ln\left(h_\theta(x)\right), \text{ quando y = 1}$$

$$Custo\left(h_\theta(x), y\right) = -\ ln\left(1\ -\ h_\theta(x)\right), \text{ quando y = 0}$$

Quando y =1, o valor do custo se aproxima de 0 na medida que $h_\theta(x)$ se aproxima de 1 e por outro lado o custo cresce até o infinito conforme $h_\theta(x)\ se\ aproxima\ de\ 0$. Na figura 3.5, ilustramos a dependência do valor da função de custo em relação ao valor da probabilidade de $h_\theta(x) = P(Y = 1)$.

Analogamente quando y =0, o valor do custo se aproxima de 0 na medida que $h_\theta(x)$ se aproxima de 0 e por outro lado o custo cresce até o infinito conforme $h_\theta(x)\ se\ aproxima\ de\ 1$. Na figura 3.6, ilustramos a dependência do valor da função de custo em relação ao valor da probabilidade de $h_\theta(x) = P(Y = 0)$.

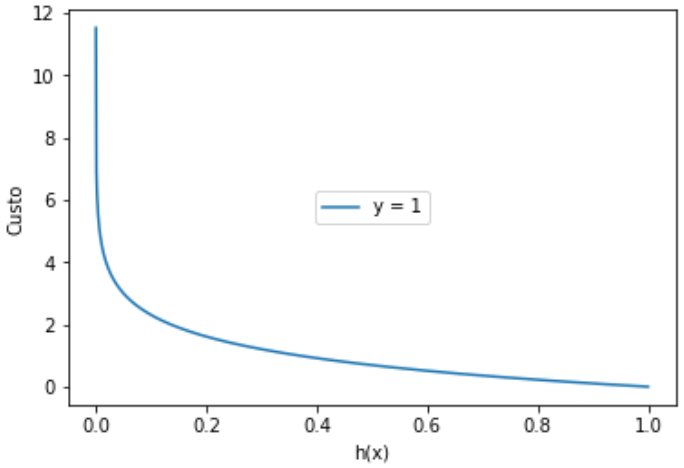


Figura 3.5 Função Custo quando y = 1

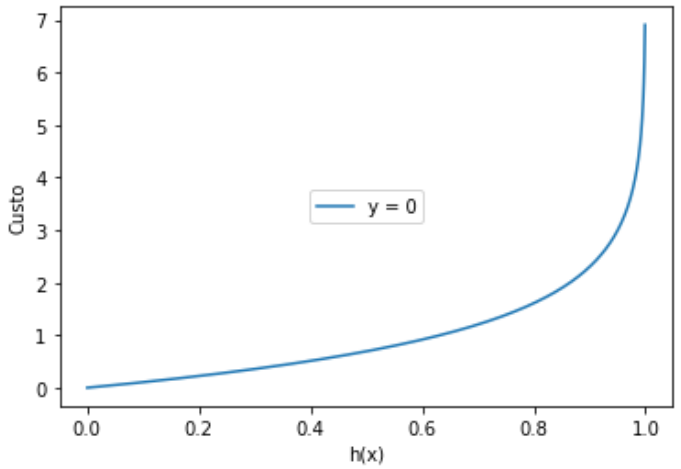


Figura 3.5 Função Custo quando y = 0

No ajuste do modelo buscamos estimar os valores do vetor de parâmetros β de forma a minimizar a função de custo. O vetor β que minimiza a função nem sempre é possível de se obter analiticamente, mas podemos utilizar métodos numéricos de estimação. Neste contexto um método de destaque na modelagem de redes neurais é o backpropagation que é um algoritmo de gradiente descendente que obedece a seguinte ordem de execução: a última camada é a primeira a ter os parâmetros atualizados e de forma sequencial a penúltima até a primeira camada atualizam os parâmetros [5].

4 RESULTADOS

Nesta seção descrevemos cada uma das etapas do trabalho desde a tratativa inicial das imagens até as técnicas de modelagem implementadas, seus parâmetros e as estatísticas de avaliação de cada modelo.

4.1 Tratamento das Imagens

Como descrito na seção 2.2.3, as imagens utilizadas neste trabalho são de imagens tomográficas de pulmões de pacientes acometidos pela doença Covid-19, e de pacientes saudáveis. O banco de imagens disponibilizado pelos autores Xie P. et al, é composto de imagens de diferentes tamanhos. O primeiro procedimento foi a conversão das imagens para RGB e o redimensionamento das imagens para o menor tamanho encontrado no banco que foi de 115 de altura e 98 de largura. Depois de convertidas e redimensionadas as imagens foram lidas e cada matriz de pixels das imagens foi então vetorizada e colocada em uma única matriz onde para cada linha temos os dados de uma determinada imagem sendo no total 749 linhas. A primeira coluna da matriz é a variável respostas da sendo 1 para imagens com Covid-19 e 0 para sem Covid-19 e as demais colunas são os pixels de cada uma das imagens. Sendo cada linha uma imagem diferente.

Após a construção da matriz total de imagens os dados foram normalizados e base foi dividida aleatoriamente em duas bases. A de treinamento com 80% dos dados, e a de teste com 20% dos dados.

4.2 Modelos ajustados

O primeiro modelo ajustado foi o modelo de KNN para o qual testamos o número de folds de 1 a 10. A medida de distância utilizada foi a distância euclidiana. Para cada ajuste foi feita a validação cruzada para seleção do melhor k (fold) e a medida de comparação na validação cruzada foi a acurácia. Para o melhor valor de k nesta etapa então ajustamos o

modelo final e o aplicamos na base de teste. As estatísticas do melhor modelo na base de teste foram as seguintes.

Modelo KNN (k=1)	
Accuracy	0.77
Precision	0.82
Recall	0.69

Tabela 4.1: Estatísticas de Desempenho Modelo KNN

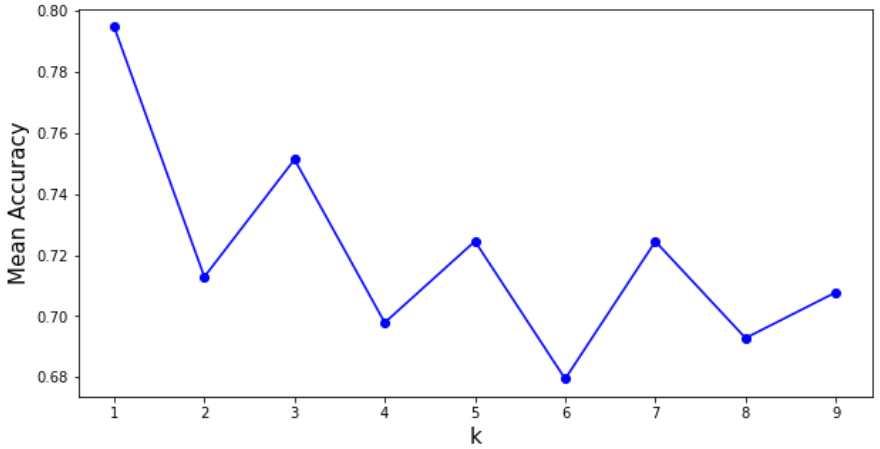


Gráfico 4.1: Acurácia em função das interações

Posteriormente o modelo de regressão logística com penalização também utilizando a validação cruzada e com número máximo de interações para otimização do modelo igual a 1000. As estatísticas para esse modelo foram as seguintes.

Modelo Reg. Log.	
Accuracy	0.69
Precision	0.73
Recall	0.59

Tabela 4.2: Estatísticas de Desempenho Modelo Regressão Logística

O terceiro modelo ajustado foi o de Support Vector Machine SVM e as estatísticas obtidas para a base de teste foram as seguintes.

Modelo SVM.	
Accuracy	0.69
Precision	0.7
Recall	0.62

Tabela 4.3: Estatísticas de Desempenho Modelo Support Vector Machine

O quarto modelo implementado foi o Xgboost com taxa de aprendizado igual a 0.1 e número de estimadores igual a 10 teve os seguintes resultados na base de teste.

Modelo Xgboost	
Accuracy	0.73
Precision	0.75
Recall	0.66

Tabela 4.4: Estatísticas de Desempenho Modelo Xgboost

O quinto modelo estudado foi o modelo de Redes Neurais com a seguinte estrutura. Uma camada densa com 128 neurônios e função de ativação Relu, segunda camada com 64 neurônios e função de ativação Relu, terceira camada com 32 neurônios e função de ativação Relu e a última camada com 1 neurônio e ativação sigmoid. As estatísticas obtidas para a base de teste foram.

Modelo RN - 1	
Accuracy	0.7
Precision	0.76
Recall	0.57

Tabela 4.5: Estatísticas de Desempenho Modelo Redes Neurais 1

O sexto modelo estudado foi o modelo de Redes Neurais com uma estrutura bem mais simplificada descrita a seguir. Uma camada densa com 20 neurônios e função de ativação Relu, segunda camada também densa com 5 neurônios e função de ativação Relu e a última camada também densa com 1 neurônio e ativação sigmoid. As estatísticas obtidas para a base de teste foram.

Modelo RN - 2	
Accuracy	0.69
Precision	0.72
Recall	0.6

Tabela 4.6: Estatísticas de Desempenho Modelo Redes Neurais 2

Abaixo ilustro o gráfico de erro em função das épocas de treinamento das redes.

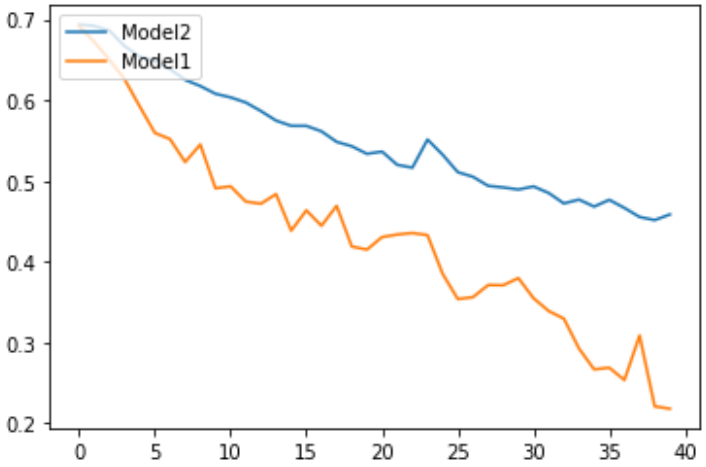


Gráfico 4.2: Treinamento: Loss para modelos de Rede Neurais em função das épocas

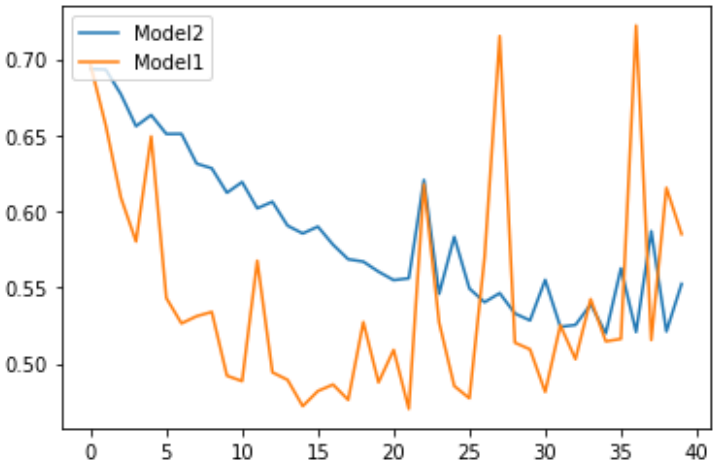


Gráfico 4.3: Validação: Loss para modelos de Rede Neurais em função das épocas

O sétimo e último modelo implementado foi uma Rede Neural Convolucional com a seguinte estrutura 3 camadas convolucionais com respectivamente 32, 16 e 8 neurônios e função de ativação Relu, intercaladas com camadas MaxPooling de tamanho 2 e uma última

camada densa com 2 neurônios e função de ativação softmax. As estatísticas para este modelo foram as seguintes.

Modelo CNN	
Accuracy	0.65
Precision	0.61
Recall	0.82

Tabela 4.7: Estatísticas de Desempenho Modelo Redes Neurais Convolucionais

5 CONCLUSÕES

Neste estudo avaliamos várias técnicas de machine learning para modelagem de dados de imagens tomográficas e o modelo que sobressaiu aos demais foi o KNN com acurácia de 0.77 na base de teste seguido do Xgboost com acurácia de 0.73 na base de teste. Todos os outros modelos testados apresentaram resultados semelhantes, com excessão da CNN que apresentou acurácia mais baixa de 0.65.

O resultado de acurácia mais baixo no modelo de CNN pode ser explicado pelo fato de a base de estudo ser uma base pequena o que afeta diretamente o desempenho de um modelo mais complexo como é o caso da CNN. Por esse motivo uma alternativa interessante é a aumentação de dados, gerando novas unidades amostrais de estudo por amostragem dos pixels das imagens disponíveis. Outra possibilidade de melhoria é a utilização de algoritmos pré-treinados para construção do modelo como foi feito pelos autores Xie P. et all que disponibilizaram a base de dados deste trabalho.

REFERÊNCIAS

1. Wu D.; Wu T.; Liu Q., Yang Z. The SARS-CoV-2 outbreak: What we Know. **International Journal of Infectious Diseases**, v.94, p. 44-48, Mar. 2020.
Disponível em: https://www.sciencedirect.com/science/article/pii/S1201971220301235

2. Yang X., He X., Zhao J., Zhang Y., Zhang S., Xie P. COVID-CT Dataset: A CT Image Dataset about COVID-19. **Cornell University**, Jun. 2020.
Disponível em: https://arxiv.org/abs/2003.13865
Fonte de dados: https: https://arxiv.org/pdf/2003.13865.pdf

3.Inui S.; Fujikawa A.; Jitsu M.; Kunishima N.; Watanabe S.; Suzuki Y.; Umeda S.; Uwabe Y. Chest CT Findings in Cases from the Cruise Ship "Diamond Princess" with Coronavirus Disease 2019 (COVID-19). **Radiology:Cardiothoracic Imaging** v.2, no.2, Mar. 2020.
Disponível em: https://pubs.rsna.org/doi/10.1148/ryct.2020200110

4. Izbicki R. Screening Procedure. R notebook using data from Diagnosis of Covid-19 and its clinical spectrum.
Disponível em: https://www.kaggle.com/rizbicki/screening-procedure

5. Izbicki R.; Santos T. M. dos. Aprendizado de máquina: Uma abordagem estatística. 1 edição. 2020. 266 páginas. **ISBN: 978-65-00-02410-4**
Disponível em: http://www.rizbicki.ufscar.br/ame/

6. Cover T. M.; Hart P. E. Nearest Neighbor Pattern Classification. **IEEE Transactions on Information Theory** v. IT-13, no. 1, Jan. 1967.
Disponível em:
https://pdfs.semanticscholar.org/a3c7/50febe8e72a1e377fbae1a723768b233e9e9.pdf?_ga=2.256092995.952590364.1593559408-1963073612.1593559408

7. Cortes C.; Vapnik V. Support-Vector Networks. **Machine Learning** 20, 273-297, Sep. 1995.
Disponível em: https://link.springer.com/content/pdf/10.1007/BF00994018.pdf

8. Vieira L.; Emery E.; Adriolo A. Covid-19 laboratory diagnosis for clinicians. **São Paulo Medical Journal** , Jun. 2020.
Disponível em: https://doi.org/10.1590/1516-3180.2020.0240.14052020

9. Pengfei S.; Xiaosheng L.; Chao X.; Wenjuan S.; Bo P. Understanding of COVID-19 basead on current evidence. **Journal of Medical Virology**, Feb. 2020.
Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1002/jmv.25722

10. Cover T. M., Hart P. E. Nearest Neighbor Pattern Classification. **IEEE Transactions on Information Theory**, vol. IT-13, N. 1, Jan. 1967
Disponível em: http://ssg.mit.edu/cal/abs/2000_spring/np_dens/classification/cover67.pdf