Técnicas de Machine learning para predição de absenteísmo às consultas médicas

JEFERSON DIONEI SEHNEM[1*]

RESUMO

A falta às consultas médicas são um dos principais desperdícios de recursos nas organizações de saúde. Esse problema não é exclusivo do Brasil, tendo grande impacto ao redor do mundo, como no serviço de saúde público do Reino Unido onde o desperdício anual é na casa de £ 1 bilhão. Para evitar ou diminuir esse desperdício as organizações buscam vários tipos de soluções para esse problema. Uma dessas possíveis soluções é a utilização de técnicas de Machine Learning para a detecção dos pacientes faltantes. No estudo utilizamos cinco algoritmos de Machine Learning, sendo eles Redes Neurais, Random Forest, Regressão Logística, XG boost e Adaboost. O que teve o melhor desempenho foi às Redes neurais com o AUC de 67% e com uma sensibilidade de 85%.

Palavras-Chave: Desperdício de recursos. Absenteísmo de pacientes. Solução. Aprendizagem de máquina. Algoritmos.

ABSTRACT

The lack of medical appointments is one of the main wastes of resources in health organizations. This problem is not unique to Brazil, having a great impact around the world, as in the UK public health service where annual waste is in the region of £1 billion. To avoid or reduce this waste, organizations seek various types of solutions to this problem. One of these possible solutions is the use of Machine Learning techniques to detect missing patients. In the study we used five Machine Learning algorithms, namely Neural Networks, Random Forest, Logistic Regression, XG boost and Adaboost. The one that performed the best was Neural Networks with an AUC of 67% and a sensitivity of 85%.

KEYWORDS: *Waste of resources. Patient absenteeism. Solution. Machine learning. Algorithms.*

[1]Códigos do estudo: https://drive.google.com/file/d/16rDdFlW5Fyhfw8xF_qJvvd0xMlCJkbew/view?usp=sharing

1 INTRODUÇÃO

Uma das grandes dificuldades dos sistemas de saúde no mundo são os pacientes que marcam uma consulta mas não comparecem. Este problema é recorrente trazendo transtornos para pacientes com doenças graves, pois aumenta o tempo de espera para o atendimento. Também, ocasiona um desperdício de recursos, pois no horário em que o paciente não comparece os médicos ficam subutilizados.

Dantas et al (2018) realizou uma revisão na literatura em 105 artigos científicos, com dados da África, das Américas, da Europa, da Ásia e da Oceania, sobre quais fatores têm mais impacto nas taxas de consultas perdidas. Identificou que as características dos pacientes que mais contribuem para o não comparecimento são a idade do paciente, onde pacientes mais jovens faltam mais, status socioeconômico mais baixo, local de residência distante da clínica e o fato de não possuir nenhum seguro de saúde privado. Outros fatores preponderantes para a falta de consultas são pacientes que fazem uso de medicamentos psiquiátricos, fazem uso de tabaco, drogas ou álcool.

Para Rodrigues et al (2020) as taxas de não comparecimentos às consultas variam de 5% a 55% entre os países, sendo a taxa média mundial em torno 23%. No Sistema de Saúde Brasileiro as taxas variam de região para região, com São Paulo tendo taxas de 52%, João Pessoa 48,3%, Ceará 32,17% e Espírito Santo chegando 38,5% em alguns tipos de especialidades.

No Reino Unido o absenteísmo é um dos grandes problemas do Serviço Nacional de Saúde (NHS), entre 2017 e 2018, 8 milhões de consultas foram perdidas gerando um custo aproximado de £ 1 bilhão por ano, este valor corresponderia a 257.000 substituições de quadril ou 990.000 operações de catarata. Nos Estados Unidos, segundo Gier (2019) o custo estimado de perda com a falta às consultas gira em torno de $150 bilhões, com a taxa de não comparecimento chegando a 30% em todo o país.

Segundo o GOVERNO DO ESTADO DE SANTA CATARINA (2019) um em cada três pacientes em Santa Catarina faltam às consultas marcadas, totalizando um total de 40000 pacientes faltantes no ano de 2018. Para conscientizar a população sobre este problema, o Ministério Público de Santa Catarina (2019),

lançou a campanha “SUS sem falta” que tenta demonstrar a importância de comparecer, ou comunicar previamente à falta nas unidades de saúde.

Verificando o problema causado pelo absenteísmo, o objetivo Geral do estudo é executar técnicas de *Machine Learning* no conjunto de dados chamado não comparecimento às consultas médicas, para a detecção dos pacientes faltantes. Os objetivos específicos serão: realizar uma análise exploratória para detecção de padrões nos dados, criar modelos de *Machine Learning* com os algoritmos de Redes Neurais, *Random Forest*, Regressão Logística, *XG boost* e *Adaboost* e aplicar técnicas de avaliação nos algoritmos e identificar qual tem o melhor poder preditivo.

O artigo está organizado da seguinte forma. A seção 2 conta com a fundamentação teórica. A seção 3 descreve a metodologia do trabalho . A seção 4 apresenta os resultados obtidos do estudo. A seção 5 apresenta a conclusão do estudo.

2 FUNDAMENTAÇÃO TEÓRICA

Com o crescimento exponencial da coleta de dados, as organizações observaram que é possível tirar vantagens competitivas e se destacar de suas concorrentes ao tirar *insights* deste amontoado de dados. É com este objetivo que surge a Ciência de Dados, uma área interdisciplinar voltada para a extração de conhecimento dos dados. Segundo Fawcett e Provost (2018, p.32): “Essa ampla disponibilidade de dados levou ao aumento do interesse em métodos para extrair informações úteis e conhecimento a partir de dados — o domínio de *data science*.”

Dentro da Ciência de Dados uma área que se destaca é a *Machine Learning*. O termo *Machine Learning* foi definido primeiramente por Arthur Samuel em seu artigo “*Some Studies in Machine Learning Using the Game of Checkers*”, onde neste artigo demonstrou como construiu um algoritmo que aprendeu a jogar damas. A sua definição para *Machine Learning* foi a seguinte: “campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados.” SAMUEL(1959). Basicamente *Machine Learning* tem a ver com processo pelo qual os computadores desenvolvem conhecimento através dos padrões contidos nos dados. Para Geron (2019) o uso de *Machine Learning* reduz tempo e aumenta a

precisão dos resultados. Se tivéssemos que programar todas as linhas de código manualmente seu programa provavelmente se tornaria uma longa lista de regras complexas — com uma manutenção muito difícil, em contrapartida, a utilização de algoritmos de *Machine Learning* o aprendizado acontece automaticamente, com o algoritmo descobrindo padrões não observados nos dados.

A pesquisa da aplicação de *Machine Learning* ocorre em vários setores da economia, na saúde ela é pesquisada tanto na parte administrativa, com objetivo de melhorar os processos, como na parte final do atendimento, como um auxiliar ao médico no diagnóstico de doenças. Neste contexto, alguns trabalhos de destacam, Hannun e Andrade (2019) utilizaram *Machine Learning* para prever se o paciente com doença renal terá compatibilidade com o órgão transplantado, Santos et al (2019) utilizaram *Machine Learning* para predizer quais idosos morreram num intervalo de até 5 anos, Barros et al (2020) usaram *Machine Learning* para predizer quais pacientes com doenças mentais têm mais probabilidade de praticar suicidio.

Olhando para a aplicação de *Machine Learning* na detecção do absenteísmo, os trabalhos são variados com métodos e tecnologias diferenciadas. Srinivas (2020) usou como características preditoras o histórico anterior de chegadas tardias, idade e nível de compromisso que indica se uma visita é agendada antes ou depois de um feriado nacional. Os dados foram obtidos em duas clínicas de especialidades diferentes (ENT e WH) localizadas em um centro médico regional na Pensilvânia, EUA. Comparando 4 algoritmos obteve como melhor desempenho o valor de AUC de 90,4% com o algoritmo Máquinas de Aumento de Gradiente.

Ferro et al (2020) estudou o comportamento do não comparecimento entre pacientes de baixa renda em comunidades carentes de Bogotá, Colômbia. Utilizaram algoritmos como Redes Neurais e *Random Forest* e descobriram que estatísticas como renda e crime na vizinhança afetam as probabilidades dos pacientes de não comparecer às consultas.

Daghistani et al (2020) propôs uma estrutura de big data para identificar o não comparecimento de pacientes ambulatoriais por meio de engenharia de recursos e aprendizado de máquina (MLlib) na plataforma *Spark*. Com 2.011.813 de dados de consultas ambulatoriais, utilizou 5 algoritmos e tendo como melhor resultado o algoritmo *Gradient Boosting* com a precisão de 79% e o ROC[2] de 81%.

[2] Curva ROC é um gráfico para analisar o comportamento do modelo, no eixo x tem-se a taxa de falsos positivo (FP/(FP+TN)) e no eixo y tem-se a taxa de verdadeiro positivo(TP/(TP+FN)). Quanto

Dashtban e Li (2019) aplicaram redes neurais em dados hospitalares com mais de 1 milhão de registros. Utilizaram uma ampla gama de fatores de saúde, meio ambiente e economia social, com o modelo obtendo bons resultados nos conjuntos de dados, com a acurácia de 0,69% e AUC[3] de 0.71%.

Chen et al (2021) utilizou dados de registros eletrônicos de saúde secundários (EHR) de uma clínica de oftalmologia pediátrica de Portland. O melhor algoritmo foi, *XGBoost*, que teve um AUC de 0,90 para prever o não comparecimento.

ROBAINA et al(2020) concluiu que alguns fatores como tempo entre a consulta e o agendamento e o tipo de pagador(seguro privado apresentou uma taxa de não comparecimento de 9% em comparação com 15% em pacientes com seguro governamental) são fatores essenciais para o absenteísmo.

3 METODOLOGIA

O trabalho tem como finalidade realizar uma pesquisa aplicada, através do método descritivo. A abordagem utilizada será a quali-quantitativa que utiliza tanto o método estatístico como a abordagem teórica. Para Gisele (2019) a utilização da pesquisa quali-quantitativa é vantajosa quando os problemas da pesquisa são complexos e as outras abordagens não fornecem as respostas necessárias.

O método de pesquisa utilizado no trabalho foi o hipotético-dedutivo. Para Carlos (2019) o método hipotético-dedutivo é praticado quando os conhecimentos disponíveis sobre determinado assunto são insuficientes para a explicação de um fenômeno. Para tentar explicar a dificuldade expressa no problema, são formuladas hipóteses. Das hipóteses formuladas, deduzem-se consequências observáveis, que deverão ser falseadas. Falsear significa tentar tornar falsas as consequências deduzidas das hipóteses. Enquanto no método dedutivo procura-se a todo custo confirmar a hipótese, no método hipotético-dedutivo, ao contrário, procuram-se evidências empíricas para derrubá-la.

---

mais para a cima e a esquerda melhor o modelo é na predição, se o modelo estiver abaixo da linha pontilhada a predição é puramente aleatória, não tendo nenhuma habilidade na identificação das classes.

[3] A área abaixo da curva ROC é chamada de AUC (área abaixo da curva) onde um classificador perfeito tem um AUC igual a 1, enquanto um classificador com AUC igual a 0.5 ou menor é puramente aleatório.

Foi realizada uma pesquisa bibliográfica em artigos científicos no site da CAPES sendo selecionado a base de dados *WEB OF SCIENCE*. Também foi utilizado a base de dados da EBSCO. A busca se deu por meio das palavras chaves "*Machine Learning*", sendo adicionada uma segunda palavra chave "*missing appointments*".

3.1 Algoritmos

3.1.1 *Random Forest*

O algoritmo *Random Forest* foi desenvolvido por Breiman (2001) e faz parte da técnica denominada *Ensemble Learning*, que combina uma coleção de modelos para fazer uma única previsão. Segundo Hastie, Tibshirani e Friedman (2009), *Random Forest* são baseados em várias Árvores de Decisão, onde cada árvore é criada utilizando o método chamado *Bagging*(*Bootstrap Aggregating*). O método *bagging* realiza o treinamento de cada árvore em subconjuntos diferentes de dados, que são obtidas através de uma reamostragem com reposição dos dados de treinamento.

Após a seleção dos dados é feita a seleção randomicamente das features que formam cada árvore, tendo cada árvore características diferentes uma da outra, uma forma de evitar a correlação entre elas. Segundo Geron (2019, p.193):

> O algoritmo Floresta Aleatória introduz uma aleatoriedade extra ao desenvolver árvores; em vez de buscar a melhor característica ao dividir um nó , ele a busca entre um subconjunto aleatório dessas características, resultando em uma grande diversidade da árvore, que troca um alto viés por uma baixa variância geralmente produzindo um melhor modelo no geral.

Após o treinamento, a predição da classe é feita pelo voto da maioria de todas as árvores.

3.1.2 *Boosting*

Os algoritmos chamados de *Boosting* também fazem parte do método de aprendizado chamado *Ensemble Learning*, segundo Geron (2019) a ideia da maioria dos algoritmos de *Boosting* é treinar sequencialmente os previsores tentando corrigir o erro do seu antecessor. Os principais algoritmos são o *Adaboost* e o *Gradient Boosting (XG Boost)*.

O *Adaboost* foi desenvolvido por Freund e Schapire (1996), é baseado em um algoritmo fraco, geralmente sendo usado as árvores de decisão. Segundo Geron(2019) o algoritmo funciona da seguinte forma: primeiro treinamos um classificador de primeira base, após isso utilizamos para fazer as previsões nos dados de treinamento, o peso relativo das instâncias classificadas erroneamente é aumentado, o próximo classificador é treinado utilizando os dados com pesos e os pesos atualizados, assim sucessivamente.

Segundo Kuhn e Johnson (2018) esse aumento do peso das instâncias classificadas incorretamente, faz com que instâncias difíceis de classificar recebem cada vez maiores pesos até que o algoritmo identifique um modelo que classifique corretamente essas amostras. Portanto, cada iteração do algoritmo é necessária para aprender um aspecto dos dados, concentrando-se em regiões que contêm amostras de difícil classificação.

Ao fim do treinamento, os algoritmos mais precisos terão os maiores pesos e os algoritmos menos precisos terão pesos menores, para fazer as previsões, o *AdaBoost* simplesmente calcula as previsões de todos os algoritmos e os pondera usando os pesos dos algoritmos. A classe do previsor é aquela que recebe a maioria dos votos ponderados.

O *Gradient Boosting* foi desenvolvido por Friedman (2001) em seu artigo "*Greedy function approximation: A gradient boosting machine*" é diferentemente do *Adaboost* que tenta dar pesos maiores para os algoritmos que mais acertam, ele tenta ajustar um novo algoritmo aos erros residuais dos algoritmos anteriores. Segundo Geron (2019) primeiro treina-se um algoritmo, obtém-se o erro residual (a diferença entre o que foi previsto e os valores reais), utiliza-se este erro residual como a resposta para a criação do próximo algoritmo, obtém-se o novo erro residual, constrói um novo algoritmo em cima deste erro residual, e atualiza-se os erros residuais, assim sucessivamente.

Segundo Kuhn e Johnson (2018) o algoritmo começa com a função de perda (por exemplo, erro quadrático médio para regressão) e um modelo de aprendiz fraco, o algoritmo busca encontrar um modelo aditivo que minimize a função de perda. O algoritmo é normalmente inicializado com a melhor estimativa da resposta (a média para regressão), o gradiente é calculado e um modelo é então ajustado aos resíduos para minimizar a função de perda. O atual modelo é adicionado ao modelo anterior, e o procedimento continua por um número de iterações especificado pelo usuário.

3.1.3 Regressão Logística

A Regressão Logística surge do desejo de modelar a probabilidade de classe através de uma função linear. Ela é comumente usada para estimar a probabilidade de uma instância pertencer uma classe, onde se a probabilidade estimada for maior que 50%, então o modelo prevê que a instância pertence a essa classe, se for menor que 50%, a probabilidade é que não pertença.

Segundo Hair (2009) a estimação dos coeficientes da regressão logística, ao contrário da regressão múltipla que utiliza o método dos mínimos quadrados, é efetuada pelo uso da máxima verossimilhança. No lugar de minimizar os desvios quadrados (mínimos quadrados), a regressão logística maximiza a probabilidade de que um evento ocorra. O ajuste da estimação do modelo dá-se pelo valor -2 vezes o logaritmo da verossimilhança (-2LL), sendo que quanto menor este valor, melhor o modelo.

Segundo Geron (2019), assim como um modelo de Regressão Linear, um modelo de Regressão Logística calcula uma soma ponderada das características de entrada, mas, em vez de gerar o resultado diretamente como o modelo de Regressão Linear, gera a logit desse resultado. A *logit* é uma função sigmóide (formato em S), que mostra os resultados em um intervalo entre 0 e 1 e está definida na equação 3.1:

Equação 3.1: Função *Logit*

$$y = \frac{1}{1 + e^{-(b0+b1*x1)}}$$

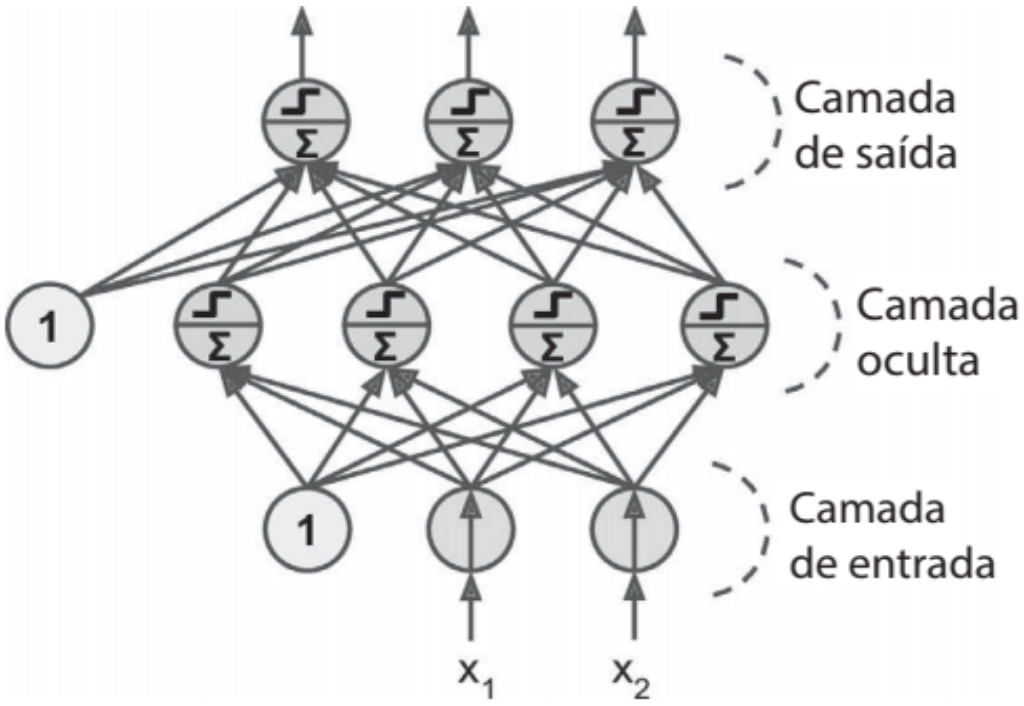
Para ajustar um modelo de regressão logística, é necessário estimar os parâmetros b0 e b1 do modelo, a partir dos dados de amostra. A estimação por máxima verossimilhança, permite encontrar os estimadores dos parâmetros do modelo, que tem maior probabilidade de replicar o padrão de observações, nos dados.

3.1.4 Redes Neurais

As Redes Neurais são um dos mais importantes algoritmos da atualidade, sendo usado tanto em problemas comuns como classificação e regressão, ou em problemas mais complexos como detecção de imagens. Ele é desenvolvido tendo como inspiração o cérebro humano, onde existem vários neurônios conectados se comunicando através de sinais.

A arquitetura de uma rede neural artificial simples é composta por uma camada de entrada que contém os dados, uma ou mais camadas ocultas (quando uma rede neural possui duas ou mais camadas ocultas, ela recebe o nome de rede neural profunda), onde a informação é passada através de cada camada, com a saída da camada anterior fornecendo entrada para a próxima camada. Por fim a camada de saída que contém o resultado do algoritmo. A figura 1 apresenta a arquitetura de uma rede neural com uma camada oculta.

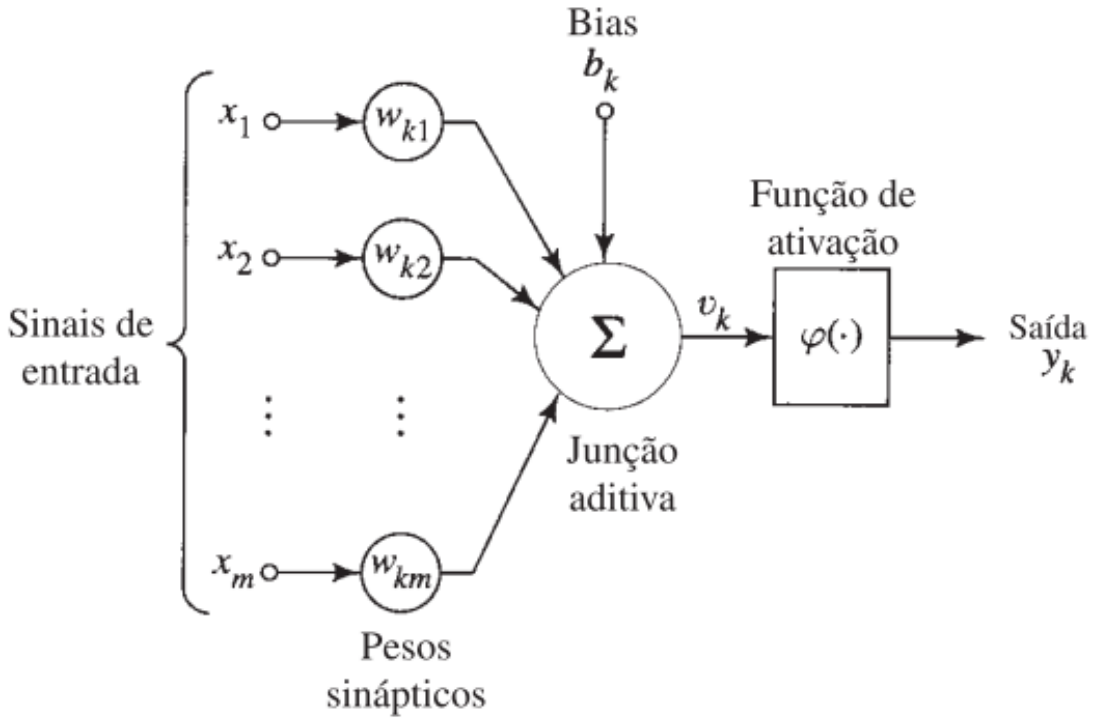
Figura 1: Arquitetura de uma rede neural de uma única camada oculta



Fonte: Geron (2019)

Dentro de cada camada existem vários neurônios, segundo Haykin (2007) um neurônio é uma unidade de informação que é fundamental para a operação de uma rede neural. É composto por três elementos básicos: o primeiro é um conjunto de sinapses ou elos de conexão, cada um caracterizada por um peso ou força própria, podendo estar em um intervalo que inclui valores positivos e negativos, a segunda é um somador para somar os sinais de entrada, ponderadas pelas respectivas sinapses dos neurônio e a terceira uma função de ativação para restringir a amplitude de saída de um neurônio. A figura 2 demonstra a representação de um neurônio artificial.

Figura 2: Representação de um Neurônio Artificial



Fonte:Haykin (2007)

Segundo FUNÇÃO DE ATIVAÇÃO, as funções de ativação são o elemento essencial para as redes neurais, ela que vai decidir se um neurônio vai ser ativado ou não, se a informação que o neurônio está recebendo é relevante para a informação fornecida ou deve ser ignorada. A função de ativação faz a transformação não-linear nos dados de entrada, tornando-o capaz de aprender e executar tarefas mais complexas. A figura 2 apresenta um neurônio a mais chamado de bias, que é uma forma de regularização, e que tem como função aumentar ou diminuir a entrada líquida, de forma que um neurônio apresenta saída não nula ainda que todas as suas entradas sejam nulas.

Um avanço importante na história das Redes Neurais se deu com a criação do algoritmo *Backpropagation*. Foi desenvolvido por Rumelhart, Hinton e Williams (1986) em seu artigo "*Learning representations by back-propagating errors*", onde forneceu um método computacional eficiente para o treinamento de *perceptron* de várias camadas. O algoritmo consiste em duas fases: a primeira fase é chamada de *forward pass* ou propagação, segundo o ALGORITMO *BACKPROPAGATION* o propósito é a de propagar nossas entradas (os dados de entrada) através da rede

aplicando uma série de *dot products* (multiplicação entre os vetores) e ativações até chegarmos à camada de saída da rede (ou seja, nossas previsões).

A segunda fase é chamada de *backward pass* ou retropropagação, onde o algoritmo mede o erro, a diferença da resposta da rede neural e a resposta verdadeira, então passa por cada camada da direita para a esquerda para medir a contribuição do erro em cada conexão e finalmente, ajusta os pesos da conexão para reduzir o erro. Segundo Geron (2019, p.268):

> …ele mede o erro de saída da rede (isto é, a diferença entre a saída desejada e a saída real da rede) e calcula o quanto cada neurônio contribuiu para o erro de cada neurônio de saída na última camada oculta. Em seguida, passa a medir a quantidade dessas contribuições de erro provenientes de cada neurônio na camada oculta anterior, e assim por diante até o algoritmo alcançar a camada de entrada. Esta passagem reversa mede eficientemente o gradiente de erro em todos os pesos de conexão na rede, propagando para trás o gradiente de erro na rede (daí o nome do algoritmo).

Para o ALGORITMO *BACKPROPAGATION*, ele é o algoritmo-chave que faz o treinamento de modelos profundos algo computacionalmente tratável. Com o Algoritmo *Backpropagation* foi possível acelerar o treinamento das redes neurais, um modelo que levava anos para ser treinado agora leva algumas horas ou dias.

3.2.1 Conjunto de dados

Os dados utilizados neste projeto são de domínio público e estão disponíveis no site do Kaggle[4], originários do Brasil. A um total de 110.527 dados com 14 atributos e uma variável resposta binária com nome *No-Show*, que indica se o paciente faltou ou não. A primeira coluna é *PatientId* que é a identificação do paciente, a segunda é *AppointmentID* que é a identificação da consulta, a terceira é *Gender* que se é homem ou mulher, quarta é a *ScheduledDay* que é o dia do agendamento da consulta, quinta é *AppointmentDay* que é o dia da consulta, sexta é Age que é a idade do paciente, sétima é *Neighbourhood* que é o bairro que a consulta será realizada.

A oitava coluna é *Scholarship* uma variável binária que indica se o paciente é elegível a receber o bolsa família, nona é *Hipertension* se o paciente tem hipertensão, décima é Diabetes se o paciente tem diabetes, décima primeira

[4] https://www.kaggle.com/joniarroba/noshowappointments/discussion/42835

Alcoholism se o paciente tem o transtorno de alcoolismo, décima segunda é *Handcap* que indica se o paciente tem alguma deficiência variando de 0 que não tem até 4 que é uma deficiência grave e a última é *SMS_received* se o paciente recebeu alguma mensagem de aviso da consulta.

3.2.2 Pré-processamento dos dados

O pré-processamento foi composto pelas seguintes ações: mudança dos nomes das colunas, mudança de tipos de dados e a criação de novas colunas com informações mais significativas. O primeiro passo realizado foi a mudança dos nomes das colunas: *PatientId* foi modificada para Id paciente, *AppointmentID* foi para Id consulta, *Gender* para sexo, *ScheduledDay* para data agendamento, *AppointmentDay* para data consulta, *Age* para idade, *Neighbourhood* para bairro, *Scholarship* para bolsa familia, *Hipertension* para hipertensão, *Diabetes* para diabetes, *Alcoholism* para alcoolismo, *Handcap* para deficiencia, *SMS_received* para sms recebido e *No-Show* para não comparecimento.

Após isso houve a divisão do conjunto de dados em treino e teste, com 70% sendo para o treino e 30% para o teste. As colunas data consulta e data agendamento foram convertidas em *datetime*, após isso foi extraída a diferença entre o dia de marcação da consulta e a consulta. Em seguida foi criada a coluna Tempo de espera, através de uma categorização com as seguintes regras: se o valor for igual ou menor que 5 o tempo de espera é rápido, maior que 5 e menor ou igual a 15 o tempo de espera é médio, maior que 15 e menor ou igual a 60 o tempo de espera é demorado e por fim maior que 60 é muito demorado. Ainda nas colunas data consulta e data agendamento foram criadas mais duas colunas, a primeira Dia da consulta que é o dia da semana que ocorreu a consulta e Marcação da consulta que é o dia da semana que ocorreu a marcação da consulta. Na coluna sexo foi modificado o F por 1 e o M por 0. A mesma substituição ocorreu na coluna não comparecimento onde Yes foi substituído por 1 e No foi trocado por 0. Na coluna idade todas as linhas com idades negativas foram removidas do conjunto de dados. Com as idades foi criada a coluna Idade, que contém as categorizações com as seguintes regras: menor ou igual a 12 é criança, maior que 12 e menor ou igual a 18 é adolescente, maior que 18 e menor ou igual a 25 é jovem adulto, maior que 25 e menor ou igual a 60 é adulto e maior que 60 idoso.

Após estes pré-processamentos foram removidas algumas colunas, sobrando as seguintes 12: bairro, bolsa familia, Hipertensão, Diabetes, Alcoolismo, deficiencia, sms recebido, Marcação da consulta, Dia da consulta, Tempo de espera, Sexo, Idade. A seguir as colunas bairro, Marcação da consulta, Dia da consulta, Tempo de espera, Idade foram transformadas em *OneHotEncoder*, tendo o conjunto de dados final 109 colunas.

3.2.3 Desbalanceamento dos dados

Um problema encontrado no conjunto de dados foi a existência de uma discrepância entre quem não compareceu às consultas e quem compareceu, com apenas 20% dos dados sendo de pacientes que compareceram. Se os algoritmos receberem este conjunto de dados, haverá uma grande probabilidade de existir um viés para identificar a classe de maior predominância, não tendo nenhum efeito na predição dos pacientes faltantes. Para corrigir este contratempo utiliza-se um método de amostragem ("*sampling"*), que tenta igualar a diferença entre as classes através de uma reamostragem.

Existem dois tipos de amostragem, o primeiro chama-se *Over-sampling* que foca nas classes minoritárias, aumentando esta classe através de remamostragem até se igualar a classe majoritária. O segundo chama-se *Under-sampling*, que foca na classe de maior frequência, eliminando aleatoriamente entradas da classe de maior ocorrência. Para resolver o problema no conjunto de dados utilizou-se o método de *Over-sampling* o que acarretará o aumento do conjunto de treinamento de 77368 para 123336 instâncias.

3.3 Métricas de avaliação

As métricas mais usadas para avaliação de algoritmos são acurácia, sensibilidade e especificidade. Segundo Lavrac (1999) acurácia mede quanto o modelo acertou de todas as respostas, a sensibilidade mede quão bem a classe verdadeira o modelo acertou e especificidade que é o complemento da sensibilidade e mede quão bem a classe negativa o modelo acertou.

Para avaliar quão boa foi a predição do modelo, existem dois indicadores, o PPV (valor preditivo positivo) e NPV (valor preditivo negativo). Segundo Kuhn e Johnson (2018), o PPV indica quantas respostas o modelo disse que eram verdadeiras realmente são verdadeiras e o NPV indica quantas respostas o modelo disse que são negativas realmente são negativas.

Outras métricas importantes são a curva ROC e o AUC. Para Bradley (1997), Curva ROC é um gráfico para analisar o comportamento do modelo, no eixo x tem-se a taxa de falsos positivo (FP/(FP+TN)) e no eixo y tem-se a taxa de verdadeiro positivo(TP/(TP+FN)). Quanto mais para a cima e a esquerda melhor o modelo é na predição, se o modelo estiver abaixo da linha pontilhada a predição é puramente aleatória, não tendo nenhuma habilidade na identificação das classes. A área abaixo da curva ROC é chamada de AUC (área abaixo da curva) onde um classificador perfeito tem um AUC igual a 1, enquanto um classificador com AUC igual a 0.5 ou menor é puramente aleatório.

3.4 Construção dos modelos Preditivos

Ao todo foram testados cinco algoritmos *Redes Neurais*, *Random Forest*, *Regressão Logística, XG boost* e *Adaboost*. Para encontrar os melhores hiperparâmetros dos algoritmos foi utilizado um método de otimização chamado *Random Search*.

O *Random Search* procura os melhores hiperparâmetros aleatoriamente, para isso se imputa uma lista de hiperparâmetros do algoritmo e o *Random Search* avalia algumas combinações tentando identificar qual traz o melhor poder de predição. Os parâmetros do estimador usados para aplicar esses métodos são otimizados por pesquisa de validação cruzada.

Após vários teste os hiperparâmetros dos modelos de machine learning que obtiveram os melhores valores no Random Search foram os seguintes: *Random Forest* (criterion=entropy, max_depth=100, max_features=8, splitter=random), *XGBoost*(learning_rate=1.2, max_features=log2, n_estimators=500), Regressão Logística(solver=liblinear) e *AdaBoostClassifier(algorithm='SAMME', learning_rate=0.0001, n_estimators=10)*.

Para as *Redes neurais,* os testes foram feitas manualmente tendo a estrutura definida com 4 camadas, as 3 primeiras com funções de ativação relu e a quarta com a função de ativação sigmóide. O otimizador é o Descida do Gradiente com learning rate de 0.01, com função de perda binary cross entropy e a métrica de avaliação foi Accuracy.

4 RESULTADOS

4.1 Análise descritiva

Foi realizada inicialmente uma análise descritiva dos dados. Em um total de 110.527 registros a idade média dos pacientes foi de 37 anos com um desvio padrão de 23,11, o sexo feminino foi o mais presente nos conjunto de dados representando 65%.

Quadro 1: Características descritivas do conjunto de dados

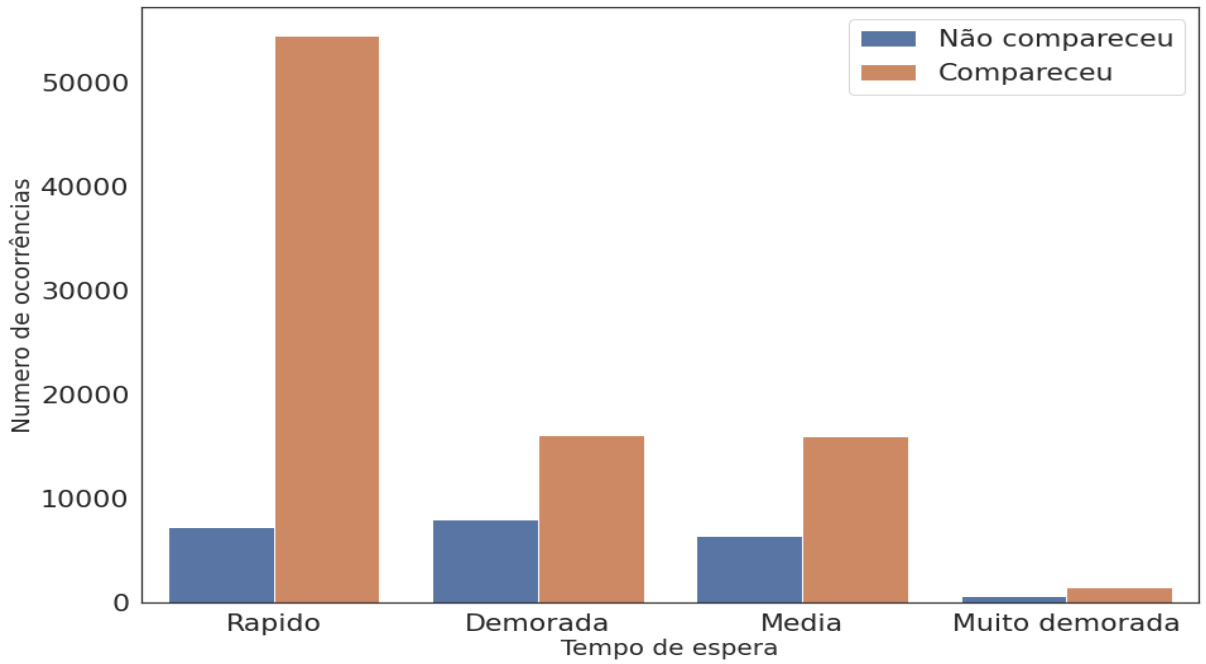
Colunas	**Não Compareceu(P%)**	**Compareceu(P %)**	**Total**
Sexo			
Feminino	14594 (64,90%)	57246 (65,39%)	71840
Masculino	7725 (35,10%)	30962 (34,61%)	38687
Bolsa família			
0	19741 (88,45%)	79925 (90,61%)	99666
1	2578 (11,55%)	8283 (9,39%)	10861
Hipertensão			
0	18547 (83,1%)	70179 (79,56%)	88726
1	3772 (16,9%)	18029 (20,44%)	21801
Diabetes			
0	20889 (93,59%)	81695 (92,62%)	102584
1	1430 (6,41%)	6513 (7,38%)	7943

Alcoolismo			
0	21642 (96,97%)	85525 (96,96%)	107167
1	677 (3,03%)	2683 (3,04%)	3360
Deficiência			
0	21912 (98,18%)	86374 (97,9%)	108286
1	366 (1,64%)	1676 (1,9%)	2042
2	37 (0,17%)	146 (0,166%)	183
3	3 (0,01%)	10 (0,011%)	13
4	1 (0,0000%)	2 (0,000%)	3
SMS recebido			
0	12535 (56,16%)	62510 (70,87%)	75045
1	9784 (43,84%)	25698 (29,13%)	35482
Dia da consulta			
Segunda	4690 (21,01%)	18025 (20,43%)	22715
Terça	5152 (23,08%)	20488 (23,22%)	25640
Quarta	5093 (22,82%)	20774 (23,56%)	25867
Quinta	3338 (14,96%)	13909 (15,77%)	17247
Sexta	4037 (18,09%)	14982 (16,99%)	19019
Sábado	9 (0,04%)	30 (0,03%)	39
Marcação da consulta			
Segunda	4561 (20,44%)	18524 (21,00%)	23085
Terça	5291 (23,70%)	20877 (23,67%)	26168
Quarta	4879 (21,86%)	19383 (21,97%)	24262
Quinta	3700 (16,58%)	14373 (16,29%)	18073
Sexta	3887 (17,42%)	15028 (17,04%)	18915
Sábado	1 (0,00004%)	23 (0,03%)	24

Fonte: Primária (2022).

Uma grande diferença observada entre os pacientes que compareceram e não compareceram foi que, os pacientes que não compareceram tiveram mais consultas categorizadas como demoradas, com 35,79%, já os pacientes que compareceram, a uma maior concentração como consultas rápidas, com 61,84%. O tempo médio de espera para a consulta foi de 10,53 com um desvio padrão de 15,07, os pacientes faltantes tiveram uma média de 15,9, contra 9,17 dos que compareceram.

Figura 3: Tempo de espera para a consulta

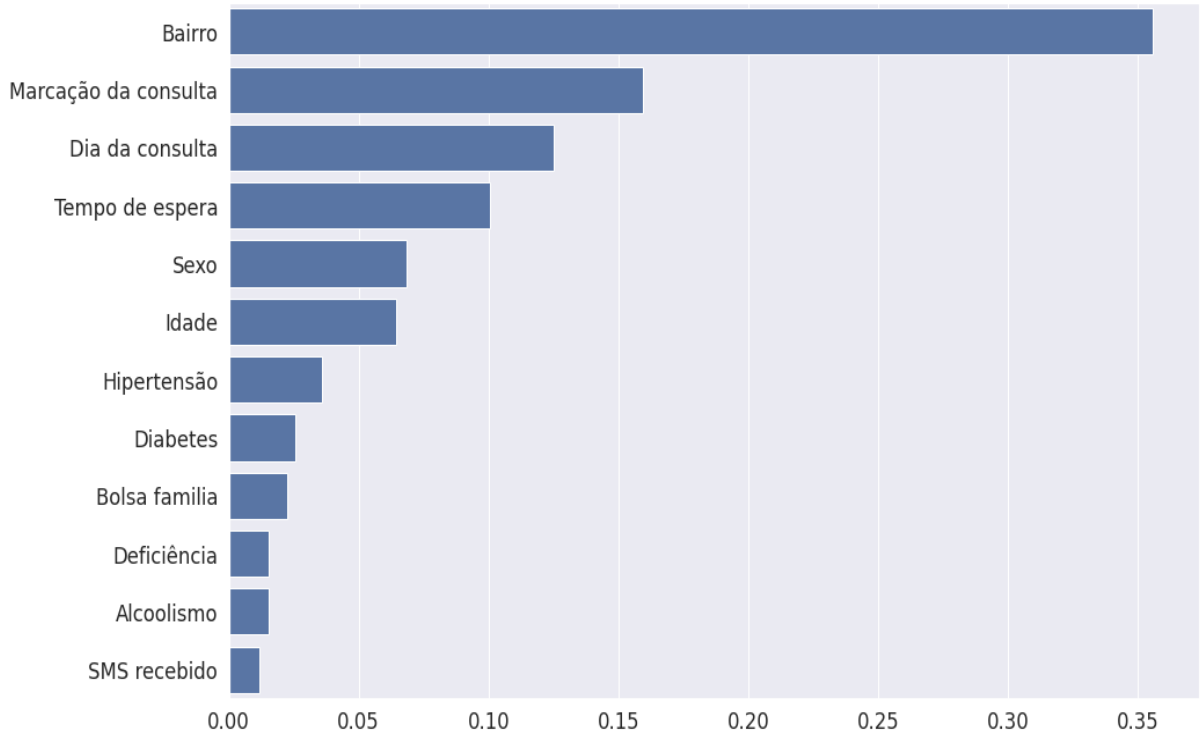


Fonte: Primária (2022).

4.2 Importância das variáveis

A classificação da importância das variáveis é encontrada através do cálculo do ganho de informação. Para Garcia (2003) ela mede quanta informação um atributo nos dá sobre a classe, quanto maior for o valor do ganho de informação mais significativo o atributo terá para o modelo. A figura 4 demonstra como cada variável contribui para o modelo.

Figura 4:Importância das variáveis para o modelo preditivo



Fonte: Primária (2022).

Os atributos que mais contribuíram para os modelos foram o bairro, dia da marcação da consulta, dia da consulta e tempo de espera, já os que tiveram menor poder de predição foram sms recebido, alcoolismo, deficiência e Bolsa Família.

4.3 Resultados dos modelos

A tabela 2 apresenta o resultado dos algoritmos usando os dados de teste como avaliação. As Redes Neurais foram melhores em três métricas de avaliação, obtendo o melhor resultado na Área Abaixo da Curva com 0.67, na sensibilidade com 0.85 e no NPV com 0.90. O *Random Forest* foi o melhor na Acurácia tendo o maior número de acertos com 0.70.

Tabela 2: Resultados dos modelos utilizando os dados de teste

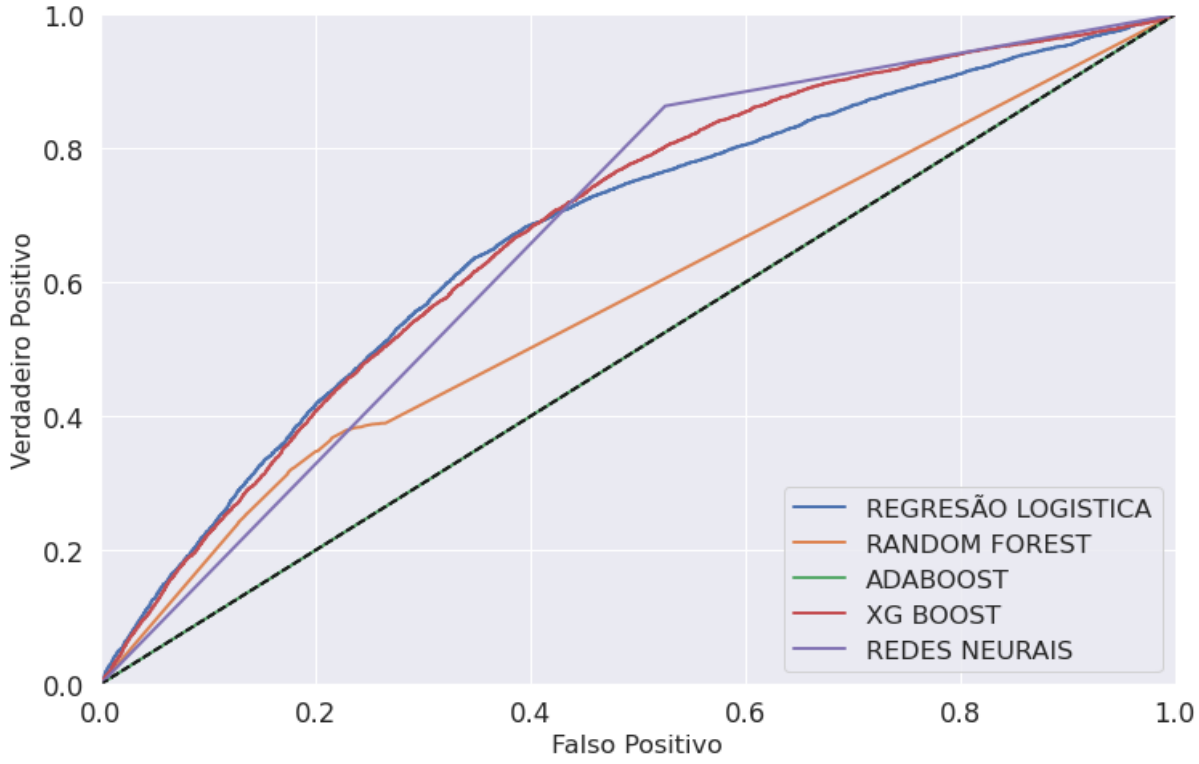
Algoritmos	**AUC**	**Sensibilidade**	**Especificidade**	**Acurácia**	PPV	NPV
Random	0,59	0,41	0,61	0,70	0,31	0,84

Forest						
Redes Neurais	0,67	0,85	0,48	0,56	0,29	0,90
Regressão Logística	0,64	0,66	0,63	0,63	0,30	0,88
XG boost	0,65	0,64	0,66	0,65	0,32	0,88
Adaboost	0,64	0,67	0,58	0,63	0,31	0,67

Fonte: Primária (2022).

Para entender melhor a eficiência dos modelos, a Figura 3 apresenta a Curva ROC para os cinco algoritmos.

Figura 5: Curva ROC



Fonte: Primária (2022).

A Rede Neural foi a que obteve o melhor resultado na curva ROC, indicada pelo maior AUC com valor de 0.67. Regressão Logística e Adaboost se comportaram identicamente e o pior modelo foi o Random Forest.

Por fim o melhor modelo foi às Redes Neurais, mesmo tendo uma baixa acurácia, conseguiu a maior sensibilidade e com isto, a capacidade de melhor identificar os pacientes que iriam faltar.

5 CONSIDERAÇÕES FINAIS

Este estudo teve como objetivo executar técnicas de Machine Learning no conjunto de dados chamado não comparecimento às consultas médicas, para a detecção dos pacientes faltantes. Para alcançar os resultados, aplicou-se tratamentos nos dados, como mudanças dos tipos de dados, criação de novas colunas com maiores poderes preditivos e um método de amostragem ("*sampling"*) para equilibrar as classes.

Na análise descritiva dos dados percebemos que a maioria das características não tinham uma diferença entre quem compareceu e não. Há uma única variável com uma discrepância entre os dois, foi o tempo de espera para a consulta, onde quem compareceu tinha um tempo de espera maior categorizada como rápida.

Olhando para a importância das variáveis, percebemos que a característica que mais teve impacto para a diferenciação entre as classes foi o bairro onde a consulta será realizada. Isso pode ser um indicativo que pessoas que moram mais longe dos locais de consulta tenham dificuldade de se locomover e desistam de comparecer às consultas. Muitos outros fatores podem influenciar o não comparecimento do paciente que não puderam ser inferidos a partir do conjunto de dados analisados.

Foi realizado o treinamento de 5 algoritmos, sendo eles Redes Neurais, *Random Forest*, Regressão Logística, *XG boost* e *Adaboost,* sendo encontrando os melhores hiperparâmetros dos modelos pelo método de *Random Search.* Após analisar as métricas de desempenho, o modelo de Redes neurais mostrou-se uma boa escolha para ser o modelo final para o conjunto de dados disponível. Ele obteve a melhor taxa de sensibilidade de 0,85 e o maior AUC com 0,65.

Os resultados obtidos com os algoritmos possibilitam o desenvolvimento de classificadores eficientes para detecção do absenteísmo. Este estudo pode servir como referência de uso de técnicas de aprendizado de máquina para o sistema de

saúde público e privado. A disponibilidade de informações adicionais sobre pacientes e consultas médicas pode ajudar a melhorar a capacidade do modelo de aprender novos comportamentos.

Para futuros trabalhos com esse conjunto de dados, seria interessante a inclusão de mais características para a melhora do modelo. Como exemplo a incorporação de duas variáveis, o clima no dia das consultas e se a consulta é realizada por órgãos governamentais ou entidades privadas, tendo como meta a análise de como o paciente se comporta quando ele que paga a consulta.

REFERÊNCIAS

ALGORITMO BACKPROPAGATION**. Parte 1 - Grafos Computacionais e Chain Rule**. Deep Learning Book. Disponível em: <https://www.deeplearningbook.com.br/algoritmo-backpropagation-parte1-grafos-computacionais-e-chain-rule/>. Acesso em: 31 Mar. 2022.

ALGORITMO BACKPROPAGATION. **Parte 2 - Treinamento de Redes Neurais**. Deep Learning Book. Disponível em: <https://www.deeplearningbook.com.br/algoritmo-backpropagation-parte-2-treinamento-de-redes-neurais/>. Acesso em: 31 Mar. 2022.

ALMUHAIDEB, Sarab *et al*. **Prediction of hospital no-show appointments through artificial intelligence algorithms**. Annals of Saudi Medicine, v. 39, n. 6, p. 373–381, 2019.

FERRO, David Barreira *et al*. **Improving healthcare access management by predicting patient no-show behaviour**. Decision Support Systems, v. 138, p. 113398, 2020.

BARROS, Jorge *et al*. **Suicide detection in Chile: proposing a predictive model for suicide risk in a clinical sample of patients with mood disorders**. Revista Brasileira de Psiquiatria, v. 39, n. 1, p. 1–11, 2016.

BATOOL, Tasneem *et al*. **Predicting Hospital No-Shows Using Machine Learning**. *In*: 2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS). [s.l.]: IEEE, 2021. Disponível em: <http://dx.doi.org/10.1109/iotais50849.2021.9359692>. Acesso em: 6 Jan. 2022.

BERGSTRA, James; BENGIO, Yoshua. 2012. **Random search for hyper-parameter optimization**. J. Mach. Learn. Res. 13, null (3/1/2012), 281–305.

BERTOZZO, Richard. Aplicação de Machine Learning em Dataset de consultas médicas do SUS. 2019. Monografia(Grau de Bacharel em Sistemas de Informação). Universidade Federal de Santa Catarina. Florianópolis. 2019.

BRADLEY, Andrew P. **The use of the area under the ROC curve in the evaluation of machine learning algorithms**. Pattern Recognition, v. 30, n. 7, p. 1145–1159, 1997.

BREIMAN, Leo (2001). **Random Forests**. Machine Learning 45 (1): 5–32. doi:10.1023/A:1010933404324.

CHEN, Jimmy et al. **Application of Machine Learning to Predict Patient No-Shows in an Academic Pediatric Ophthalmology Clinic**. AMIA. Annual Symposium proceedings. AMIA Symposium vol. 2020 293-302. 25 Jan. 2021

DA COSTA, Thiago Martini *et al*. **The impact of short message service text messages sent as appointment reminders to patients' cell phones at outpatient clinics in São Paulo**, Brazil. International Journal of Medical Informatics, v. 79, n. 1, p. 65–70, 2010.

DAGHISTANI, Tahani *et al*. **Predictors of outpatients' no-show: big data analytics using apache spark**. Journal of Big Data, v. 7, n. 1, 2020.

DANTAS, Leila F *et al*. **No-shows in appointment scheduling – a systematic literature review**. Health Policy, v. 122, n. 4, p. 412–421, 2018.

DASHTBAN, Muhammad; LI, Weizi. **Deep Learning for Predicting Non-attendance in Hospital Outpatient Appointments**. *In*: Proceedings of the

Annual Hawaii International Conference on System Sciences. [s.l.]: Hawaii International Conference on System Sciences, 2019. Disponível em: <http://dx.doi.org/10.24251/hicss.2019.451>. Acesso em: 6 Jan. 2022.

ENGLAND, NHS. **NHS England » NHS to trial tech to cut missed appointments and save up to £20 million**. 14 de outubro de 2018. Disponível em: <https://www.england.nhs.uk/2018/10/nhs-to-trial-tech-to-cut-missed-appointments-and-save-up-to-20-million>. Acesso em: 6 Jan. 2022.

FAN, Guorui *et al*. **Machine learning-based prediction models for patients no-show in online outpatient appointments**. Data Science and Management, v. 2, p. 45–52, 2021.

FAWCETT, Tom; PROVOST, Foster. **Data Science para Negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados**. Rio de Janeiro: Alta Books Editora, 2018.

FREUND, Yoav; SCHAPIRE, Robert E. **Experiments with a New Boosting Algorithm**. Machine Learning: Proceedings of the Thirteenth International Conference, p. 148–156, 1996.

FRIEDMAN, Jerome H. **Greedy function approximation: A gradient boosting machine**. The Annals of Statistics, v. 29, n. 5, 2001.

FUNÇÃO DE ATIVAÇÃO. Deep Learning Book. Disponível em: <https://www.deeplearningbook.com.br/funcao-de-ativacao/>. Acesso em: 30 Mar. 2022.

GARCIA, S. C. **O uso de árvores de decisão na descoberta de conhecimento na área da saúde**. Dissertação (Mestre em Ciência da Computação) –Universidade Federal do Rio Grande do Sul. Porta Alegre, p. 88. 2003.

Géron, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Rio de Janeiro.Alta Books.2019.

GIER, Jamie. **Missed appointments cost the U.S. healthcare system $150B each year**. Hea! thcare innovation. Disponivel: <https://www.hcinnovationgroup.com/clinical-it/article/13008175/missed-appointments-cost-the-us-healthcare-system-150b-each-year>. Acesso em: 26 de 2021.

GIL, Antônio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GOVERNO DO ESTADO DE SANTA CATARINA. **UM A CADA TRÊS PACIENTES FALTA A CONSULTAS OU EXAMES AGENDADOS EM SC** . Site oficial do governo do estado de Santa Catarina. Disponível em: <https://www.saude.sc.gov.br/index.php/noticias-geral/todas-as-noticias/1641-noticias-2019/10720-um-a-cada-tres-pacientes-falta-a-consultas-ou-exames-agendados-em-sc> . Acesso em: Julho 2021.

GUORUI, Fan et al. **Machine learning-based prediction models for patients no-show in online outpatient appointments**, Data Science and Management, Volume 2, 2021, Pages 45-52, ISSN 2666-7649, (https://www.sciencedirect.com/science/article/pii/S2666764921000175)

HANNUN, Pedro Guilherme Coelho; ANDRADE, Luis Gustavo Modelli. **The future is coming: promising perspectives regarding the use of machine learning in renal transplantation**. **Brazilian Journal of Nephrology**, v. 41, n. 2, p. 284–287, 2019.

HAIR, Joseph F *et al*. **Análise multivariada de dados - 6 ed.** São Paulo: Bookman Editora, 2009.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome H. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Stanford: Springer, 2009.

HAYKIN, Simon. **Redes Neurais: Princípios e Prática**. Porto Alegre: Bookman Editora, 2007.

INCZE, Eduard *et al*. **Using machine learning tools to investigate factors associated with trends in 'no-shows' in outpatient appointments**. Health & Place, v. 67, p. 102496, 2021.

KUHN, Max; JOHNSON, Kjell. **Applied Predictive Modeling**. New York: Springer, 2018.

LAVRAČ, Nada. **Selected techniques for data mining in medicine**. Artificial Intelligence in Medicine, v. 16, n. 1, p. 3–23, 1999

LOZADA, Gisele; NUNES, Karina da S. **Metodologia Científica**. Grupo A, 2019. 9788595029576. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788595029576/. Acesso em: 19 Jul 2021

LUO, Li *et al*. **Machine learning for identification of surgeries with high risks of cancellation**. Health Informatics Journal, v. 26, n. 1, p. 141–155, 2018.

MINISTERIO PUBLICO DE SANTA CATARINA. **SUS sem falta é tema de nova campanha do MPSC** . Disponível em: <https://www.mpsc.mp.br/noticias/sus-sem-falta-e-tema-de-nova-campanha-do-mpsc> . Acesso em: Julho 2021.

ROBAINA, Joey A *et al*. **Predicting no-shows in paediatric orthopaedic clinics**. BMJ Health & Care Informatics, v. 27, n. 1, p. e100047, 2020.

RODRIGUES, Jonathan Grassi *et al*. **Impacto de los mensajes de texto para reducir el absentismo en consultas especializadas: un estudio aleatorio**. Revista Cubana de Información en Ciencias de la Salud, v. 31, n.3 e1566, sept. 2020 . Disponível em <http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2307-21132020000300014&lng=es&nrm=iso>. accedido en 13 jul. 2021. Epub 30-Oct-2020

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. **Learning representations by back-propagating errors**. Nature, v. 323, n. 6088, p. 533–536, 1986.

SAMORANI, Michele; BLOUNT, Linda Goler. **Machine Learning and Medical Appointment Scheduling: Creating and Perpetuating Inequalities in Access to Health Care**. American Journal of Public Health, v. 110, n. 4, p. 440–441, 2020.

SAMUEL, A. L. **Some Studies in Machine Learning Using the Game of Checkers**. IBM Journal of Research and Development, v. 3, n. 3, p. 210–229, 1959.

SANTOS, Hellen Geremias dos et al. **Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo**, Brasil. Cad Saúde Pública 2019; 35(7):e00050818. Cadernos de Saúde Pública, v. 36, n. 1, 2020.

SCHLEDER, Gabriel R.; FAZZIO, Adalberto. Machine **Learning na Física, Química, e Ciência de Materiais: Descoberta e Design de Materiais**. Revista Brasileira de Ensino de Física, v. 43, n. suppl 1, 2021.

SOARES, Jorge. **Uma breve viagem pela Inteligência Artificial**. Revista de Ciências da Computação, *[s. l.]*, v. 14, p. 1–34, 2019. Disponível em: http://search.ebscohost.com/login.aspx?direct=true&db=foh&AN=141373309&lang=pt-br&site=ehost-live. Acesso em: 28 jun. 2021.

SRINIVAS, Sharan. **A Machine Learning-Based Approach for Predicting Patient Punctuality in Ambulatory Care Centers**. **International Journal of Environmental Research and Public Health**, v. 17, n. 10, p. 3703, 2020.

TERMO DE APROVAÇÃO

O aluno Jeferson Dionei Sehnem, regularmente matriculado(a) no Curso de Pós-

Graduação - Especialização em " Ciência de Dado" apresentou o artigo **"**Técnicas de Machine learning para predição de absenteísmo às consultas médicas", obtendo do Avaliador o conceito (*) "_____" (cf. Res. 1/01 – CNE/CES, Parecer nº 908/98 – CES, Res. 21/02 – CEPE/UNINILLE e Res.

001/01 – CEE, Art. 45, incisos I, II e III.)

Professor Avaliador: _____

***Apta ou Inapta**

.