UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Departamento de Sistemas de Computação

Análise de hemogramas para diagnóstico do Sars-Cov-2: uma abordagem baseada em Aprendizado de Máquina

Vinícius Molina Garcia

São Carlos – SP

Análise de hemogramas para diagnóstico do Sars-Cov-2: uma abordagem baseada em Aprendizado de Máquina

Vinícius Molina Garcia

Orientador: Francisco Aparecido Rodrigues

Monografia referente ao projeto de conclusão de curso dentro do escopo da disciplina Projeto de Formatura I (SSC0670) do Departamento de Sistemas de Computação do Instituto de Ciências Matemáticas e de Computação – ICMC-USP para obtenção do título de Engenheiro(a) de Computação.

Área de Concentração: Aprendizado de Máquina

USP – São Carlos
Junho de 2021

“Viva poeticamente e encontrará a felicidade.”

- Edgar Morin

Dedicatória

Dedico este trabalho à minha família: Rita, Júnior e Sofia. Pelo incansável apoio, suporte e companheirismo ao longo de todos estes anos.

Agradecimentos

Agradeço ao caos do universo.

Agradeço a cada membro da famigerada República Romana, que fez da Antônio Álvaro Zuim um recanto de equilíbrio para a saúde mental, um lar são-carlense que gerou diversos checkpoints, discussões filosóficas e amigos para uma vida toda.

Agradeço a cada docente que apoiou minhas ideias, devaneios e iniciativas empreendedoras. Dentre eles gostaria de enfatizar os agradecimentos ao Daniel Smania (ICM), por incitar cada aluno às maravilhas da matemática, ao João Renato (IFSC), por levar para a sala de aula a energia da Física, com brilho e ânimo contagiantes, ao Vanderlei Bagnato (IFSC), por acreditar em mim desde o primeiro semestre de graduação, à Janaína Mascarenhas (EESC), por todas as palavras ditas, os conselhos dados, e a oportunidades geradas, ao André Ponce (ICMC), por ter me acolhido, apoiado, acreditado e ensinado, e ao meu orientador Francisco Rodrigues (ICMC), por toda sua dedicação ao ensino, por toda a paciência e todo o suporte que tem me dado em minhas aventuras empreendedoras.

Carta de Motivação

Escrevo aqui uma breve carta elucidando os motivos para a escolha de tal tema e para não tomar muito do vosso tempo, serei breve com minhas intenções. Meu nome é Vinícius Molina Garcia, e tudo começou quando eu nasci.

Com 6 meses de idade eu fiz minha primeira de três cirurgias para corrigir uma deficiência na perna chamada pé torto congênito. Esta foi a primeira pedra no meio do caminho. E ela me serviu como motivação para superar as dificuldades – mesmo que intrínsecas a mim como esta deficiência – para me dedicar às coisas que são importantes para mim.

Mas tinham outras pedras no meio do caminho. Aos 14 anos fui diagnosticado com câncer na mão, um sarcoma sinovial bifásico. Fiz mais duas cirurgias, quimioterapia e todo o processo de acompanhamento. Todo este processo foi feito no Hospital de Amor de Barretos, instituição com a qual mantenho um vínculo de enorme carinho e respeito pelo trabalho. Com esta pedra eu aprendi que a vida realmente acaba! Ver os meus colegas do hospital morrendo um a cada sessão de quimio, foi difícil, mas me incentivou a viver com intensidade e significado. Buscar coisas grandes que fizessem valer a pena o tempo em que eu estiver por aqui.

E para concretizar estes sonhos grandes, eu tentei entrar na melhor universidade do país, para estar perto de pessoas que são referências na minha área. Com isso, eu entrei em Engenharia Aeronáutica na USP de São Carlos. E para minha surpresa, esta foi mais uma pedra no meio do caminho. Além de não ter gostado muito da área aeronáutica, não vi muito propósito naquilo. Insistindo neste caminho, eu enfrentei a depressão pela primeira vez. Precisei de muita coragem para abandonar um dos cursos mais concorridos da USP. Entretanto, esta foi uma das pedras mais importantes, pois com ela eu aprendi que é preciso seguir o coração. O que me fez parar na Engenharia de Computação.

Descobri que o empreendedorismo é uma forma de concretizar minhas ideias e levar as tecnologias para a sociedade. Eu fui colecionando estas pedras até aqui para construir um propósito. Retomando o que eu disse, toda minha vontade de resolver problemas e o

envolvimento com a área da saúde começaram quando eu nasci. Todas essas vivências me moldaram de forma profunda.

Talvez ainda precise de algumas sessões de psicanálise para entender todos os fatores, mas como Freud diz em seu conceito do *nachträglich*, só ligamos todos os pontos a posteriori. Essas experiências me tornaram um rebelde perante o status quo, querendo sempre fazer as coisas de uma forma diferente e inovar perante os problemas que forem aparecendo. Fiquei muito ligado à área social e, principalmente, à área da saúde. Não obstante, construí dois negócios de impacto social na saúde, com os quais conquistei uma série de prêmios e tive a honra de liderar equipes multidisciplinares com pessoas incríveis.

Minha formação e experiências até o momento, têm me permitido construir tecnologias que colaborem para a área da saúde. Este trabalho é uma forma de levar para o mundo os benefícios que a tecnologia pode provocar na vida das pessoas.

Muito obrigado,

#BoraSalvarVidas

Resumo

Os erros médicos matam 6 pessoas por hora no Brasil, o que equivale a 4320 mortes evitáveis todos os meses. Nos EUA, os erros médicos já são a terceira principal causa de morte, e 2,6 milhões de pessoas morrem todos os anos em todo o mundo devido a esta causa (Fioravanti, 2020). Desta forma, estes eventos estão sendo cada vez mais evidenciados e debatidos, sendo mais bem delimitados e entendidos. Estes debates são nevrálgicos para a criação e instauração de programas de melhoria de processos no público e no privado, com o objetivo de reduzir a chamada iatrogenia, a qual refere-se a um estado de doença, efeitos adversos ou complicações causadas por ou resultantes do tratamento médico. O combate aos danos e mortes causados por erros evitáveis durante o tratamento médico torna-se, então, um desafio mundial (Schmidt, Pesquisa FAPESP).

Dentre as principais causas destes erros estão a falta de informação (não de conhecimento) e recursos. Em muitos casos, o médico não possui todas as informações em mãos para tomar a melhor decisão para o paciente, tampouco sabe como o quadro deste evoluirá, podendo intervir de forma preventiva. Partindo de uma definição bem interessante sobre valor para o paciente, cunhada pelo professor de Harvard Michael Porter, temos que valor é o resultado (*outcome*) obtido pelo paciente em razão dos custos para obtê-lo. Atualmente, os resultados obtidos pelos pacientes (de forma bem mais acentuada na rede pública) são baixos, dada as falhas no processo diagnóstico (e.g. "tudo é virose") e da impossibilidade de fazer intervenções prévias por dificuldade no monitoramento ou predição da saúde do paciente.

Assim, este trabalho propõe uma plataforma assistida por Inteligência Artificial que processa informações de exames simples, baratos e de amplo acesso (como o hemograma completo) para fornecer relatórios inteligentes que auxiliam no processo decisório durante o diagnóstico e prognóstico.

Índice

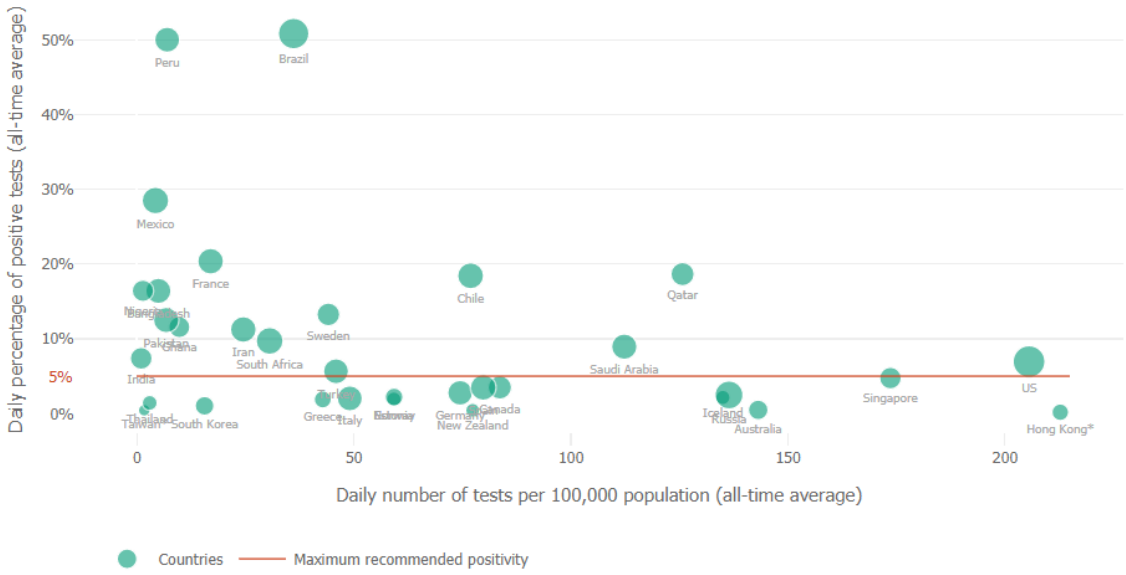
CAPÍTULO 1: INTRODUÇÃO

1.1. Contextualização e Motivação

Com origem em Wuhan, na China, o mundo se viu infligido pela pandemia do novo coronavírus Covid-19, o qual causa doença respiratória que se expressa de forma grave em algumas pessoas (Zhu, N. et al, 2020; Perlman, S., 2020). De acordo com critérios estabelecidos pela OMS, com o intuito de avaliar se um país está conseguindo combater à pandemia, a razão entre o número de casos positivos e o total de testes realizados mostra o quanto a vigilância epidemiológica está conseguindo acompanhar o espalhamento da doença, um valor de sucesso para tal medida é de 5% (OMS).

Comparado com outros países, o Brasil é um dos que apresenta pior performance sob a ótica do indicador supracitado (Site Worldometer; Hospital John Hopkins}. Esta deficiência no processo de testagem, leva a uma enorme subnotificação da doença (Veiga E Silva, L. et al, 2020).

Figura 1. Relação entre testes positivos e número de testes diários compilada pelo Hospital John Hopkins



Fonte: Hospital John Hopkins

Segundo o Observatório Covid-19 da Fiocruz, é nevrálgica a realização de testes diagnósticos para que se possa conhecer e controlar a epidemia, bem como avaliar o impacto das medidas preventivas. É a partir da articulação entre a vigilância epidemiológica e a atenção básica de saúde que se traça a estratégia para controle da doença, estabelecendo o momento e a dimensão dos investimentos que devem ser feitos. Este déficit no diagnóstico, deixa o sistema de saúde despreparado, pois não se sabe a infraestrutura hospitalar necessária para atender a demanda, fazendo com que todas as ações sejam corretivas.

Alguns grupos mostraram-se mais suscetíveis a terem uma expressão mais grave da doença, tendo maiores taxas de admissão em UTIs. Pessoas idosas e aquelas com alguma comorbidade (e.g. diabetes, doenças vasculares ou cardiovasculares, câncer ou quadros de imunossupressão) foram as principais componentes destes grupos de risco (CDC; Chen, N. et al, 2020; Huang, C. et al, 2020). Nestes pacientes, onde a gravidade clínica é aparente, torna-se evidente a necessidade por internação e monitoramento destes pacientes. Entretanto, esta decisão é dificultada quando os sinais clínicos assumem valores mais tênues, distantes de cenários extremos.

Isto porque, ao longo da evolução do curso natural da doença, muitos relatos começaram a sugerir que ao longo da segunda semana as probabilidades de piora no quadro clínico do indivíduo. Nestes relatos, quando houve associação com pneumonia, o desconforto respiratório surgiu, em média, 8 dias após o início dos sintomas (variando de 5 a 13 dias).

Por conseguinte, surge a demanda de um sistema para auxiliar nos diagnósticos dos pacientes suspeitos para coronavírus e que também faça uma predição dos possíveis desfechos e prognósticos do quadro do paciente. Sistema o qual atua como suporte à decisão clínica, auxiliando na questão da liberação do paciente aparentemente estável em oposição ao monitoramento do mesmo.

O impacto de um sistema digital para auxiliar no processo diagnóstico e prognóstico é ainda mais acentuado ao se considerar como a desigualdade social - extremamente presente no Brasil - afeta o acesso a uma saúde de qualidade para enfrentamento da doença (Malta, M. et al, 2020). A tecnologia da informação tem capacidade plena de democratizar o acesso à saúde, levando um atendimento de qualidade para qualquer local do país que tenha acesso à internet.

Dentre estes sistemas, uma subcategoria de grande expressão é a Inteligência Artificial, relacionada com o uso de tecnologia da informação e representações computacionais para se aproximar do comportamento humano, ou seja, se refere a máquinas e programas que são criados para realizar tarefas de modo automatizado e, ao mesmo tempo, aprendem racionalmente a se aprimorarem. Essas máquinas, para alcançar este objetivo, lidam com uma massa infindável de dados, fazendo com que os algoritmos por trás de seus sistemas aprendam com o tempo e, por meio de erros e acertos, refinam seus protocolos e eficácia.

A utilização da Inteligência Artificial tem avançado de uma maneira surpreendente em termos de aplicabilidade nos mais diversos setores, muito devido à sua capacidade de facilitar a rotina das pessoas, aumentar a assertividade e precisão das decisões e reduzir custos dos processos. No setor da saúde, a penetração de tal tecnologia vem empoderando os profissionais da saúde e gestores de instituições do setor, sendo uma ferramenta nevrálgica no apoio à tomada de decisão.

A medicina diagnóstica, englobando exames laboratoriais, exames de imagem e até mesmo a grande área da Patologia, é uma das áreas de maior destaque da aplicação da tecnologia para apoio à tomada de decisão. Na radiologia, os sistemas computacionais assistidos por Inteligência Artificial já são agentes presentes no dia a dia para suporte ao diagnóstico por imagem e à decisão terapêutica (Santos, M. K. et al, 2019). Logo, identificar o posicionamento complementar da inteligência artificial na consulta médica é um desafio fundamental para o futuro (Powell, J., 2019).

Loh resume os últimos meses de pesquisa em saúde em Inteligência Artificial (IA), em diferentes especialidades médicas, e se discute os pontos fortes e os desafios atuais relacionados a essa tecnologia emergente além de enfatizar que os médicos, principalmente aqueles em funções de liderança, precisam estar cientes da rapidez com que a IA está avançando na saúde, para que estejam aptos para liderar a mudança necessária para sua adoção pelo sistema de saúde (Loh, E., 2018).

Os principais problemas de saúde estudados nas pesquisas de IA são: câncer, depressão, doença de Alzheimer, insuficiência cardíaca e diabetes. Além disso, redes neurais artificiais, *Support Vector Machine* (SVM) e redes neurais convolucionais têm o maior impacto na área de saúde. Esta análise fornece uma visão geral abrangente da pesquisa

relacionada à IA conduzida neste campo específico, o que ajuda pesquisadores, formuladores de políticas e profissionais a entender melhor o desenvolvimento destas pesquisas e possíveis implicações na prática, já que o emprego desta tecnologia na área da saúde deve ser feita de forma cuidadosa e por meio de ferramentas corretas e confiáveis, uma vez que os dados clínicos tendem a ser confusos, incompletos e potencialmente tendenciosos, ressaltando, igualmente, os desafios advindos da heterogeneidade de dados (Ghassemi et al., 2019).

Dessa forma, tendo em vista o cenário atual de pandemia e o crescimento do emprego da Inteligência Artificial na área médica, surge o questionamento: como ela poderia ser empregada para conter a propagação do coronavírus? McCall aponta que os casos confirmados de doença coronavírus 2019 (Covid-19) excedem os da síndrome respiratória aguda grave (SARS) e que tanto o Covid-19 quanto o SARS se espalham pelos continentes, infectam animais e humanos e usam mecanismos semelhantes para entrar e infectar a célula. Na linha de frente, a resposta tática ao Covid-19 é semelhante à do SARS, mas existe uma grande diferença: nos 17 anos desde o SARS, surgiu uma nova ferramenta poderosa que poderia ser instrumental para manter esse vírus dentro de limites razoáveis - a saber, Inteligência Artificial (McCall, B., 2020).

Pode-se mencionar o valor na aplicação da IA ao atual surto de Covid-19, por exemplo, na previsão da localização do próximo surto. Esta aplicação é efetivamente o que a empresa canadense Blue Dot tentou fazer e, como tal, foi amplamente divulgada como a primeira organização a revelar a notícia do surto no final de dezembro. Várias outras aplicações de IA que surgiram em resposta à última epidemia incluem BenevolentAI e Imperial College London, que relatam que um medicamento aprovado para artrite reumatoide, o Baricitinib, pode ser eficaz contra o vírus, enquanto a Insilico Medicine, com sede em Hong Kong, anunciou recentemente que seus algoritmos de IA haviam projetado seis novas moléculas que poderiam impedir a replicação viral (McCall, B., 2020).

A Inteligência Artificial também pode ser explorada, em termos da pandemia atual do coronavírus, como uma ferramenta cujo objetivo seria melhorar o diagnóstico da doença, tendo em vista inclusive a quantidade considerável de falsos negativos, o que colabora com a continuidade da propagação do vírus. Em Mei et al. (2020), eles descrevem que, para o diagnóstico da doença coronavírus, um teste de reação em cadeia da polimerase da transcriptase reversa específico do vírus SARS-CoV-2 (RT-PCR) é usado rotineiramente.

No entanto, este teste pode levar até 2 dias para ser concluído, havendo necessidade, igualmente, de testes em série para descartar a possibilidade de resultados falsos negativos e atualmente há uma escassez de kits de teste RT-PCR, ressaltando uma urgência de métodos alternativos para métodos rápidos e diagnóstico preciso de pacientes com suspeita de Covid-19 (Mei, X. et al, 2020). A tomografia computadorizada (TC) de tórax é um componente valioso na avaliação de pacientes com suspeita de infecção por SARS-CoV-2. No entanto, a TC sozinha pode ter valor preditivo negativo limitado para descartar a infecção por SARS-CoV-2, sendo assim, neste estudo, eles usaram algoritmos de IA para integrar os achados da TC de tórax aos sintomas clínicos, histórico de exposição e testes laboratoriais para diagnosticar rapidamente os pacientes positivos para Covid-19.

Nessa mesma linha de encontrar uma maneira para facilitar o diagnóstico destes pacientes, Jamshidi et al. desenvolvem alguns métodos de *Deep Learning* (DL), incluindo *Generative Adversarial Networks* (GANs), *Extreme Learning Machine* (ELM) e *Long / Short Term Memory* (LSTM). Ele delineia uma abordagem de bioinformática integrada cuja principal vantagem dessas plataformas baseadas em IA é acelerar o processo de diagnóstico e tratamento da doença Covid-19 (Jamshidi, M. B., et al, 2020).

1.2. Objetivos

Este trabalho tem como objetivo fomentar as discussões acerca da presença de algoritmos de Inteligência Artificial nas decisões tomadas por profissionais da saúde. Com isso, iniciar algumas reflexões sobre os profissionais-centauro, que conseguem trabalhar juntamente com os sistemas computacionais, unindo seus conhecimentos e seu lado humano com a capacidade analítica e velocidade dos sistemas computacionais para ofertar o melhor atendimento e cuidado para os pacientes.

Assim, este trabalho objetiva a criação de um sistema inteligente que analisa exames simples, baratos e de amplo acesso, como o hemograma, para classificar pacientes com suspeita de estarem infectados com Covid-19. Partindo de ampla análise literária para avaliar os melhores algoritmos classificadores para problemas binários, este trabalho pondera sobre métricas para avaliação de performance, formas de exploração dos dados e construção do conjunto de treino, assim como a interpretabilidade dos modelos.

Esta última exerce papel fundamental na penetração dos algoritmos inteligentes para apoio a decisões médicas, uma vez que os profissionais da saúde precisam atuar em conjunto com os sistemas digitais para tomar uma decisão. Assim, quando saímos de modelos que são considerados “caixa-preta” e passamos a oferecer interpretações sobre a decisão tomada pelo modelo, conseguimos aumentar a adesão desta abordagem.

1.3. Organização do Trabalho

Este trabalho está organizado em 4 capítulos e um apêndice, sendo este capítulo introdutório o primeiro deles. O Capítulo 2 versa sobre os principais conceitos utilizados neste trabalho, abordando uma visão geral sobre os algoritmos classificadores, métricas para avaliação de performance e formas de interpretar as decisões dos modelos. O Capítulo 3 introduz um framework para desenvolvimento de projetos de *Machine Learning*, o qual orientará a apresentação do que foi desenvolvido neste trabalho, bem como os resultados alcançados pelo projeto. No Capítulo 4, são feitas observações acerca do atingido, pontuadas limitações e possibilidades de melhora, bem como alguns aspectos relacionados à relação entre o trabalho e a formação do autor. Por fim, no Apêndice A, é comentado sobre a plataforma web desenvolvida para que a modelagem feita nesse trabalho pudesse ser utilizada na prática.

CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA

2.1. Considerações Iniciais

Neste capítulo tratar-se-á sobre alguns dos principais conceitos, algoritmos e métricas relacionadas a problemas de classificação binária. Para isso, é apresentada a fundamentação teórica levantada da literatura, tratando dos principais classificadores para este tipo de problema, as principais métricas para avaliação de performance dos diferentes modelos – que consequentemente são utilizadas para escolher um modelo que irá para produção – e formas de interpretar as classificações feitas por um modelo.

2.2. Conceitos e Técnicas Relevantes

2.2.1. Algoritmos Classificadores

Para o bom desenvolvimento do projeto, foi feita uma revisão bibliográfica dos principais algoritmos de classificação, com foco no problema da classificação binária. Este problema consiste em receber um conjunto de variáveis de entrada e atribuir a amostra a uma de duas categorias disponíveis. Geralmente, o *output* do algoritmo é um valor que indica a certeza que o algoritmo tem de que a amostra pertence à categoria positiva.

Desta forma, é estabelecido um valor limiar tal que, se o valor de certeza for maior do que o valor limiar, então a amostra pertence à classe positiva. Caso contrário, ela pertence à classe negativa. Dito isso, podemos entender melhor algumas características relevantes de cada um dos algoritmos.

Regressão Logística

A Regressão Logística é uma técnica recomendada para situações nas quais a variável dependente é de natureza dicotômica, que é o caso deste trabalho. Este algoritmo é um clássico e possui aplicações nos mais diversos setores, além de ser altamente interpretável. É preciso tomar um certo cuidado em sua utilização a respeito de outliers e variáveis correlacionadas (Kumar et al.,1995).

Árvore de Decisão

As Árvores de Decisão são modelos baseados em quanto uma variável agrega de informação para a tomada de decisão. Para realizar tal avaliação, os algoritmos podem usar critérios como o de Entropia ou o Coeficiente de Gini. Ou seja, em um dado nó da árvore, ela busca entender qual variável vai trazer mais informação para a tomada de decisão e, então, atribuí maior importância para tal atributo (Izenman, A. J., 2008).

Random Forest

O *Random Forest*, também chamado de Floresta Aleatória, é uma composição de Árvores de Decisão, desta forma apresenta uma robustez para ruídos e também para variáveis colineares. Sua performance tende a melhorar com o aumento de exemplos disponíveis. É preciso tomar cuidado com a preparação das variáveis antes de mostrar para este algoritmo, pois a importância das variáveis pode ser afetada pela escala e número de categorias. Por fim, é importante tomar atenção na hora de analisar os resultados do *Random Forest*, uma vez que por fazer um *ensambling* aleatório, ele pode atribuir grande importância para uma variável e pouca para outra, ao passo que ambas possuem importâncias semelhantes (Ali et al., 2012; Strobl et al., 2007).

Ada Boost

O Ada Boost é um método baseado em meta-aprendizagem, criado inicialmente para aumentar a eficiência de classificadores binários. Ele usa uma abordagem iterativa para aprender com os erros dos piores classificadores e os transforma em bons classificadores. Esta abordagem em identificar os classificadores ruins atribuí ao *Ada Boost* uma boa interpretabilidade. Além disso, é um algoritmo com grande resistência a *overfitting*, o qual é uma grande fonte de problemas para soluções de *Machine Learning* (Jayaprada, S. et al, 2021).

Support Vector Machine (SVM)

O SVM é um modelo que independe da dimensionalidade do conjunto de dados e apresenta bom desempenho em pequenos conjuntos de dados. Para garantir boa performance, é preciso ter uma boa escolha para o *kernell*, o que é um processo difícil. Em

produção, a utilização deste algoritmo pode ser lenta, dependendo do número de vetores de suporte. Por fim, vale ressaltar que ele pode apresentar problemas de convergência em conjuntos grandes de dados (Li et al., 2007).

Tabela 1 - Comparação entre algoritmos testados

Algoritmo	Prós	Contras
Regressão Logística	- Altamente interpretável - Algoritmo clássico com uso disseminado	- Sensível a outliers - Sensível a variáveis correlacionadas - Interações complexas entre variáveis precisam ser explicitamente construídas
Árvore de Decisão	- Funciona com variáveis numéricas e categóricas - Fácil de entender e interpretar - Requer pouco pré-processamento dos dados - A escolha das *features* importantes acontece naturalmente	- Tende a produzir resultados com *overfitting* caso não tenha uma prunagem na profundidade
Random Forest	- Robusto a ruídos - Performance tende a melhorar com o aumento de exemplos disponíveis	- Importância das variáveis do modelo pode ser afetada pela escala e número de categorias
AdaBoost	- É rápido e versátil - Não exige conhecimento prévio dos classificadores fracos - Não tem hiperparâmetros para tunar (exceto por T)	- É vulnerável a ruído uniforme - Quando os classificadores fracos são muito fracos, pode levar a baixas margens de performance
SVM	-Independe da dimensionalidade do conjunto de dados -Bom desempenho em datasets pequenos	-A melhor escolha de kernell é difícil -A predição em produção pode ser lenta dependendo do número de vetores de suporte -Pode apresentar problemas de convergência em conjuntos grandes de dados

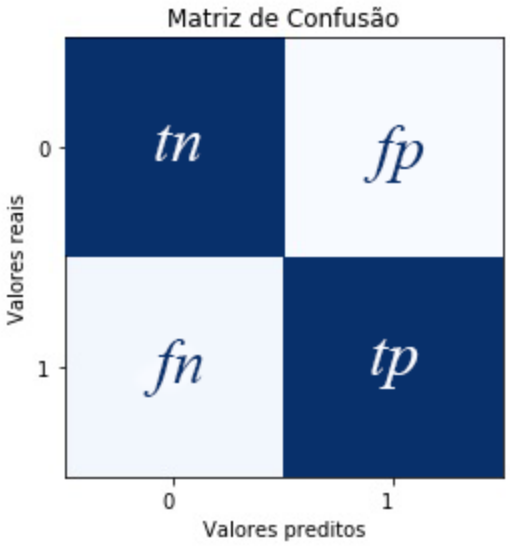
2.2.2. Métricas para avaliação de modelos

Não somente treinar modelos, é fulcral saber como avaliar se a performance do mesmo está adequada. Para cada situação, uma métrica diferente pode ser mais interessante, por isso, esta sessão se dedica a explorar um pouco das possíveis métricas para analisar um modelo.

Matriz de Confusão (FP/FN)

A Matriz de Confusão é uma forma bem comum de mostrar os verdadeiros positivos (tp), os verdadeiros negativos (tn), os falsos positivos (fp) e os falsos negativos (fn). Ela mostra estes valores na forma de uma matriz, na qual o eixo Y mostra os valores reais da amostra, enquanto que o eixo X mostra qual foi o valor atribuído pelo algoritmo. Desta forma, conseguimos entender como está o erro do modelo, e se ele está pior em uma das classes. Geralmente, tal métrica é mostrada no formato da Figura 2.

Figura 2. Matriz de confusão ilustrando os quadrantes de valores preditos e valores reais



Fonte: reprodução própria

Especificidade

A especificidade mede quantas amostras negativas foram de fato classificadas como negativas. No caso deste trabalho, ela mostra quantas amostras não infectadas pelo Covid-19 foram classificadas pelo modelo como negativas. Em termos matemáticas, podemos definir esta métrica como a taxa de verdadeiros negativos (TRN) de acordo com a equação 1.

$$TRN = \frac{tn}{tn + fp} \quad (1)$$

Sensibilidade (Recall)

A sensibilidade é a métrica análoga à especificidade, porém, para o caso positivo. Ela mede quantas amostras positivas foram de fato classificadas como positivas. Ela também pode ser chamada de taxa de verdadeiros positivos (TRP) e calculada de acordo com a equação 2.

$$TRP = \frac{tp}{tp + fn} \quad (2)$$

Otimizar os modelos focando no *recall* é uma forma de descobrir o máximo de positivos possível.

Acurácia

A acurácia avalia quantas observações, tanto positivas quanto negativas, foram classificadas de forma correta. O problema desta métrica é que ela pode ser muito enviesada para conjuntos de dados desbalanceados, de forma que se o algoritmo chutar tudo para a classe majoritária, ele obtém uma boa acurácia. Está métrica pode ser calculada de acordo com a equação 3.

$$ACC = \frac{tp + tn}{tp + fn + tn + fp} \quad (3)$$

Precisão

A precisão (valor preditivo positivo), que pode gerar certa confusão com o *recall*, avalia quantas das observações preditas como positivas são de fato positivas, ou seja, seu objetivo é ter certeza de que, se uma amostra foi classificada como positiva, ela realmente é positiva. A diferença é sutil, repare na alteração do fator *fn* para *fp* no denominador na equação 4, em relação à equação 2.

$$PPV = \frac{tp}{tp + fp} \quad (4)$$

F_β-Score

Está métrica é uma forma de ponderar entre *recall* e precisão em uma métrica, ou seja, queremos descobrir o maior número de amostras positivas, ao mesmo tempo que queremos ter certeza de que as classificadas como positivas serão de fato positivas. Ela é calculada pela equação 5.

$$F_\beta = (1 + \beta^2)\frac{PPV * TRP}{\beta^2 * PPV + TRP} \quad (5)$$

Repare que a escolha do β é o que vai ditar o que queremos priorizar. Quanto mais nos importamos com o *recall* ao invés de precisão, maior tem que ser o valor de β. Para o nosso caso, não podemos deixar nenhum positivo escapar, então priorizamos o *recall*, escolhendo um $\beta = 2$.

Área sob a Curva ROC

Para entender o que significa a área sob a curva ROC (AUC-ROC) é preciso entender o que de fato é a curva ROC. Ela nada mais é do que um gráfico que mostra a relação entre as taxas de verdadeiros positivos (TPR) e taxas de falsos positivos (FPR). Para cada *threshold* estabelecido para o modelo, pode-se traçar uma curva no gráfico, a fim de ter uma comparação de qual é o melhor a ser escolhido. Outra abordagem é para mostrar a AUC-ROC para diferentes folhas em um método de validação cruzada (estratégia utilizada neste trabalho). Assim, a AUC-ROC nada mais é do que calcular a área que está abaixo da curva traçada no gráfico (Bradley, A.P., 1997).

2.2.3. Interpretação da resposta dos classificadores

Valores de Shapley

Uma predição pode ser explicada ao assumirmos que a predição é uma aposta em que cada *feature* é um jogador que participa da mesma. Os valores de Shapley, que surgiram na Teoria dos Jogos de Coalizão, nos diz qual é a probabilidade de cada uma das *features* ganhar esta aposta. Baseado na sua cooperação para atingir um resultado final, cada jogador

(*feature*) recebe uma premiação proporcional. Em termos técnicos, o valor de Shapley é a contribuição marginal média para o valor de uma característica em todas as coalizões possíveis (Shapley, L. S., 1953).

O autor recomenda fortemente que os leitores interessados explorem a biblioteca SHAP do Python, criada por Scott Lundberg e disponível em: https://github.com/slundberg/shap. Esta biblioteca traz formas de explicar modelos em árvore, de gradiente, baseados em *kernel*, lineares, por partição, permutação, amostragem e até mesmo modelos de Aprendizagem Profunda.

2.3. Trabalhos Relacionados

Muitos trabalhos feitos na área da saúde buscam correlacionar algum biomarcador com uma condição fisiológica da saúde do paciente. Muitas vezes são realizadas coletas de dados direcionadas e a motivação parte de alguma suspeita biológica do fenômeno. Em (Foy, B. H. et al, 2020) os autores buscam uma forma de associar a largura de distribuição de glóbulos vermelhos (RDW) no sangue com a chance de mortalidade por Covid-19. Este tipo de análise desperta no autor a curiosidade sobre a potencialidades de fazer tal analises de forma multivariada e com o auxílio de técnicas computacionais, em particular as de Aprendizado de Máquina.

Podemos enumerar algumas pesquisas importantes envolvendo tecnologia da informação atrelada à saúde, dentre elas a de Liu et al., (2018) sobre o uso de imagens multimodais e IA para diagnóstico e prognóstico dos estágios iniciais da doença de Alzheimer; a de Erfurth et al., (2018), sobre o emprego da IA em retina; de Huang et al., (2020), sobre a Inteligência Artificial no diagnóstico e prognóstico do câncer; de Vaishya et al., (2020), sobre as aplicações de IA para a pandemia de COVID-19; de Brinati et al. (2020), sobre a detecção de infecção por COVID-19 em exames de sangue de rotina com aprendizado de máquina e a de Fatima & Pasha (2017), que trata de uma pesquisa de algoritmos de aprendizado de máquina para diagnóstico de doenças, além de inúmeras outras publicações.

2.4. Considerações Finais

Conclui-se então que a área da saúde pode se beneficiar extremamente de análises envolvendo grandes quantidades de dados e as ciências estatísticas e computacionais pode oferecer as ferramentas para realizar este tipo de análise. No capítulo seguinte, o autor apresenta o trabalho realizado para aplicar algoritmos de *Machine Learning* para diagnosticar Covid-19 a partir de dados de hemograma.

CAPÍTULO 3: DESENVOLVIMENTO DO TRABALHO

3.1. Considerações Iniciais

Neste capítulo, será descrito em detalhes o projeto, com foco na parte da análise, modelagem, avaliação e interpretação dos dados. No Apêndice A, é possível encontrar uma descrição breve de plataforma web que foi desenhada para que os modelos criados a partir deste projeto pudessem ser utilizados na prática.

Serão discutidas as principais estratégias utilizadas na limpeza dos dados, bem como aprendizados que foram adquiridos por meio de entrevistas com pessoas da área – em particular de hematologistas e epidemiologistas – a fim de agregar para o desenvolvimento deste projeto conhecimentos técnicos da área da saúde que podem auxiliar no processo de modelagem.

Ao final, serão levantadas todas as dificuldades encontradas no decorrer do projeto, bem como todas as limitações mapeadas até o momento. Espera-se que ao ler este trabalho, o leitor sinta-se encorajado e entusiasmado a replicá-lo, explorar os dados por si e trazer novas contribuições para esta área tão empolgante que são os diagnósticos assistidos por aprendizado de máquina.

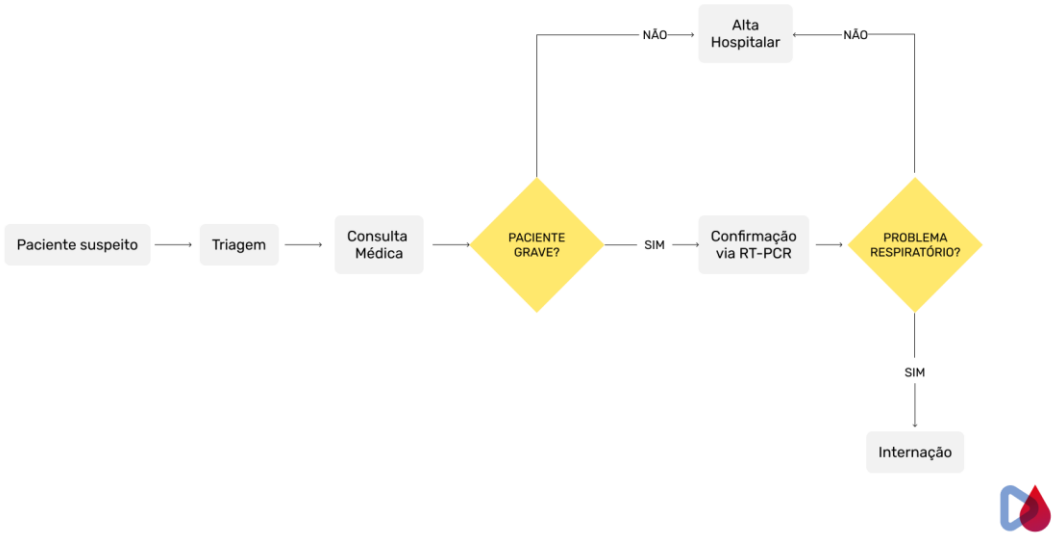
3.2. Projeto

É mais do que comprovado que a Inteligência Artificial na saúde veio para ficar e no cenário da pandemia ela tem se mostrado extremamente útil. Muitos desenvolvimentos foram feitos no aspecto diagnóstico do coronavírus a partir de exames de imagem, principalmente de Tomografia Computadorizada (TC) de tórax. Entretanto, de acordo com o DataSUS, este exame custa cerca de R$136,70 para o sistema público de saúde, além de ser um exame muito empregado em outras doenças, como no caso de câncer, não sendo todo município (ou região) que tem acesso a um aparelho do tipo.

Neste cenário, a motivação deste trabalho é partir de um exame simples, barato e de amplo acesso, que é o hemograma completo. O DataSUS mostra que o custo de um

hemograma para o sistema público é de R$4,10, mostrando um gasto bem inferior em relação à tomografia e, principalmente, em relação ao rt-PCR (exame o qual nem sempre está disponível). Além disso, mesmo que não exista um laboratório no município capaz de processar a amostra de sangue, esta pode ser encaminhada para o centro mais próximo e o paciente continua tendo acesos. Por fim, o sistema de saúde brasileiro já é apto para realizar exames de hemograma em grandes quantidades, sendo que no ano de 2019 foram realizados 67 milhões de exames de sangue, e não é um procedimento que requer treinamento ou especialização extra para os profissionais da área.

Figura 3. Fluxo tradicional de atendimento a um paciente suspeito para Covid-19



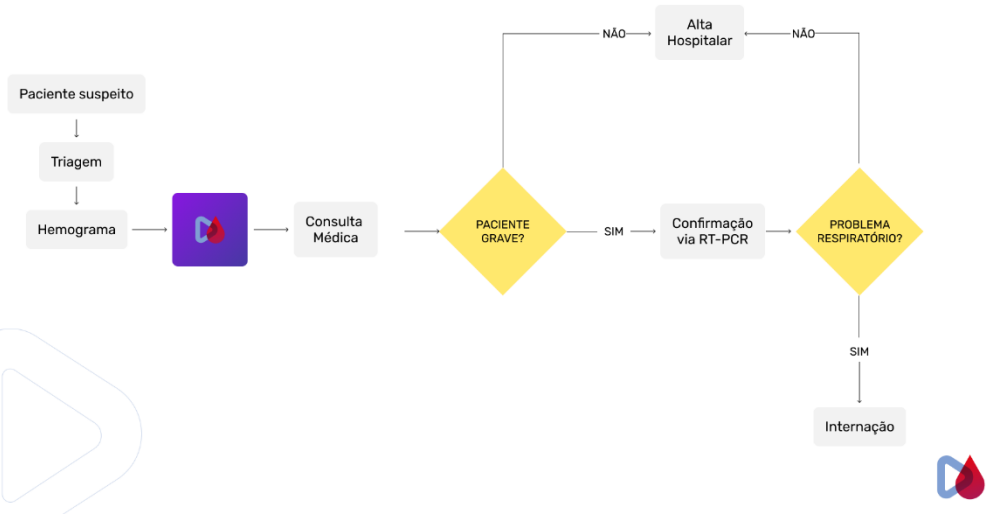
Fonte: reprodução própria

Na Figura 3, podemos observar que as informações disponíveis para o médico na avaliação clínica são praticamente inexistentes. Em um cenário com insuficiência de recursos diagnósticos "padrão-ouro", o médico não tem parâmetros para dar alta hospitalar com segurança, tampouco para solicitar exames mais aprofundados, pois em palavras presidenciais, o paciente pode ter apenas uma "gripezinha". Neste contexto, muitos pacientes são encaminhados para casa sem terem realizado ao menos um exame, aumentando as taxas de readmissão, complicação, óbitos e - principalmente - o contágio da doença.

Levando em conta todos os fatores supracitados, este projeto propõe um sistema que foi chamado de DiagoNow, criado para empregar Inteligência Artificial no processo diagnóstico e prognóstico, processando exames simples e baratos para descobrir padrões

emergentes e fornecer muita informação e inteligência para auxiliar na tomada de decisão dos médicos. No cenário do Covid-19, os algoritmos processam o resultado de um exame de hemograma, inserido no processo de triagem, como mostrado na Figura 4.

Figura 4. Fluxo priorizado proposto pela DiagoNow para atendimento de paciente suspeito de Covid-19

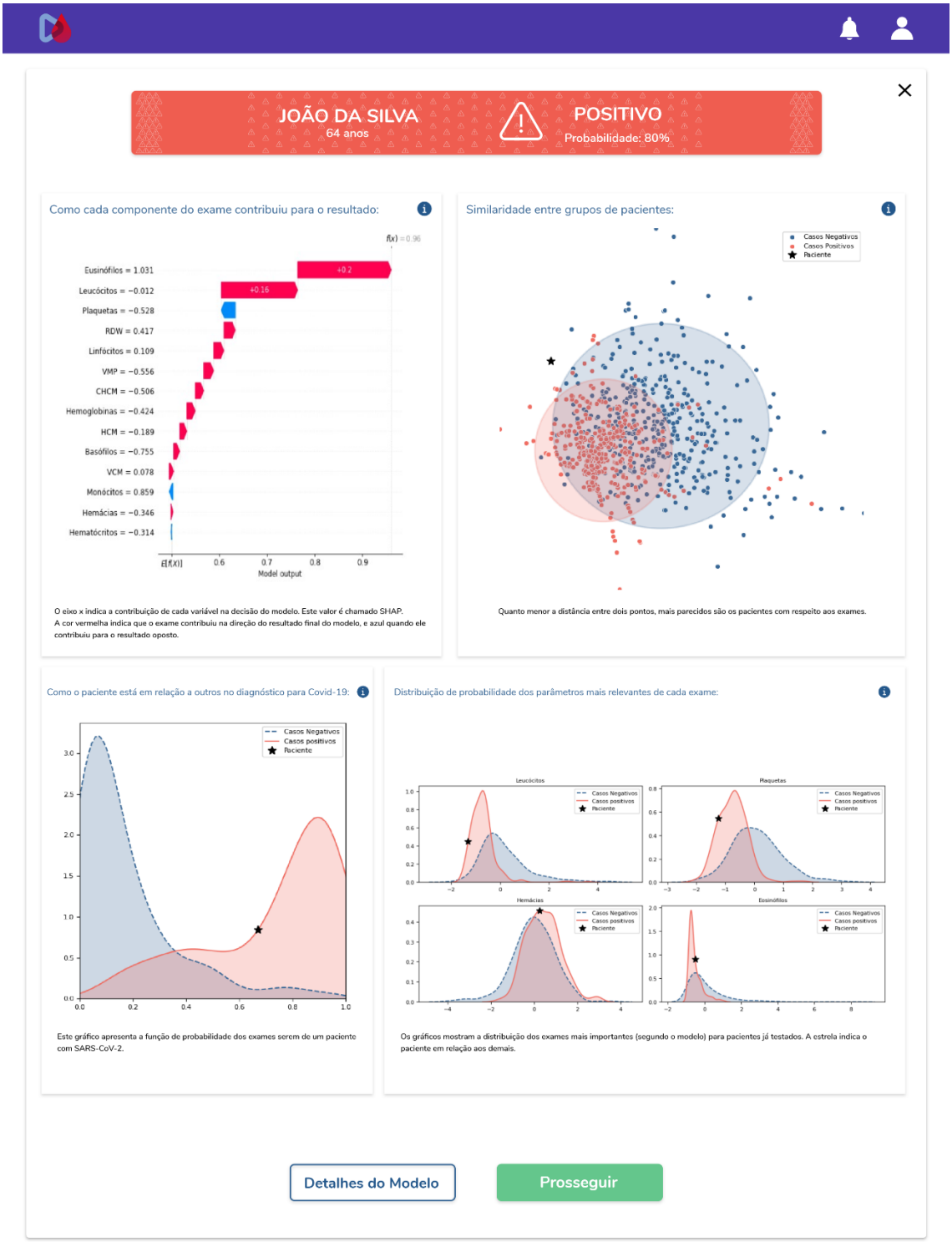


Fonte: reprodução própria

O resultado do exame de hemograma é inserido na plataforma, seja manualmente ou por integração via API com o prontuário eletrônico, a qual processa a requisição e retorna dois relatórios para o médico (Figuras 5 e 6), a partir dos quais o profissional pode tomar uma decisão mais assertiva e embasada. Desta forma, os pacientes recebem um atendimento de maior qualidade, os médicos possuem indicadores de risco para auxiliar na tomada de decisão, os recursos do hospital no processo de testagem são otimizados e o acesso a uma saúde de qualidade e ao processo de testagem é democratizado, garantindo que mais pessoas consigam averiguar melhor seu estado de saúde.

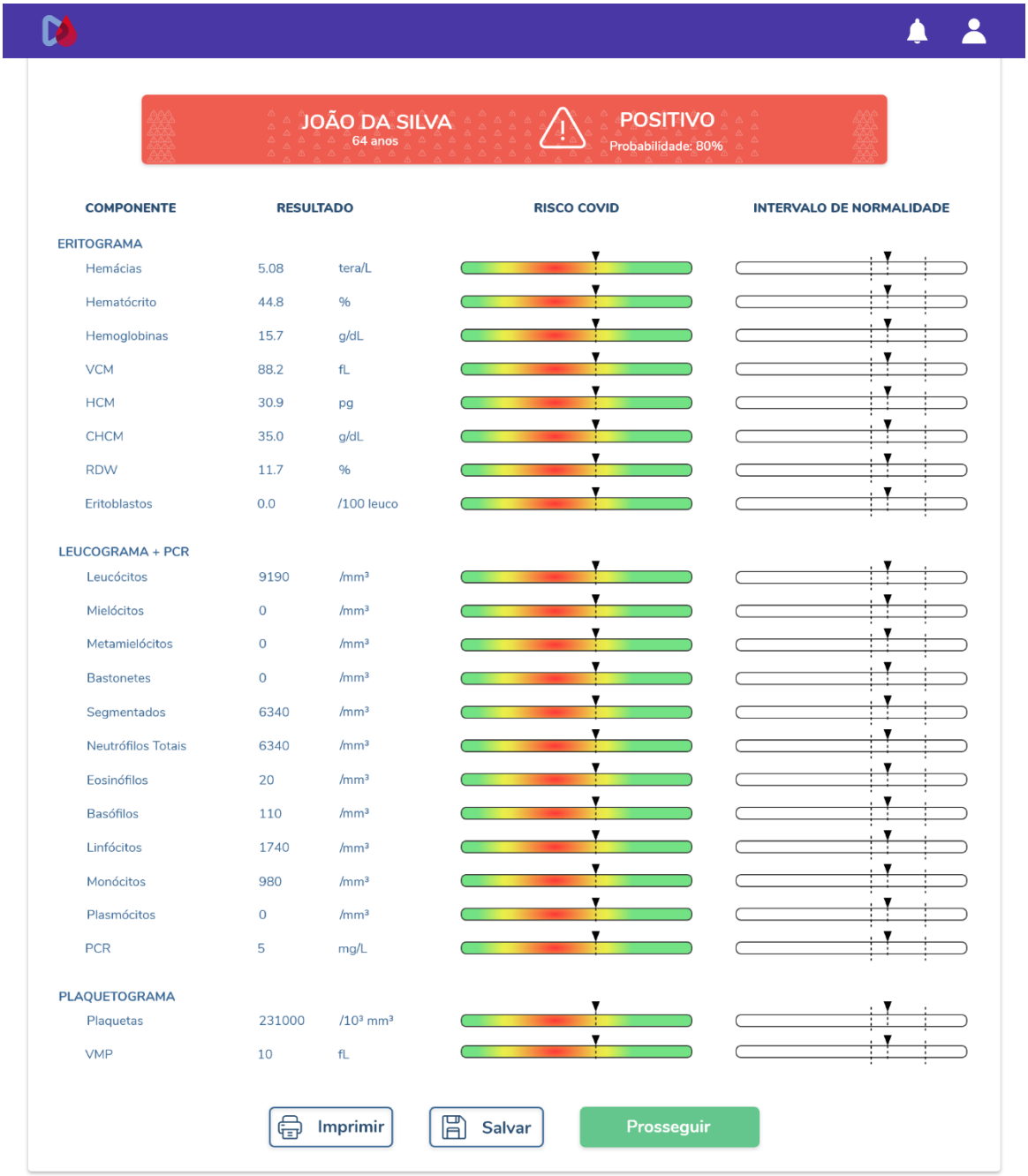
A arquitetura, funcionamento e alguns diagramas sobre a estruturação da plataforma web são apresentadas no Apêndice A, uma vez que o escopo deste trabalho está focado no desenvolvimento dos modelos preditivos. Para realizar este projeto de Machine Learning, o autor se embasou em um *framework* proposto por Aurélien Géron, em seu livro "*Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*" (Géron, A., 2019).

Figura 5. Relatório explicando os principais fatores que levaram o modelo a realizar a classificação da amostra em questão



Fonte: reprodução própria

Figura 6. Visão alternativa para um hemograma comum, mostrando os gráficos de calor relatando a probabilidade de infecção



Fonte: reprodução própria

Os passos deste *framework* servirão como base para a apresentação das atividades realizadas na próxima sessão. Esta lista de passos proposta por Aurélien para guiar projetos de *Machine Learning* possui oito passos principais, os quais são:

1. Entender o problema e veja o panorama geral
2. Obter os dados
3. Explorar os dados para ganhar *insights*
4. Preparar os dados para que os padrões escondidos possam ser melhor expostos aos algoritmos de *Machine Learning*
5. Explorar diferentes modelos e liste os melhores
6. Fazer um *fine-tune* nos modelos e combine-os entre si para chegar em uma solução ótima
7. Apresentar a solução
8. Colocar o modelo em produção, monitora e dê manutenção ao sistema

3.3. Descrição das Atividades Realizadas

A motivação para a realização deste projeto surgiu ao perceber uma escassez de testes para o Covid-19 em meados de 2020. Tal escassez, além de dificultar o controle da pandemia, estava distribuída de forma desigual pelo país, uma vez que o teste rt-PCR está disponível pela rede privada desde o início da pandemia. Entretanto, o valor médio encontrado para o teste, considerado o "padrão-ouro" é cerca de R$180,00, sendo inacessível para grande parcela da população.

Desta forma, o objetivo do trabalho se tornou identificar uma forma de utilizar exames simples, baratos e de amplo acesso para realizar o diagnóstico do vírus. Analisando soluções do mercado, como a Smart Blood Analytics, e outros trabalhos envolvendo diagnósticos com exames laboratoriais, bem como interagindo com profissionais da área de saúde, o autor optou por utilizar o hemograma completo para fazer as predições. O hemograma, mais conhecido como exame de sangue, é um dos exames mais realizados no mundo, pois ao analisar as séries de analitos do sangue, ele fornece diversos insights que vão desde indicar quadro de sepse, evidenciar a presença de infecção virais e bacterianas, auxiliar na percepção de quadro anêmico, entre outros, tornando-o um forte aliado nos diagnósticos diferenciais.

O resultado do hemograma completo pode ser dividido em três grandes blocos:

- Eritograma (série vermelha)
- Leucograma (série branca)
- Plaquetograma (contagem e volume médio das plaquetas)

3.3.1. Obtenção dos dados

Os primeiros desenvolvimentos deste trabalho foram realizados a partir de uma base de dados fornecida pelo Hospital Israelita Albert Einstein, em um desafio na plataforma Kaggle no início de Abril de 2020, chamado "*Diagnosis of COVID-19 and its clinical spectrum".* Estes dados possibilitaram que o autor adquirisse familiaridade com o domínio do problema, bem como desenvolvesse uma solução inicial para o desafio.

Entretanto, uma das características desta base é que ela foi disponibilizada ao público após uma série de transformações – não disponíveis ao público. Sem acesso às funções utilizadas para transformar o *dataset*, inviabilizou-se sua aplicação em um cenário real.

A próxima estratégia surgiu com uma iniciativa da Fapesp para disponibilizar um conglomerado de dados de grandes instituições de saúde do país, relacionadas ao Covid-19. Este repositório foi denominado COVID-19 Data Sharing/BR, disponível em: https://repositoriodatasharingfapesp.uspdigital.usp.br. O repositório contém dados do Grupo Fleury (GF), do Hospital Sírio Libanês (HSL), do Hospital Israelita Albert Einstein (HAE), da Beneficência Portuguesa de São Paulo (BP) e do Hospital das Clínicas da Faculdade de Medicina da USP (HC).

De forma geral, com algumas particularidades, cada instituição forneceu dados anonimizados, respeitando-se todas as questões de privacidade dos dados e aprovações dos respectivos Comitês de Ética. Estes dados contemplam dados demográficos tais quais: Chave de Identificação do Paciente, Sexo, Data de Nascimento, País, Estado, Cidade e CEP. Também trazem dados sobre exames laboratoriais, em que cada linha da base corresponde a um analito analisado. Para cada analito (linha), é disponibilizada a Chave de Identificação do Paciente, uma Chave de Identificação do Atendimento, a Data da Coleta, a Origem do Pedido (isto é, sem qual setor da instituição foi feito o pedido do exame), Descrição do

Exame, Descrição do Analito, Descrição do Resultado, a Unidade do Resultado e o Valor de Referência para este Analito. Adicionalmente, os dados oferecidos pela BP também contêm informações sobre o desfecho clínico do paciente, os quais permitem a realização de outros trabalhos abrangendo a análise de prognósticos.

Ao entrar em contato com as bases de dados, pode-se levar um susto com a imensa quantidade de dados faltantes (esparsidade), a qual acontece em uma média de 90%. Alguns analitos estão quase que totalmente faltantes, com colunas apresentando mais de 99% de *missing values*, ou sem nenhum valor. Apesar de gerar estranhamento em um primeiro momento, este fato é completamente entendido observando a forma de captura destes dados.

Decisões tomadas por profissionais de saúde são um processo complexo, quando os médicos veem um paciente pela primeira vez com uma queixa aguda (por exemplo, o aparecimento recente de febre e sintomas respiratórios), colhem o histórico médico, realizam um exame físico, e baseiam as suas decisões nesta informação. Solicitar ou não testes laboratoriais, e quais deles solicitar, está entre estas decisões, e não existe um conjunto padrão de testes que sejam pedidos a cada indivíduo ou a uma condição específica. Isto dependerá das queixas, dos resultados do exame físico, do histórico médico pessoal (por exemplo, doenças atuais e previamente diagnosticadas, medicamentos em uso, cirurgias prévias, vacinação), hábitos de vida (por exemplo, tabagismo, uso de álcool, exercício físico), histórico médico familiar, e exposições prévias (por exemplo, viagens, profissão).

O conjunto de dados reflete a complexidade da tomada de decisões durante os cuidados clínicos de rotina, em oposição ao que acontece num ambiente de investigação mais controlado, e espera-se, portanto, uma maior escassez de dados.

Não somente em questões de esparsidade, os dados em saúde são bem heterogêneos. Modelos treinados na base de uma instituição e testados em uma base de outra apresentaram performance irrisória, da mesma forma que modelos treinados em dados de diversas instituições apresentaram performance pior do que modelos mais especializados. Estas análises não estão contempladas no escopo deste projeto e não serão mostradas em detalhes. Entretanto, o autor interagiu com hematologistas do Grupo Fleury, que forneceram *insights* importantes sobre tais diferenças. Elas acontecem muito por conta do perfil de cada instituição, o HSL por exemplo, tem um perfil mais hospitalar, com pacientes mais graves, em estado de internação, enquanto que o Grupo Fleury apresenta dados com perfil mais

ambulatorial, que são os pacientes que se apresentaram à emergência e não se agravaram. Outro fator que corrobora para a variabilidade dos resultados interinstitucionais é a diferença entre os equipamentos utilizados para fazer a contagem de células. O hemograma é um exame sensível e possui alta variabilidade no resultado.

Desta forma, para o escopo deste trabalho optou-se por analisar os dados disponibilizados pelo Grupo Fleury, que são datados de agosto de 2020. Esta base apresenta uma vantagem perante às outras no quesito de balanceamento das classes, além disso, ela conta com dados de hemograma de 2019, fator o qual possibilitou a formação da classe negativa de maneira mais robusta. Vamos entender então como foi feita a separação das classes e tirar alguns insights dos dados.

3.3.2. Análise Exploratória

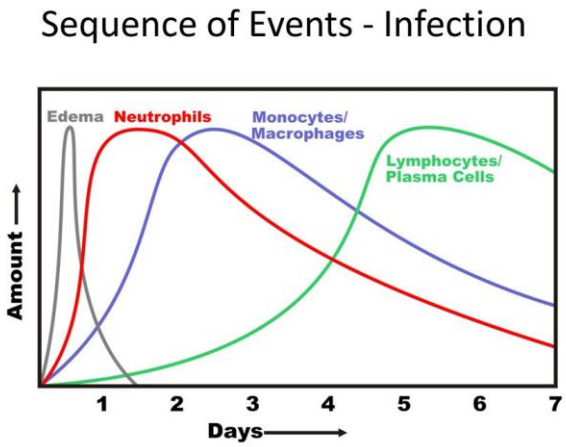
Estamos lidando com um problema de classificação binária, isto é, o algoritmo precisa receber um conjunto de entradas e fornecer como saída uma pontuação de previsão, que indica a certeza que o sistema possui de que uma dada observação pertence a uma das duas classes (positiva ou negativa). A decisão de qual classe a observação pertencerá cabe ao limite de classificação, que também costuma ser chamado de *threshold*. Caso a observação possua uma pontuação maior do que este limiar, então ela pertence à classe positiva, caso contrário, pertence à classe negativa. A escolha deste valor limiar é muito importante, pois impactará no número de classificações corretas (verdadeiros positivas ou verdadeiros negativas) e de incorretas (falsos positivos e falsos negativos). A sessão "2.2.2. Métricas para avaliação de modelos" traz uma visão geral sobre as principais métricas que serão utilizadas na sessão "3.3.4. Modelagem" para avaliar os modelos.

Além de um problema de classificação binária, modulamos este problema como aprendizado supervisionado, uma vez que existem informações no conjunto de dados que são confiáveis o suficiente para serem uma variável *target*. Adotamos duas estratégias diferentes para montar a classe positiva e a negativa.

Começando pela classe positiva, escolhemos a variável "Covid 19, Detecção por PCR" como variável alvo, isto porque este teste apresenta alta sensibilidade e especificidade analítica, muito próxima ou igual a 100% (ROCHA, M. A., 2021). Desta forma, um resultado positivo para o teste assegura com margem de confiança satisfatória a presença do vírus.

Assim, para montar a classe positiva do conjunto de treino, foram selecionados todos os pacientes que apresentou resultado positivo para "Covid 19, Detecção por PCR" **e também** realizaram um exame de hemograma no mesmo dia. Aqui é muito importante o fator temporal, uma vez que o nível de células do sangue varia temporalmente após uma infecção, como mostrado na Figura X.

Figura 7. Evolução temporal de alguns analitos do sangue em um quadro de infecção



Fonte: J. Matthew Velkey em notas de aula intituladas Cell Injury, Death, Inflammation, and Repair para a Duke University

Esta condição foi o maior limitante no volume de dados, uma vez que das 19.372 amostras que apresentaram rt-PCR positivados, apenas 544 haviam feito um exame de sangue no mesmo dia. Uma estratégia para maximizar o volume de dados da classe positiva, considerar amostrar que fizeram um hemograma com um intervalo de 3 dias do rt-PCR, por exemplo.

Já sobre a classe negativa, deparou-se com uma alta probabilidade de falsos negativos ao considerar os resultados negativos na variável "Covid 19, Detecção por PCR". Isto acontece porque existe um período de incubação viral, que é o intervalo de tempo entre a exposição ao vírus e o início dos sintomas, o qual dura em torno de 5 dias. Neste período, a carga viral é baixa e o teste molecular (RT-PCR) apresenta baixa sensibilidade, o que gera altos níveis de falsos negativos. No período de incubação, a taxa de falso negativos é cerca de 68%, indo para 38% após o primeiro dia dos sintomas e para 20% após o terceiro dia

(Khouri, R., 2021). Assim, existe um grande ruído nos dados devido à metodologia de aplicação do teste, levando o autor a buscar outras alternativas para a construção da classe negativa.

A alternativa escolhida pelo autor foi utilizar as amostras de hemogramas coletadas anteriormente à pandemia, isto é, no ano de 2019, de forma a ter uma maior garantia da não presença do vírus. Na base de dados fornecida pelo Grupo Fleury, este fator não foi um limitante, uma vez que existem 12.910 amostras com estas características.

3.3.3. *Feature Engineering*

Com a montagem anterior das classes, obteve-se um conjunto de dados com as seguintes covariáveis:

Demográficas

- Sexo
- Idade

Eritograma

- Concentração de Hemoglobina Corpuscular (CHCM)
- Eritrócitos
- Hematócrito
- Hemoglobina
- Hemoglobina Corpuscular Média (HCM)
- RDW
- VCM

Leucograma

- Leucócitos
- Basófilos
- Basófilos (%)
- Eosinófilos
- Eosinófilos (%)
- Linfócitos
- Linfócitos (%)
- Monócitos

- Monócitos (%)
- Neutrófilos
- Neutrófilos (%)

Plaquetograma

- Plaquetas
- Volume Plaquetário Médio (VPM)

Vale salientar algumas particularidades que foram observadas nestas variáveis. Em primeiro lugar, sobre as variáveis do Leucograma, percebe-se que todas elas, com exceção dos Leucócitos, possuem sua grandeza em unidade absoluta e em valores relativos (%). Isto ocorre, porque as células brancas são subtipos das células leucocitárias, então são avaliados em razão do número absoluto pelo número de leucócitos. Desta forma, a soma de todas as covariáveis da série branca precisa resultar no valor total de leucócitos. Notou-se algumas amostras que não respeitava esta condição, as quais foram excluídas (quando a diferença era de mais de 5%). Estas diferenças, de acordo com especialistas de análises laboratoriais, são provocadas pela forma como são feitas as contagens de células; as máquinas analisam uma determinada área da lâmina e fazem uma projeção da contagem para o volume como um todo, por serem análises automatizadas, elas podem apresentar discrepâncias.

Nesta etapa, também foi feita uma verificação de amostras outliers, a qual indicou algumas amostras registradas com unidades erradas (valores com muitas ordens de grandeza de diferença).

Também é importante ressaltar que a escolha das variáveis nesta etapa pode impactar a aplicabilidade do sistema na prática. Em conversas com mais de 30 médicos atuantes da rede pública e trabalhando na linha de frente da pandemia, descobriu-se que a variável Volume Plaquetário Médio (VPM) nem sempre é citada no resultado do exame, pois depende muito da precisão e qualidade do equipamento que processou a amostra. Assim, dependendo do foco de aplicação do modelo, esta variável pode não ser uma boa escolha.

Recorrendo à literatura para buscar alguma outra informação que pudesse auxiliar na predição, encontrou-se alguns biomarcadores inflamatórios representados por razões entre outros analitos do sangue (Yang, A. et al, 2020). Podemos citar como os principais:

- A razão neutrófilo-linfócito (NLR)

- A razão linfócito-monócito (LMR)
- A razão plaqueta-linfócito (PLR)
- e a razão RNL derivada (d-NLR)

Assim, estas quatro novas *features* foram criadas para poder ter uma base comparativa de seu ganho de performance. Para meios de comparação, foi feito um classificador *RandomForest*, realizando um *RandomUnderSampling* na base, o qual obteve os seguintes resultados:

Figura 8. Matrizes de confusão comparando a adição das novas features

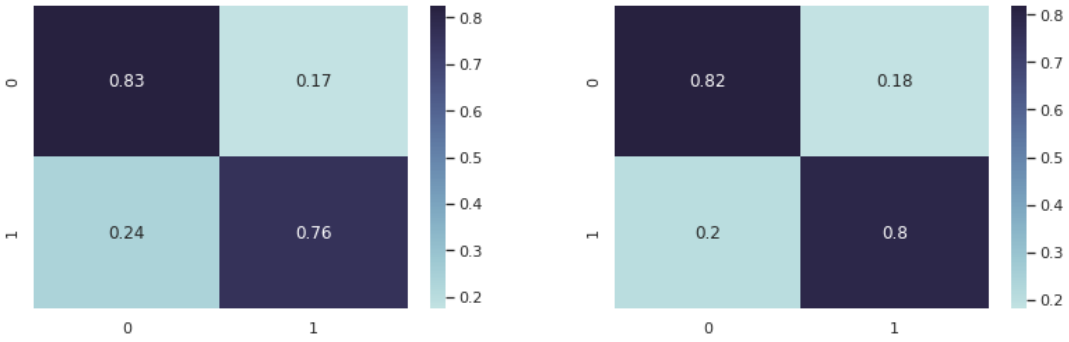


Figura 8.a. Sem "features" adicionais

Figura 8.b. Com "features" adicionais

Fonte: reprodução própria

Apesar de muito pequeno, a adição destas novas *features* conseguiu melhorar um pouco o resultado dos falsos negativos. Neste ponto, vale uma reflexão sobre o que é mais importante para o nosso modelo e como vamos mensurar se é um bom resultado ou não. Pensemos... no cenário da Covid-19, uma amostra negativa significa que o paciente não está infectado. Se ele for classificado corretamente, o médico poderá dar alta para o mesmo sem nenhum problema, e se ele for classificado incorretamente, ele vai ser submetido a um teste mais adequado. Reparem que, não é ideal que ocorram erros, mas o prejuízo de errar um negativo (dar falso positivo) é a realização de um teste mais caro, o que prejudica a entrega de valor deste trabalho, gera desconfianças no modelo e proporciona incômodos para o paciente.

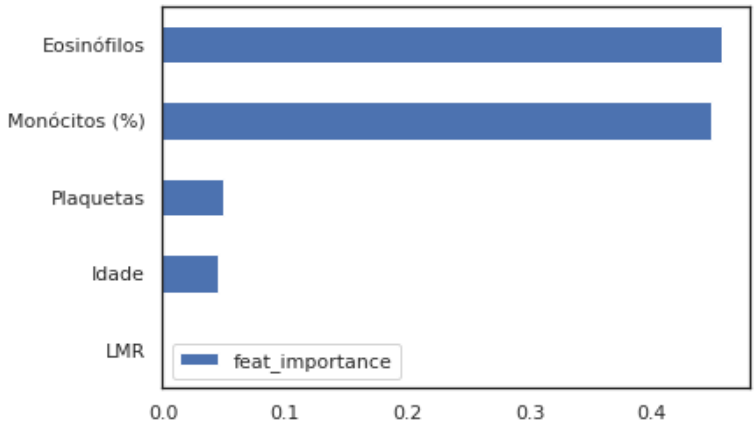
Entretanto, vamos analisar o caso positivo. Se ele é classificado corretamente, as medidas de contenção e cuidados preventivos serão tomadas, o mesmo ficará sob observação

e de quarentena. Porém, caso a classificação seja feita incorretamente o paciente é liberado, ele estará transmitindo o vírus sem ao menos saber, podendo infectar muitas outras pessoas, além de poder ter um agravamento em seu quadro clínico. Desta forma, o objetivo maior deste trabalho é reduzir ao máximo as taxas de falsos-negativos. Por isso, mesmo que ínfima, a adição das *features* provocou uma redução na métrica objetivo e fazem sentido para alcançar o objetivo do trabalho.

Uma métrica deveras interessante para trabalharmos é o F_β Score, definido como a média harmônica entre precisão e *recall*, dando um peso β para o recall (taxa de positivos verdadeiros). Desta forma, conseguimos garantir que nosso modelo reduz com maior proporção os falsos-negativos, ao mesmo tempo em que não começa a classificar tudo como positivo para não errar. Neste trabalho, vamos adotar um valor $\beta = 2$, obtendo, assim, o F2-Score.

Para ter algum *insight* sobre a importância de cada uma das *features* foi feita uma análise com base em uma *Decision Tree* simples, com profundidade máxima de 3, o que resultou que os Eosinófilos, os Monócitos (%), as Plaquetas e a Idade são as quatro variáveis de maior importância.

Figura 9. Importância das variáveis

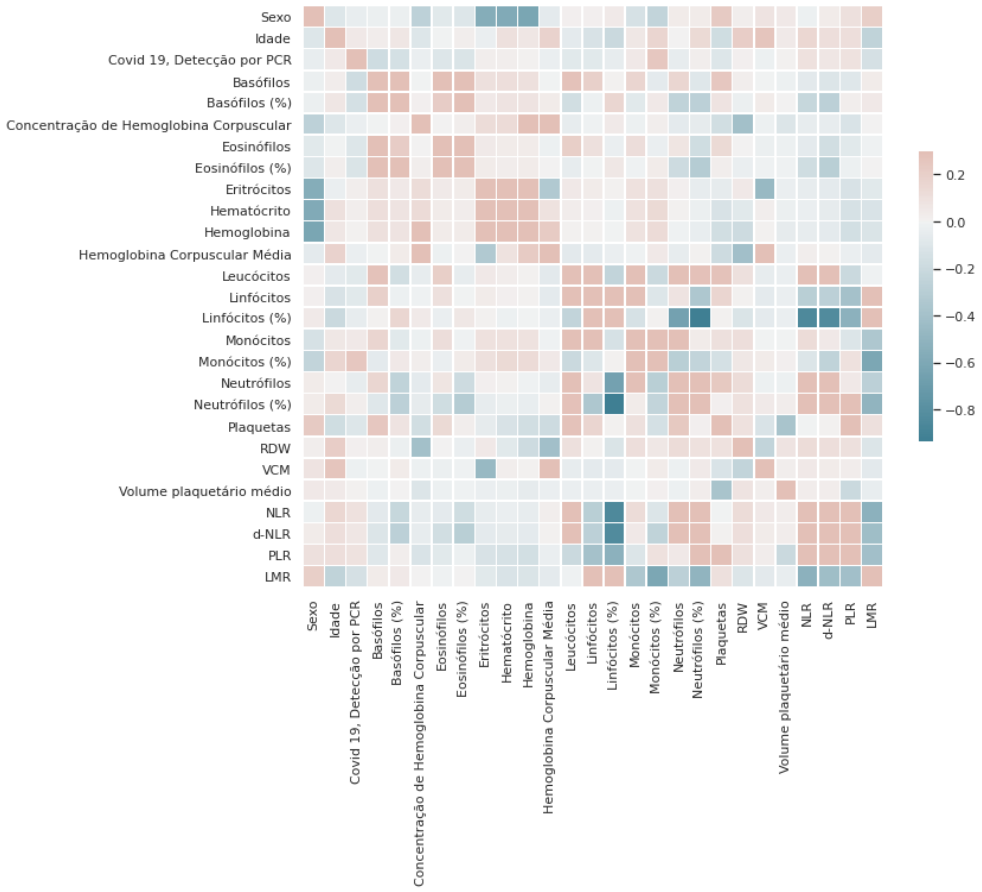


Fonte: reprodução própria

Por fim, é válida uma análise sobre a correlação entre as variáveis, como mostrada na Figura 10. Neste trabalho, estamos utilizando modelos que são derivações de Árvores de

Decisão, as quais são algoritmos robustos para colinearidade entre as variáveis. Desta maneira, não foram removidas variáveis correlacionadas.

Figura 10. Matriz de correlação das covariáveis do hemograma



Fonte: reprodução própria

3.3.4. Modelagem

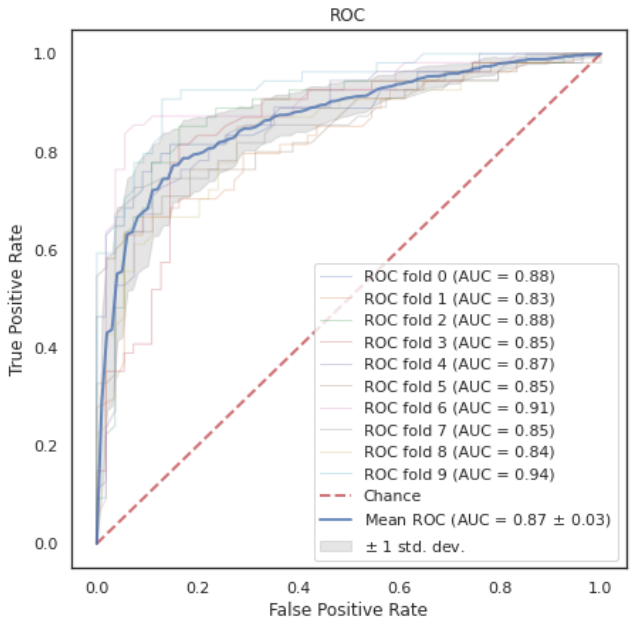
Após a limpeza dos dados e engenharia de features, foi hora de começar a modelagem. Partindo de um *Random Under Sampler* para lidar com a questão dos balanceamentos dos dados, foram testados cinco algoritmos diferentes: Regressão Logística, Árvore de Decisão, Floresta Aleatória, AdaBoost e SVM. Para todas as análises foi feita uma *cross-validation* com 10 *folds.* A comparação entre os algoritmos é feita na Tabela 1, a qual mostra a área sob a curva ROC, o recall e a acurácia, todos medidos em um conjunto de teste de 20% do tamanho da base.

Tabela 2. Comparativo entre diferentes modelos de classificação, utilizando um método de k-folhas para validação cruzada

	Logistic Regression	**Decision Tree**	**Random Forest**	**AdaBoost**	**SVM**
ROC-AUC	0.782	0.741	0.824	0.796	0.645
Recall	0.761	0.807	0.826	0.798	0.550
Accuracy	0.801	0.680	0.822	0.794	0.733

A partir destes resultados, o foco passou a ser tunar os resultados alcançados pelo algoritmo de *Random Forest*. Para isso, foi utilizado método do *Grid Search Cross Validation*, o qual avalia todas as combinações de hiperparâmetros definidas pelo autor. Para fazer a definição inicial do grid, é possível utilizar uma busca aleatória, para estreitar o espaço de busca. Após testar 288 combinações diferentes para o algoritmo de RandomForest, chegou-se em um resultado final, que pode ser visualizado pelo gráfico da Figura X.

Figura 11. Gráfico contendo curvas ROC para uma validação cruzada de 10 folhas



Fonte: reprodução própria

3.3.5. Interpretação das classificações

A última sessão trouxe um resultado interessante na predição de casos de Covid-19 utilizando um hemograma, o qual pode ser um passo inicial interessante para a utilização de algoritmos de *Machine Learning* na tomada de decisões clínicas. Entretanto, para que tal abordagem possa realmente ser adotada na prática, é preciso que haja uma sinergia entre a forma de pensar do profissional de saúde com as informações fornecidas pelo modelo. É improvável que o profissional na linha de frente vai simplesmente acatar uma predição sem entender o que está acontecendo por trás.

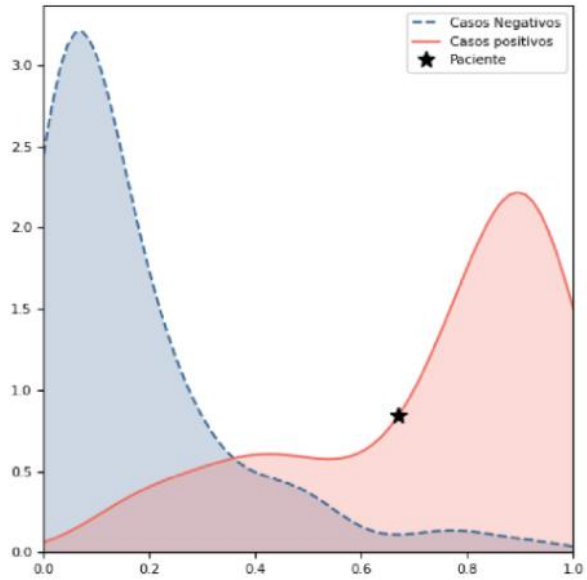
É por isso que a interpretabilidade dos resultados se torna um fator tão importante quanto a predição em si. Neste cenário, os algoritmos que fogem da lógica da "caixa-preta" começam com uma vantagem inicial, porém, já existe um vasto material na literatura sobre este assunto, para os mais variados algoritmos. Nesta sessão, vamos apresentar algumas visualizações propostas, com ênfase para os valores de Shapley.

Vamos começar do básico: curvas de distribuição de probabilidade. Os profissionais da saúde estão acostumados com terminologias estatísticas, e, partindo disto, foram propostos dois gráficos baseados nas distribuições de probabilidades, um deles de forma geral (o qual mostra a separabilidade que o modelo proporcionou aos dados), como na Figura 12, e outro focado nas *features* que foram mais relevantes para uma dada classificação, mostrado na Figura 13.

Para construir o gráfico da Figura 12, foi utilizada uma informação do modelo implementado utilizando Python, com a biblioteca *sklearn*, chamada ***"predict_proba"***. Ela fornece o valor da certeza dada pelo algoritmo para a classe positiva e para a classe negativa, assim, com base no *threshold* definido, sabemos também a qual classe aquela amostra pertence. Desta maneira, tendo as classes e as probabilidades, torna-se trivial o plot deste gráfico.

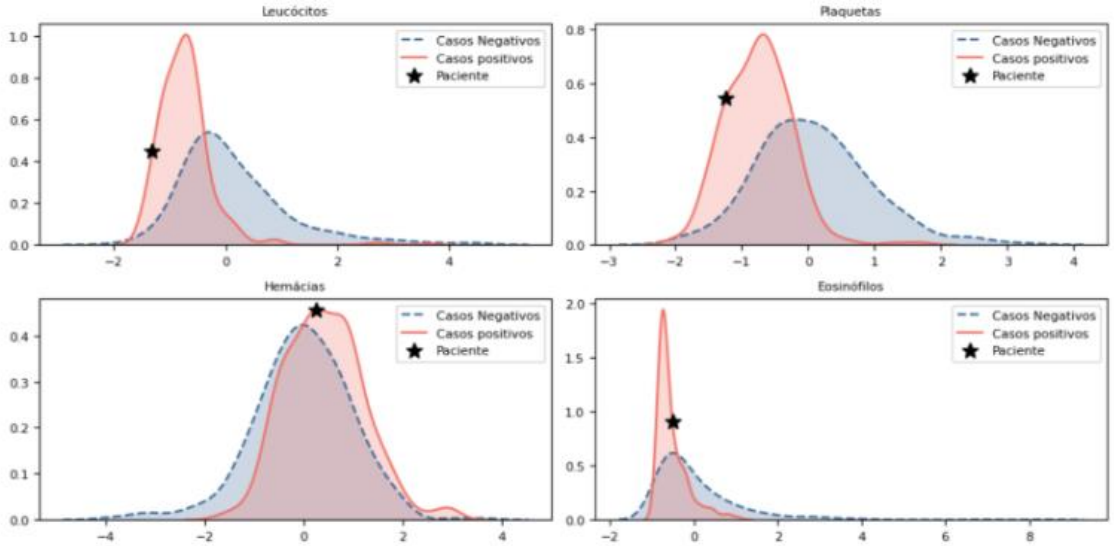
Já o gráfico da Figura 13 tem um desafio um pouco maior. E tem duas alternativas para fazê-lo: escolher as quatro componentes principais do modelo como um todo, ou então escolher as quatro que mais contribuíram para a classificação desta amostra. Ao escolher a última estratégia, é possível utilizar os valores de Shapley para tal.

Figura 12. Gráfico de explicação do modelo mostrando a separabilidade dos dados e a posição do paciente que está sendo classificado no momento.



Fonte: reprodução própria

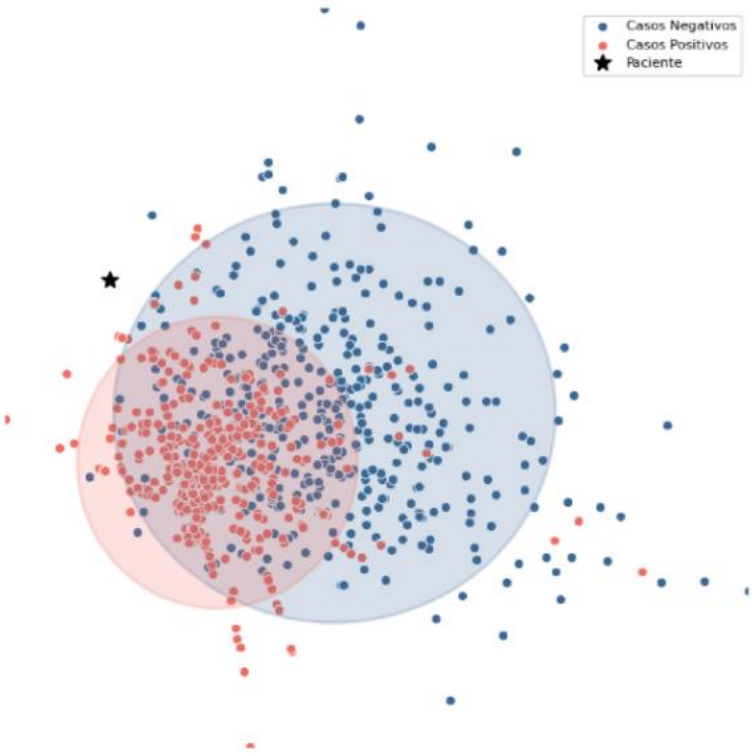
Figura 13. Gráfico mostrando a distribuição de probabilidades de cada uma das quatro variáveis principais para a decisão da classificação.



Fonte: reprodução própria

Outra visualização sobre o resultado foi uma tentativa de mostrar onde está o paciente em relação a todos os outros que já foram classificados pela plataforma, de forma a obtermos um *scatterplot* reduzido às 2 componentes principais, mostrado na Figura 14.

Figura 14. Scatterplot mostrando os pacientes classificados pela plataforma utilizando as duas componentes principais.

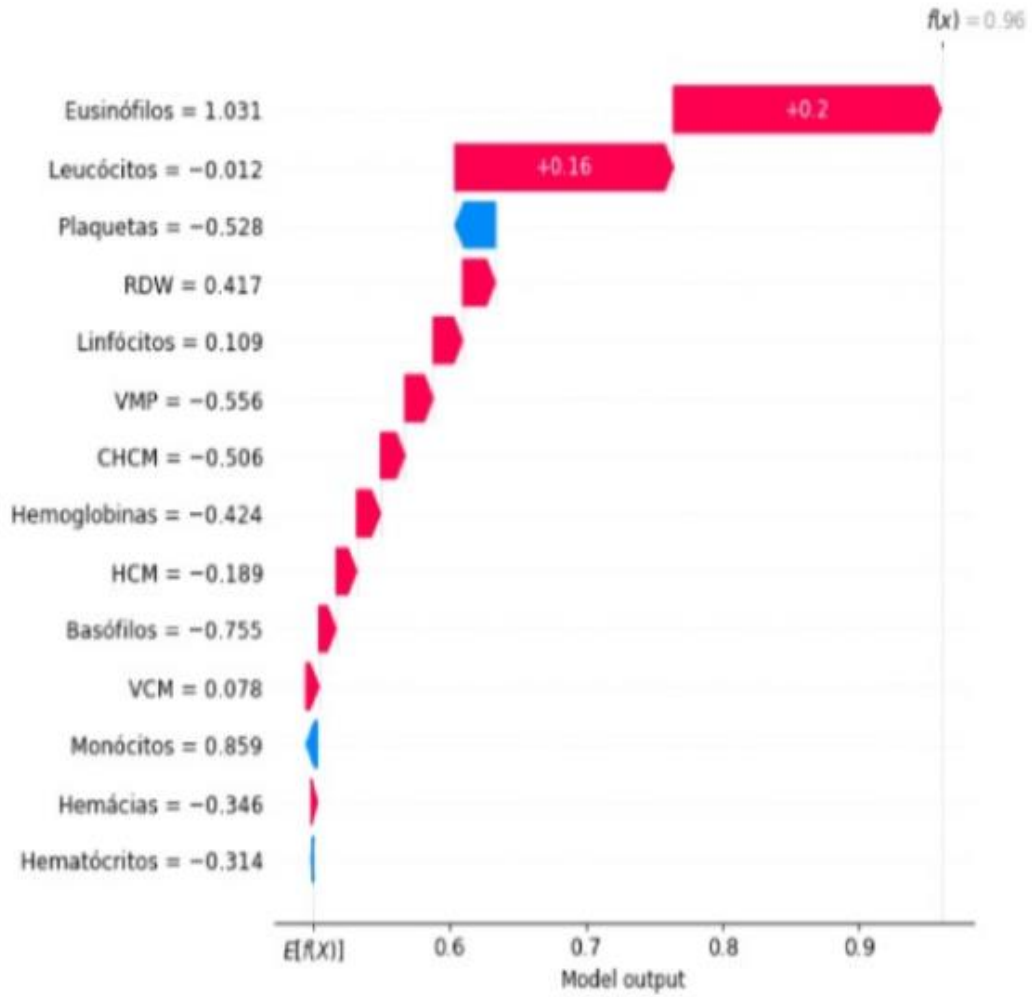


Fonte: reprodução própria

Para gerar o gráfico da Figura 14, foram utilizados algoritmos de *Principal Component Analysis* (PCA).

Em testes de usabilidade com médicos da linha de frente, estes três gráficos não foram muito bem recebidos, de forma a não agregarem para a tomada de decisão. Em contra partida, o último gráfico funcionou muito bem, o qual utiliza o conceito dos valores de Shapley, e para uma dada classificação resulta em um gráfico como o da Figura 15.

Figura 15. Gráfico mostrando os valroes de Shapley para a classificação de uma determinada amostra.



Fonte: reprodução própria

3.4. Resultados Obtidos

A partir deste trabalho foi obtido um classificador binário, baseado no algoritmo Random Forest e validado utilizando o método da validação cruzada em 10 folhas. Como o conjunto de dados era desbalanceado, foi utilizado o método de *undersampling* para que o conjunto de treino possuísse a mesma quantidade de amostras de cada uma das classes, a fim

de não apresentar nenhum viés ao algoritmo. Assim, com uma amostra considerável (N = 1088) foram obtidos os seguintes resultados mostrados na Tabela 3.

Tabela 3. Resultados para modelo Random Forest

ROC-AUC	0.870 ± 0.030
F2-Score	0.774 ± 0.063
Sensibilidade	0.763 ± 0.061
Especificidade	0.836 ± 0.033

3.5. Dificuldades e Limitações

O desenvolvimento deste trabalho englobou uma grande variedade de atividades, necessitando de entrevistar profissionais da saúde, buscar parcerias com instituições do setor, trabalhar com grandes volumes de dados, buscar soluções para modelagem, desenhar uma plataforma que pudesse servir os modelos criados para os usuários finais e até mesmo a perspectiva de negócios. Neste processo, a sensibilidade de se trabalhar com questões de saúde foi um fator decisivo. Foi necessário entender mais sobre Ética em Pesquisa com Seres Humanos, sobre Ensaios Clínicos Randomizados, Estudos Diagnósticos Transversais Retrospectivos e outras práticas que, apesar de comum na saúde, está distante do dia a dia dos profissionais da Engenharia.

Além dos aprendizados em metodologia de pesquisa em saúde, este trabalho colocou o autor em contato com as dificuldades de se levar um modelo de *Machine Learning* para a prática. Ao realizar parcerias com instituições de saúde para testar o modelo, o autor viu a performance cair consideravelmente, mostrando a complexidade de se trabalhar com dados de saúde. Logo, a heterogeneidade entre os dados de diferentes instituições foi um grande desafio. Outro fator, ainda relacionado com o fato de os dados serem muito heterogêneos, é a complexidade dos sistemas biológicos. Para conseguir explicar hipóteses e fenômenos que apareceram durante à modelagem, foi necessário um entendimento maior dos acontecimentos médico-biológicos. Por conseguinte, o autor sugere fortemente que estudos

do gênero sejam realizados por grupos multidisciplinares, englobando profissionais da saúde e da computação, para que todas as hipóteses possam ser exploradas em sua completude.

Vale ressaltar algumas limitações e sugestões de melhoria deste trabalho:

- **Aprofundar as relações entre as instituições**: o autor reconhece que uma análise interessante a ser feita é a respeito dos dados das diferentes instituições presentes no repositório de dados da Fapesp. Alternando entre os conjuntos de treino e teste, e buscando novas estratégias para formar as classes positivas e negativas. Desta forma, podem surgir modelos mais generalizáveis do que aqueles desenvolvidos em uma única instituição.

- **Outros componentes do sangue e/ou outros exames**: o autor focou o trabalho em dados de um hemograma completo (i.e., Eritograma, Leucograma e Plaquetograma). Entretanto, com base em evidências da literatura, outras componentes do sangue podem auxiliar nas predições, podemos citar algumas que apresentaram correlações com o vírus, como os Dímeros-D (Thachil, J., 2020), a Proteína C-Reativa e a Ferritina (Jacinto, D. M. et. al, 2020). Além disso, a incorporação de outros exames na análise, como o de urina, podem enriquecer os resultados.

- **Análise de imputação e tratativas de desbalanceamento**: por ser uma base de dados esparsa, este problema abre oportunidades para estudos mais aprofundados de imputação de dados e estratégias para lidar com conjuntos de dados desbalanceados. O autor optou por estratégias de *undersampling*, o que pode ser uma limitação do trabalho, já que a quantidade de dados é reduzida por tal estratégia. Sugere-se investigações futuras com outras estratégias, explorando o *oversampling* com técnicas como o SMOTE ou *random oversampling*.

- ***Backtest* para outras síndromes gripais:** por ausência de dados, este trabalho não conseguiu comprovar se os modelos estavam realmente diagnosticando Covid-19 ou simplesmente um processo inflamatório padrão.

> Assim, uma limitação do projeto – e sugestão de trabalho futuro – é conseguir associar os dados relacionados à Covid-19 juntamente com o de outras síndromes gripais, para que o modelo possa ter maior robustez nestes casos.

Em suma, apesar de limitado em termos de aplicabilidade prática nos hospitais – ainda requer muitas validações – este trabalho colabora para as discussões sobre o uso do Aprendizado de Máquina no apoio à tomada de decisões clínicas. Neste sentido, o autor encoraja fortemente que novos trabalhos sejam realizados com esta linha de pensamento. Cada vez mais os profissionais da saúde estão mais abertos para a tecnologia e àqueles que cooperaram sinergicamente com as máquinas serão os profissionais de maior diferencial no futuro.

3.6. Considerações Finais

Neste capítulo discutiu-se o desenvolvimento deste trabalho, bem como os resultados obtidos. Foi proposto um modelo de classificação satisfatório, seu desempenho foi avaliado, suas limitações, expostas, e foram discutidas as principais dificuldades e limitações do trabalho desenvolvido. No capítulo seguinte, serão apresentadas as conclusões deste trabalho.

CAPÍTULO 4: CONCLUSÃO

4.1. Contribuições

Trabalhos como este fornecem ao aluno uma perspectiva extremamente ampla, pois ao lidar com uma base de dados real, para resolver um problema real e pertinente, muitos fatores de complexidade precisam ser explorados e superados para que o trabalho atinja um resultado satisfatório. Foi possível, ao desenvolver este trabalho, adquirir conhecimentos profundos na área de Aprendizado de Máquina, ganhar experiência durante a exploração dos dados e enfrentar problemas que são rotineiros na vida de um profissional do setor.

Além disso, o desenvolvimento deste projeto despertou no autor a ânsia por empreender e transformar desenvolvimentos como este em tecnologias que ajudem a salvar vidas e cheguem à sociedade através do empreendedorismo. Não só restrito ao trabalho presente, o conteúdo apresentado aqui se tornou uma empresa, a qual propiciou crescimento profissional relevante, uma sorte de premiações e contatos nevrálgicos para a caminhada profissional de alguém que anseia por trabalhar com tecnologia e saúde.

4.2. Trabalhos Futuros

Neste trabalho foram utilizados dados laboratoriais, sobretudo de hemogramas, para realizar a predição de infecção pelo Covid-19. Porém, para além destes, dados clínicos, como os de sintomas, contato e situação de internação são dados que podem aumentar o poder preditivo deste tipo de aplicação. Além disso, não só com foco no diagnóstico, outros trabalhos podem ser desenvolvidos com o intuito de prever o prognóstico dos pacientes, isto é, dado um conjunto de dados sobre um paciente no instante t, como o quadro clínico deste paciente se comportará no instante $t + n$? Será que ele irá deteriorar? Vai precisar de internação? Receberá alta?

Outros trabalhos que podem ser desenvolvidos a partir da ideia deste trabalho é lidar com exames de imagem, como a Tomografia Computadorizada de Tórax, que se mostrou um exame importantíssimo para o diagnóstico do vírus. Podem também combinar dados laboratoriais e de imagem para aumentar a capacidade preditiva.

Por fim, este tipo de trabalho não está restrito ao Covid-19, então deixo aqui uma provocação para o leitor: em quais outros cenários você imagina que trabalhos assim possam ser aplicados no setor da saúde?

REFERÊNCIAS

ALI, Jehad et al. Random forests and decision trees. **International Journal of Computer Science Issues (IJCSI)**, v. 9, n. 5, p. 272, 2012.

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.

BRINATI, Davide et al. Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. **Journal of medical systems**, v. 44, n. 8, p. 1-12, 2020.

Centers for Disease Control and Prevention (CDC). **Coronavirus**. Disponível em https://www.cdc.gov/coronavirus/index.html.

CHEN, Nanshan et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. **The Lancet**, v. 395, n. 10223, p. 507-513, 2020.

FATIMA, Meherwar et al. Survey of machine learning algorithms for disease diagnostic. **Journal of Intelligent Learning Systems and Applications**, v. 9, n. 01, p. 1, 2017.

FIORAVANTI, Carlos. Um diagnóstico do erro médico, **Revista Pesquisa Fapesp**, 2020.

FOY, Brody H. et al. Association of red blood cell distribution width with mortality risk in hospitalized adults with SARS-CoV-2 infection. **JAMA Network Open**, v. 3, n. 9, p. e2022058-e2022058, 2020.

GÉRON, Aurélien. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems**. O'Reilly Media, 2019.

GHASSEMI, Marzyeh et al. Practical guidance on artificial intelligence for health-care data. **The Lancet Digital Health**, v. 1, n. 4, p. e157-e159, 2019.

HUANG, Chaolin et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. **The lancet**, v. 395, n. 10223, p. 497-506, 2020.

IZENMAN, Alan Julian. Modern multivariate statistical techniques. **Regression, classification and manifold learning**, v. 10, p. 978-0, 2008.

JACINTO, D. M. et al. Descrição Das Alterações Do Hemograma Correlacionados À Proteína C Reativa (Pcr) E Ferritina Em 7942 Pacientes Com Covid-19. **Hematology, Transfusion and Cell Therapy**, v. 42, p. 529, 2020.

JAMSHIDI, Mohammad et al. Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. **IEEE Access**, v. 8, p. 109581-109595, 2020.

JAYAPRADA, S.; JAYALAKSHMI, G.; KANYAKUMARI, L. Fast Hybrid Adaboost Binary Classifier For Brain Tumor Classification. In: **IOP Conference Series: Materials Science and Engineering**. IOP Publishing, 2021. p. 012016.

JOHN Hopkins Hospital, How Does Testing in the U.S. Compare to Other Countries? - **JohnsHopkins Coronavirus Resource Center**.

KUMAR, Akhil; RAO, Vithala R.; SONI, Harsh. An empirical comparison of neural network and logistic regression models. **Marketing Letters**, v. 6, n. 4, p. 251-263, 1995.

LI, Rui; BHANU, Bir; KRAWIEC, Krzysztof. Hybrid coevolutionary algorithms vs. SVM algorithms. In: **Proceedings of the 9th annual conference on Genetic and evolutionary computation**. 2007. p. 456-463.

LIU, Xiaonan et al. Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. **Translational Research**, v. 194, p. 56-67, 2018.

LOH, Erwin. Medicine and the rise of the robots: a qualitative review of recent advances of artificial intelligence in health. **BMJ Leader**, p. leader-2018-000071, 2018.

MALTA, Monica et al. Coronavirus in Brazil: The heavy weight of inequality and unsound leadership. **EClinicalMedicine**, v. 25, 2020.

MATTHEW, J. V., Cell Injury, Death, Inflammation, and Repair. Lecture notes. **Duke University**. Disponível em: https://slideplayer.com/slide/4382692/

MCCALL, Becky. COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread. **The Lancet Digital Health**, v. 2, n. 4, p. e166-e167, 2020.

MEI, Xueyan et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. **Nature medicine**, v. 26, n. 8, p. 1224-1228, 2020.

PERLMAN, Stanley. Another decade, another coronavirus. 2020.

POWELL, John. Trust me, I'ma chatbot: How artificial intelligence in health care fails the Turing test. **Journal of medical Internet research**, v. 21, n. 10, p. e16222, 2019.

ROCHA, M. A., Covid-19: Pesquisador da Fiocruz tira dúvidas sobre testes de Covid-19. **Portal Fiocruz**. Disponível em: https://portal.fiocruz.br/noticia/covid-19-pesquisador-da-fiocruz-tira-duvidas-sobre-testes-de-covid-19. Acesso em: 21, junho de 2021.

SANTOS, Marcel Koenigkam et al. Inteligência artificial, aprendizado de máquina, diagnóstico auxiliado por computador e radiômica: avanços da imagem rumo à medicina de precisão. **Radiologia Brasileira**, v. 52, n. 6, p. 387-396, 2019.

SCHMIDT-ERFURTH, Ursula et al. Artificial intelligence in retina. **Progress in retinal and eye research**, v. 67, p. 1-29, 2018.

SHAPLEY, Lloyd S. Stochastic games. **Proceedings of the national academy of sciences**, v. 39, n. 10, p. 1095-1100, 1953.

STROBL, C., Boulesteix, AL., Zeileis, A. *et al.* Bias in random forest variable importance measures: Illustrations, sources and a solution. ***BMC* Bioinformatics 8,** 25 (2007).

TRIMAILLE, Antonin et al. D-Dimers Level as a Possible Marker of Extravascular Fibrinolysis in COVID-19 Patients. **Journal of Clinical Medicine**, v. 10, n. 1, p. 39, 2021.

VAISHYA, Raju et al. Significant role of modern technologies for COVID-19 pandemic. **Journal of Industrial Integration and Management**, v. 6, n. 2, 2021.

VEIGA E SILVA L, de Andrade Abi Harb MDP, Teixeira Barbosa Dos Santos AM, de Mattos Teixeira CA, Macedo Gomes VH, Silva Cardoso EH, et al. COVID-19 mortality underreporting in Brazil: analysis of data from government internet portals. **J Med Internet** Res 2020 Aug 18;22(8):e21413J Med Internet Res 2020;22(8):e21413

WORLDOMETER. **Coronavirus Cases**, 2020. Disponível em: https://www.worldometers.info/coronavirus/country/brazil/

YANG, A. P., Liu, J. P., Tao, W. Q., & Li, H. M. (2020). The diagnostic and predictive role of NLR, d-NLR and PLR in COVID-19 patients. ***International immunopharmacology***, *84*, 106504.

ZHU N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. **N Engl J Med**. 2020;382:727-33.

APÊNDICE A – A Plataforma Web

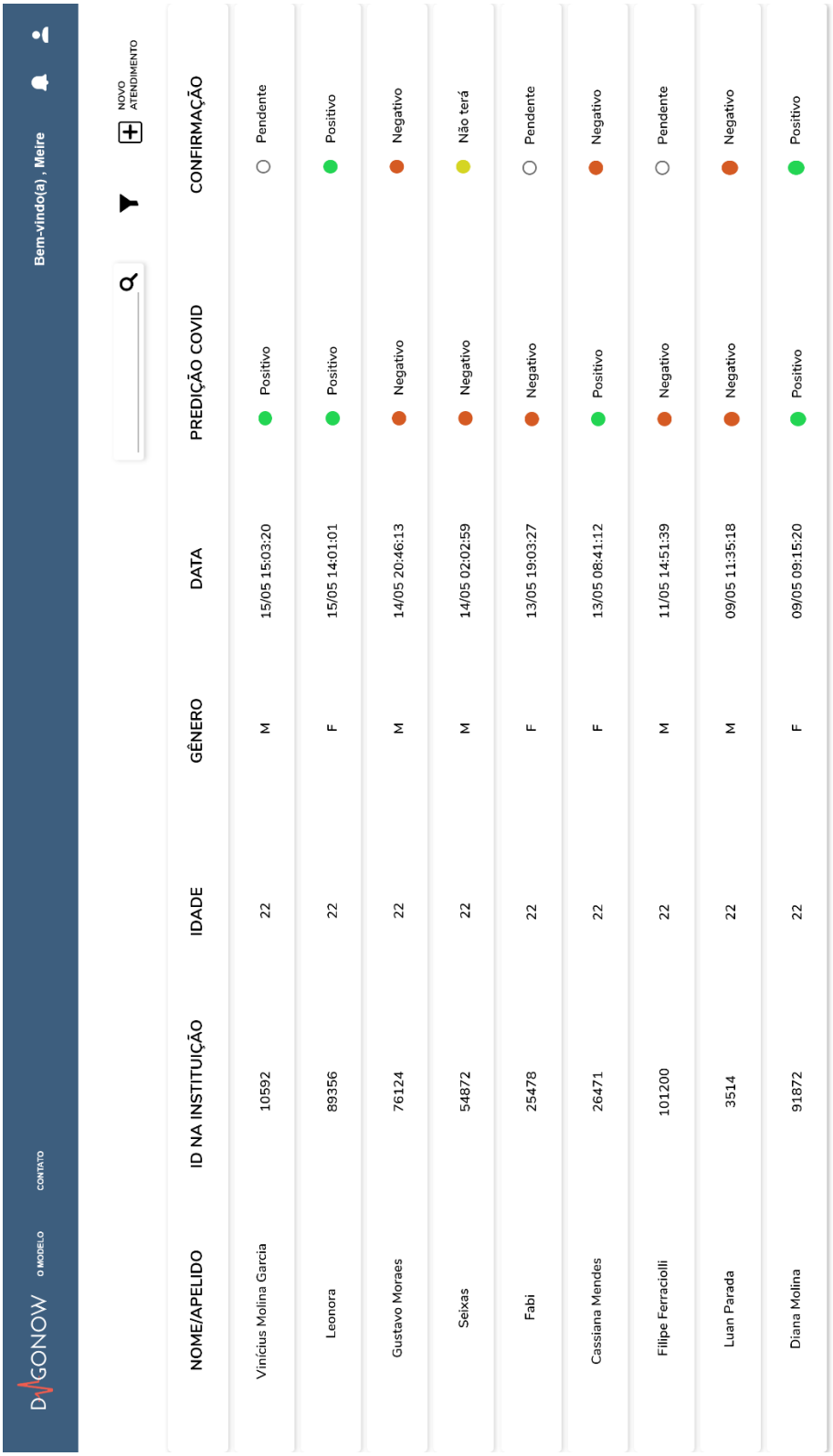
Para que os modelos de *Machine Learning* chegassem até os usuários finais, foi desenvolvida uma plataforma web, que faz parte do grande projeto denominado Diagonow. Nesta plataforma, foi feito um sistema de permissionamento para fornecer tokens de autorização entre os diferentes tipos de usuário:

- Paciente
- Enfermagem
- Médicos
- Admin
- Super

Os pacientes podem ver todos os seus atendimentos, os profissionais de enfermagem têm acesso aos resultados de todos os pacientes que atendeu, a equipe médica tem acesso a todos os resultados de todos os pacientes da instituição. Já os *admins* têm todas as permissões anteriores e mais permissões cadastrais. Já os usuários do tipo "super" são os mantenedores do sistema.

Ao realizar login no sistema, o usuário visualiza uma lista dos atendimentos dentro de sua permissão, como mostrado na Figura 16. Nesta tela, ele também consegue visualizar informações sobre o paciente, como idade, um identificador interno da unidade de saúde (e.g. número do prontuário) e questões relacionadas ao atendimento, como a data e qual foi a classificação realizada pelo modelo. Também tem um campo disponível para que o usuário possa confirmar o resultado do exame, caso seja feito um teste mais específico. Desta forma, a plataforma continua sendo alimentada com novos dados, abrindo possibilidades para aprendizado por reforço. Vale salientar, que na estruturação atual, os modelos não são atualizados em tempo real. Os dados são salvos em banco e, periodicamente, são replicados para uma base no BigQuery da Google, onde os cientistas de dados podem estudar os novos dados e aprimorar os modelos de forma mais robusta.

Figura 16. Tela inicial da plataforma web, mostrando uma lista de atendimentos já realizados.

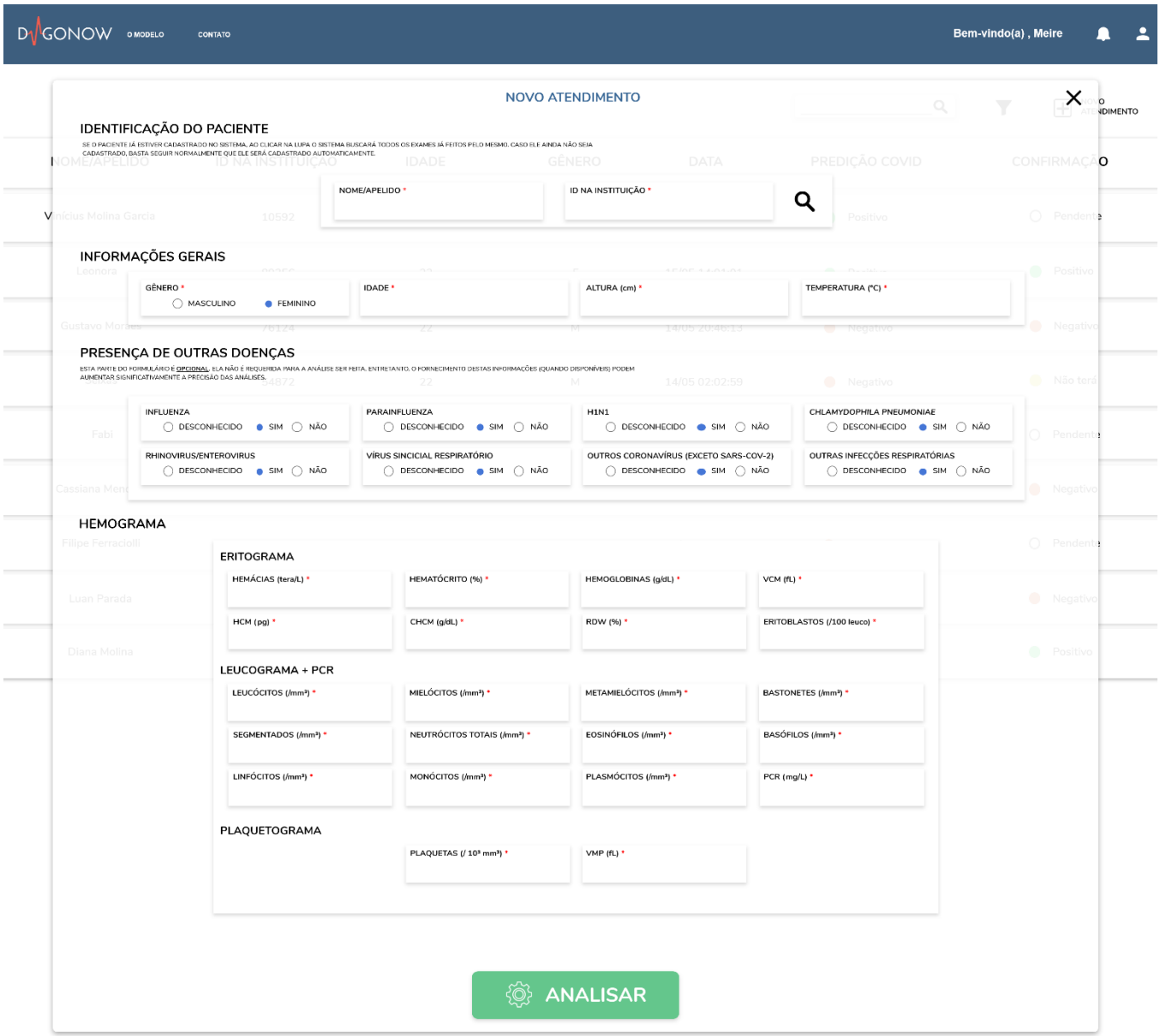


NOME/APELIDO	ID NA INSTITUIÇÃO	IDADE	GÊNERO	DATA	PREDIÇÃO COVID	CONFIRMAÇÃO
Vinícius Molina Garcia	10592	22	M	15/05 15:03:20	Positivo	Pendente
Leonora	89356	22	F	15/05 14:01:01	Positivo	Positivo
Gustavo Moraes	76124	22	M	14/05 20:46:13	Negativo	Negativo
Seixas	54872	22	M	14/05 02:02:59	Negativo	Não terá
Fabi	25478	22	F	13/05 19:03:27	Negativo	Pendente
Cassiana Mendes	26471	22	F	13/05 08:41:12	Positivo	Negativo
Filipe Ferraciolli	101200	22	M	11/05 14:51:39	Negativo	Pendente
Luan Parada	3514	22	M	09/05 11:35:18	Negativo	Negativo
Diana Molina	91872	22	F	09/05 09:15:20	Positivo	Positivo

Fonte: reprodução própria

Ao selecionar a opção de “Novo Atendimento” o usuário tem acesso a um formulário para inserir as informações de um novo atendimento, como na Figura 17.

Figura 17. Tela para registrar um novo atendimento na plataforma.

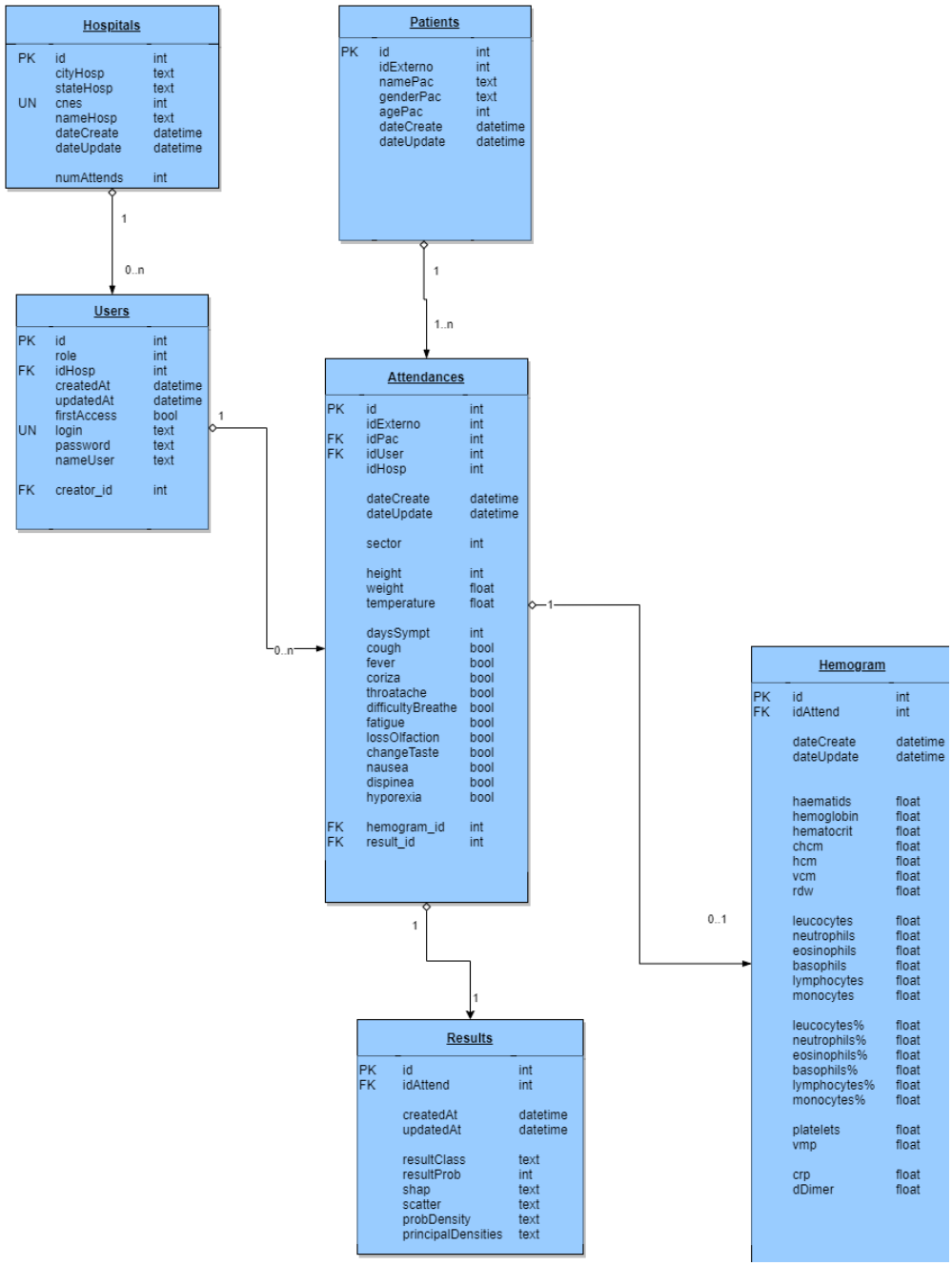


Fonte: reprodução própria

Após clicar em analisar, o usuário recebe um relatório explicando os principais aspectos da classificação realizada pelo modelo, que pode ser visto nas Figura 5 e 6. Adicionalmente, vale-se comentar de dois aspectos do sistema: a modelagem do banco de dados e a arquitetura desenhada.

Quanto ao primeiro fator, foi estruturado um banco de dados relacional, com a tecnologia PostgreSQL, hospedado na RDS da AWS. Sua estrutura de tabelas pode ser vista na Figura 18.

Figura 18. Modelagem do banco de dados da plataforma

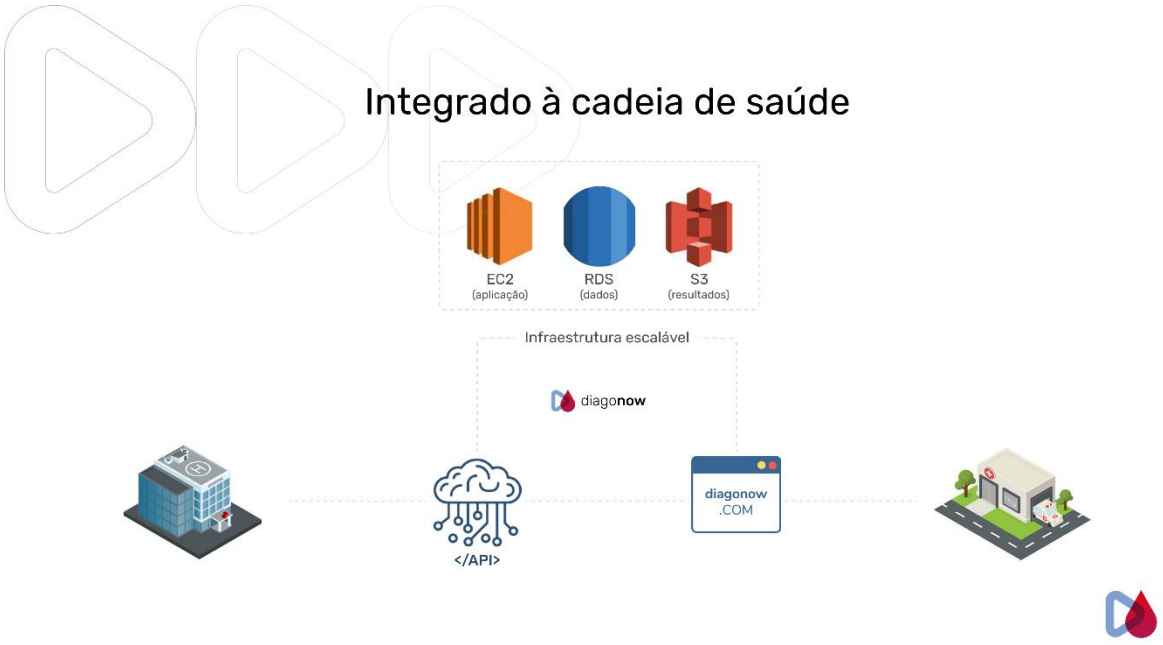


Fonte: reprodução própria

Outro fator é a arquitetura em cloud para conferir escalabilidade ao sistema. Optou-se por utilizar os serviços da AWS, principalmente a EC2 e o S3. O primeiro é onde está hospedada a aplicação e o segundo foi configurado para armazenar arquivos relacionados a PDFs e fotos de hemogramas, em cima dos quais poderia ser feito um processamento de imagem para extração do conteúdo. Ou também para receber exames de imagem em um contexto futuro.

Além da inserção de informações via formulário, mostrada na Figura X, também foi estruturada uma API que permite todas as operações que podem ser feitas pelo *frontend* da aplicação, além de poder realiza-las em lotes. Assim, tem-se uma arquitetura ilustrada pela Figura 19.

Figura 19. Visão ilustrativa da arquitetura integrada à cadeia de saúde



Fonte: reprodução própria