UNIVERSIDADE FEDERAL DE MINAS GERAIS
Programa de Pós-graduação em Inovação Tecnológica - PPGIT

Carlos Anderson Oliveira Silva

Sistema de suporte à decisão baseado em Inteligência Artificial para predição de doenças arteriais coronárias

Belo Horizonte

2022

Carlos Anderson Oliveira Silva

Sistema de suporte à decisão baseado em Inteligência Artificial para predição de doenças arteriais coronárias

Tese apresentada ao Programação de Pós-graduação em Inovação Tecnológica da Universidade Federal de Minas Gerais, como requisito para a obtenção do grau de Doutor em Inovação Tecnológica/Ciência da Computação.

Orientador: Prof. Dr Cristiano Leite de Castro
Co-orientador: Prof. Dr Michel Bessani

Belo Horizonte

2022

Ficha Catalográfica

S586s
2022
T

Silva, Carlos Anderson Oliveira.
Sistema de suporte à decisão baseado em inteligência artificial para predição de doenças arteriais coronárias [manuscrito] / Carlos Anderson Oliveira Silva. 2022.
1 recurso online (84 f. : il., gráfs., tabs., color.) : pdf.

Orientador: Cristiano Leite de Castro.
Coorientador: Michel Bessani.

Tese (doutorado) – Universidade Federal de Minas Gerais – Departamento de Química (Programa de Pós-Graduação em Inovação Tecnológica).
Bibliografia: f. 76-84.

1. Inovações tecnológicas – Teses. 2. Inteligência artificial – Teses. 3. Aprendizado do computador – Teses. 4. Coronariopatias – Teses. 5. Sistemas de suporte de decisão – Teses. 6. Clínica médica – Processo decisório – Teses. 7. Redes neurais (Computação) – Teses. 8. Síndrome das apneias do sono – Teses. 9. Coração – Doenças – Diagnóstico – Teses. 10. Python (Linguagem de programação de computador) – Teses. 11. Software – Desenvolvimento – Teses. I. Castro, Cristiano Leite de, Orientador. II. Bessani, Michel, Coorientador. III. Título.

CDU 043

Elaborada por Sérgio Ferreira da Silva – CRB6-2719.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Programa de pós-graduação em Inovação Tecnológica

“SISTEMA DE SUPORTE À DECISÃO BASEADO EM INTELIGÊNCIA ARTIFICIAL PARA PREDIÇÃO DE DOENÇAS ARTERIAIS CORONÁRIAS”.

CARLOS ANDERSON OLIVEIRA SILVA, Nº DE REGISTRO 2017771010.

Tese **Aprovada** pela Banca Examinadora constituída pelos Professores Doutores:

Professor Doutor Cristiano Leite de Castro (Orientador)
(PPG em Inovação Tecnológica da UFMG)

Professor Doutor Michel Bessani (Coorientador)
(Departamento de Engenharia Elétrica da UFMG)

Professor Doutor Carlos Dias Maciel
(Universidade de São Paulo - USP-São Carlos)

Professor Doutor Thiago Souza Rodrigues
(Centro Federal de Educação Tecnológica de Minas Gerais - CEFET-MG)

Professor Doutor Henrique Resende Martins
(Departamento de Engenharia Elétrica da UFMG)

Professor Doutor Wagner Meira Junior
(PPG em Inovação Tecnológica da UFMG)

Professor Doutor Ado Jório de Vasconcelos
Coordenador do PPG em Inovação Tecnológica da UFMG

Belo Horizonte, 17 de novembro de 2022.

Documento assinado eletronicamente por **Carlos Dias Maciel**, **Usuário Externo**, em 17/11/2022, às 15:09, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Michel Bessani**, **Professor do Magistério Superior**, em 17/11/2022, às 15:16, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Thiago de Souza Rodrigues**, **Usuário Externo**, em 18/11/2022, às 11:20, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Cristiano Leite de Castro**, **Professor do Magistério Superior**, em 18/11/2022, às 14:17, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Henrique Resende Martins**, **Professor do Magistério Superior**, em 05/12/2022, às 11:18, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Wagner Meira Junior**, **Professor do Magistério Superior**, em 26/01/2023, às 16:48, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ado Jorio de Vasconcelos**, **Coordenador(a)**, em 23/02/2023, às 10:31, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1903051** e o código CRC **267D0AEB**.

Referência: Processo nº 23072.267119/2022-36 SEI nº 1903051

Agradecimentos

O caminho percorrido para se obter o grau de doutor é longo e árduo, mas muito edificante. Não iniciamos este percurso no ato da matrícula no programa de doutorado, mas durante toda a vida acadêmica. Durante esse tempo, foi necessário ter muito empenho e dedicação nas lutas, que foram diárias e se deram, ainda, num período em que a ciência foi tantas vezes injustamente desacreditada e sucateada em nosso país. Resistência e persistência. E, graças a Deus e às boas pessoas que me cercaram, apoiaram e incentivaram, este sonho se tornou realidade. São para esses que escrevo os meus agradecimentos.

Primeiramente, agradeço a Deus. Sem Ele, nada somos e nada seremos. A todos os meus professores. Desde a minha infância até os que tive no meu doutoramento. Vocês são inspirações para muitos, inclusive para mim.

Agradeço ao meu pai, Noé Caetano, pelos ensinamentos de honra e caráter, que sempre estiveram presentes nos seus exemplos. À minha mãe, Maria Aparecida, sertaneja, amorosa e de baixa escolaridade, que, mesmo sem saber ao certo o significava o *stricto sensu* para minha vida, sempre me apoiou e acreditou nas minhas escolhas. Aos meus irmãos, Emanoel, Jéssica e Joyce, fontes de amor, incentivo e força. Aos irmãos que a vida me presenteou, João Leandro, Igor Oliveira e Rafael González: obrigado pelo companheirismo, incentivo, pelas companhias e bons conselhos que, sem dúvidas, me fizeram chegar até aqui.

Ao meu orientador, Prof. Dr. Cristiano Leite; ao meu co-orientador, Prof. Dr. Michel Bersani; aos mestres Prof. Dr. Ruben Sinisterra, Prof.^{a} Dr.^{a} Liliana Otero e Prof. Dr. Antônio Pádua de Braga: obrigado pela oportunidade, apoio, correções e ensinamentos. À minha querida amiga e secretária do PPGIT, Eni Rocha; aos companheiros de laboratório LITC da UFMG: obrigado pelo acolhimento, apoio e carinho.

Aos amigos Tiago Moreira, Edmilson Barbosa, Jardel Martins, Gil Gobira e Eduardo Juneo, que abriram as portas das suas casas para me receber num período exaustivo de minha caminhada, onde precisei conciliar o Doutorado, a docência e os quase $800km$ que separavam meu trabalho e a UFMG.

Aos colegas e amigos do Instituto Federal do Norte de Minas Gerias, agradeço pela compreensão do meu esforço e apoio com flexibilização de horários. Aos colegas e amigos do Instituto Federal Baiano, sou-lhes grato pela recepção e apoio no fechamento deste trabalho. Aos amigos da Colômbia, aos amigos do Brasil, e a todos aqueles que fizeram parte, de forma direta ou indireta, deste trabalho, o meu sincero e entusiasmado MUITO OBRIGADO!

Aos meus professores, minha filha, meus pais, irmãos e sobrinhos.

Resumo

A aplicação de aprendizagem de máquina tem se tornado cada vez mais comum em diversas áreas profissionais. Dentre as várias possibilidades que essa tecnologia permite, utiliza-la como ferramenta de predição e suporte à tomada de decisão vem se mostrando bastante promissora. Entretanto, a característica "caixa preta" de alguns modelos tem inviabilizado a utilização dessa tecnologia, principalmente na área médica. Os profissionais de saúde precisam de clareza sobre os fatores que indicam um diagnóstico. Afinal, um diagnóstico errado pode levar a perca da vida de um paciente. Utilizando uma base de dados com mais de 560 mil registros de consultas médicas em pacientes, este trabalho propõe uma metodologia que construa modelos de aprendizagem de máquina explicáveis para predição de diagnósticos de Fibrilação Auricular, Enfermidade Coronária e Apneia do Sono, incorporando dados históricos estruturados e não estruturados dos pacientes. Aprendizado Fracamente Supervisionado é usado para rotular os dados não estruturados, XGBoost é usado para predição e o método SHAP é usado para explicar a predição. Por fim, toda proposta é implementada em um *software web* escrito em linguagem de programação Python. Os resultados são promissores, além da capacidade de predição acurada, a explicação da predição destaca características históricas do paciente com maior impacto no processo de tomada de decisão do diagnóstico sugerido.

Palavras-chave: Aprendizado de Máquina. Doenças Coronárias. Sistema de Apoio à Decisão Clínica. Aprendizado de máquina explicável. Aprendizado Fracamente Supervisionado.

Abstract

The application of machine learning has become increasingly common in various professional areas. Among the various possibilities that this technology allows, using it as a prediction and decision-making tool has been very promising. However, the "Black-Box" feature of some models has made the use of this technology unviable, especially in the medical field. Healthcare professionals need clarity about the factors that indicate a diagnosis. After all, a wrong diagnosis can lead to a patient's life. Using a database with more than 560,000 records of medical appointments in patients, this work proposes a methodology that builds explainable machine learning models for the diagnosis prediction of auricular fibrillation, coronary sickness, and sleep apnea, incorporating patient's structured and unstructured historical data. Weak supervision is used to label the unstructured data, XGBoost is used for prediction, and the SHAP method is used to explain the prediction. Finally, the methodology is implemented in Web Software written in the Python programming language. The results are promising, in addition to the accurate prediction capability, the prediction explanation highlights the patient's historical characteristics with a higher impact on the decision-making process of the suggested diagnostic.

Keywords: Machine Learning. Coronary Diseases. Clinical Decision Support System. Explainable Machine Learning. Weak Supervision.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AS	Apneia do Sono
BoW	*Bag of Words*
CBOW	*Continuous Bag of Words*
CDSS	*Clinical Decision Support Systems*
DL	*Deep Learning*
EC	Enfermidade Coronária
EHR	*Electronic Health Record*
F1	F1-*score*
FA	Fibrilação Auricular
I.A.	Inteligência Artifical
ICD-10th	Classificação Internacional de Doenças (10ª versão)
KDD	*Knowledge Discovery in Databases*
LF	*Label Function*
LPI	Lei de Propriedade Industrial
ML	*Machine Learning*
MLP	Rede Neural de Múltiplas Camadas
RNA	Redes Neurais Artificiais
ROC	*Receiver Operating Characteristic*
SME	*Subject Matter Experts*
SV	*Shapley Values*
UAC	Área sob a curva ROC

Sumário

1 Introdução

A inovação voltada ao mercado da saúde é um campo em expansão com expectativas promissoras. Pesquisadores de todo o mundo se valem dos diversos recursos tecnológicos disponíveis para aprimorarem as suas pesquisas e melhorarem a perspectiva e qualidade de vida da população mundial. Um grande desafio destes pesquisadores é a longevidade aliada à qualidade de vida. Entender o desenvolvimento de doenças para diagnosticá-las no tempo mais precoce possível tem um impacto direto na diminuição de riscos de morte. Sob a perspectiva de diminuir os números de mortes e, consequentemente, aumentar a longevidade da população, uma ação intuitiva é atacar as doenças com maior incidência de óbitos. E por essa ótica, não há doença que mais mate no mundo que as doenças cardiovasculares (OMS, 2022).

Enfermidades cardiovasculares é um grupo de doenças do coração e dos vasos sanguíneos. Enfermidade Coronariana (EC), Fibrilação Auricular (ou Atrial - FA) e Cardiopatia Congênita são exemplos dessas doenças. Elas são a principal causa de mortes no mundo. Mais se morre por essas doenças do que por qualquer uma outra em todo o planeta. Ademais, mais de três quartos dessas mortes ocorrem em países de baixa e média renda, como Brasil e Colômbia (OPAS/OMS, 2022).

Tanto a FA quanto a EC compartilham fatores de risco associados, tais como Obesidade, Hipertensão e Diabetes Mellitus (MOVAHED MEHRTASH HASHEMZADEH, 2005). Além disso, Apneia do Sono (AS) não diagnosticada pode contribuir para o aumento da incidência de FA e EC (LINZ et al., 2018).

Uma parceria entre a Universidade Federal de Minas Gerais (UFMG), situada na cidade Belo Horizonte, Brasil, a *Pontificia Universidad Javeriana* (PUC-Co), situada em Bogotá, Colômbia, o *Instituto del Corazón de Bucaramanga* (ICB-Co), situado nas cidades de Bucaramanga e Bogotá, ambas na Colômbia e a Universidade de Calgary (UC-Ca), situada em Calgary, Canadá propõe a aplicação de novas ferramentas de análise, utilizando *Big Data*, *PheWAS* e *GWAS* para avançar no conhecimento dos mecanismos fisiopatológicos da Fibrilação Auricular e Apneia do Sono. Os objetivos do projeto fruto dessa parceria são: desenvolver novas tecnologias de vigilância em medicina; propor novos protocolos de assistência à saúde que contribuam para o diagnóstico e prevenção precoces de Fibrilação Auricular e Apneia do Sono; identificar marcadores de risco de FA e AS que permitam reduzir os custos de saúde pública causados pelo sub-diagnóstico e tratamento dessas doenças e suas comorbidades.

Dentre os recursos disponíveis para os pesquisadores, há uma base dados que possui mais de 560 mil registros de consultas médicas em pacientes que foram diagnosticados com diversas enfermidades. Dentre elas Fibrilação Auricular, Apneia do Sono e Enfermidades Coronárias. Essa base é composta com informações históricas de pacientes que foram atendidos no *Instituto*

del Corazón de Bucaramanga e no *Hospital Universitario San Ignacio* de Bogotá (HUSI-Co), ambos na Colômbia. Dentre as informações dos pacientes que fazem parte da base de dados, pode-se citar peso, altura, idade, sexo, sintomas apresentados, histórico do uso de tóxico ou medicamentos, antecedentes cirúrgicos, familiares, dentre outros.

A eficácia na detecção de FA e EC pode ser aprimorada por modelos de previsão de risco baseados em aprendizado de máquina criados a partir de registros eletrônicos de saúde (do inglês, *Electronic Health Records* - EHR), conforme relatado em estudos anteriores (LIP et al., 2010; ALONSO et al., 2013; SALIBA et al., 2016; LI et al., 2019; HULME et al., 2019). De fato, a previsão do risco clínico dos pacientes é uma ferramenta importante para prever ocorrências críticas na área médica, incluindo readmissão hospitalar (CARUANA et al., 2015; MAHMOUDI et al., 2020; DU et al., 2020; DU et al., 2021), início de doença crônica (CHOI et al., 2017; BRISIMI et al., 2018), mortalidade intra-hospitalar (SZEGEDY et al., 2016; ENDO et al., 2018; WANG et al., 2020; BRAJER et al., 2020), entre outros. Atacar estes problemas de saúde a partir de dados históricos e demográficos de pacientes tem despertado o interesse de muitos pesquisadores nos últimos anos. Não apenas por ser importante em perspectivas clínicas, mas devido ao desafio de lidar com EHR que geralmente possui dados esparsos, ausentes e não estruturados (ZHANG et al., 2019; REN et al., 2019; TIWARI et al., 2020).

Embora estudos tenham relatado sucesso na previsão do risco de doenças cardiovasculares a partir do EHR, uma série de dificuldades são encontradas quando se opta por trabalhar com dados dessa natureza. Dentre algumas das dificuldades podemos citar a alta dimensionalidade, a falta de padronização na captura, o não preenchimento de todos os dados disponíveis, a omissão de informações relevantes por parte do paciente, ou ainda os custos associados ao processamento dos dados. Frente a estes obstáculos, em muitas clínicas médicas e hospitais estes dados podem não ser aproveitados de maneira eficiente, desperdiçando uma relevante fonte de conhecimento acerca do desenvolvimento de doenças ou ainda de um melhor entendimento e caracterização de eventuais especificidades de uma determinada população.

Essas bases de dados, em seu estado bruto, são comumente encontradas com os dados em duas formas: estruturados e não estruturados. Os dados estruturados são aqueles com escopo e estrutura bem definidas (idade, sexo, índice de massa corporal e outros), enquanto os não estruturados são registros feitos a mão e que relatam de maneira direta aquilo que o médico avalia ser importante sobre o paciente (descrições textuais sobre registros de saúde da família, abuso de tóxico, sintomas e outros). Embora estudos tenham relatado sucesso na predição de risco de doenças cardiovasculares a partir dos EHRs, uma quantidade significativa de dados não estruturados têm sido geralmente negligenciados na construção dos modelos devido as dificuldades e custos associados ao processamento desses tipos de dados (TAYEFI et al., 2021).

Lidar com dados não estruturados não é uma tarefa trivial. Para se utilizar os modelos computacionais neste tipo de dados, se faz necessário mudar a sua representação de textual para uma computacionalmente compreensível. Na literatura há trabalhos que propõem soluções para

este problema baseado em Inteligência Artificial (I.A.) (TOUTANOVA et al., 2015; XIAO et al., 2017; ZAPPONE et al., 2019; GURUNATH et al., 2021). Dentre estes, uma abordagem que se apresenta promissora na literatura para atacar esse problema de conversão de reapresentação dos dados textuais é a *Embedding* (LE; MIKOLOV, 2014; TOUTANOVA et al., 2015; KIM et al., 2019; GURUNATH et al., 2021). Essa abordagem é fundamentalmente uma forma de organização que liga a compreensão humana do conhecimento de forma significativa à compreensão de uma máquina (GURUNATH et al., 2021) e com isso possibilita a utilização destes dados como vetor de entrada em modelos de predição de riscos.

O potencial de usar os dados não estruturados juntamente com variáveis estruturadas/tabulares é enorme, especialmente com o advento das técnicas de *Embedding* textual baseadas em redes neurais de aprendizado profundo (do inglês, *Deep Learning*). Nesse contexto, os autores Shickel et al. (2017a) relataram que o uso de *Embeddings* textuais clínicos melhorou o desempenho preditivo em comparação com os modelos convencionais e, portanto, potencialmente tem inúmeras aplicações em ambientes de saúde onde informações complexas e heterogêneas requerem representação sucinta.

A *Deep Learning* (DL) é uma abordagem que utiliza Redes Neurais Artificiais (RNA) com muitas camadas de neurônios matemáticos para aprender e reconhecer padrões. Ela tem um grande poder de convergência em diversas aplicações. Contudo, os seus altos índices de acurácia não têm sido suficientes para convencer médicos e profissionais da saúde a serem adotadas como ferramentas de suporte a decisão clínica. Isso por que, segundo Rush, Celi e Stone (2019a), os modelos DL tem um característica de serem “caixas pretas" e não oferecem uma transparência de como se chegou até os resultados obtidos. Ou seja, tem-se uma entrada de dados no modelo (*input*) e um resultado é obtido (*outcome*) e nenhuma explicabilidade dos fatores que levaram a obtenção daquela resposta.

A complexidade desses modelos e a consequente falta de explicação (transparência) sobre suas decisões têm imposto dificuldades para sua aceitação como ferramentas de apoio em um ambiente clínico. Assim, DL e outros modelos de aprendizado de máquina devem apresentar, além de previsões acuradas, explicações sobre o processo de inferência.

Os sistemas de apoio à decisão clínica (do inglês, *Clinical Decision Support Systems* - CDSS) não são projetados para substituir especialistas médicos, mas sim para ajudá-los a diagnosticar doenças com base no conhecimento empírico (NAZARI et al., 2018). Esses sistemas devem ser seguros e confiáveis (SHORTLIFFE; SEPÚLVEDA, 2018), fornecendo suporte à decisão clínica baseado em evidências pelo menos semelhante ao padrão de atendimento.

Em face a base de dados com informações históricas de pacientes atendidos no ICB-Co e no HUSI-Co e tomada como uma relevante fonte de informação e conhecimento a cerca das doenças diagnosticas naqueles indivíduos, este trabalho levanta as seguintes hipóteses: É possível rotular e classificar os relatos não estruturados de antecedentes médicos dos pacientes? É possível predizer o desenvolvimento das doenças com base nos dados clínicos histórico? É

possível desenvolver uma metodologia que construa modelos de predição de desenvolvimento de Fibrilação Auricular, Enfermidade Coronária e Apneia do Sono incorporando dados históricos estruturados e não estruturados de pacientes? É possível desenvolver uma ferramenta de suporte a decisão médica com boa acurácia e explicabilidade das decisões tomadas?

Para que essas perguntas sejam respondidas de forma adequada, faz-se necessário lidar com problemas que são usualmente encontrados em bases de dados médicas: dados faltantes, informação não estruturada na forma de texto, preenchimento não-padronizado de campos feito por diferentes agentes no processo de coleta, explicabilidade da tomada de decisão de modelos de predição, entre outras coisas.

1.1 Justificativa

A Fibrilação Auricular afeta cerca de 34 milhões de pessoas em todo o mundo, sendo que os pacientes com FA tem maior risco de graves consequências para a saúde, incluindo morte e Acidente Vascular Cerebral (TISON et al., 2018). A Apneia do Sono é a anormalidade respiratória mais comum durante o sono. O diagnóstico e o tratamento da AS em pacientes com FA requerem uma estreita colaboração interdisciplinar entre eletrofisiológicas, cardiologistas e especialistas em sono (LINZ et al., 2018).

A AS não diagnosticada pode contribuir para o aumento da incidência de FA e EC (LIP; BEEVERS, 1995; MIYASAKA et al., 2006), pois a EC está presente em $17\% - 46,5\%$ dos pacientes com FA (HOHNLOSER et al., 2009; AFFIRM Investigators et al., 2002). Tanto a FA quanto a EC compartilham fatores de risco associados, como Obesidade, AS, Hipertensão e Diabetes Mellitus. A incorporação de sistemas de aprendizado de máquina aos EHRs de pacientes diagnosticados com AS, EC e FA pode ser útil para conhecer o comportamento de dados fisiológicos e as relações entre essas doenças na população. Uma preocupação com o aprendizado de máquina aplicado à saúde é descrito como o fator "caixa preta" (RUSH; CELI; STONE, 2019b). De fato, os algoritmos podem distinguir pacientes com FA, EC e AS, mas a contribuição dos fatores envolvidos no diagnóstico de cada doença não são indicadas por estes modelos.

A AS é reconhecida como fator de risco para Enfermidades Cardiovasculares, mas é amplamente sub-diagnosticada. Fragmentação do sono, Hipóxia intermitente repetitiva e inflamação sistêmica têm sido propostas como fatores de risco para EC e FA em pacientes com AS (LIP; BEEVERS, 1995), mas os mecanismos fisiopatológicos envolvidos no risco de doença cardiovascular nesses pacientes não foram elucidados.

De acordo com Academia Americana de Medicina do Sono (AASM) [1] 136 bilhões de dólares é o custo anual representado pela perda de produtividade no trabalho, acidentes de trânsito, acidentes de trabalho e custos relacionados ao tratamento de Enfermidades Cardiovasculares e outras comorbidades associadas à Apneia do Sono. Outro estudo que mostra o impacto negativo dessas doenças sob o aspecto econômico foi realizado na Austrália em 2013 e apontou que 5.1 bilhões de dólares foram o impacto produzido por distúrbios do sono na economia daquele país (TUNG; ANTER, 2016).

Até onde sabemos, não há estudos anteriores que usem EHR para desenvolver modelos de aprendizado de máquina para prever o risco de AS. Além disso, há uma falta de estudos explorando a fusão de dados de características físicas, diagnósticos anteriores e dados não estruturados sobre os antecedentes dos pacientes para desenvolver modelos de previsão de diagnóstico.

[1] https://aasm.org/cost-sleepiness-pricey-ignore/

Visto isso, este trabalho justifica-se pelo impacto social e econômico que as suas contribuições produzirão para a sociedade.

1.2 Trabalhos Correlatos

Outros trabalhos na literatura que utilizam métodos similares aos propostos neste trabalho colaboram com os avanços alcançados até aqui. Vários deles sobre o diagnóstico de doenças utilizando EHR, e alguns específicos de doença cardíaca. Em específico para FA, pode-se citar, por exemplo, CHARGE-AF (ALONSO et al., 2013), CHA2DS2-VASc (LIP et al., 2010; SALIBA et al., 2016), C2HEST (LI et al., 2019) e EHR-AF (HULME et al., 2019).

Um trabalho que a partir de técnicas de aprendizagem de máquina (do inglês, *Machine Learning*) extrai conhecimento em base de dados médicas é o desenvolvido por Batbaatar, Pham e Ryu (2020). Neste, os autores tem por objetivo projetar um modelo de classificação, capturar informações semânticas de uma grande quantidade de documentos afim de extrair a complexidade do câncer dos pacientes e analisar tendências e tópicos característicos a cada câncer. Os métodos empregados são redes neurais, Aprendizado Fracamente Supervisionado (do inglês, *Weak Supervision*) e Modelagem de Tópicos. Os resultados mostram que é possível extrair com eficiência estruturas complexas de conhecimento sobre o câncer, de acordo com as características e tópicos relacionados.

Brajer et al. (2020) fazem uso de *Machine Learning* (ML) para avaliar os dados dos prontuários de 75.247 pacientes e prospectar, em função do tempo, a evolução a óbito dos pacientes. Para avaliar o desempenho da técnica foi utilizada a curva ROC (do inglês, *Receiver Operating Characteristic*). Esse é um método gráfico para avaliação, organização e seleção de sistemas de diagnóstico e/ou predição. Os índices obtidos oscilaram entre 0.84 e 0.89. Os autores argumentam que o modelo demonstrou boa discriminação na identificação de pacientes com alto risco de mortalidade hospitalar e que pode ser usado para melhorar a tomada de decisões clínicas e operacionais.

Pickhardt et al. (2020) compara a capacidade prognóstica de biomarcadores identificados via algoritmos baseados em características e aprendizado profundo com parâmetros clínicos dos pacientes a fim de prever a ocorrência de eventos cardiovasculares e tempo de sobrevida na fase adulta. São avaliados 9.223 pacientes e a métrica de desempenho utilizada é a curva ROC. Com um intervalo de confiança de 95%, combinados todos os biomarcadores preditados, os índices alcançados oscilam entre 0.76 e 0.86. Na conclusão do trabalho, os autores discutem que biomarcadores identificados via algoritmo por meio automatizado podem superar os parâmetros clínicos estabelecidos para estratificação de risco pré-sintomática para futuros eventos graves adversos.

Hulme et al. (2019) fundiu EHR de múltiplas instituições de saúde e identificou 412.085

indivíduos entre 45 e 95 anos sem diagnóstico de Fibrilação Auricular (de 2000 a 2014). Eles compararam modelos com o objetivo de avaliar o desempenho frente a tarefa de predição de risco de FA para os próximos cinco anos, em que 14.334 pacientes foram diagnosticados com essa doença. O modelo que obteve o melhor desempenho foi EHR-AF, alcançando uma área sob a curva ROC (AUC) de 0.77.

Outro estudo comparativo de métodos preditores de risco foi publicado por Khurshid et al. (2020). Usando registros eletrônicos de mais de quatro milhões de pacientes, os autores compararam o desempenho dos métodos EHR-AF, CHARGE-AF, C2HEST e CHA2DS2-VASc aplicado à previsão de risco de desenvolver FA. As variáveis de entrada do modelo foram sexo, idade, raça, fumante, altura, peso, pressão arterial, hipertensão, hiperlipidemia, insuficiência cardíaca, doença coronariana, doença valvular, acidente vascular cerebral prévio, doença arterial periférica, doença renal crônica e hipotireoidismo. A métrica escolhida para avaliação foi AUC e o melhor modelo foi EHR-AF com 0.808.

Kim e Kang (2017) propôs um método para a previsão de risco de EC baseado em um modelo de rede neural e correlação de características. As variáveis de entrada escolhidas para o treinamento de modelo foram idade, sexo, índice de massa corporal (IMC), colesterol total, colesterol HDL, pressão arterial sistólica, pressão arterial diastólica, triglicéride, hemoglobina, doença da tireoide, hepatite do tipo B, hepatite do tipo C, cirrose, tabagismo e diabetes. Todos estes, dados estruturados. O desempenho do modelo foi avaliado com a AUC reportando 0.749 como o melhor resultado.

Os estudos acima mencionados mostraram que a predição de risco das doenças cardíacas com EHR é possível. No entanto, como mencionado por Khurshid et al. (2020), embora tais estimativas de risco possam ser viáveis e precisas, elas exigem uma validação adicional observando as características, os sintomas e a história clínica dos pacientes. Além disso, esses estudos não levaram em conta todos os registros dos diagnósticos passados dos pacientes. Nosso estudo, por sua vez, considera todas as informações disponíveis, incluindo dados textuais não estruturados sobre a história da saúde da família. Nossas estimativas baseadas em modelos não se destinam a substituir a opinião do especialista, mas sim apoiar sua decisão.

Shickel et al. (2017a) também considerou a fusão de dados não estruturados e estruturados. Ao aplicar uma técnica de incorporação em um banco de dados de mais de 2,7 milhões de EHR de internações do sistema nacional do Reino Unido e, com mais de 13 milhões de termos, eles construíram um modelo para aprender palavras de diagnósticos e procedimentos. Em seguida, os vetores de características (resultado da *Embendding* textual) foram usados como entrada de uma máquina de vetor de suporte linear (do inglês, *Support Virtual Machine* - SVM) (HEARST et al., 1998) que foi treinada para prever insuficiência cardíaca. A métrica empregada para avaliar o desempenho preditivo foi AUC (LI; FINE, 2010), e o modelo proposto alcançou 0.6965.

Por fim, os pesquisadores Shickel et al. (2017b) revisaram a aplicação de técnicas de DL

em tarefas clínicas com base nos dados de EHR. Eles encontraram uma variedade de métodos e estruturas de *Deep Learning* aplicados a tipos distintos de problemas de saúde, incluindo extração de características, aprendizagem de representação, predição de resultados, fenotipagem e identificação de doença. Além disso, eles trazem uma discussão sobre como explicar os modelos de aprendizagem de máquina. A transparência do modelo é de extrema importância para aplicações clínicas, uma vez que a tomada de decisão correta pode ser a diferença entre a vida e a morte, e os profissionais de saúde devem ser capazes de entender e confiar nas previsões e recomendações feitas por este tipo de sistema.

1.3 Objetivos

1.3.1 Objetivo geral

Desenvolver uma metodologia que construa modelos interpretáveis de predição de diagnósticos de Fibrilação Auricular, Enfermidade Coronária e Apneia do Sono e que incorpora dados não estruturados em sua construção;

1.3.2 Objetivos específicos

Como desdobramento do objetivo geral, pretende-se atingir os seguintes objetivos específicos:

- Aplicar técnicas de *Embedding* textuais e Aprendizado Fracamente Supervisionado com vistas a rotular relatos não estruturados de antecedentes médicos de pacientes;
- Realizar a predição de diagnóstico de AS, EC e FA com base em dados estruturados e não estruturados históricos das consultas dos pacientes;
- Apresentar uma explicação humanamente interpretável para as predições realizadas em macro e micro perspectivas;
- Desenvolver um software de suporte a decisão clínica com a tecnologia desenvolvida.

1.4 Contribuições

Diante das questões apresentadas até aqui, este trabalho apresenta duas metodologias de aprendizado, uma não supervisionada e outra supervisionada, que agrega dados não estruturados e estruturados para predizer o diagnóstico de AS, EC e FA. Um esquema de *Embedding* textual e um modelo de rotulação Fracamente Supervisionado (ou *Weak Supervision*) são usados para categorizar dados não estruturados com relatos médicos históricos dos pacientes em formato textual. Além disso, como uma melhor interpretabilidade do resultados obtidos leva a um aumento na adoção dos CDSS, a proposta busca um melhor equilíbrio entre precisão e explicabilidade. A metodologia se baseia em modelos de predição complexos, mas também agrega uma abordagem de teoria dos jogos, conhecida como SHAP (*SHapley Additive exPlanations*), para explicar os resultados do modelo em nível de grupo e individual (por paciente). Por fim, todos estes modelos são implementados em um software de suporte à decisão clínica médica que poderá ser implantado em clínicas médicas e hospitais, públicos e privados.

As principais contribuições deste trabalho são:

- aplicação de *Embeddings* textuais e Aprendizado Fracamente Supervisionado para construção de conhecimentos em dados não estruturados de pacientes;
- explicação do processo de tomada de decisão do modelo de predição de diagnóstico proposto em perspectivas macro e micro;
- desenvolvimento de um software de suporte à decisão médica e que está em fase de registro junto a Coordenadoria de Transferência e Inovação Tecnológica (CTIT) da UFMG;
- publicação do artigo "*Interpretable Risk Models for Sleep Apnea and Coronary Diseases from Structured and Non-Structured Data*"na revista *Expert Systems With Applications* [2] (Fator de impacto JCR: 8.665)
- artigo em prepração para submissão no *International Journal of Cardiology* (Fator de impacto JCR: 4.039)

[2] Silva, Carlos Anderson Oliveira, et al. "Interpretable Risk Models for Sleep Apnea and Coronary Diseases from Structured and Non-Structured Data." Expert Systems with Applications, vol. 200, Aug. 2022, p. 116955. DOI.org (Crossref), https://doi.org/10.1016/j.eswa.2022.116955.

2 Referencial Teórico

Desde a descoberta do fogo pelos nossos ancestrais *Homo sapiens* até o mais sofisticado sistema de *chipset* utilizado nos potentes computadores espalhados por *Data Centers* em todo o planeta, a inovação e a tecnologia têm sido um grande aliado no desenvolvimento e, consequente, melhoria na condição de vida da humanidade.

Segundo Ayub e Bacic (2020), a inovação, criação de novas combinações de bens ou serviços ou métodos de produção, é uma das marcas da sociedade moderna. No entanto, a criação de novas inovações não se dão pelo acaso. No desenvolvimento de um produto ou serviço é importante ter-se o mais amplo conhecimento, teórico e prático, para que se aumente as chances de êxito na sua concepção. E esse conhecimento pode se referir a pessoas, materiais, contextos, ferramentas, expertise, dentre outros. Segundo Burgelman, Christensen e Wheelwrigth (2013), existem vários tipos de inovação na literatura. Dentre elas estão as incrementais, radicais e as de arquitetura. Eles as descrevem assim:

> As inovações incrementais preveem adaptação, refinação e aprimoramento dos produtos e serviços existentes, assim como os sistemas de produção e distribuição. As radicais incluem categorias novas de produtos e serviços e/ou de sistemas de produção e distribuição. As de arquitetura referem-se a reconfigurações do sistema de componentes que constituem o produto.

Como pode ser observado na definição dos autores, inovação pode ter vários tipos e basear-se em diversos fatores. Inovar é um ciclo abrangente. Todos estes fatores no processo de inovação consomem tempo e requerem pesquisas até chegar o seu desenvolvimento. Contudo, pesquisas e desenvolvimento tem custo e necessita de investimentos das mais variadas grandezas.

Conforme discorre Ayub e Bacic (2020), para inovar há custos elevados, sendo preciso ter expectativas de lucratividade para se investir em desenvolvimento. À luz destes custos na inovação, Hall e Rosenberg (2010) reforça que o lucro conferido pelo direito de exclusividade é uma recompensa para se gerar a inovação. Se qualquer um puder replicar uma inovação, as recompensas seriam reduzidas a ponto de não haver mais incentivo para inovar.

Com vistas a proteger a propriedade das invenções, cada país dispõe de leis específicas. Essa proteção pode-se se dar de vários modos, como por exemplo, patentes, marcas, desenhos industriais, dentre outros. No Brasil, por exemplo, a lei de nº 9.279, de 14 de maio de 1996, conhecida como Lei de Propriedade Industrial (LPI), regulamenta os direitos e obrigações relativos à propriedade industrial (SOARES, 1997). Em tese, a proteção permite que as empresas se apropriem dos lucros gerados pelas tecnologias inovadoras ao impedir sua exploração por terceiros (AYUB; BACIC, 2020).

Contudo, a proteção à propriedade industrial é apenas um dos 3 grupos que envolvem

a propriedade intelectual. Além desse, têm-se o direito autoral e a proteção *Sui Generis*. Esse, compreende proteção a topografias de circuitos integrados, cultivares e conhecimentos tradicionais, enquanto aquele abrange obras literárias, artísticas e científicas, programas de computador, descobertas científicas e direitos conexos. Assim sendo, e como discute Jungmann e Bonetti (2010), a propriedade intelectual não se resume apenas em objetos e em suas cópias, mas na informação ou no conhecimento refletido, sendo, portanto, um ativo intangível.

Visto isso, é possível entender que a propriedade intelectual vai além da já conhecida patente de produtos ou serviços. Ela abrange os direitos relativos à produção ou ao conhecimento literário, artístico, científico, tradicional, etc.

O conhecimento, também referido por expertise ou *know-how*, como forma de propriedade intelectual é discutido na literatura. Para Visconti e Weis (2020) *know-how* é, no sentido mais amplo, o monitoramento da garantia da qualidade e do reparo das inovações, que é referente também à redução de defeitos e melhorias constantes.

Este *know-how* pode gerar monetização através de contratos que, conforme discute Almeida (2019), não geram patenteamento, mas que é possível documentar e negociar o descritivo técnico de testes, caminhos críticos, conhecimentos específicos, expertises e técnicas que são de conhecimento no momento da sua negociação. Dito isso, é possível concluir que o conhecimento sobre determinados conteúdos, produtos ou inovações, de um modo geral, são importantes e geram uma valoração intrínseca às companhias ou pessoas que a detêm.

Definições elementares, mas que se intersectam em parte e causam alguma confusão em seu entendimento, são as de dado, informação e conhecimento. Dados são itens elementares, captados e armazenados sem valor semântico. Informação representa os dados processados com significado e contexto bem definidos. Conhecimento corresponde a um padrão ou conjunto de padrões cuja formulação pode envolver e relacionar dados e informações (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

O grande volume de dados gerado pelas áreas do conhecimento, seja na academia, no comércio ou na indústria, acabam por acumular um registro histórico de ações e características individuais ou coletivas que são uma importante fonte de conhecimento. Entretanto, análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de tecnologias da informação.

A descoberta de conhecimento em base de dados é referida na literatura pelo acrônimo KDD (do inglês, *Knowledge Discovery in Databases*). Ela é composta por etapas operacionais que visam a preparação dos dados para que seja possível extrair conhecimento. Popular, a expressão Mineração de Dados (do inglês, *Data Mining*), é uma tarefa de KDD (AZEVEDO, 2019; HYVÖNEN; RANTALA et al., 2019; CHEN et al., 2020).

Segundo Goldschmidt, Passos e Bezerra (2015) antes dos termos KDD e Mineração de Dados se tornarem populares, muitos pesquisadores em áreas como Estatística, Aprendizado de Máquina, Reconhecimento de Padrões, Inteligência Computacional, dentre outras, trabalhavam

sobre os mesmos tipos de problemas sem uma integração entre uma e outra área. KDD reúne todas estas disciplinas sob a premissa de que bancos de dados guardam mais informação do que dados neles armazenados.

2.1 Aprendizado de máquina

Para Goldschmidt, Passos e Bezerra (2015), a Mineração de Dados é a etapa mais importante do KDD e nela há um conceito que merece ser destacado: o aprendizado. Esse conceito, no contexto da KDD, também referido como aprendizado indutivo, é definido como a capacidade que determinados algoritmos têm de aprender a partir de exemplos.

A definição de aprendizado é ampla e flutua nas diversas áreas do conhecimento. Haykin (1999) discute que o processo de aprendizado depende do ponto de vista e por isso é difícil uma definição precisa a respeito. Mas, ainda segundo ele, no âmbito do aprendizado de máquina, pode-se definir como um processo pelo qual parâmetros são adaptados de acordo com o respectivo contexto e tendo na maneira de modificação destes parâmetros o tipo de aprendizado.

A indução é a forma de inferência lógica que permite, a partir do conhecimento sobre um conjunto finito de amostras, predizer eventos ainda desconhecidos. Monard e Baranauskas (2003) lembram que foi através da indução que Arquimedes descobriu a primeira lei da hidrostática e o princípio da alavanca, que Kepler descobriu as leis do movimento planetário e que Darwin descobriu as leis da seleção natural das espécies. Contudo, estes mesmos autores chamam a atenção sobre as limitações e como essa metodologia é utilizada no aprendizado de máquina.

> Apesar da indução ser o recurso mais utilizado pelo cérebro humano para derivar conhecimento novo, ela deve ser utilizada com cautela, pois se o número de exemplos for insuficiente, ou se os exemplos não forem bem escolhidos, as hipóteses obtidas podem ser de pouco valor. [...] O aprendizado indutivo pode ser dividido em supervisionado e não-supervisionado. No aprendizado supervisionado é fornecido ao algoritmo de aprendizado, ou indutor, um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido. Já no aprendizado não supervisionado, o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos ou *clusters* (MONARD; BARANAUSKAS, 2003).

Lima, Pinheiro e Santos (2016) dizem que o aprendizado de máquina é uma linha de pesquisa que tem por objetivo estudar e desenvolver métodos computacionais para obtenção de sistemas capazes de adquirir conhecimento de forma automática. A construção deste se dá com o relacionamento de variáveis de entrada e saída a partir de dados amostrados.

Para Russell e Norvig (2016), quando o problema de aprendizado tem uma variável de

saída que pode assumir valores contínuos, trata-se de um problema de regressão. Já quando os possíveis valores são discretos, trata-se de um problema de classificação de padrões. Ainda segundo estes autores, o processo de aprendizado pode se dar de três modos: não supervisionado, supervisionado ou semi supervisionado.

2.1.1 Aprendizado não supervisionado

No aprendizado não supervisionado, o agente aprende padrões de entrada, mesmo não sendo fornecido nenhum *feedback* sobre a saída esperada. A tarefa mais comum no aprendizado não supervisionado é o agrupamento. Neste, o algoritmo tem por meta procurar padrões com base nos dados de entrada para poder inferir relações de similaridade daquele respectivo conjunto de dados (RUSSELL; NORVIG, 2016).

2.1.2 Aprendizado supervisionado

A tarefa de aprendizado supervisionado é a seguinte: dado um conjunto de dados χ composto por N pares, representado por

$$\chi = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)...(x_N, y_N)\},$$

onde cada y_i foi gerado por uma função desconhecida $y = f(x)$. O objetivo então consiste em descobrir uma função h que se aproxime da função verdadeira f. A função h é uma hipótese (RUSSELL; NORVIG, 2016).

Como pode ser visto, no aprendizado supervisionado os valores esperados na saída do algoritmo são conhecidos $(y_1, y_2, ..., y_N)$. Em outras palavras, o aprendizado supervisionado pode ser descrito como um mapeamento da entrada para a saída de um conjunto de dados através do aprendizado de uma função. Conforme descreve Russell e Norvig (2016), às vezes essa função é estocástica, e o que deve ser aprendido é a distribuição de probabilidade condicional $P(Y|x)$.

São muitas as técnicas disponíveis na literatura para atacar problemas de aprendizado supervisionado. O método kNN (acrônimo do termo inglês *k-Nearest-Neighbor*) é um delas. Simples, mas eficaz, segundo Guo et al. (2003), ela propõe uma classificação baseada na probabilidade de uma amostra fazer parte de um determinado grupo considerando apenas a similaridade entre aquele ponto e os seus vizinhas. Supondo a aplicação desta técnica para reconhecimento de um dado padrão de teste (desconhecido), sua classificação é realizada da seguinte maneira:

1. calcula-se a similaridade entre o padrão desconhecido e todos os padrões já conhecidos (padrões da fase de treinamento);

2. verifica-se quais classes pertencem os k padrões mais próximos;

3. classifica-se fazendo a associação do padrão de teste à classe que for mais frequente entre os k padrões mais próximos daquele novo padrão desconhecido.

A ideia é classificar um indivíduo desconhecido inserido no espaço de amostras com base nos rótulos atribuídos aos k vizinhos mais próximos (HAYKIN, 1999).

2.1.3 Aprendizado semi supervisionado

No semi supervisionado, a informação de saída esperada ora é disponibilizada, ora não. Ou ainda, os valores de saída possuem algum tipo de ruído que possa comprometer a fidedigna representação das expectativas do modelo a ser construído. Russell e Norvig (2016) discute que este tipo de aprendizado é um *continuum* entre à abordagem supervisionado e não supervisionado.

2.2 *Aprendizado Fracamente Supervisionado*

O aprendizado supervisionado é poderoso, mas a quantidade de dados necessária para a consolidação de uma base de dados robusta para produzir resultados de classificação significativos, pode exigir tempo e esforço humano manual. Tempo por demandar que os dados sejam rotulados um a um até conseguir produzir uma condição da base totalmente rotulada. Esforço manual humano por, se tratando de bases de dados extensas, uma atividade cansativa e repetitiva. Ainda, essas dificuldades são agravadas quando se trabalha com base de dados especializadas. Afinal, a rotulação manual restringe-se à capacidade de um corpo de especialistas no assunto, chamados na literatura de SMEs (do inglês, *Subject Matter Experts*. Suponha uma base de dados médica onde apenas médicos de uma determinada especialidade tenha competência técnica para classificar a ocorrência de um determinado evento dado os seus registros textuais?

Para amenizar esses esforços, a comunidade de aprendizado de máquina explorou o aprendizado semi supervisionado e métodos como *bootstrapping* (RILOFF; SHEPHERD, 1999), *Expectation Maximization* (DEMPSTER; LAIRD; RUBIN, 1977), *Feature Discovery* (MILLER; GUINNESS; ZAMANIAN, 2004), *Generalized Expectation* (MANN; MCCALLUM, 2010), dentre outros, foram propostos.

A *Deep Learning* permitiu que uma parte considerável desse esforço fosse evitado. Este tipo de aprendizado é tido como particularmente eficaz para problemas que possuem alta dimensionalidade e variância de entrada, como textos e imagens. No entanto, segundo Ratner et al. (2019), o aprendizado profunda tem um grande custo inicial: esses métodos precisam de

conjuntos de treinamento maciços de exemplos rotulados para aprender - geralmente dezenas de milhares a milhões para atingir o desempenho preditivo máximo.

Mais um ponto que é destaque negativo desses métodos de rotulação baseados em DL, é a relação entre acurácia e cobertura de rotulação sobre os dados. Apesar das abordagens de aprendizado profundo conseguirem uma boa cobertura de rotulação sobre os dados a serem rotulados, elas não apresentam uma acurácia tão elevada, se comparada com a assertividade obtida pela classificação feita por um especialista. Em contra partida, um SME demandaria um tempo significativamente maior para conseguir uma cobertura de rotulagem próxima a gerada por esses métodos (RATNER et al., 2019).

Com vistas à equalizar essa complexa relação na rotulação de dados com alta dimensionalidade e boa acurácia, surge a técnica de Aprendizado Fracamente Supervisionado (do inglês, *Weak Supervision*) (CACHAY; BOECKING; DUBRAWSKI, 2021; PLATANIOS et al., 2020; CHEN et al., 2020; RATNER et al., 2020; RATNER et al., 2019; BACH et al., 2017). A técnica faz o uso de funções rotuladoras (do inglês, *Label Funtions* - LF) que identificam padrões heurísticos sobre um determinado assunto nos dados não estruturados e, por meio de mecanismos matemáticos, que vão de inferência probabilística até redes neurais, geram uma rotulação para os dados com cobertura e acurácia significativas. Em vez de rotular os dados manualmente, as funções rotuladoras realizam essa atividade de maneira programática, indicando os eventuais pontos interesse nos dados, ou se abstendo quando a heurística empregada não for conclusiva.

Para um melhor entendimento da aplicação dessa técnica, tomamos a seguinte situação como exemplo: em posse de uma base de dados contendo um grande volume de comentários de usuários avaliando diversos produtos à venda no *Mercado Livre*, temos como objetivo construir um modelo que possa classificar avaliações *SPAM*, *NÃO SPAM* ou *ABSTENÇÃO*. Para essa construção, na fase de treinamento, necessitamos de dados rotulados para construir o conhecimento do decisor. Mas rotular isso manualmente é uma tarefa lenta e custosa. Para este caso podemos recorrer à abordagem do Aprendizado Fracamente Supervisionado através das LFs.

A construção das heurísticas das LFs podem ser feitas através de uma simples busca textual do termo de interesse, até à aplicação de uma sofisticada rede neural. Dada a natureza pedagógica deste exemplo, vamos demonstrar através da busca de termos ou expressões chaves. Contudo, a ideia é exatamente a mesma para as outras abordagens.

Uma característica dos comentários *SPAM* é a publicação de links no meio do texto. Assim, para indicar o rótulo de *SPAM* nos textos, podemos criar LFs que sejam responsáveis por identificar termos ou expressões como “http", “www", “.com.br", “clique aqui", “acesse o *link*" e outros mais. Em contra-partida, funções rotuladoras de comentários *NÃO SPAM* podem pesquisar por termos como “custo-benefício", “satisfeito", “recomendo" e outros. Quando a pesquisa da LF não encontra o valor de interesse, o retorno é *ABSTENÇÃO*. Ou seja, as funções sempre retornam os valores 0 para a classe Falsa (*NÃO SPAM*), 1 para a classe Verdadeira (*SPAM*) e

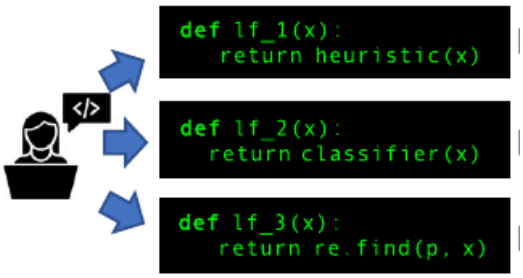

Figura 1 – Exemplo de construção de *Label Fuctions* utilizando abordagens diferentes. Fonte: Ratner et al. (2019)

-1 para dúvida da classificação (*ABSTENÇÃO*). A Figura 1 mostra exemplos genéricos de LFs utilizando abordagens distintas.

Em termos gerais, podemos descrever o funcionamento do método como, sendo x_i uma dada amostra de um conjunto de dados X de dimensão m, existe a ela um rótulo verdadeiro $y_i \in \{-1, 1\}$ associado. A partir de n LFs ($\lambda_1, \lambda_2, \lambda_3, ..., \lambda_n$), aplicadas em x_i, as saídas ($\Lambda_{i1}, \Lambda_{i2}, \Lambda_{i3}, ..., \Lambda_{in}$) são produzidas. O domínio de cada Λ_{ij} é $\{0, 1, -1\}$, correspondendo a Falso, Verdadeiro ou Abstenção. O objetivo é estimar um modelo probabilístico que gere as saídas da *Label Function* $\Lambda \in \{-1, 0, 1\}^{m \times n}$ (BACH et al., 2017).

Para isso, assume-se que as saídas são condicionalmente independentes dos rótulos verdadeiros e que a relação Λ e y é orientada por

$$\phi_j^{Acc}(\Lambda_i, y_i) = y_i \Lambda_{ij} \tag{2.1}$$

com o parâmetro ϕ_j^{Acc} indicando a acurácia de cada LF λ_j. O modelo de probabilidade condicional é descrito por:

$$p_\theta(\Lambda, Y) \propto \exp\left(\sum_{i=i}^{m}\sum_{j=i}^{n} \theta_j^{Acc}\phi_j^{Acc}(\Lambda_i, y_i)\right) \tag{2.2}$$

onde $Y = y_i, y_2, ...y_m$.

A estimativa do parâmetro θ se dá pela minimização log negativo da probabilidade marginal de $p_\theta(\bar{\Lambda})$ por meio de uma matriz observada pela saída das *Label Functions* $\bar{\Lambda}$:

$$\text{argmin}_\theta - \log \sum_Y p_\theta\left(\bar{\Lambda}, Y\right) \tag{2.3}$$

Depois de ajustar os parâmetros do modelo generativo, a distribuição entre os rótulos gerados e verdadeiros pode ser estimada e usada para treinar um modelo discriminativo, minimizando a perda em relação a sua distribuição (BACH et al., 2017). A *Label Function* não precisa ter precisão perfeita, mas sim, representar um padrão que o usuário deseja enviar ao modelo

e que é mais fácil de codificar via função de rotulação do que manualmente (RATNER et al., 2016).

Ratner et al. (2019) propuseram um sistema que permitia que os usuários treinassem modelos com boas convergência e acurácia sem rotular manualmente nenhum dado de treinamento. Para tal, foi necessário apenas que os usuários escrevessem funções rotuladoras. Neste mesmo trabalho eles disponibilizam uma biblioteca em Python chamada *Snorkel* [1]. Os autores concluem o seu trabalho dizendo que os seus experimentos demonstraram que o método reduziu significativamente o custo e a dificuldade de treinar modelos de aprendizado de máquina e se aproximando da qualidade de grandes conjuntos de treinamento rotulados à mão.

Tempo depois o mesmo grupo de pesquisadores (RATNER et al., 2019) propuseram o *MeTaL* [2]. Naquele trabalho eles argumentam que havia uma limitação quando o aprendizado atacava problemas de múltiplas tarefas ou multi granularidade e, adicionalmente, com a nova proposta, era possível realizar a inferência de rotulações desconhecidas até então. Isso só se tornou possível graças aos grafos de tarefas que são utilizados para parametrizar e arquitetar, dinamicamente, uma rede neural de múltiplas camadas acopladas na saída do modelo. Os resultados apresentados mostram uma melhora de até 7% em relação ao estado da arte.

Já no ano de 2020 outro grupo de pesquisa (PLATANIOS et al., 2020) apresentou uma outra abordagem para atacar problemas dessa natureza. Eles argumentam que nenhum dos métodos anteriormente propostos resolvem efetivamente a agregação de rótulos na presença de alta subjetividade. Argumentam que, para se tornarem mais eficazes, esses métodos precisam como, por exemplo, usar metadados e outros tipos de informação que possam estar disponíveis sobre as instâncias de dados. Além do uso dos metadados, nesse método a rede neural utilizada possui mais camadas que a proposta por Ratner et al. (2019). Com essa abordagem, os autores conseguiram uma melhora na qualidade da rotulação gerada sobre os resultados do *Snorkel* e *MeTal* de até 25%. Mas, ainda segundo os autores, a melhora se fez no treinamento do modelo para prever vários rótulos relacionados simultaneamente. Para quando não é o caso, os desempenhos são próximos aos dos modelos anteriores.

2.2.1 Modelo de múltiplas tarefas

Um dos maiores desafios do Aprendizado Fracamento Supervisionado é combinar múltiplas fontes de rotulação para diversas tarefas. Com a construção de várias LFs, naturalmente, teremos vários geradores de rótulos. E este número por ser expressivo. Vejamos, voltando ao exemplo da Seção anterior (rotular comentários na avaliação dos produtos do *Mercado Livre*), teremos funções rotuladoras com a tarefa de rotular comentários como SPAM, e outras para etiquetar como *NÃO SPAM*. Frente a esse problema de múltiplas LFs e com tarefas distintas, a

[1] http://snorkel.stanford.edu
[2] github.com/HazyResearch/metal

solução encontrada é aplicar pequenas restrições a alguns rótulos com o objetivo de diminuir o ruído gerado com a combinação dessas tarefas (RATNER et al., 2020; RATNER et al., 2019; RATNER et al., 2019; RATNER et al., 2016).

Segundo Ratner et al. (2019), estas restrições podem ser representadas por meio de um grafo que indica as relações e restrições entre as tarefas. Esse esquema é referido na literatura como *Task Graph* (G_{task}).

Sendo $\mathbf{Y} = [Y_1, Y_2, ..., Y_t]^T$ um vetor de dimensão t com os valores amostrados a partir das funções λ_t; t é o número de tarefas a serem rotuladas; $Y_i = \{1, 2, ...k_t\}$ são os valores amostrados para a tarefa i e k_t é a cardinalidade da cada tarefa (RATNER et al., 2019). Através do grafo G_{task} é descrito a lógica do relacionamento entre as tarefas, definindo um conjunto de valores factíveis para um conjunto de rótulos γ, tendo $\mathbf{Y} \in \gamma$.

Para uma melhor compreensão, pode-se simular uma situação tomando o G_{task} da Figura 2 como referência. Dado um valor de entrada x_i do tipo texto, as tarefas a seguir têm as seguintes finalidades:

- Y_1: identificar se o texto refere-se a uma empresa ou a uma pessoa;
- Y_2: identificar o sexo de uma pessoa;
- Y_3: identificar se é uma empresa de tecnologia.

Dessa forma, os valores possíveis para Y_1 são *PESSOA* ou *EMPRESA*, para Y_2 são *MASCULINO* ou *FEMININO* ou *N/A* e para Y_3 são *SIM* ou *NÃO* ou *N/A*. As cardinalidades dessas tarefas são 2, 3 e 3, respectivamente.

Logo, e ainda supondo que x_i refere-se a descrição de uma pessoa do sexo feminino, as *Label Functions* poderão produzir a rotulação $\mathbf{Y} = [\textit{PESSOA}, \textit{FEMININO}, \textit{N/A}]^T$ sendo estes valores factíveis em γ. Caso esse processo retorne algo indicando que a saída é $\mathbf{Y} = [\textit{EMPRESA}, \textit{FEMININO}, \textit{SIM}]^T$, essa rotulação é ruidosa, pois se Y_1 = *EMPRESA* não faz sentido Y_2 ser *FEMININO* ou *MASCULINO*. Em outras palavras, este é um resultado não factível pois Y_2 é uma dependente condicional de Y_1.

O modelo intermediário de rotulação faz uso de probabilidade condicional para inferir os rótulos dos textos. As saídas das LFs são as variáveis observadas na inferência. O modelo final por sua, é constituído por uma Rede Neural de Múltiplas Camadas (MLP). A camada de saída dessa é composta por t neurônios. Cada neurônio tem uma função de ativação do tipo $softmax$ com k_t saídas (RATNER et al., 2019). A Figura 2 mostra o esquema desse método.

O número de neurônios na camada de entrada não segue uma métrica bem definida como a camada de saída. Ela está diretamente ligada com a dimensionalidade dos dados a serem analisados, normalmente gerados por uma função de *Embedding* (HAYKIN, 1999).

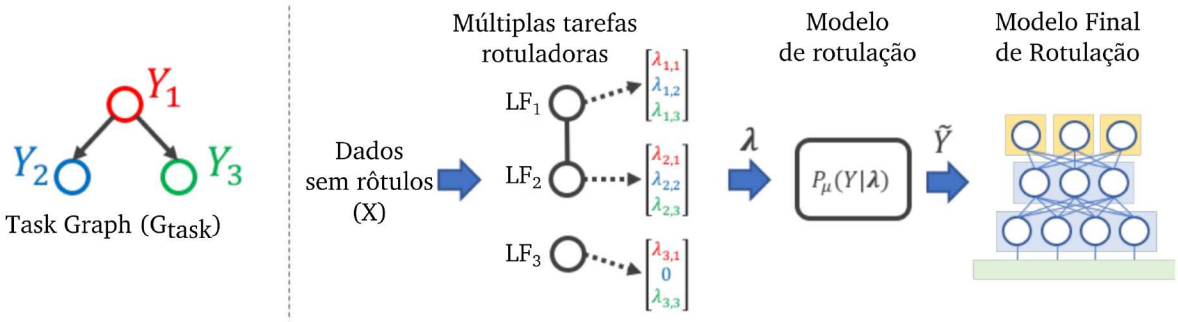

Figura 2 – Esquema do aprendizado de múltiplas tarefas. As restrições dos rótulos são definidas por um *Task Graph*. Um modelo de rotulação intermediário é construído com base nas *Label Functions*. A saída desse é dada como a saída esperada do modelo final, que por sua vez, e com respeito as restrições do G_{task}, rotulam os dados utilizando uma Rede Neural de Múltiplas Camadas. Fonte: Traduzido e adaptado de Ratner et al. (2019)

2.3 *Embedding* Textual

Embedding Textual (ou *Word Embendding*) é uma técnica de mapeamento de palavras em uma representação de valores numéricos. Essa transformação é necessária, pois muitos modelos de aprendizado de máquina conseguem entender apenas a representação vetorial numérica (TOUTANOVA et al., 2015; XIAO et al., 2017; GURUNATH et al., 2021).

Um exemplo intuitivo de como podemos representar um texto em vetores de valores numéricos, é através da técnica de transformação *One-Hot Enconding* (OKADA; OHZEKI; TAGUCHI, 2019). Ela consiste em atribuir valores binários para indicar a existência, ou não, de uma palavra que faz parte de um vocabulário em um determinado texto. Assim sendo, poderíamos representar vários textos com sequências de 0 e 1, conforme pode ser visto na Figura 3. Entretanto, essa abordagem pode se tornar computacionalmente impossível, pois quando se tratar de vocabulários extensos, a representação exigirá vetores com milhares de dimensões. Além disso, essa codificação não possui indicação das relações ocultas entre as palavras.

A *Embedding* Textual é diferente dessa abordagem ingênua, pois consegue representar frases e palavras em vetores numéricos e não binários. Além disso, fornece uma relação semântica oculta entre as palavras (BIRUNDA; DEVI, 2021). Por exemplo, as palavras “rei", “casa" e “rainha" são distribuídas no espaço vetorial n-dimensional. Cada dimensão daquele espaço vetorial tem uma escala, que normalmente vai de -1 a 1(KIM et al., 2019). Essa escala indica a relação dos termos com cada dimensão. Ou seja, se uma das n dimensões for, por exemplo, “Gênero", quanto mais próximo de -1, mais feminino, e quanto perto de 1, mais masculino é aquele termo. Outra dimensão pode ser “Realeza": quanto mais próximo de -1, aquele termo está relacionado com plebeu; mais perto de 1, mais relacionado com monarquia; e assim por diante. A Figura 4 ilustra como seria uma representação vetorial n-dimensional desse exemplo.

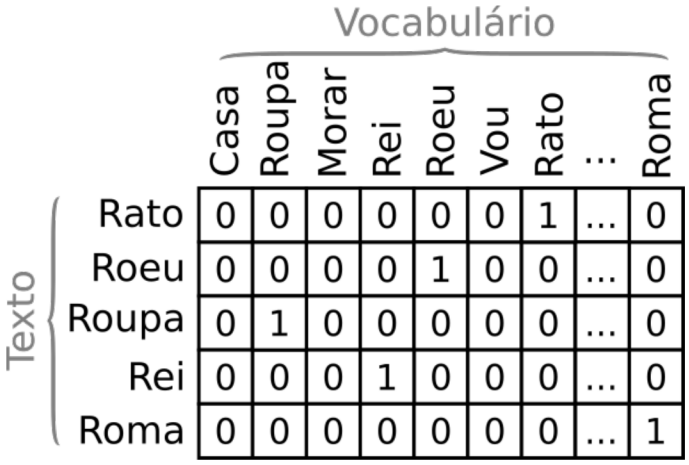

Figura 3 – Esquema de representação de textos através da transformação *One-Hot Enconding*.

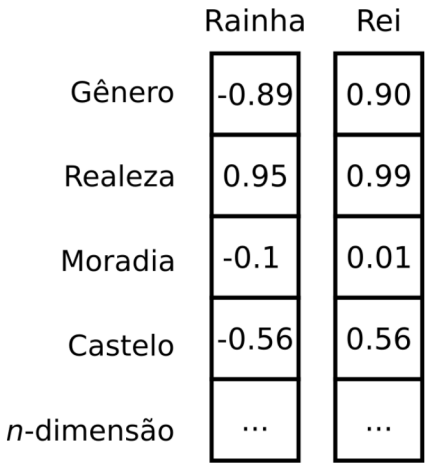

Figura 4 – Exemplo de representação de palavras em um vetor resultante de um *Embedding* Textual.

Para definir o valor de n e quais as dimensões mais relevantes dado o contexto da aplicação, há várias abordagens disponíveis na literatura. Entre elas, as baseadas em PCA (MAĆKIEWICZ; RATAJCZAK, 1993), Matriz de co-ocorrência (CHEN et al., 2020), Grafos (XIAO et al., 2017), Redes Neurais Artificiais (GURUNATH et al., 2021; KIM et al., 2019; LE; MIKOLOV, 2014), entre outras. Contudo, aqueles que vem apresentando resultados promissores nesse meio é o com uso de Redes Neurais (BIRUNDA; DEVI, 2021; GURUNATH et al., 2021; ZAPPONE et al., 2019).

Na literatura as três principais categorias de *Word Embeddings* são: *Traditional Word Embedding*, *Static Word Embeddings* e *Contextualized Word Embedding* (GURUNATH et al., 2021).

2.3.1 *Traditional Word Embedding*

Essa classificação de *Embedding Textual* é baseada na quantidade de vez que determinado termos aparecem em um conjunto de documentos, identificando a ocorrência e a co-ocorrência de cada palavras. São exemplos dessa categoria os métodos *Bag Of Words* (BoW) (SETHY; RAMABHADRAN, 2008; EL-DIN, 2016) e *Term Frequency-Inverse Document Frequency* (TFI-DF) (KADHIM et al., 2014).

O BoW, em sua tradução literal para o português, o "saco de palavras", recebe esse nome porque não considera a ordem ou a estrutura que as palavras estão dispostas no texto, mas sim se ela é parte ou ainda qual a frequência com que aparece nele. Ele é divido em dois passos: montagem do vocabulário e cálculo da ocorrência das palavras (SETHY; RAMABHADRAN, 2008).

O primeiro passo, também chamado como "Tokenização", consiste em identificar os termos ("Tokens") presentes nos textos ou documentos a serem processados, com isso, havendo a montagem do vocabulário. Agora no segundo passo, são contabilizadas as ocorrências das palavras dentro dos textos/documentos gerando o vetor final de representação. Como pode ser percebido, a abordagem do *Bag of Words* não faz nenhuma ponderação da importância dos termos no *corpus* (conjunto de documentos). Ela apenas contabiliza quantas vezes o termo ocorre e utiliza a concatenação desses valores para representar o texto. Diante desse problema, foi proposto na literatura o *Term Frequency-Inverse Document Frequency* (TFI-DF) (KADHIM et al., 2014).

Baseado no *Bag of Words*, o TFI-DF utiliza métodos estatísticos para medir o quão importante um termo é em um documento. Este valor é obtido multiplicando as métricas *Term Frequency* (TF) e *Inverse Document Frequency* (IDF), sendo essas definidas por

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{2.4}$$

e

$$IDF(t,D) = \ln \frac{N}{n_t} \tag{2.5}$$

onde, $f_{t,d}$ é o número ocorrências do termo t no documento d, $\sum_{t' \in d} f_{t',d}$ é o número de termos do documento d, D é o *corpus*, N é o número total de documentos e n_t o número de documentos que o termo t aparece. Logo:

$$\text{TF-IDF}(t,d,D) = TF(t,d) \times IDF(t,D). \tag{2.6}$$

Assim sendo, segundo o TF-IDF, quanto mais ocorrências uma palavra tiver em um documento, mais importante ela tende a ser. Porém, ela deve permanecer frequente ao longo dos

documentos que compõe o *corpus* para sua importância seja mantida. Caso isso não ocorra, o termo é penalizado (TRSTENJAK; MIKAC; DONKO, 2014).

2.3.2 *Static Word Embeddings*

Static Word Embeddings é uma abordagem que fornece probabilidade de ocorrência das palavras e mapeia todas elas em um vetor. Normalmente esses vetores são densos e tem uma dimensionalidade menor do que o tamanho do vocabulário. Os métodos que recebe esta classificação também tem a característica de não alterarem a suas representações em frases diferentes. Ou seja, palavras que aparecem em contextos semelhantes, terão representações vetoriais semelhantes (GURUNATH et al., 2021).

O Word2Vec, proposto por Mikolov et al. (2013), é um exemplo de método dessa classificação de *Embedding* Textual. Através de uma Rede Neural Artificial com duas arquiteturas distintas, esse método é capaz de predizer o termo faltante no meio de uma oração, ou ainda, dado a palavra, em qual contexto ela está inserida. As arquiteturas propostas para a RNA são *Continuous Bag of Words* (*CBOW*) e *Skip-Gram*. Elas são mostradas na Figura 5.

Na arquitetura *CBOW*, as representações de contexto (ou palavras ao redor) são combinadas para prever a palavra no meio. Enquanto no modelo *Skip-Gram*, a representação distribuída da palavra de entrada é usada para prever o contexto. Por exemplo, utilizando a arquitetura *CBOW*, dado como entrada no modelo a frase “Com o (?) vou lavar o cabelo"; o método é capaz de inferir que, dado o contexto, a palavra faltante é “*shampoo*"; se utilizada a arquitetura *Skip-Gram* apenas com a palavra “*shampoo*", a rede será capaz de inferir as palavras "lavar" e "cabelo"(GE; MOH, 2017).

A arquitetura *CBOW* tem como vetor de entrada as palavras que compõe a frase e delimitam o contexto. Para predizer o termo faltante, é calculado na saída a média da projeção de todos os termos presentes no contexto da camada de entrada. No modelo *Skip-Gram* acontece o processo inverso. Dada uma palavra no vetor de entrada, ela observa em quais os contextos aquela palavra tem maior projeção e indica o contexto (MIKOLOV et al., 2013).

2.3.3 *Contextualized Word Embedding*

Por fim, mas não menos importante, temos a classificação de métodos que atuam com a abordagem *Contextualized Word Embedding*. Essa é baseada na contextualização de uma palavra específica, onde palavras semelhantes terão representações contrastantes. Tais representações irão variar dinamicamente de acordo o contexto que elas estão inseridas. Ou seja, essa representação pode capturar várias propriedades de sintaxes e semânticas de palavras em diversos contextos linguísticos (LIU; KUSNER; BLUNSOM, 2020). Alguns métodos dessa classificação que estão

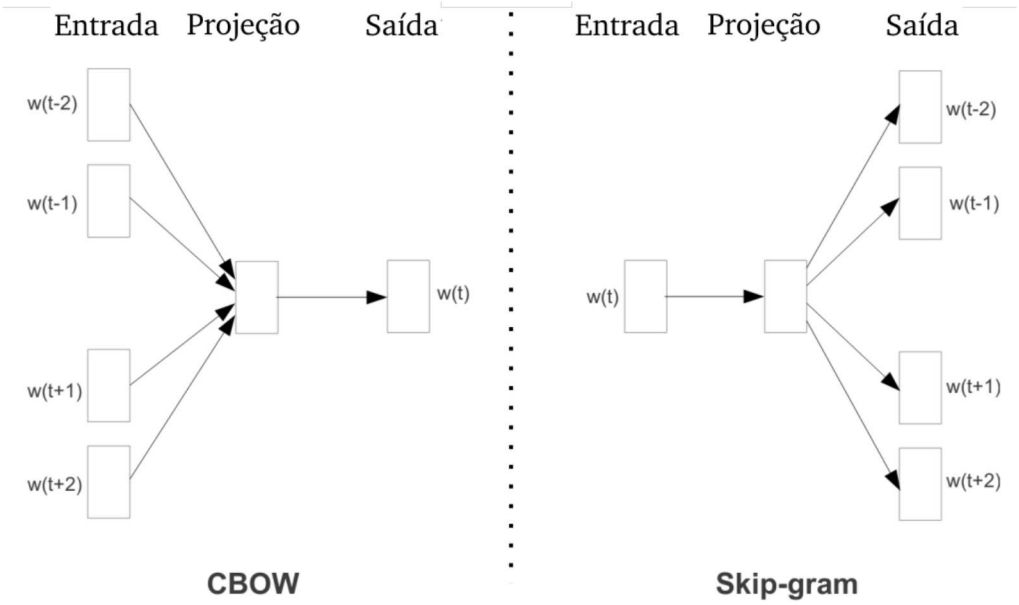

Figura 5 – Arquiteturas *Continuous Bag of Words* (*CBOW*) e *Skip-Gram* do Word2Vec. Fonte: Adaptado de Mikolov et al. (2013).

disponíveis na literatura são o ELMo (SARZYNSKA-WAWER et al., 2021), GPT-2 (RADFORD et al., 2019) e BERT (DEVLIN et al., 2018).

Todos eles utilizam Redes Neurais Artificiais com múltiplas camadas para realizar as transformações e projeções dos termos. Contudo, para entendê-los, é preciso que antes se saiba dois conceitos importantes: *Attention* e *Transformers*.

A técnica *Attention* possibilita que as palavras tenham um vetor de pesos com percentuais de associação que indicam a proximidade dos termos e assim ampliando o alcance do contexto (LUONG; PHAM; MANNING, 2015). Por exemplo, na frase “Maria mudou-se para uma casa maior, pois ela [...]", a palavra “ela" poderia se referir tanto a “Maria" quanto a “casa". Contudo, por está mais próximo do termo “casa" outros métodos convencionais de *Embedding* irão associar apenas ao contexto da casa. Através da técnica de *Attention*, é possível dar pesos ponderados pela distância até os termos “Maria" e “casa".

Transformer é uma arquitetura de RNA baseadas em mecanismos de *Attention* (WU et al., 2016). Sua arquitetura é composta por codificadores e decodificadores, onde cada codificador possui uma camada de *Attention* e uma camada de RNA *Feed Forward*; e cada decodificador possui as mesmas camadas, porém no meio delas possui uma outra camada de *Attention* que ajuda o decodificador a focar em partes relevantes do vetor de entrada.

O *Bidirectional Encoder Representations from Transformers* (BERT), proposto por (DEVLIN et al., 2018), é um dos modelos baseados em *Transformers* mais populares atualmente. A sua popularidade se deu por ele utilizar apenas codificadores da arquitetura *Transformer* e consegui processar sequências de textos em ambos sentidos (direita para esquerda e esquerda para direita), diferentemente da proposto original dessa arquitetura que processa da esquerda

para a direita. Daí a origem de seu nome, que em uma tradução livre para o português significa "Representações de Codificador Bidirecionais a partir de *Transformers*". Em outras palavras, o BERT é capaz de produzir probabilidades condicionais para um elemento desconhecido no meio de uma sequência de entrada com base nos elementos que o antecedem e o sucedem. Essa abordagem é conhecida como MLM (acrônimo do termo em inglês, *Masked Language Modeling*), e rendeu ao modelo grande sucesso, além de diversos resultados estado-da-arte em tarefas de Processamento de Linguagem Natural na época de publicação (BIRUNDA; DEVI, 2021).

2.4 Explicabilidade com *SHAP*

A Explicabilidade da tomada de decisão de modelos é especialmente importante em aplicações na área médica ou carros autônomos, onde a confiança do decisor nas características corretas deve ser garantidas (BOJARSKI et al., 2017). Por essa razão, aplicações de modelos lineares ou árvores de decisão simples tem muita aplicação nessas áreas, pois oferecem explicação intuitiva e de fácil interpretação. Contudo, não são os modelos mais acurados disponíveis na literatura. Os modelos de aprendizado profundo são que apresentam melhores resultados, entretanto, com baixa ou complexa interpretabilidade (MONTAVON; SAMEK; MÜLLER, 2018).

Modelos mais interpretáveis são mais confiáveis, fáceis de depurar, ajuda a promover melhorias no processo de coleta de dados e geração de atributos, além de dar melhor apoio e suporte a tomadas de decisão (MCGOVERN et al., 2019). Segundo Montavon, Samek e Müller (2018), há dois conceitos importantes na área de explicabilidade de modelos que muitas vezes são confundidos entre si: Interpretação e Explicação. Interpretação é o mapeamento de um conceito abstrato em um domínio que o humano pode entender (por exemplo: imagens, textos, gráficos, entre outros.); Explicação é a coleção de características do domínio interpretável, que contribuíram para um determinado exemplo poder produzir uma decisão.

Entre as abordagens disponíveis na literatura que tem por objetivo a explicação do modelo preditor, temos o SHAP (*SHapley Additive exPlanations*) (LUNDBERG; LEE, 2017; AAS; JULLUM; LØLAND, 2019). Uma abordagem que é baseada em *Shapley Values* (SV), que foi proposto por Shapley (1953) (Prêmio Nobel de Economia no ano de 2012)[3], no contexto da teoria dos jogos.

Dado um jogo colaborativo qualquer, os *Shapley Values* consistem em quantificar a contribuição marginal de cada jogador para o resultado final alcançado pela equipe. No contexto da aplicação dessa teoria para a explicabilidade de modelos computacionais de predição, os "jogadores" são as variáveis de entrada do modelo e o "resultado alcançado" pela equipe é a saída

[3] https://www.nobelprize.org/prizes/economic-sciences/2012/summary/

do modelo. De forma mais intuitiva, a contribuição de cada jogador é a média de sua contribuição em relação a todas as permutações de times que não o incluem.

Para um melhor entendimento desta teoria, supomos que uma equipe de um jogo qualquer, que compete em duplas, é composta pelos jogadores A, B e C. Assumindo que todos esses já atuaram em alguma partida, podemos calcular a contribuição marginal de cada um deles. Com vistas à didática do exemplo, vamos calcular a contribuição do jogador A. Sabe-se que todas as vezes que a equipe atuou com os jogadores B e C juntos, e sem a participação do jogador A, obtiveram média de 80 pontos; quando o jogador A atuou com B, e sem C, obtiveram média de 110 pontos; quando C e A aturam juntos e sem B, não pontuaram. Logo, a contribuição marginal de A para a equipe será dada pela soma de todas médias obtidas por ele quando participou das partidas, menos a média dos pontos obtidos pelo time sem a participação dele. Ou seja, 110 - 80 = 30;

Em termos matemáticos, considerando um modelo com M variáveis de entrada, cujo objetivo seja maximizar uma métrica de resposta e dado que $S \subseteq M = \{1, \ldots, m\}$ seja um subconjunto das variáveis $|S|$ (também chamada de coalizão ou aliança), então temos:

$$\sum_{S \subseteq M \{j\}} \frac{|S|!(M - |S| - 1)!}{M!}(v(S \cup \{j\}) - v(S)), \quad j = 1, \ldots, m \tag{2.7}$$

onde $(v(S \cup \{j\}) - v(S))$ é a contribuição marginal da variável de entrada j quando é adicionada ao subconjunto S. A média ponderada de todos os subconjuntos de variáveis S que não contêm a variável j fornece sua contribuição (AAS; JULLUM; LØLAND, 2019; SHAPLEY, 1953).

Shapley (1953) estabelece 4 propriedades para que haja o cálculo de uma contribuição justa. São elas:

1. **Eficiência**: a soma das contribuições é igual a diferença da predição com a predição média;
2. **Simetria**: se duas variáveis contribuem igualmente em todas as alianças, o *Shapley Value* delas será o mesmo;
3. **Neutralidade**: se uma variável não contribui na predição dentro de qualquer aliança, o *Shapley Value* deve ser zero;
4. **Aditividade**: para jogos com contribuições combinada, pode-se somar os *Shapley Values*;

3 Materiais e Métodos

Este capítulo descreve os materiais e os métodos utilizados no desenvolvimento de uma *framework* que realiza a fusão de dados estruturados e não estruturados para construir modelos de predição de diagnósticos de três doenças: Fibrilação Auricular, Enfermidade Coronária e Apneia do Sono.

A Figura 6 mostra uma visão geral da *framework* proposta. Primeiramente, os EHRs são filtrados e agrupados para compor os conjuntos de dados a serem aprendidos pelos modelos preditivos. Nesse processo, os registros dos paciente são concatenados para formar um único vetor no conjunto de aprendizagem. Isso é realizado para cada doença de interesse, gerando um conjunto de aprendizado para cada doença: EC, AS ou FA.

3.1 Base de Dados e pré-processamento

A base de dados utilizada neste estudo possui registros de mais de 22.000 indivíduos. Todos eles pacientes do *Instituto del Corazón de Bucaramanga*, na Colômbia. Cada tupla (registro) desta base corresponde a uma consulta médica e contém as seguintes características: número de identificação do paciente (ID), idade, sexo, peso, altura, data, classificação internacional de doenças em sua 10ª revisão (ICD-10th) das doenças diagnosticadas, e descrições textuais sobre histórico de saúde familiar, tóxico (álcool, tabagismo ou abuso de substâncias), histórico cirúrgico, alergias, histórico patológico, além de um campo denominado "geral" que contém descrições textuais de sintomas, exames e procedimentos do paciente. Em alguns casos, é possível observar mais de uma doença diagnosticada na mesma data ou ainda dados repetidos.

Frente a isso, foi realizada uma limpeza nos dados eliminando registros redundantes e não coesos (valores fora dos parâmetros normais) com a finalidade de remover *outliers*. Após este pré-processamento, obteve-se um total de mais de +560.000 registros de 19.818 pacientes, com mais de 150 doenças diagnosticadas.

No grupo de EC, foram incluídos os pacientes que apresentavam obstrução de pelo menos 50% de uma artéria coronária principal, diagnosticada por cinecoronariografia (SOLIMENE; RAMIRES, 2003). No grupo de FA, foram incluídos pacientes com diagnóstico confirmado por eletrocardiograma de 12 variações. A AS foi diagnosticada por polissonografia e incluiu pacientes com IAH (Índice de Apneia-Hipopnéia) > 5 associado a uma dessaturação de oxi-hemoglobina de 4 (MARTINS; TUFIK; MOURA, 2007). 90% dos pacientes incluídos com diagnóstico de AS tinham Apneia Obstrutiva do Sono (AOS) e 10% tinham apneia do sono central. A AS foi classificada em três níveis: leve, quando o IAH está entre 5 e 15, moderada para IAH entre 15 até 30 e grave quando IAH é maior que 30.

A Seção 3.2 detalha a preparação dos conjuntos de dados para aprendizagem e construção do modelo de diagnóstico. Os dados textuais não estruturados presentes nos conjuntos de aprendizagem são pré-processados para gerar suas variáveis categóricas correspondentes. Este processo, detalhado na Seção 3.2.1, conta com uma técnica de *Embedding* textual e um modelo de rotulação baseado em Aprendizado Fracamente Supervisionado (RATNER et al., 2016). Simultaneamente, os dados estruturados são pré-processados para lidar com valores ausentes. Operações estatísticas são aplicadas para inferir estes dados e relacionadas aos registros clínicos (por exemplo: idade, peso, altura e outros). Além disso, os registros de doenças de um determinado paciente, previamente diagnosticadas, são agrupados no nível categórico da ICD-10th (por exemplo: Z33) e convertidos em variáveis *dummy* (booleanas). A Seção 3.2.2 apresenta os detalhes do processamento de dados estruturados.

Figura 6 – Diagrama com a visão geral da *framework* proposta.

A fusão dos dados estruturados e não estruturados resulta em variáveis de entrada numéricas e categóricas que descrevem a condição clínica do paciente e o histórico médico, pessoal e familiar. Os modelos então são treinados para prever o diagnóstico das doenças de interesse. Para o treinamento de aprendizagem supervisionada foi utilizado o método XGBoost (CHEN et al., 2015). Sobre este modelo treinado, usamos o método SHAP (AAS; JULLUM; LØLAND, 2019) para explicar a sua tomada de decisão. Detalhes sobre o XGBoost e SHAP são fornecidos nas Seções 3.3, e 3.5, respectivamente.

3.2 Preparação dos conjuntos de dados

Os modelos preditivos propostos neste trabalho consideram apenas informações pré-diagnóstico de uma determinada doença de interesse (AS, EC ou FA). Assim, as datas das consultas médicas com seus respectivos diagnósticos são consideradas ao montar os conjuntos de dados de aprendizagem. Para cada paciente, o vetor de características que é usado como entrada para o modelo preditivo é formado pela concatenação de todos os registros antes da data do primeiro diagnóstico da doença alvo. A Figura 7 ilustra uma sequência de consultas de um paciente arbitrário que foi diagnosticado com uma das doenças de interesse (por exemplo, FA) no instante de tempo T4. O vetor de características representando este paciente para o modelo que prevê FA é uma agregação de dados clínicos, informações não estruturadas e histórico de diagnósticos que ocorreram até T4. Informações adicionais ocorridas após este instante são descartadas.

Figura 7 – *Timeline* ilustrando a sequência de consultas e diagnósticos de um paciente qualquer.

A rotulação do conjunto de aprendizado ocorre concomitantemente com a concatenação dos registros para formar os vetores de características. Os pacientes que durante sua linha do tempo foram diagnosticados com a doença de interesse são considerados como pertencentes à classe positiva (rótulo $Y = 1$), e os pacientes que não foram diagnosticados, pertencerão à classe negativa (rótulo $Y = -1$).

Formalmente, cada modelo preditivo estima a probabilidade condicional $P(Y = 1|\lambda = \mathbf{\Lambda})$, onde $\mathbf{\Lambda}$ é um vetor de características que representa um paciente. Como há uma disparidade na frequência de ocorrência das doenças AS, EC e FA, os conjuntos de aprendizagem são desbalanceados, com maior predominância de diagnósticos negativos nos conjuntos de dados, principalmente para AS. As porcentagens de desequilíbrio para cada doença nos conjuntos de aprendizagem são: $7\%/93\%$ para FA, $31\%/69\%$ para EC e $3\%/97\%$ para AS.

3.2.1 Dados não estruturados

Cada registro no banco de dados apresenta diversos campos textuais referentes ao histórico do paciente e histórico de saúde da família inserido por um médico durante uma consulta. O desafio é transformar esses campos textuais em variáveis de entrada relevantes para o diagnóstico de doenças. Não é incomum, por exemplo, encontrar textos conflitantes sobre a toxicidade de um paciente. Ao acompanhar os registros de pacientes na base de dados, encontra-se expressões referindo-se ao mesmo indivíduo e colhidas em momentos diferentes, contudo conflitantes, como “não referido", “fumo pesado", “não faz uso de fumo" e “álcool casual". Outro complicador é que o paciente pode omitir informações relevantes (sobre o uso de álcool, tabagismo ou substâncias ilícitas, por exemplo) por vergonha, culpa ou motivos pessoais diversos.

Para superar a ambiguidade e a falta de padronização presente nos registros textuais, empregamos um método que consiste em *Embedding* textual via Doc2Vec (LE; MIKOLOV, 2014; ZHONG; GAO; YI, 2018; KIM et al., 2019) e Aprendizado Fracamente Supervisionado (RATNER et al., 2019). A entrada do método consiste em dados textuais brutos, por exemplo,

"um maço de cigarros por dia, não referido, tabagismo intenso", relacionado a um fator específico, como o histórico de tabagismo do paciente, e a saída consiste em um rótulo (ou categoria) indicando neste caso, se ou não o paciente usou tal substância tóxica. Essa transformação de uma sequência de texto em uma ou mais variáveis categóricas ocorre devido ao uso de heurísticas baseadas em regras, conhecidas na literatura de Aprendizado Fracamente Supervisionado como funções de rotulação (do inglês, *Label Functions* - LFs).

Os fatores relevantes que o método deve identificar foram indicados por um especialista e são mostrados nas Tabelas 1, 2 e 3. Há um total de 61 perguntas relacionadas a patologias, cirurgia, família, tóxico e alergia. Usando a framework Metal[1], funções de rotulação foram programadas para cada fator. A ideia básica das LFs é injetar informações de domínio em um nível superior para gerar rótulos automaticamente sem a necessidade de rotular manualmente milhares de dados. Há uma variedade de LFs programáticas que variam de reconhecimento de padrões a pesquisas de palavras-chave (RATNER et al., 2019). Voltando ao nosso exemplo de "fumante", cada função de rotulação recebe como entrada uma sequencia de texto, como *"um maço de cigarros por dia, não referido, fumo pesado, etc."*, e então produz uma saída com um dos rótulos *True* ou *False* ou *Abstain* (abstenção). Uma LF típica pesquisará por palavras-chave como "maço de cigarros" ou "fumante"para assinalar o rótulo *True* caso alguma delas estejam contidas na sequência do texto.

Depois de aplicar as LFs em todas as sequencias de texto, o resultado é uma matriz de etiquetas em que cada linha corresponde a uma sequencia de texto e cada coluna corresponde a saída de uma respectiva função de rotulagem. Como as LFs têm precisões e correlações desconhecidas, seus rótulos de saída podem se sobrepor e conflitar. Um modelo de rótulo, chamado de *Label Model*, é então usado para converter os rótulos produzidos pelas LFs em um único rótulo por sequência de texto (RATNER et al., 2019). Por fim, as sequências de texto rotuladas e resultantes são usadas para treinar uma rede neural que é conhecida como modelo final (do inglês, *End Model*). Espera-se que este modelo final possa generalizar eventuais lacunas de cobertura das funções de rotulação e do modelo de rótulo. O modelo final escolhido foi uma Rede Neural de Múltiplas Camadas (MLP) com 3 camadas ocultas com 100, 50 e 10 neurônios, respectivamente, e uma camada de saída com 2 neurônios. Esta arquitetura é recomendada em (RATNER et al., 2019). A MLP foi implementada via módulo SkLearn do Python[2]. Para cada tópico (agrupadores das questões listadas nas Tabelas 1, 2 e 3) haverá um modelo final.

O vector de entrada da rede neural tem dimensão igual a 100 e corresponde ao comprimento do vetor de saída do Doc2Vec (ZHONG; GAO; YI, 2018). A técnica do Doc2Vec associa cada sequência de texto com um vetor de características usando similaridade por cosseno. Ele identifica a semelhança entre documentos com base nas frases ou palavras que os compõem. O conjunto de treinamento para o modelo final é composto pelas amostras baseadas no *Embedding*

[1] github.com/HazyResearch/metal
[2] scikit-learn.org/stable/

textual e seus rótulos correspondentes gerados através do modelo de rotulação. Para avaliar a precisão dos modelos finais, dependemos de um conjunto de dados construído através de rotulagem manual feita por um profissional da área médica.

Para que isso fosse possível, foi desenvolvido em linguagem PHP[3] uma ferramenta que permite o especialista classificar relatos da base de dados. Nela, o usuário, a partir dos relatos dos pacientes registradas pelo médico, responde as perguntas listadas nas Tabelas 1, 2 e 3 com *Sim*, *Não* ou *Inconclusivo*. Essas respostas são exclusivas entre si. A ferramenta realiza uma amostragem randomizada na base de dados por grupo de interesse. Nessa tarefa são descartadas os relatos já rotulados. A rotulação feita pelo especialista é comparada com a feita pelo método Aprendizado Fracamente Supervisionado, definindo assim a sua assertividade. Os resultados dessa validação são apresentados na Seção 4.1.

A Figura 8 mostra uma tela da ferramenta desenvolvida e utilizada. As questões estão no idioma Espanhol porquê os dados e o especialista são nativos deste idioma.

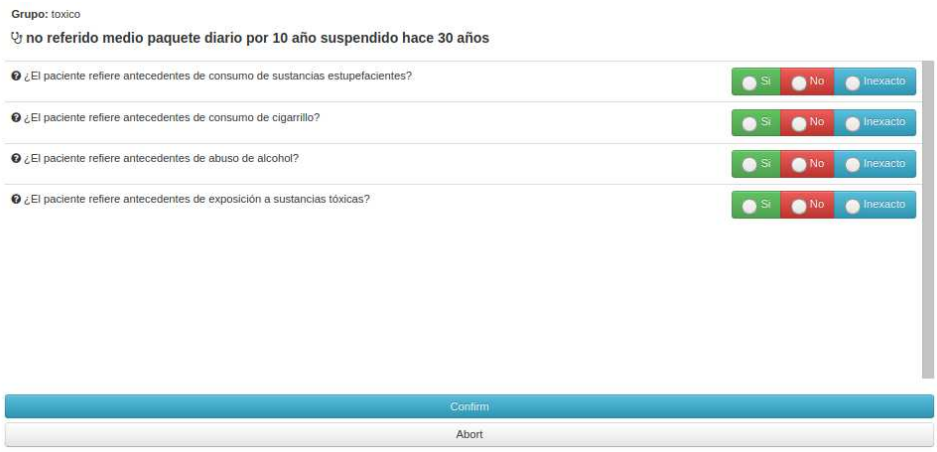

Figura 8 – Tela da aplicação desenvolvida para a rotulagem dos dados não estruturados.

A Figura 9 apresenta o esquema de geração de variáveis de entrada a partir de dados textuais. Um especialista define as perguntas relevantes que o método de Supervisão Fraca deve identificar (por exemplo, sobre tabagismo, álcool, hipertensão, etc.). As funções de rotulação são construídas para cada fator com informações do domínio específico. Os modelos de rotulação (*Label Models*) são aplicadas a todas as sequências de texto para gerar os rótulos de cada fator. As sequências de texto são transformadas em vetor de características via Doc2Vec. O modelo de rede neural é treinado a partir do conjunto de treinamento rotulado (vetores de características e rótulos) visando generalizar o conhecimento do modelo de rotulação.

[3] www.php.net

Tabela 1 – Perguntas sobre Família definidas por um especialista e respondidas via método de Aprendizado Fracamente Supervisionado.

ID	Pergunta
F87	Paciente relata histórico de pai com insuficiência mitral?
F86	Paciente relata histórico de mãe com insuficiência mitral?
F85	Paciente relata histórico de pai com doença renal crônica?
F84	Paciente relata histórico de mãe com doença renal crônica?
F83	Paciente encaminhou pai com apneia do sono?
F82	Paciente encaminhou mãe com Apneia do Sono?
F81	Paciente refere pai com Fibrilação Atrial?
F80	Paciente encaminhou mãe com Fibrilação Atrial?
F79	Paciente refere pai com infarto agudo do miocárdio?
F78	Paciente refere mãe com infarto agudo do miocárdio?
F77	Paciente refere pai com hipertensão?
F76	Paciente encaminhou mãe com hipertensão?
F62	Paciente refere histórico de tios com alguma doença?
F61	Paciente refere histórico de avô com alguma doença?
F60	Paciente refere histórico de avó com alguma doença?
F59	Paciente refere histórico de irmãos com alguma doença?
F58	Paciente refere histórico de pai com doença?
F57	Paciente refere histórico de mãe com alguma doença?

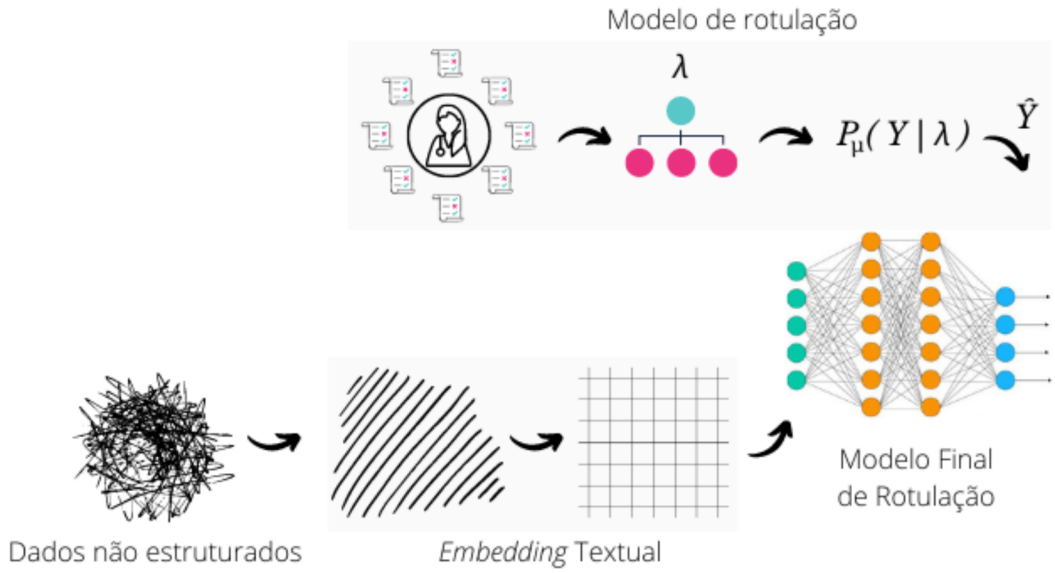

Figura 9 – Esquema para a geração das variáveis de entrada dos dados textuais.

3.2.2 Dados estruturados

Os dados estruturados referem-se a idade, sexo, peso, altura, código internacional das doenças pré-diagnosticadas e a identificação do paciente. Como alguns registros do banco de dados original apresentaram inconsistência ou falta de dados, a imputação de novos valores foi aplicada com base em outros registros do paciente. Por exemplo, se um paciente em sua quinta

Tabela 2 – Perguntas sobre Patologias definidas por um especialista e respondidas via método de Aprendizado Fracamente Supervisionado.

ID	Pergunta
P33	Paciente relata histórico de câncer?
P32	Paciente relata histórico de morte súbita e reanimação?
P31	Paciente relata histórico de reabilitação cardíaca?
P30	Paciente relata histórico de convulsões?
P29	Paciente é obeso?
P25	Paciente relata histórico de dislipidemia?
P24	Paciente relata histórico de doença cerebrovascular?
P23	Paciente relata histórico de vertigem?
P22	Paciente relata histórico de miocardite?
P21	Paciente relata histórico de infarto agudo do miocárdio sem doença coronariana?
P20	Paciente relata histórico de cardiopatia congênita?
P19	Paciente relata histórico de doença valvar?
P17	Paciente relata dispneia?
P13	Paciente relata histórico de doenças neurológicas?
P12	Paciente relata histórico de doenças psiquiátricas?
P11	Paciente relata histórico de distúrbios de coagulação?
P10	Paciente relata histórico de doenças autoimunes?
P9	Paciente relata histórico de doenças pulmonares?
P8	Paciente relata histórico de tratamento medicamentoso para alguma doença?
P7	Paciente relata histórico de tratamento cirúrgico para alguma doença?
P6	Paciente relata histórico de apnéia do sono?
P5	Paciente relata histórico de diabetes?
P4	Paciente relata histórico de fibrilação atrial?
P3	Paciente relata histórico de doença coronariana?
P2	Paciente relata histórico de hipotireoidismo?
P1	Paciente relata histórico de hipertensão arterial?

consulta médica não tivesse a idade informada, com base nos dados anteriores, sua idade foi imputada. Funções estatísticas, como moda e mediana, também foram aplicadas para agrupar diferentes valores de altura e valores de peso que aparecem em vários registros para o mesmo paciente. Em seguida, valores de peso e altura foram transformados em índice de massa corporal (IMC). A variável idade foi agrupada pela década de nascimento.

Os registros da ICD-10th foram processados de acordo com o nível hierárquico proposto pela OMS [4]. Em seguida, estes registros foram pivotados e convertidos em variáveis binárias com valores 0 indicando “não diagnosticado"e 1 “diagnosticado".

[4] https://www.who.int/standards/classifications/classification-of-diseases

Tabela 3 – Perguntas sobre Cirurgia, Alergias e Tóxico definidas por um especialista e respondidas via método de Aprendizado Fracamente Supervisionado.

	ID	Pergunta
Cirurgia	C75	Paciente relata cirurgia pancreática?
	C74	Paciente refere cirurgia bariátrica?
	C56	Paciente relata histórico de cirurgia por traumatismo craniano?
	C55	Paciente relata histórico de cirurgia pulmonar?
	C54	Paciente relata histórico de cirurgia para câncer?
	C53	Paciente relata histórico de cirurgia cardiovascular?
	C52	Paciente relata histórico de cirurgia cardíaca?
Alergias	A72	Paciente relata histórico de algum tipo de alergia?
	A71	Paciente relata histórico de alergia alimentar?
	A70	Paciente relata histórico de alergia a medicamentos?
	A69	Paciente relata histórico de dermatite?
	A68	Paciente relata histórico de rinite alérgica?
	A67	Paciente relata histórico de asma?
Tóxico	T66	Paciente relata histórico de exposição a substâncias tóxicas?
	T65	Paciente relata um histórico de abuso de álcool?
	T64	Paciente relata histórico de consumo de cigarro?
	T63	Paciente relata histórico de consumo de substâncias entorpecentes?

3.3 Modelo e métricas de avaliação

A junção de dados estruturados e não estruturados resultou em um total de 701 variáveis candidatas (todas numéricas e categóricas) a serem usadas como entradas dos modelos preditivos baseados em XGBoost (CHEN et al., 2015) para o diagnóstico de AS, EC e FA. Para avaliar o impacto das variáveis sobre a previsão final, elas foram agrupadas nas seguintes categorias: (i) físicos, (ii) histórico de diagnósticos e, (iii) dados textuais (não estruturados). Consequentemente, os modelos XGBoost foram avaliados sobre 7 diferentes configurações de variáveis de entrada, representando as combinações dos grupos de variáveis, como pode ser visto na Tabela 4. Os significados das siglas “F",“H"e “N"são dados físicos, histórico de diagnósticos e dados não estruturados, respectivamente.

Para avaliar a capacidade preditiva dos modelos, os conjuntos de dados foram divididos de maneira estratificada em subconjuntos de treinamento/teste em uma relação $70/30$. A seleção do modelo (ajustes nos hiper-parâmetros, taxa de aprendizagem e número de folhas do XGBoost) foi realizada usando a técnica de *grid-search* com validação cruzada *K-Fold* e valor de K=5. Para lidar com o desbalanceamento das classes, cada subdivisão de treinamento foi sinteticamente equilibrada através da técnica de super-amostragem SMOTE (WANG et al., 2006). Como métrica de avaliação, foi escolhida a área sob a curva ROC (AUC), sensibilidade e especificidade, conforme recomendado na literatura (LALKHEN; MCCLUSKEY, 2008). Essas métricas foram calculadas sobre os subconjuntos de teste de maneira independente e de acordo as configurações listadas na Tabela 4.

Tabela 4 – Configurações das variáveis de entrada dos modelos XGBoost

Grupo	Dados Físicos	Histórico de diagnósticos	Dados não estruturados
H	Não	Sim	Não
N	Não	Não	Sim
F	Sim	Não	Não
H+N	Não	Sim	Sim
F+N	Sim	Não	Sim
F+H	Sim	Sim	Não
F+H+N	Sim	Sim	Sim

3.4 Interpretação da decisão do modelo

Depois de obter os resultados gerados pelo XGBoost, a técnica SHAP (do termo em inglês, *SHapley Additive exPlanations*) (LUNDBERG; LEE, 2017), baseada na teoria dos jogos, foi aplicada para interpretar as decisões tomadas pelo modelo. Com essa técnica, é possível avaliar quanto cada variável de entrada contribui, positivamente ou negativamente, para o resultado de uma previsão nas perspectivas macro (global) e micro (local). Em resumo, o SHAP calcula a importância de uma variável de entrada comparando o que o modelo prevê com e sem essa variável. Ele compara o valor médio do SHAP do conjunto de dados com os valores individuais, mostrando a contribuição marginal de cada variável na previsão de uma determinada amostra de entrada.

3.5 Software de suporte a decisão

Após desenvolvimento das técnicas e modelos de diagnóstico apresentados até aqui, essas tecnologias foram incorporadas em um software desenvolvido em linguagem de programação Python 3.6. A arquitetura de desenvolvimento é baseada em requisição e resposta sob um protocolo de transferência de hipertexto (do inglês, *Hypertext Transfer Protocol* - HTTP). O banco de dados adotado é o MySQL 5.7 e o framework de desenvolvimento é o Django 3.0.6.

A ferramenta é divida em módulos e cada módulo tem uma finalidade definida. O módulo de login gerencia a autenticação e níveis de autorização de acesso a todos os módulos do sistema. O de importação permite a que a instituição de saúde importe os dados de seus pacientes a partir de arquivos do tipo *xls* ou *xlsx*. O módulo histórico do paciente permite que o usuário, através do documento de identificação do paciente, possa acessar todo o histórico clínico dele através de gráficos de *timeline*, respostas geradas pelo Aprendizado Fracamente Supervisionado e detalhamento do prontuário em cada consulta.

O módulo de importação recebe um arquivo no formato *xls* ou *xlsx*, faz *upload* dele e processa-o validando a estrutura e composição dos dados no arquivo. Caso tudo esteja de acordo, eles são persistidos num banco de dados. Caso não, é indicado num *log* quais ações

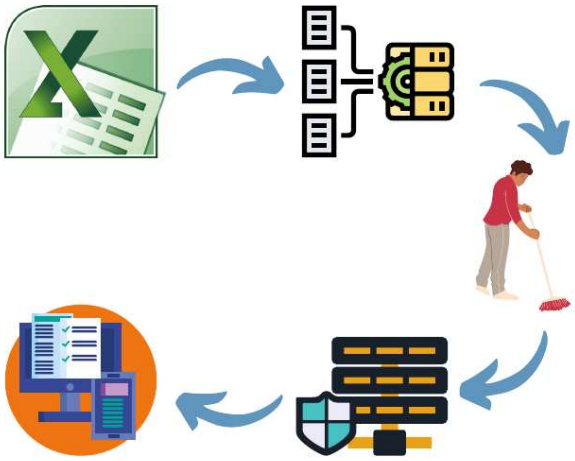

Figura 10 – Diagrama do esquema de funcionamento do módulo de importação de arquivos do *software*.

são necessárias para a importação ocorrer com sucesso. Caso a validação dos dados ocorra com sucesso, é realizada uma limpeza nos dados removendo eventuais dados repetidos ou incompletos. Por fim, estes são submetidos aos modelos de inteligência desenvolvidos para gerar o conhecimento e predição de diagnósticos dos indivíduos via sistema de suporte a decisão. A Figura 10 mostra o esquema do módulo de importação até a informação ficar disponível para o usuário da ferramenta.

4 Resultados

Nesta seção apresentamos os resultados do método de Aprendizado Fracamente Supervisionado aplicado aos dados não estruturados dos pacientes; os modelos de diagnóstico das doenças AS, EC e FA usando diferentes configurações de entrada de dados; a interpretação das saídas desses modelos com base na abordagem SHAP; e o *software* desenvolvido para suporte à decisão médica.

4.1 Rotulação com Aprendizado Fracamente Supervisionado

O modelo de rotulação deve ser capaz de imitar as respostas dos especialistas. Como a rotulação manual é uma tarefa muito cara, apenas uma pequena amostra de dados, correspondente àqueles que foram respondidos pelo especialista, foi usada para avaliar o modelo de rotulação. Considerando todos os registros não estruturados do banco de dados, os especialistas classificaram aproximadamente 30% de cada um dos tópicos (Família, Patologias, Cirurgia, Alergias e Tóxico).

Tabela 5 – *F1-score* (F1) alcançados pelo método de Aprendizado Fracamente Supervisionado com a abordagem de múltiplas tarefas.

Tópico	**ID**	**F1**	**ID**	**F1**	**ID**	**F1**
Família	F58, F77, F79, F81, F83, F85 e F87	0.904	F59	0.935	F60	0.967
	F57, F76, F78, F80, F82, F84 e F86	0.904	F61	0.967	F62	0.967
Patologias	P1	0.784	P2	0.338	P3	0.603
	P4	0.460	P5	0.568	P6	0.569
	P7	0.630	P8	0.627	P9	0.669
	P10	0.638	P11	0.638	P12	0.649
	P13	0.629	P17	0.649	P19	0.645
	P20	0.654	P21	0.695	P22	0.655
	P30	0.637	P23	0.659	P24	0.655
	P25	0.546	P29	0.618	P31	0.654
	P32	0.659	P33	0.637		
Cirurgia	C53 e C52	0.966	C54	0.980	C55	0.993
	C56	0.993	C74	0.993	C75	-
Alergias	A67, A68, A69, A70, A71 e A72	0.996				
Tóxico	T66, T65, T64 e T63	0.831				

A Tabela 5 mostra os resultados obtidos pela técnica. O método foi capaz de reproduzir a rotulação feita pelo especialista com uma precisão que variou de $33,8\%$ a $99,6\%$. Algumas questões são agrupadas no tópico porquê existe uma relação de dependência/restrição entre elas.

Tabela 6 – Perguntas sobre Família definidas por um especialista e respondidas via método de Aprendizado Fracamente Supervisionado.

ID	Pergunta
F87	Paciente relata histórico de pai com insuficiência mitral?
F86	Paciente relata histórico de mãe com insuficiência mitral?
F85	Paciente relata histórico de pai com doença renal crônica?
F84	Paciente relata histórico de mãe com doença renal crônica?
F83	Paciente encaminhou pai com apneia do sono?
F82	Paciente encaminhou mãe com Apneia do Sono?
F81	Paciente refere pai com Fibrilação Atrial?
F80	Paciente encaminhou mãe com Fibrilação Atrial?
F79	Paciente refere pai com infarto agudo do miocárdio?
F78	Paciente refere mãe com infarto agudo do miocárdio?
F77	Paciente refere pai com hipertensão?
F76	Paciente encaminhou mãe com hipertensão?
F62	Paciente refere histórico de tios com alguma doença?
F61	Paciente refere histórico de avô com alguma doença?
F60	Paciente refere histórico de avó com alguma doença?
F59	Paciente refere histórico de irmãos com alguma doença?
F58	Paciente refere histórico de pai com doença?
F57	Paciente refere histórico de mãe com alguma doença?

Essas relações são estabelecidas pelos *Task Graphs* mostrados na Figura 11, e como pode ser visto, para o tópico *Tóxico*, a questão T65 (Paciente relata um histórico de abuso de álcool?), T64 (Paciente relata histórico de consumo de cigarro?) e T63 (Paciente relata histórico de consumo de substâncias entorpecentes?) é condicionada à resposta de T66 (Paciente relata histórico de exposição a substâncias tóxicas?); a pergunta C52 (Paciente relata histórico de cirurgia cardíaca?), do tópico *Cirurgia*, é condicionada a resposta da questão C53 (Paciente relata histórico de cirurgia cardiovascular?); para o tópico *Alergias*, a pergunta A72 (Paciente relata histórico de algum tipo de alergia?) é condicionante das questões A67 (Paciente relata histórico de asma?), A68 (Paciente relata histórico de rinite alérgica?), A69 (Paciente relata histórico de dermatite?), A70 (Paciente relata histórico de alergia a medicamentos?) e A71 (Paciente relata histórico de alergia alimentar?); e para o tópico *Família*, a pergunta F57 (Paciente refere histórico de mãe com alguma doença?) é condicionante para as questões F76 (Paciente encaminhou mãe com hipertensão?), F78 (Paciente refere mãe com infarto agudo do miocárdio?), F80 (Paciente encaminhou mãe com Fibrilação Atrial?), F82 (Paciente encaminhou mãe com Apneia do Sono?), F84 (Paciente relata histórico de mãe com doença renal crônica?) e F86 (Paciente relata histórico de mãe com insuficiência mitral?) e a F58 (Paciente refere histórico de pai com doença?) para as perguntas F77 (Paciente refere pai com hipertensão), F79 (Paciente refere pai com infarto agudo do miocárdio?), F81 (Paciente refere pai com Fibrilação Atrial?), F83 (Paciente encaminhou pai com apneia do sono?), F85 (Paciente relata histórico de pai com doença renal crônica?) e F87 (Paciente relata histórico de pai com insuficiência mitral?).

Tabela 7 – Perguntas sobre Patologias definidas por um especialista e respondidas via método de Aprendizado Fracamente Supervisionado.

ID	Pergunta
P33	Paciente relata histórico de câncer?
P32	Paciente relata histórico de morte súbita e reanimação?
P31	Paciente relata histórico de reabilitação cardíaca?
P30	Paciente relata histórico de convulsões?
P29	Paciente é obeso?
P25	Paciente relata histórico de dislipidemia?
P24	Paciente relata histórico de doença cerebrovascular?
P23	Paciente relata histórico de vertigem?
P22	Paciente relata histórico de miocardite?
P21	Paciente relata histórico de infarto agudo do miocárdio sem doença coronariana?
P20	Paciente relata histórico de cardiopatia congênita?
P19	Paciente relata histórico de doença valvar?
P17	Paciente relata dispneia?
P13	Paciente relata histórico de doenças neurológicas?
P12	Paciente relata histórico de doenças psiquiátricas?
P11	Paciente relata histórico de distúrbios de coagulação?
P10	Paciente relata histórico de doenças autoimunes?
P9	Paciente relata histórico de doenças pulmonares?
P8	Paciente relata histórico de tratamento medicamentoso para alguma doença?
P7	Paciente relata histórico de tratamento cirúrgico para alguma doença?
P6	Paciente relata histórico de apnéia do sono?
P5	Paciente relata histórico de diabetes?
P4	Paciente relata histórico de fibrilação atrial?
P3	Paciente relata histórico de doença coronariana?
P2	Paciente relata histórico de hipotireoidismo?
P1	Paciente relata histórico de hipertensão arterial?

O tópico *Alergias*, cuja finalidade foi reproduzir a rotulação feita pelo especialista das questões A72 (Paciente relata histórico de algum tipo de alergia?), A71 (Paciente relata histórico de alergia alimentar), A70 (Paciente relata histórico de alergia a medicamentos?), A69(Paciente relata histórico de dermatite?), A68 (Paciente relata histórico de rinite alérgica?) e A67 (Paciente relata histórico de asma?), obteve o maior *F1-score* (99.6%).

Os valores da métrica *F1-score* mais baixas foram observadas para perguntas no tópico da *Patologias*, especificamente para as perguntas “Paciente relata histórico de hipotireoidismo?"(ID P2) e “Paciente relata histórico de fibrilação atrial?"(ID P4). A métrica da pergunta C75 está vazia porque todos os dados rotulados pelo especialista tiveram a mesma resposta "abster-se"ou "imprecisa", o que tornava inviável o cálculo da métrica. Infelizmente, isso mostra que essa pergunta criada pelos especialistas não pôde ser respondida com base nos dados disponíveis. Os registros do banco de dados que se referem a este tópico têm uma grande quantidade de dados ausentes ou incompletos, principalmente sobre cirurgia pancreática, objeto de questionamento da pergunta de ID C75. Esse tipo de situação tende a inviabilizar o processo de inferência do

Tabela 8 – Perguntas sobre Cirurgia, Alergias e Tóxico definidas por um especialista e respondidas via método de Aprendizado Fracamente Supervisionado.

	ID	Pergunta
Cirurgia	C75	Paciente relata cirurgia pancreática?
	C74	Paciente refere cirurgia bariátrica?
	C56	Paciente relata histórico de cirurgia por traumatismo craniano?
	C55	Paciente relata histórico de cirurgia pulmonar?
	C54	Paciente relata histórico de cirurgia para câncer?
	C53	Paciente relata histórico de cirurgia cardiovascular?
	C52	Paciente relata histórico de cirurgia cardíaca?
Alergias	A72	Paciente relata histórico de algum tipo de alergia?
	A71	Paciente relata histórico de alergia alimentar?
	A70	Paciente relata histórico de alergia a medicamentos?
	A69	Paciente relata histórico de dermatite?
	A68	Paciente relata histórico de rinite alérgica?
	A67	Paciente relata histórico de asma?
Tóxico	T66	Paciente relata histórico de exposição a substâncias tóxicas?
	T65	Paciente relata um histórico de abuso de álcool?
	T64	Paciente relata histórico de consumo de cigarro?
	T63	Paciente relata histórico de consumo de substâncias entorpecentes?

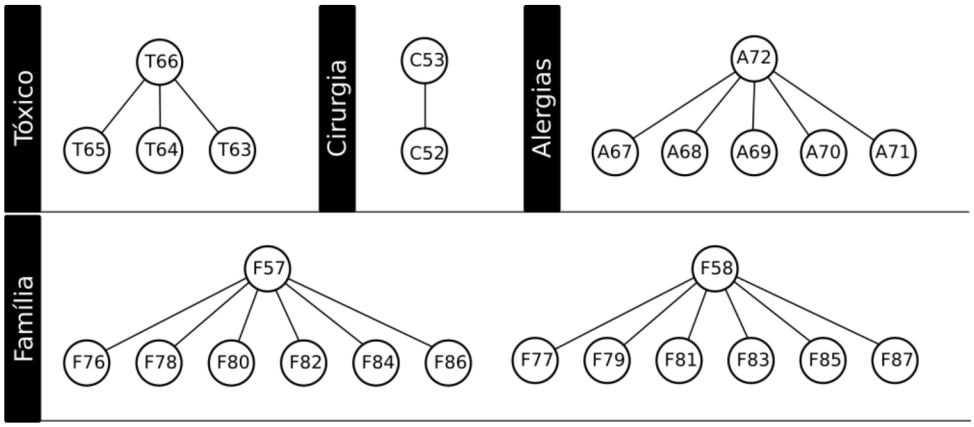

Figura 11 – *Task Graphs* das questões com as relações de dependência/restrição entre elas.

modelo de rotulação.

Além disso, os registros do banco de dados têm irregularidades e falta de padronização na escrita. Os profissionais que registram as informações sobre a saúde ou histórico do paciente usam vários formatos de escrita. Essa falta de padronização resulta em interpretações de dados mais desafiadoras e prejudica o processo de aprendizado do modelo. Por exemplo, para um determinado paciente que tem um registro do tópico *Tóxico* em uma dada consulta como “não referido", não é incomum ver outros registros apontando o contrário, como “fumante pesado". Essas ocorrências tendem a aumentar a incerteza no processo de decisão. No entanto, a técnica de Aprendizado Fracamente Supervisionado demonstrou ser robusta a esses eventos de ruído e apresentou boa precisão em geral. A Tabela 9 ilustra alguns resultados de rotulação alcançados

Tabela 9 – Exemplo de registro da base de dados com respeito ao uso de tóxicos e seus respectivos resultados de rotulação estimados pelo método de Aprendizado Fracamente Supervisionado para as perguntas T66 (Paciente relata histórico de exposição a substâncias tóxicas?), T65 (Paciente relata um histórico de abuso de álcool?), T64 (Paciente relata histórico de consumo de cigarro?) e T63 (Paciente relata histórico de consumo de substâncias entorpecentes?). Os valores 0, 1 e -1 indicam *Não*, *Sim* e *Impreciso*, respectivamente.

Paciente	**Uso de tóxico**	**T66**	**T63**	**T64**	**T65**
4557	não referido ex-fumante fumante pesado	1	-1	1	-1
345	não referido ex-fumante álcool ocasionalmente	1	-1	1	1
1876	não referido ex-fumante vai fazer 8 meses consumo de álcool vai fazer 10 anos	1	-1	1	-1
8146	maços diários suspenso a 8 dias não referido	1	-1	-1	-1
8275	várias não referido	0	-1	-1	-1

pela técnica, mesmo com essas dubiedades de registros.

Após a aplicação do método Aprendizado Fracamente Supervisionado, 61 novas variáveis foram geradas, cada uma representando as questão dispostas das Tabelas 6, 7 e 8.

4.2 Resultados dos modelos preditivos

As Tabelas 11, 12 e 13 apresentam os resultados dos modelos preditivos baseados no XGBoost para as diferentes configurações de variáveis de entrada que são descritas na Tabela 10 para as doenças EC, FA e AS, respectivamente. As métricas utilizadas para avaliar os modelos foram AUC, Sensibilidade e Especificidade.

Tabela 10 – Configurações das variáveis de entrada dos modelos XGBoost

Grupo	**Dados Físicos**	**Histórico de diagnósticos**	**Dados não estruturados**
H	Não	Sim	Não
N	Não	Não	Sim
F	Sim	Não	Não
H+N	Não	Sim	Sim
F+N	Sim	Não	Sim
F+H	Sim	Sim	Não
F+H+N	Sim	Sim	Sim

Como pode ser visto na Tabela 11, a previsão para EC atingiu os melhores valores de AUC para o arranjo F+H+N (0.83) com 0.74 e 0.76 de Sensibilidade e Especificidade, respectivamente. As configurações H, H+N e F+H também tiveram bom desempenho, mostrando que o histórico de doenças diagnosticadas do paciente é importante para a predição. O melhor resultado de Sensibilidade foi obtido no grupo H e a melhor Especificidade no H+N.

Tabela 11 – Resultados do XGBoost para as diferentes configurações de entrada das variáveis para diagnóstico de EC.

Grupo	AUC	Sensibilidade	Especificidade
H	0.80	**0.78**	0.67
N	0.58	0.46	0.67
F	0.67	0.62	0.64
H+N	0.81	0.67	**0.79**
F+N	0.69	0.58	0.69
F+H	0.82	0.72	0.77
F+H+N	**0.83**	0.74	0.76

Tabela 12 – Resultados do XGBoost para as diferentes configurações de entrada das variáveis para diagnóstico de FA.

Grupo	AUC	Sensibilidade	Especificidade
H	**0.72**	**0.96**	0.25
N	0.50	0.49	0.51
F	0.59	0.43	**0.72**
H+N	0.70	0.70	0.59
F+N	0.60	0.55	0.61
F+H	0.70	0.70	0.59
F+H+N	0.70	0.72	0.55

Os melhores resultados de cada métrica foram alcançados para as configurações F+H+N, H e H+N, respectivamente. Esses resultados refletem dois aspectos importantes: I) o histórico médico do paciente desempenha um papel importante na predição de EC; II) a possibilidade de considerar as informações não estruturadas processadas pela abordagem de Aprendizado Fracamente Supervisionado aumentou a capacidade de previsão do modelo para doença de EC.

A Tabela 12 apresenta os resultados para a predição de diagnóstico de FA. O grupo H resultou no melhor resultado para métricas de AUC (0.72) e Sensibilidade (0.96), mas foi no grupo F que a Especificidade atingiu o maior valor, 0.72. Semelhante ao caso da predição de EC, o histórico de diagnósticos de um paciente tem importância central para predição de FA com Sensibilidade de 96%.

Além disso, os dados não estruturados parecem ter uma papel importante para equilibrar a Sensibilidade e Especificidade do modelo, ou seja, reduzindo a diferença entre erros falsos positivos e falsos negativos. É importante mencionar que o grupo F+H+N resultou nos segundos melhores valores de AUC e Sensibilidade, destacando que essa combinação para predição de FA apresenta algumas vantagens que devem ser melhor investigadas.

Para AS, conforme apresentado na Tabela 13, AUC e Sensibilidade alcançaram seu melhor desempenho no grupo H. Nota-se porém, que este modelo (baseado somente no grupo H) apresenta um número elevado de falso positivos e portanto, uma Especificidade baixa. O grupo N sozinho, e combinado com H, melhora a Especificidade do modelo sem, no entanto,

Tabela 13 – Resultados do XGBoost para as diferentes configurações de entrada das variáveis para diagnóstico de AS.

Grupo	AUC	Sensibilidade	Especificidade
H	**0.64**	**0.98**	0.07
N	0.59	0.57	**0.45**
F	0.55	0.84	0.32
H+N	0.57	0.62	0.41
F+N	0.59	0.85	0.29
F+H	0.55	0.72	0.34
F+H+N	0.50	0.81	0.31

prejudicar muito o desempenho global do modelo (AUC). Por fim, é importante ressaltar que a capacidade preditiva dos modelos de FA e AS parece ter sido prejudicada pelo elevado grau de desbalanceamento desses conjuntos de dados. Apesar do uso de uma técnica para o balanceamento artificial do conjunto de dados (SMOTE), os valores mais baixos de AUC para FA e AS, quando comparados a EC, indicam que esses modelos sofreram mais com a falta de representatividade de exemplos positivos para FA e AS.

4.2.1 Explicando as decisões do modelo

A avaliação de desempenho dos modelos de diagnóstico realizada na subseção anterior ilustrou o efeito causado pela inserção de diferentes grupos de variáveis. Nesta subseção, aprofundamos essa análise utilizando o método SHAP para interpretar decisões individuais e também coletivas tomadas pelos modelos. O SHAP permite verificar a relevância das variáveis de entrada em relação a cada decisão individualmente; e também a principal relevância para um conjunto de observações.

Uma análise baseada nos valores de SHAP foi realizada para as configurações dos modelos preditivos que apresentaram os melhores resultados das métricas para as doenças EC, FA e AS.

A Figura 12 mostra o gráfico de relevância de características do SHAP para a configuração F+H+N para o diagnóstico de Enfermidade Coronária. Como se pode observar, as variáveis Sexo, IMC, ICD-10th R07 (Dor na garganta e tórax), ICD-10th I49 (Outras arritmias cardíacas), ICD-10th I10 (Hipertensão essencial), Questão T65 (O paciente relata histórico de abuso de álcool?), ICD-10th Z13 (Exame especial de triagem para outras doenças e distúrbios), ICD-10th R00 (Taquicardia, não especificada) e ICD-10th Q21 (Malformações congênitas dos septos cardíacos) tiveram maior impacto na saídas fornecidas pelo modelo. Para a configuração H+N, as variáveis que relataram maior impacto foram ICD-10th I10, perguntas dos IDs T66 (O paciente relata histórico de exposição a substâncias tóxicas?) e ICD-10th Z95 (Presença de implantes e enxertos cardíacos e vasculares), respectivamente. Para a configuração H, as variáveis das ICDs-10th I10, ICDs-10th I49 e ICDs-10th R07 também contribuíram de maneira destacável

para as decisões do modelo.

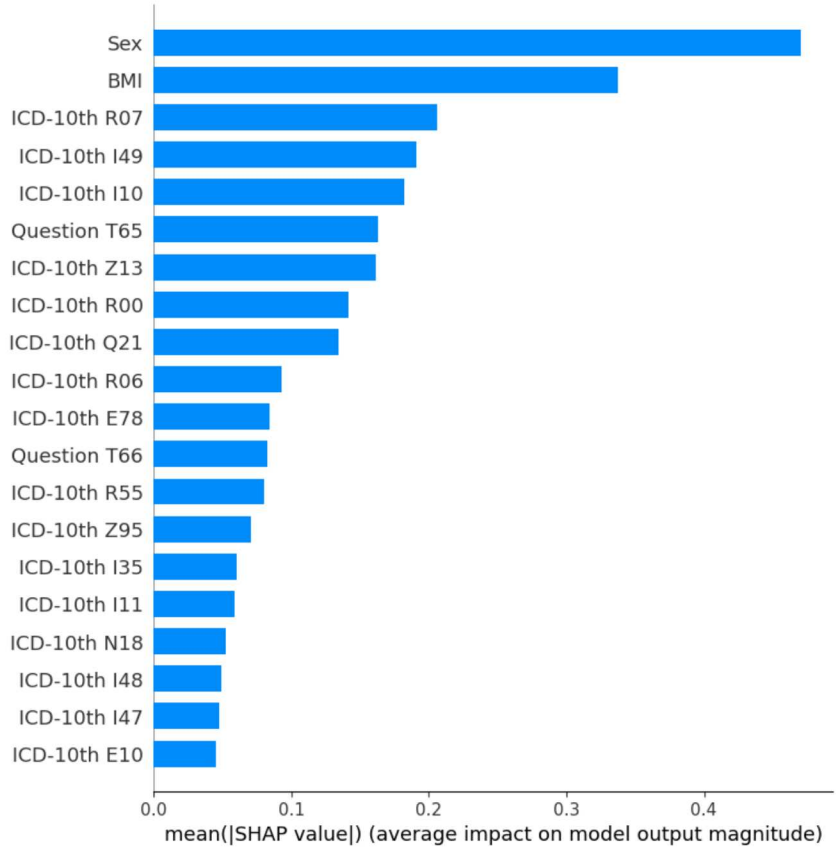

Figura 12 – Valores do SHAP para os 20 variáveis mais importantes para previsão de EC usando dados físicos, histórico de diagnósticos e dados não estruturados do paciente. Quanto maior o valor do SHAP de uma variável, maior é sua importância para a predição.

A Figura 13 mostra os valores médios de SHAP para o diagnóstico de Fibrilação Auricular utilizando apenas os dados de diagnósticos dos pacientes (H). As variáveis com maior importância nas decisões do modelo foram ICD-10th I25 (Cardiopatias Isquêmicas), ICD-10th R07, ICD-10th I42 (Cardiomiopatia), ICD-10th I50 (Insuficiência cardíaca), ICD-10th J44 (Outras doenças pulmonares obstrutivas crônicas), ICD-10th I10 e ICD-10th Z95. Utilizando os dados não estruturados e o histórico do paciente (H+N) como dados de entrada do modelo, as questões T66 (Paciente relata histórico de exposição a substâncias tóxicas?) e T65 (Paciente relata um histórico de abuso de álcool?) aparecem entre os 20 fatores que mais impactam na saída do modelo. Para o arranjo de variáveis que utiliza todos os dados disponíveis (F+H+N), Sexo e IMC são a segunda e a quarta variáveis de maior impacto na saída do modelo, respectivamente.

Para o diagnóstico de AS, e considerando apenas o histórico de diagnóstico do paciente, as variáveis ICD-10th J44 (Doença pulmonar obstrutiva crônica com infecção respiratória inferior aguda), ICD-10th E66 (Obesidade), ICD-10th I10, ICD-10th I20 (Angina instável), ICD-10th Z95 e ICD-10th I21 tem uma magnitude de impacto maior na saída do modelo. Como poder ser visto na Figura 14, o diagnóstico de Doença Pulmonar Obstrutiva Crônica com Infecção

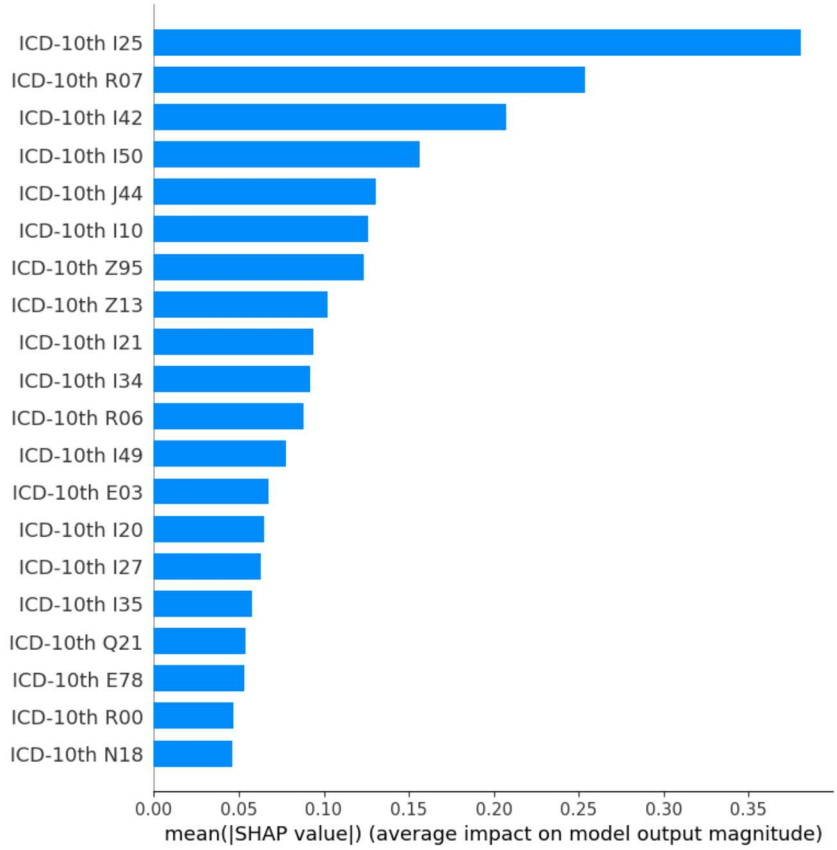

Figura 13 – Valores do SHAP para os 20 variáveis mais importantes para previsão de FA usando histórico de diagnósticos. Quanto maior o valor do SHAP de uma variável, maior é sua importância para a predição.

Respiratória Inferior Aguda (ICD-10th J44) e Obesidade (ICD-10th E66) tem um impacto destacado na saída do modelo.

A interpretação do modelo a partir de uma perspectiva global, que considera a magnitude de relevância média das variáveis, é importante para identificar padrões de uma determinada doença. No entanto, a compreensão dos fatores que levaram a um diagnóstico individual de um paciente são fundamentais para a aceitação do modelo como ferramenta de apoio à decisão. Por isso, além da explicabilidade em perspectiva macro, analisamos o decisor de maneira individualizada (por paciente).

A Figura 15 mostra a explicação dada pelo SHAP para um paciente arbitrário, ou seja, sob uma perspectiva individual. O valor de cada variável desse paciente está entre parênteses. As contribuições de cada variável para o resultado do modelo são ordenadas por sua importância modular (módulo dos pesos). Como se pode observar, as variáveis ICD-10th R55, Sexo, IMC e ICD-10th I10 contribuem negativamente para o diagnóstico. O modelo tem um valor esperado ($E[f(x)]$) igual a $0,046$, porém, por se tratar de um paciente sem diagnóstico de ICD-10th I10, que possui IMC considerado normal (menor que 25), é do sexo masculino e foi diagnosticado com ICD-10th R55, a saída do modelo foi fortemente revertida e alcançou $f(x) = -3.245$ como

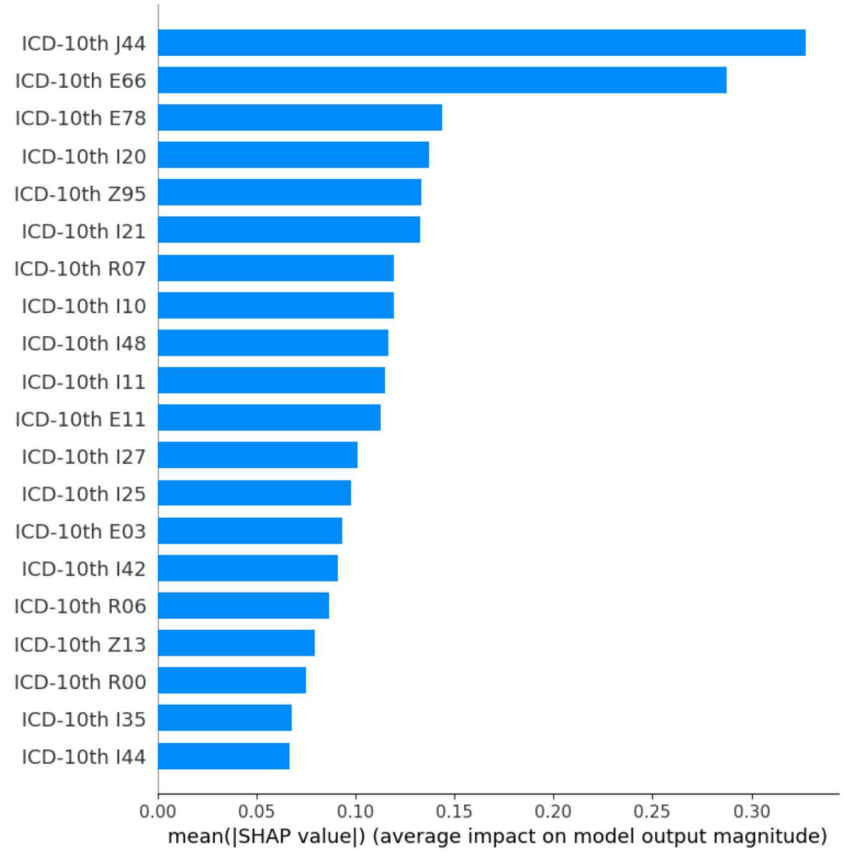

Figura 14 – Valores do SHAP para os 20 variáveis mais importantes para previsão de AS usando histórico de diagnósticos. Quanto maior o valor do SHAP de uma variável, maior é sua importância para a predição.

resultado da predição. Valores menores que zero são classificados como “não diagnóstico".

Outro exemplo de explicabilidade dada pelo SHAP é mostrado na Figura 16. Desta vez, com uma classificação positiva (“diagnóstico"). Inicialmente, o valor esperado para aquele modelo e com aquele arranjo de dados de entrada é de -0.191, mas dado que o paciente teve o diagnóstico de Distúrbios do Metabolismo de Lipoproteínas e Outras Lipidemias (ICD-10th E78), houve um forte impacto na saída do decisor, indicando que o diagnóstico deste paciente é positivo ($f(x) = 2.327$) para aquela doença.

Diante deste dois exemplos é possível observar a importância da informação que a explicabilidade do decisor em perspectiva micro apresenta. Com este recurso adicional ao modelo preditivo, é possível notar a intensificação das características individuais de um paciente, considerando que cada um deles têm as suas Especificidades. E essas, por suas vez, pode alterar, de maneira significativa, a saída do modelo de diagnósticos.

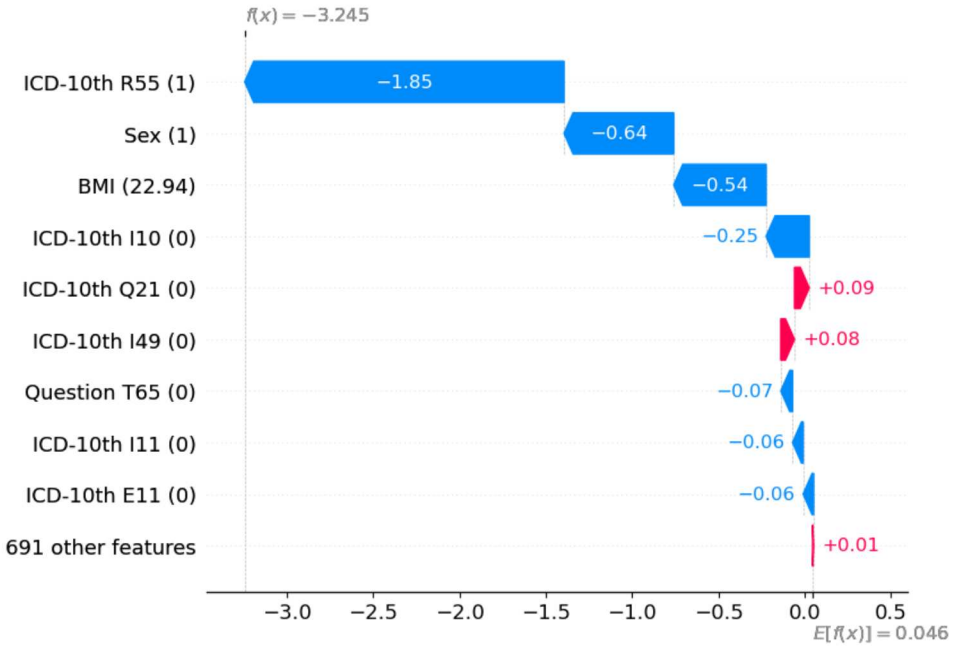

Figura 15 – Explicação dada pelo SHAP para um paciente arbitrário classificado como classe negativa. O valor de cada variável está entre parênteses.

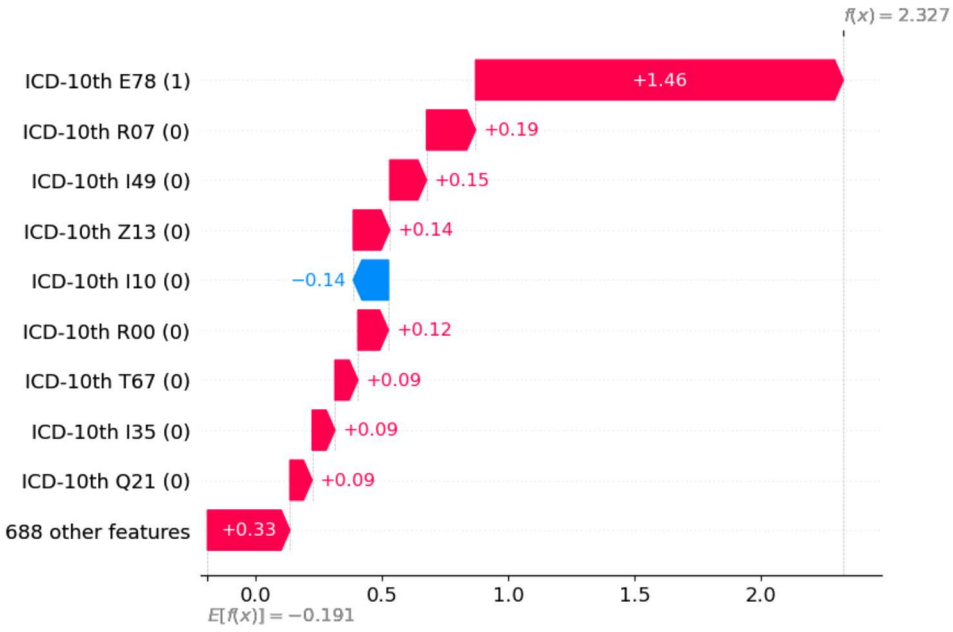

Figura 16 – Explicação dada pelo SHAP para um paciente arbitrário classificado como classe positiva. O valor de cada variável está entre parênteses.

4.3 Software de suporte à decisão

Por fim, um software foi desenvolvido para possibilitar o acesso a todas as tecnologias propostas e desenvolvidas neste trabalho. A ferramenta foi desenvolvida no idioma inglês. O acesso se dá por meio de login e senha havendo níveis de autorização dentro do sistema. A equipe

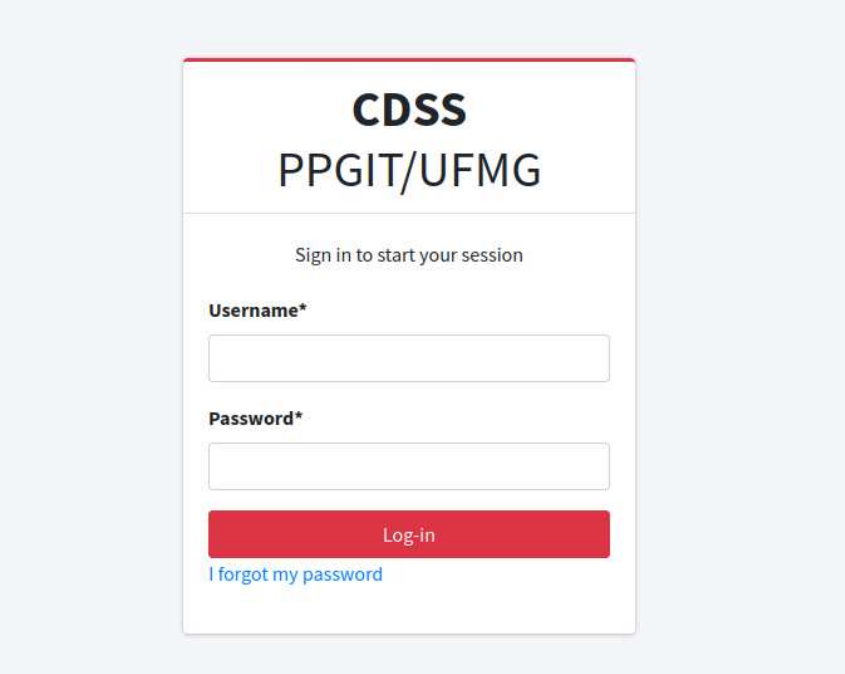

Figura 17 – Tela do *login* do sistema de suporte à decisão médica desenvolvido neste trabalho.

médica tem acesso ao histórico clinico dos pacientes e os resultados das predições e rotulações realizadas pela *framework*. Além deste perfil de acesso, há um outro nível que contempla acesso total aos módulos e foi pensado para que uma equipe de suporte (por exemplo, de Tecnologia da Informação) possa proceder com importações e execuções dos modelos de processamento dos dados. A Figura 17 mostra a tela de acesso ao sistema.

No *Dashboard* do perfil com acesso total, é possível visualizar as informações de dados importados, processados, total de pacientes e o total de registros que compõe a base dados. A Figura 18 apresenta essa tela, que ainda contem os *links* para acessar o *log* de importações e executar as rotinas de processamentos dos dados.

Com o acesso ao sistema, a equipe médica, através da tela mostrada na Figura 19, busca o paciente pelo seu número de documento. Os dados disponíveis para identificar o indivíduo além do documento são sexo e data nascimento. Essa condição de identificação foi definida pelo provedor dos dados, pois a legislação Colombiana não permite a concessão de alguns dados de identificação dos pacientes como nome ou nome da mãe. Após a localização do paciente, através do botão de *Details* é possível acessar detalhes do histórico daquele indivíduo, além de todos os dados disponíveis (prontuários, idade e outros).

A tela do histórico é composta por 4 blocos: identificação do paciente; *timeline* de diagnósticos; respostas baseadas nos dados não estruturadas; e detalhamento do prontuário médico em cada consulta. As Figuras 20, 21 e 22 mostram a tela do histórico do paciente e onde se pode visualizar os supracitados blocos.

Na Figura 20 é possível visualizar a identificação do paciente e o gráfico do tipo *timeline*

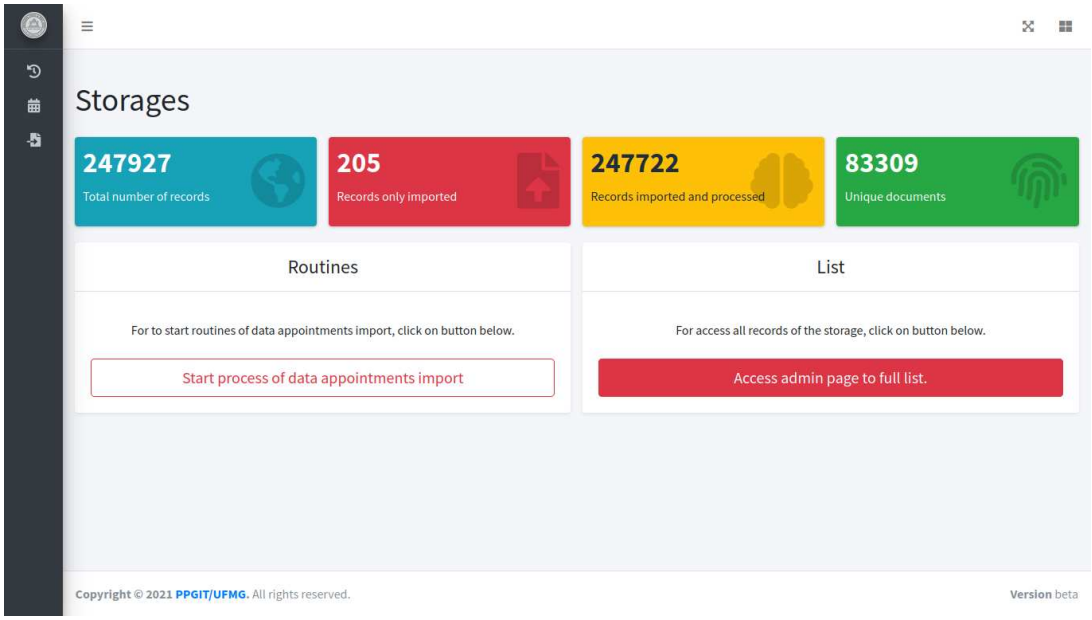

Figura 18 – *Dashboard* do sistema de suporte à decisão médica desenvolvido neste trabalho.

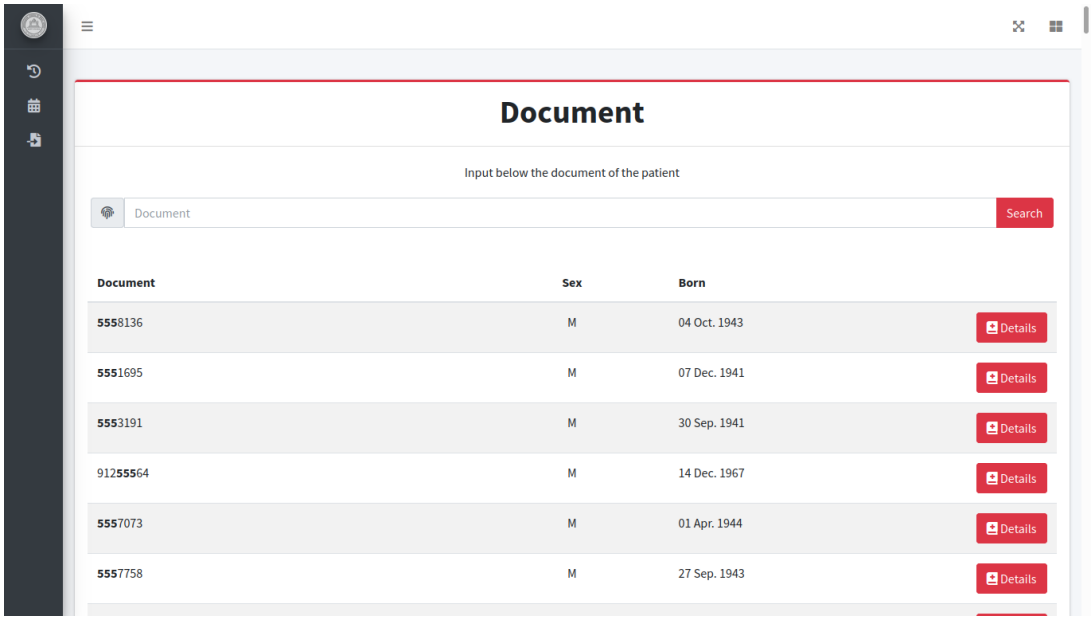

Figura 19 – Tela do sistema de busca por documento para acesso ao histórico clínico e predições de diagnóstico dos pacientes.

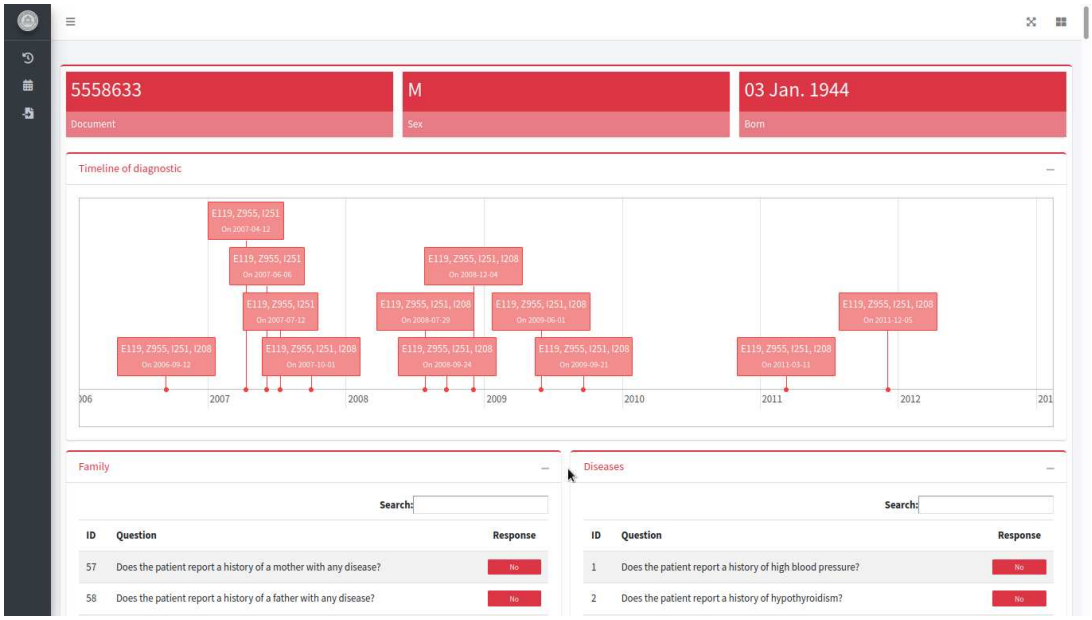

Figura 20 – Tela do histórico com a identificação do paciente e *timeline* de diagnósticos.

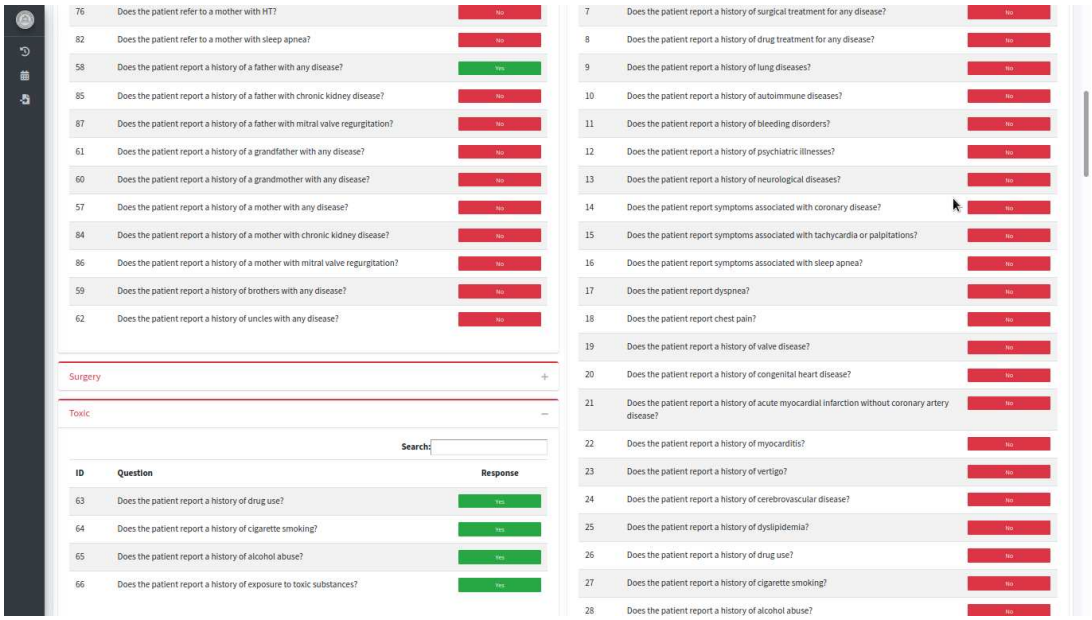

Figura 21 – Tela do histórico do paciente com as respostas produzidas pelo método de Aprendizado Fracamente Supervisionado para as questões indicadas pelo especialista.

onde se apresenta todos os diagnósticos daquele paciente em sua respectiva data. Esse recurso é interativo, podendo ser alterado através de operações com o mouse, como aproximar (*zoom in*), afastar (*zoom out*) ou ainda arrastar/mover para diferentes direções.

Na Figura 21 são apresentadas as questões indicadas pelo especialista e cujas respostas foram inferidas pelo modelo de supervisão fraca. As respostas podem ser ordenadas por ID, texto da pergunta ou pela resposta do modelo. Ademais, há um campo de filtro em cada agrupador (tópico) que permite o usuário buscar alguma pergunta específica. Quando a resposta é positiva a

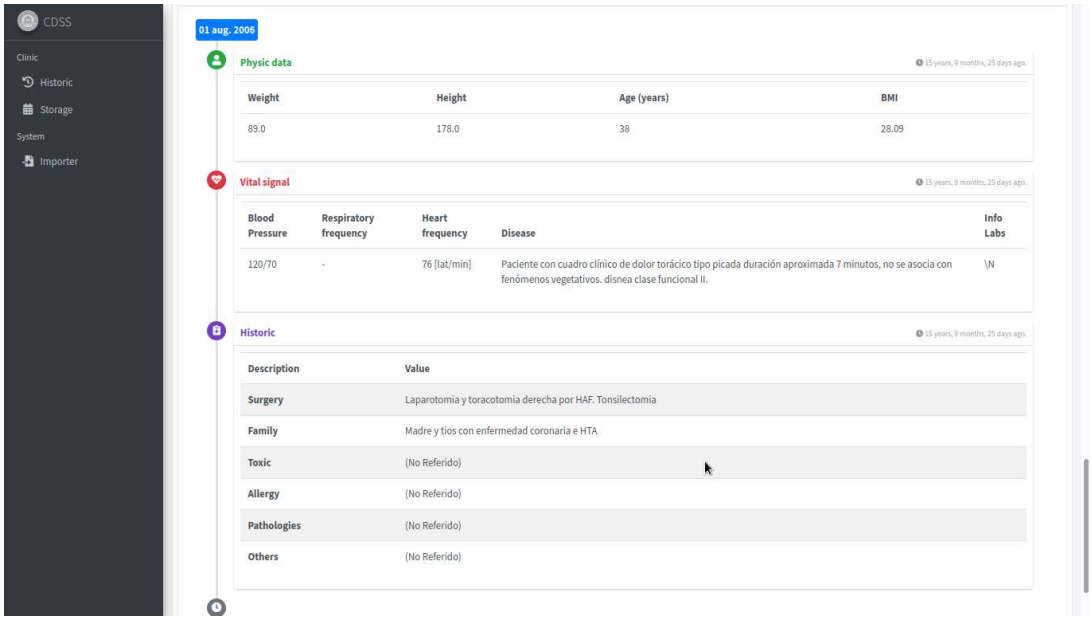

Figura 22 – Tela do histórico com o detalhamento do prontuário médico de cada consulta do paciente.

célula da coluna de resposta daquela pergunta muda para a cor verde. Caso a resposta seja “Não", permanece na cor vermelha.

Por fim, é possível visualizar os detalhes do prontuário em cada consulta do paciente (Figura 22). Além de peso (*Weight*), altura (*Height*), IMC (BMI), frequências respiratória (*Respiratory Frequency*) e cardíaca (*Heart Frequency*) e pressão arterial (*Blood Pressure*) no dia da consulta, há os dados registrados textualmente pelo médico sobre antecedentes familiares, cirurgias, tóxico, alergias, patologias e outros do paciente informações narradas por ele naquele dia. A possibilidade de visualização de todos os dados dos prontuários daquele do paciente oferece uma maior segurança para o profissional de saúde que está utilizando o sistema. Afinal, caso haja alguma dúvida sobre quais foram as informações consideradas para que o modelo produzisse respectiva saída, ele pode visualizar isso de maneira rápida e organizada.

4.4 Discussões

Os resultados obtidos nesse trabalho são promissores e com isso foi possível gerar um produto com potencial de ser usado, e possivelmente comercializado, em entidades de saúde públicas e privadas.

A qualidade dos dados disponibilizados é um fator importante no processo de análise de estimativa do diagnóstico das doenças. Processar os dados textuais não estruturados se mostrou desafiador. Isso porque se fez necessário organizar e disponibilizar toda a representação contida naqueles dados para um modelo computacional, que por sua vez pode processá-los e transformá-los em conhecimento. A escrita livre nos registros dos pacientes é um dificultador para realizar essa atividade. Entretanto, as técnicas aqui aplicadas trouxeram resultados satisfatórios com potencial para ser aplicado na prática.

As perguntas relacionadas ao histórico de *Patologias* dos pacientes, em média, apresentou a menor assertividade entre os tópicos ao tentar reproduzir as respostas do especialista via Aprendizado Fracamente Supervisionado. Isso pode ser justificado por ser o único tópico que não fez uso do suporte de um *Task Graph*, o que melhora o despenho das métricas, como reportado por Ratner et al. (2019). Além disso, por ser o grupo em que há maior incidência de registros que expressam o diagnóstico através de escritas de difícil entendimento, como por exemplo, abreviações, acrônimos e siglas.

Outro fator relevante que os resultados obtidos com a técnica de supervisão fraca mostrou, é a possibilidade de realizar uma leitura do prontuário do paciente e fornecer, de maneira organizada, amigável e informativa, as principais questões a cerca do histórico clínico e familiar dos pacientes. Isso agiliza o antedimento e facilita a leitura dos dados do indivíduo, promovendo um ganho de tempo no processo de tomada de decisão.

Os experimentos com os modelos preditivos de diagnóstico destacaram a importância das diferentes fontes de dados para a previsão de EC, AS e FA. Pode-se afirmar que, no geral, os dados não-estruturados contribuíram para aumentar o desempenho de predição para EC. Para AS e FA, apesar dos dados não estruturados não resultarem nos melhores desempenhos de predição, os experimentos dão indicativos de que esse tipo de informação pode melhorar as métricas de desempenho dos modelos preditivos como, por exemplo, reduzir o número de falsos positivos/negativos. Cabe ressaltar no entanto, que esses modelos podem ter sido prejudicados pela falta de representatividade da amostra de diagnósticos positivos dessas doenças (FA e AS) no conjunto de treinamento. Além disso, todos os modelos com bom desempenho de predição em nosso estudo tiveram como entrada o histórico de diagnóstico dos pacientes. Isso reforça que esse tipo de informação tem papel fundamental na predição de diagnóstico do paciente.

É importante esclarecer que os registros mostraram mais de 150 doenças em 19.180 pacientes, incluindo pacientes com AS que posteriormente desenvolveram EC ou FA. Além disso,

incluímos pacientes com FA que posteriormente desenvolveram EC e com FA desenvolvida após EC. A prevalência de FA em pacientes com EC relatada na literatura oscila entre 0,2%-5% (MICHNIEWICZ et al., 2018). A prevalência de AS em pacientes com FA e EC tem sido relatada em poucos estudos, mas não para uma grande população, pois o rastreamento de AS com polissonografia nesses pacientes não é comum. No entanto, a prevalência de AS relatada para pacientes com FA varia entre 32% a 70% (STEVENSON et al., 2008; HOLMQVIST et al., 2015). Contudo, uma limitação de nossa análise é que não podemos identificar a relação causal ou bidirecional para essas doenças.

Outra limitação de nosso trabalho é que foi utilizada aqui uma base de dados com pacientes que foram diagnosticados num serviço especializado (atenção secundária) e, geralmente, esses pacientes podem apresentar mais comorbidades do que os indivíduos de serviços de atendimento à saúde mais abrangente (atenção primária) ou da população geral.

Mesmo que estudos anteriores tenham relatado o uso de aprendizado de máquina em dados de EHR para classificação, diagnóstico e previsão de hospitalização futura (STEELE et al., 2018), a identificação de AS como fator de risco para EC e FA não foi suficientemente estudada. Embora, em nosso estudo, o papel da AS para o risco de FA e EC não tenha sido claramente identificado, abordagens de previsão de risco baseadas em dados interpretáveis podem fornecer suporte ao diagnóstico, permitindo o tratamento oportuno desses pacientes. Considerando que a doença cardiovascular é a principal causa de morte no planeta, a identificação de pacientes sub-diagnosticados poderá diminuir a mortalidade no mundo.

A abordagem proposta neste trabalho pode prever o diagnóstico das doenças de EC, AS e FA com um desempenho competitivo com a literatura. Os autores Hulme et al. (2019), que em seu trabalho fundiu EHR de múltiplas instituições de saúde a fim de realizar a predição de risco de FA em pacientes acima de 45 anos, obteve com o método EHR-AF uma AUC de 0.77. O método aqui proposto atingiu para o diagnostico dessa mesma doença um AUC de 0.72, e considerando pacientes acima de 18 anos. Conforme reportado na literatura, a prevalência de FA na população geral é de $0,4\%$ e aumenta com o avanço da idade; e a partir dos 50 anos, duplica a cada década (BENJAMIN et al., 1998; FILHO et al., 2003; JUSTO; SILVA, 2014). Ou seja, os dados que aqueles autores utilizaram na fase de treinamento tendem a ter, proporcionalmente, um número de amostras da classe positiva maior do que a utilizada neste trabalho, favorecendo a construção de um modelo com conhecimento mais equilibrado entre as características das duas classes.

Para a predição do diagnóstico de EC, o trabalho de Kim e Kang (2017), utilizando um modelo de rede neural e correlação de características, propôs um método para a previsão de risco de EC. O melhor índice obtido para AUC foi 0.749. Utilizando dados demográficos, histórico clínico de pacientes e dados não estruturados na predição do diagnóstico dessa doença, este trabalho conseguiu superar esse valor e atingiu um índice de AUC de 0.83.

Além de atingir desempenho competitivo com a literatura, o método proposto consegue

explicar o processo de tomada de decisão nas perspectivas macro e micro; característica fundamental para superar o fator "caixa preta" inerentes a estes modelos de predição. Esse equilíbrio entre capacidade preditiva e interpretabilidade é um aspecto importante para a aplicação clínica desse tipo de procedimento. Este trabalho também difere da literatura em outros pontos. Dentre eles, podemos citar a utilização da técnica de Aprendizado Fracamente Supervisionado para gerar novas variáveis que vão compor o vetor de entrada dos modelos a partir da fusão com as variáveis dos dados estruturados.

Analisando a Figura 12, que demonstra a interpretação geral de decisão do modelo, é possível observar que IMC e Sexo são algumas das características mais relevantes do processo. No entanto, quando vemos a Figura 15, a ordem de relevância muda, refletindo a capacidade do modelo de considerar as particularidades de cada paciente. Isso destaca a importância de uma explicação na perspectiva macro e micro para fornecer informações confiáveis em uma eventual utilização dessa *framework* em um sistema de apoio à decisão para diagnóstico de doenças. Vale notar que a interpretabilidade do modelo não significa causalidade, ela apenas mostra como o modelo tomou a decisão de obter um determinado resultado.

A explicação do modelo em uma perspectiva macro permitiu identificar os fatores que mais contribuem para o processo decisório de modo geral. Já na perspectiva micro, foi possível avaliar o diagnóstico ao nível das características específicas de um paciente.

5 Conclusão

Neste estudo, foram desenvolvidos modelos de diagnóstico capazes de explicar a tomada de decisão, com base em dados estruturados e não estruturados (extraídos de EHR) para FA, EC e AS. Os dados foram pré-processados antes do aprendizado do modelo. Os dados não estruturados foram mesclados por paciente e tipo de registro, as *stopwords* foram removidas e um método de supervisão fraca foi aplicado, transformando-os em variáveis categóricas numéricas. Os valores tabulares faltantes foram inferidos com base em registros anteriores dos pacientes. Em seguida, os dados rotulados foram mesclados com os dados estruturados e usados para compor o vetor de entrada dos modelos de predição de diagnóstico.

A Fibrilação Auricular é o distúrbio do ritmo cardíaco mais comum em todo o mundo. Enfermidade Coronariana e Apneia do Sono são conhecidos fatores de risco associados para FA. Os resultados alcançados nesse trabalho pelos modelos de predição de diagnósticos dessas doenças são competitivos com os reportados na literatura. Para além disso, apresentamos como novidade o diagnóstico acurado com um modelo interpretável dos fatores que influenciam na tomada de decisão do preditor, o que pode ser é um importante aliado na construção do conhecimento desses diagnósticos.

A construção dos modelos com o desfecho da tomada de decisão explicável pode permitir uma maior aceitação na clínica médica de ferramentas baseadas em Inteligência Artificial como suporte à decisão médica. Além da explicabilidade, outro fator que pode vir a contribuir com uma melhora na aceitação do recurso é que os métodos de aprendizado empregados conseguiu construir conhecimentos a cerca das doenças em consonância com a literatura médica. Ou seja, a explicabilidade não se dá de maneira aleatória. Por exemplo, ao explicar os fatores de maior impacto no decisor para realizar o diagnóstico de EC (Figura 12), foram apontados variáveis como IMC, Hipertensão (ICD-10th I10), Arritmias Cardíacas (ICD-10th I49), Histórico de Alcoolismo, entre outras que são reportadas na literatura médica como fatores de risco dessa doença (CARVALHO, 1992; FARIA et al., 2002; STIPP et al., 2007; MASSAROLI et al., 2018; YADAV; JADHAV, 2021). Ou ainda, ao explicar os fatores de maior impacto no processo decisório para diagnóstico de AS (Figura 14), o SHAP indica as variáveis Doenças Pulmonares Obstrutivas Crônicas (ICD-10th J44), Obesidade (ICD-10th E66), Dislipidemia (ICD-10th E78), entre outros; e todas essas reconhecidas por especialistas como fatores importantes para o desenvolvimento da enfermidade (ORNELAS et al., 2019; CARNEIRO; FONTES; TOGEIRO, 2010; MARTINS; TUFIK; MOURA, 2007; MANCINI; ALOE; TAVARES, 2000; ROBERTO; BRITO; ROGÉRIO, 2000).

Para além da explicabilidade da tomada de decisão em perspectiva macro, esse trabalho apresenta também explicabilidade da tomada de decisão do modelo em perspectiva micro (por

paciente). Essa abordagem mostrou-se importante por considerar as especificidades de cada indivíduo. Afinal, o modelo global consegue realizar uma explicação do comportamento médio do indivíduos que compõe o conjunto de dados na fase de treinamento. Contudo, cada paciente tem as suas características sociais, culturais, familiares, entre outras; que podem ser fatores relevantes no processo de diagnóstico das doenças.

As instituições de saúde ao redor do mundo possuem um grande volume de dados e muitas vezes esses dados não são utilizados para construir conhecimento. A previsão do risco de doença é uma tarefa multifacetada e complexa. Talvez, o principal motivo da dificuldade em analisar tais dados seja a não padronização na captura e armazenamento dos dados. Abordagens para lidar com esse desafio são amplamente discutidas e investigadas na literatura, e geralmente resultam em modelos que realizam a predição da doença a partir de algumas informações de um paciente. No entanto, modelos preditivos aplicados apenas para dar um resultado sem interpretabilidade terão dificuldades de serem adotados para apoiar a decisão clínica. A alta precaução dos médicos no diagnóstico exige uma tomada de decisão confiável e explicável. Portanto, a boa comunicação com estes usuários e a interpretabilidade dos modelos é tão importante quanto a boa acurácia.

Lidar com dados não estruturados não é uma tarefa trivial, contudo este trabalho mostrou que, quando considerados, as informações contidas nessas variáveis podem agregar maior acurácia na construção de modelos de predição de diagnóstico. Assim sendo, uma captura mais organizada desses dados podem contribuir para um processamento mais ágil e que pode agregar maior assertividade dos modelos.

O método de Aprendizado Fracamente Supervisionado mostrou-se eficaz na sua aplicação. Com o uso dos *Task Graphs* a sua apuração for mais precisa. A construção das *Label Functions* é uma tarefa custosa e que demanda o estudo de várias heurísticas para conseguir extrair bons resultados. Uma característica negativa dessa técnica é que as LFs baseadas em associação de padrões textuais geram um grande volume de codificação. Afinal, para cada padrão a ser buscado, uma nova função é criada. Entretanto, essa mesma característica permite uma escalabilidade do método em eventuais ampliação ou manutenção dos modelos.

As tecnologias e *software* aqui desenvolvidas é um produto de inovação que tem o potencial para ser comercializado em clínicas e hospitais publicas e privadas em toda a América Latina. Ademais, a sua arquitetura modular de desenvolvimento permite a sua escalabilidade, podendo ainda chegar à outras doenças que atingem a população latina americana. Por uma limitação temporal, a tecnologia de explicabilidade dos modelos de diagnósticos não foi embarcada no *software*.

Dos objetivos geral e específicos propostos nesse trabalho todos foram alcançados. A metodologia de construção de modelos de predição de diagnósticos que incorporam dados não estruturados em sua construção apresentou resultados compatíveis com a literatura; as técnicas de *Embedding* textual e Aprendizado Fracamente Supervisionado foram aplicadas e produziram resultados relevantes de rotulação; A predição dos diagnósticos de AS, FA e EC

utilizou das variáveis resultantes da fusão dos dados estruturados e não estruturados; as variáveis que mais impactaram nas saídas das predições foram explicadas em macro e micro perspectivas e produziram conhecimento alinhados com a literatura médica; finalmente, estes recursos foram implementados em um software que tem potencial de ser comercializado em instituições públicas e privadas de saúde de atenção secundária.

Por fim, espera-se que as tecnologias aqui desenvolvidas possam ser aplicadas no dia-a-dia da prática médica e que possa ser um importante recurso para o diagnóstico precoce dessas doenças e assim contribuir com a diminuição da incidência de mortes causadas por essas enfermidades.

5.1 Propostas de continuidade

Como continuidade das pesquisas desenvolvidas nesse trabalho, sugere-se investir nos seguintes problemas relacionados ao tema:

- Investigação de propostas de sistematização na coleta dos dados não estruturados durante as consultas médicas dos pacientes. A falta de padronização no preenchimento dos campos pode ser atacada através da utilização de tecnologias já disponíveis na literatura de desenvolvimento de *software* como, por exemplo, o Auto Completar (LOH, 2016) ou Mineração de Dados (NIEMANN; MOEHRLE; FRISCHKORN, 2017; AZEVEDO, 2019; CHEN et al., 2020);

- Extensão do modelo de predição de diagnóstico com explicabilidade utilizando bases de dados com mais dados sobre os pacientes, tais como: colesterol total, colesterol HDL, pressão arterial sistólica, pressão arterial diastólica, triglicéride, imagens bi e tri dimensionais, entre outros;

- Uma limitação deste trabalho é não explicação de causalidade das variáveis. Essa informação pode produzir resultados importantes para entender a correlação e interação entre o histórico clínico dos pacientes e as demais variáveis. Uma técnica disponível na literatura, que é um interpretável e que indica as relações de causalidades entre as *features* do conjunto de treinamento, são as Redes Bayesianas (RB) (MARQUES; DUTRA, 2002). Elas tem sido explorada na saúde graças à capacidade de modelarem situações de incerteza. Redes Bayesianas permitem modelar as variáveis do domínio de interesse na forma de uma rede, em que as variáveis são conectadas pelas relações de influência presentes entre elas; por isso, a estrutura causal da rede é explicitamente representada (MCLACHLAN et al., 2020). Assim sendo, recomendamos a expansão dessa proposta de utilizar dados resultantes da fusão de dados estruturados e não estruturados aplicados à metodologia da Redes Bayesianas;

- Considerando a predição da doenças como um fator importante para o seu diagnóstico precoce, realizar a predição de desenvolvimento da doenças em função do tempo se mostra uma boa proposta de continuidade deste trabalho. Para isso, há disponível na literatura a Análise de sobrevivência (do inglês, *Survival Analysis*); ou Análise de Sobrevida, é uma coleção de métodos estatísticos no qual a variável de resposta de interesse é o tempo que um determinado evento acontece (WANG; LI; REDDY, 2019; KLEIN et al., 2016; KLEINBAUM; KLEIN, 2010).

  Chamada de tempo de falha, este tempo pode ser, por exemplo, o tempo que um paciente foi a óbito após descobrir uma doença, bem como o tempo que levou para se curar. Os dados de sobrevida de um indivíduo i sob estudo, geralmente são representados pelo par (t_i, δ_i), onde t_i é o tempo de falha ou censura e δ_i indica a ocorrência ou censura do evento de interesse (COLOSIMO; GIOLO, 2006). Isto é:

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é um tempo de falha} \\ 0, & \text{se } t_i \text{ é um tempo censurado} \end{cases}$$

  Ainda segundo Colosimo e Giolo (2006), havendo covariáveis relacionadas com estes indivíduos, os dados podem ser representados por (t_i, δ_i, x_i), ou ainda, em caso especiais, por $(l_i, u_i, \delta_i, x_i)$, onde x_i são as covariáveis e l_i e u_i os limites de máxima e mínima do intervalo, respectivamente, amostrado para i-éssimo indivíduo.

  A função de sobrevivência é a mais usada para descrever estudos de sobrevida. Ela é usada para representar a probabilidade de um evento de interesse não falhar até um intervalo de tempo t. Em outras palavras, a probabilidade de uma amostra sobreviver ao tempo t (WANG; LI; REDDY, 2019; COLOSIMO; GIOLO, 2006; LEE; WANG, 2003). Em descrição matemática, a função de sobrevivência é denotada por:

$$S(t) = P(T \geq t) \tag{5.1}$$

  Por consequência, a função de distribuição acumulada de uma amostra não sobreviver ao tempo t é $F(t) = 1 - S(t)$, e a função de densidade de não-sobrevida ou falha, pode ser obtida por $f(t) = \frac{d}{dt}F(t)$, para casos contínuos, e $f(t) = [F(t + \Delta t) - F(t)]/\Delta t$, onde Δt denota um intervalo de tempo, para casos discretos (KLEINBAUM; KLEIN, 2010).

  Outra função muito utilizada em Análise de Sobrevivência é a função de Hazard ($h(t)$). Também citada na literatura como "Força de mortalidade" (do inglês, *Force of mortality*), "Taxa de morte instantânea" (do inglês, *Instantaneous death rate*) ou "Falha condicional" (do inglês, *Conditional failure*) (DUNN; CLARK, 2009),

  A função de Hazard não indica a chance ou a probabilidade que o evento de interesse aconteça, mas sim a probabilidade do evento ocorrer em um tempo t dado que esse não

ocorreu no instante anterior ao tempo t (WANG; LI; REDDY, 2019). Matematicamente, a função de Hazard é definida por:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)} \quad (5.2)$$

Enquanto a função de sobrevivência $S(t)$ sempre decresce ao longo do tempo, a função $h(t)$ pode ter uma mudança de comportamento. Considerando que podemos escrever $f(t) = -\frac{d}{dt}S(t)$, é possível reescrever a função de Hazard como:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}S(t) \cdot \frac{1}{S(t)} = -\frac{d}{dt}[\ln S(t)]. \quad (5.3)$$

Então, podemos reescrever a função de sobrevivência definida na Equação 5.1 como (WANG; LI; REDDY, 2019):

$$S(t) = exp(-H(t)), \quad (5.4)$$

onde $H(t) = \int_0^t h(u)du$ é a Função Cumulativa de Hazard (CHF - do inglês, *Cumulative Hazard Function*) (KLEINBAUM; KLEIN, 2010).

A Análise de Sobrevivência é dividida em métodos paramétricos, não paramétricos e semi paramétricos. Dentre os modelos semi paramétricos há o modelo de Cox (COX, 1972). O modelo de regressão de Cox permite a análise de dados provenientes de estudos de tempo de vida em que a resposta é o tempo até a ocorrência de um evento de interesse, ajustado por covariáveis (COLOSIMO; GIOLO, 2006).

Dado um conjunto de dados qualquer com indivíduos de dois grupos, 1 e 0, para obtermos o valor da proporção de falha ou risco relativo de cada grupo K, têm-se:

$$\frac{h_1(t)}{h_0(t)} = K \quad (5.5)$$

Se x_i indica qual grupo pertence o indivíduo i e $K = exp(\beta_i x_i)$, sendo β um vetor com o fator proporcional da covariável, então temos:

$$h(t) = h_0(t) \cdot \exp(\beta_i x_i) \quad (5.6)$$

ou seja,

$$h(t) = \begin{cases} h_1(t) = h_0(t)exp(\beta), & \text{if } x_i\text{= 1} \\ h_0(t), & \text{if } x_i\text{= 0} \end{cases} \quad (5.7)$$

A Equação 5.6 é o modelo Proporcional Hazard de Cox para um única variável (WANG; LI; REDDY, 2019; KLEIN et al., 2016; KLEINBAUM; KLEIN, 2010; COLOSIMO; GIOLO, 2006). Sendo reescrita em termos gerais temos:

$$\lambda(t) = \lambda_0(t) \cdot g(\boldsymbol{\beta}\mathbf{x}) \tag{5.8}$$

onde g é uma função que tem que ser especificada de modo que $g(0) = 1$.

Baseado nessa coleção de método estatísticos, é possível estudar e avaliar o risco em que AS, FA e EC poderão ser diagnosticadas nos paciente; ou seja, predizer o risco de diagnóstico das doenças em função do tempo. Para tal, espera-se que a utilização do Modelo de Cox e o cálculo do risco de desenvolvimento das doenças em função do tempo através do *Hazard Ratio* produza essa informação;

- Implementação da *framework* aqui posposta em uma plataforma *mobile*, utilizando dados *online* (por exemplo: pressão sanguínea, batimentos cardíacos, entre outros) captados por microcontroladores (ISLAM et al., 2012; SüTő; ONIGA; ORHA, 2013; RAHAMAN et al., 2019). Assim sendo, que possibilite ao usuário o alerta do risco de desenvolvimento das doenças em seu dispositivo móvel de acesso diário (celular, *smartwacth*, entre outros), e recomendando procurar ajuda médica.

Referências

AAS, K.; JULLUM, M.; LØLAND, A. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *arXiv preprint arXiv:1903.10464*, 2019. Citado 3 vezes nas páginas 40, 41 e 43.

AFFIRM Investigators et al. Baseline characteristics of patients with atrial fibrillation: the affirm study. *American heart journal*, Elsevier, v. 143, n. 6, p. 991–1001, 2002. Citado na página 20.

ALMEIDA, G. C. d. Relatório técnico da metodologia senai para valoração e negociação de propriedade intelectual. 2019. Citado na página 27.

ALONSO, A. et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the charge-af consortium. *Journal of the American Heart Association*, Am Heart Assoc, v. 2, n. 2, p. e000102, 2013. Citado 2 vezes nas páginas 17 e 21.

AYUB, N. Í.; BACIC, M. J. Patentes: justificativas econômicas e seus efeitos sobre a inovação. *Economic Analysis of Law Review*, v. 10, n. 2, p. 153–172, 2020. Citado na página 26.

AZEVEDO, A. Data mining and knowledge discovery in databases. In: *Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics*. [S.l.]: IGI Global, 2019. p. 502–514. Citado 2 vezes nas páginas 27 e 72.

BACH, S. H. et al. Learning the structure of generative models without labeled data. *Proceedings of machine learning research*, NIH Public Access, v. 70, p. 273, 2017. Citado 2 vezes nas páginas 31 e 32.

BATBAATAR, E.; PHAM, V.-H.; RYU, K. H. Multi-task topic analysis framework for hallmarks of cancer with weak supervision. *Applied Sciences*, Multidisciplinary Digital Publishing Institute, v. 10, n. 3, p. 834, 2020. Citado na página 21.

BENJAMIN, E. J. et al. Impact of atrial fibrillation on the risk of death: the framingham heart study. *Circulation*, Am Heart Assoc, v. 98, n. 10, p. 946–952, 1998. Citado na página 68.

BIRUNDA, S. S.; DEVI, R. K. A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application*, Springer, p. 267–281, 2021. Citado 3 vezes nas páginas 35, 36 e 40.

BOJARSKI, M. et al. Explaining how a deep neural network trained with end-to-end learning steers a car. *arXiv preprint arXiv:1704.07911*, 2017. Citado na página 40.

BRAJER, N. et al. Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. *JAMA Network Open*, American Medical Association, v. 3, n. 2, p. e1920733–e1920733, 2020. Citado 2 vezes nas páginas 17 e 21.

BRISIMI, T. S. et al. Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. *Proceedings of the IEEE*, IEEE, v. 106, n. 4, p. 690–707, 2018. Citado na página 17.

BURGELMAN, R. A.; CHRISTENSEN, C. M.; WHEELWRIGTH, S. C. *Gestão estratégica da tecnologia e da inovação: conceitos e soluções*. [S.l.]: AMGH Editora, 2013. Citado na página 26.

CACHAY, S. R.; BOECKING, B.; DUBRAWSKI, A. End-to-end weak supervision. *Advances in Neural Information Processing Systems*, v. 34, 2021. Citado na página 31.

CARNEIRO, G.; FONTES, F. H.; TOGEIRO, S. M. G. P. Consequências metabólicas na saos não tratada. *Jornal Brasileiro de Pneumologia*, SciELO Brasil, v. 36, p. 43–46, 2010. Citado na página 70.

CARUANA, R. et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2015. p. 1721–1730. Citado na página 17.

CARVALHO, J. J. M. Antecedentes da doença coronária: os fatores de risco. *Arq. Bras. Cardiol*, p. 263–7, 1992. Citado na página 70.

CHEN, T. et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, v. 1, n. 4, 2015. Citado 2 vezes nas páginas 43 e 49.

CHEN, Z. et al. Knowledge discovery and recommendation with linear mixed model. *IEEE Access*, IEEE, v. 8, p. 38304–38317, 2020. Citado 4 vezes nas páginas 27, 31, 36 e 72.

CHOI, E. et al. Gram: graph-based attention model for healthcare representation learning. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2017. p. 787–795. Citado na página 17.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: Editora Blucher, 2006. Citado 3 vezes nas páginas 73, 74 e 75.

COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. Citado na página 74.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977. Citado na página 30.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Citado na página 39.

DU, G. et al. Joint imbalanced classification and feature selection for hospital readmissions. *Knowledge-Based Systems*, Elsevier, v. 200, p. 106020, 2020. Citado na página 17.

DU, G. et al. Towards graph-based class-imbalance learning for hospital readmission. *Expert Systems with Applications*, Elsevier, v. 176, p. 114791, 2021. Citado na página 17.

DUNN, O. J.; CLARK, V. A. *Basic statistics: a primer for the biomedical sciences*. [S.l.]: John Wiley & Sons, 2009. Citado na página 73.

EL-DIN, D. M. Enhancement bag-of-words model for solving the challenges of sentiment analysis. *International Journal of Advanced Computer Science and Applications*, Science and Information (SAI) Organization Limited, v. 7, n. 1, 2016. Citado na página 37.

ENDO, A. et al. Prediction model of in-hospital mortality after hip fracture surgery. *Journal of orthopaedic trauma*, LWW, v. 32, n. 1, p. 34–38, 2018. Citado na página 17.

FARIA, A. N. et al. Tratamento de diabetes e hipertensão no paciente obeso. *Arquivos Brasileiros de Endocrinologia & Metabologia*, SciELO Brasil, v. 46, p. 137–142, 2002. Citado na página 70.

FILHO, A. L. et al. Diretriz de fibrilação atrial. *Arquivos Brasileiros de Cardiologia*, SciELO Brasil, v. 81, p. 2–24, 2003. Citado na página 68.

GE, L.; MOH, T.-S. Improving text classification with word embedding. In: IEEE. *2017 IEEE International Conference on Big Data (Big Data)*. [S.l.], 2017. p. 1796–1805. Citado na página 38.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. *Data Mining*. [S.l.]: Elsevier Brasil, 2015. Citado 2 vezes nas páginas 27 e 28.

GUO, G. et al. Knn model-based approach in classification. In: SPRINGER. *OTM Confederated International Conferences"On the Move to Meaningful Internet Systems"*. [S.l.], 2003. p. 986–996. Citado na página 29.

GURUNATH, R. et al. A novel approach for linguistic steganography evaluation based on artificial neural networks. *IEEE Access*, IEEE, v. 9, p. 120869–120879, 2021. Citado 4 vezes nas páginas 18, 35, 36 e 38.

HALL, B. H.; ROSENBERG, N. *Economics of innovation*. [S.l.]: Elsevier, 2010. Citado na página 26.

HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. [S.l.]: Prentice Hall, 1999. (International edition). ISBN 9780132733502. Citado 3 vezes nas páginas 28, 30 e 34.

HEARST, M. A. et al. Support vector machines. *IEEE Intelligent Systems and their applications*, IEEE, v. 13, n. 4, p. 18–28, 1998. Citado na página 22.

HOHNLOSER, S. H. et al. Effect of dronedarone on cardiovascular events in atrial fibrillation. *New England Journal of Medicine*, Mass Medical Soc, v. 360, n. 7, p. 668–678, 2009. Citado na página 20.

HOLMQVIST, F. et al. Impact of obstructive sleep apnea and continuous positive airway pressure therapy on outcomes in patients with atrial fibrillation—results from the outcomes registry for better informed treatment of atrial fibrillation (orbit-af). *American heart journal*, Elsevier, v. 169, n. 5, p. 647–654, 2015. Citado na página 68.

HULME, O. L. et al. Development and validation of a prediction model for atrial fibrillation using electronic health records. *JACC: Clinical Electrophysiology*, American College of Cardiology Foundation Washington, DC, v. 5, n. 11, p. 1331–1341, 2019. Citado 3 vezes nas páginas 17, 21 e 68.

HYVÖNEN, E.; RANTALA, H. et al. Knowledge-based relation discovery in cultural heritage knowledge graphs. In: CEUR-WS. ORG. *Digital Humanities in Nordic Countries Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*. [S.l.], 2019. Citado na página 27.

ISLAM, M. M. et al. Microcontroller based health care monitoring system using sensor network. In: *2012 7th International Conference on Electrical and Computer Engineering*. [S.l.: s.n.], 2012. p. 272–275. Citado na página 75.

JUNGMANN, D. d. M.; BONETTI, E. A. Inovação e propriedade intelectual: guia para o docente. *Brasília: Senai*, 2010. Citado na página 27.

JUSTO, F. A.; SILVA, A. F. G. Aspectos epidemiológicos da fibrilação atrial. *Revista de Medicina*, v. 93, n. 1, p. 1–13, 2014. Citado na página 68.

KADHIM, A. I. et al. Feature extraction for co-occurrence-based cosine similarity score of text documents. In: IEEE. *2014 IEEE student conference on research and development*. [S.l.], 2014. p. 1–4. Citado na página 37.

KHURSHID, S. et al. Performance of atrial fibrillation risk prediction models in over four million individuals. *Circulation: Arrhythmia and Electrophysiology*, Am Heart Assoc, 2020. Citado na página 22.

KIM, D. et al. Multi-co-training for document classification using various document representations: Tf–idf, lda, and doc2vec. *Information Sciences*, Elsevier, v. 477, p. 15–29, 2019. Citado 4 vezes nas páginas 18, 35, 36 e 44.

KIM, J. K.; KANG, S. Neural network-based coronary heart disease risk prediction using feature correlation analysis. *Journal of healthcare engineering*, Hindawi, v. 2017, 2017. Citado 2 vezes nas páginas 22 e 68.

KLEIN, J. P. et al. *Handbook of survival analysis*. [S.l.]: CRC Press, 2016. Citado 2 vezes nas páginas 73 e 75.

KLEINBAUM, D. G.; KLEIN, M. *Survival analysis*. [S.l.]: Springer, 2010. Citado 3 vezes nas páginas 73, 74 e 75.

LALKHEN, A. G.; MCCLUSKEY, A. Clinical tests: sensitivity and specificity. *Continuing education in anaesthesia critical care & pain*, Oxford University Press, v. 8, n. 6, p. 221–223, 2008. Citado na página 49.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: *International conference on machine learning*. [S.l.: s.n.], 2014. p. 1188–1196. Citado 3 vezes nas páginas 18, 36 e 44.

LEE, E. T.; WANG, J. *Statistical methods for survival data analysis*. [S.l.]: John Wiley & Sons, 2003. v. 476. Citado na página 73.

LI, J.; FINE, J. P. Weighted area under the receiver operating characteristic curve and its application to gene selection. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 59, n. 4, p. 673–692, 2010. Citado na página 22.

LI, Y.-G. et al. A simple clinical risk score (c2hest) for predicting incident atrial fibrillation in asian subjects: derivation in 471,446 chinese subjects, with internal validation and external application in 451,199 korean subjects. *Chest*, Elsevier, v. 155, n. 3, p. 510–518, 2019. Citado 2 vezes nas páginas 17 e 21.

LIMA, I.; PINHEIRO, C. A.; SANTOS, F. A. O. *Inteligência artificial*. [S.l.]: Elsevier Brasil, 2016. v. 1. Citado na página 28.

LINZ, D. et al. Associations of obstructive sleep apnea with atrial fibrillation and continuous positive airway pressure treatment: a review. *JAMA cardiology*, American Medical Association, v. 3, n. 6, p. 532–540, 2018. Citado 2 vezes nas páginas 16 e 20.

LIP, G. Y.; BEEVERS, D. G. Abc of atrial fibrillation: history, epidemiology, and importance of atrial fibrillation. *Bmj*, British Medical Journal Publishing Group, v. 311, n. 7016, p. 1361, 1995. Citado na página 20.

LIP, G. Y. et al. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, Elsevier, v. 137, n. 2, p. 263–272, 2010. Citado 2 vezes nas páginas 17 e 21.

LIU, Q.; KUSNER, M. J.; BLUNSOM, P. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*, 2020. Citado na página 38.

LOH, L. C. Autocomplete: Dr google's "helpful" assistant? *Canadian Family Physician*, The College of Family Physicians of Canada, v. 62, n. 8, p. 622–623, 2016. Citado na página 72.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems 30*. [S.l.]: Curran Associates, Inc., 2017. p. 4765–4774. Citado 2 vezes nas páginas 40 e 50.

LUONG, M.-T.; PHAM, H.; MANNING, C. D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. Citado na página 39.

MAĆKIEWICZ, A.; RATAJCZAK, W. Principal components analysis (pca). *Computers & Geosciences*, Elsevier, v. 19, n. 3, p. 303–342, 1993. Citado na página 36.

MAHMOUDI, E. et al. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *bmj*, British Medical Journal Publishing Group, v. 369, 2020. Citado na página 17.

MANCINI, M. C.; ALOE, F.; TAVARES, S. Apnéia do sono em obesos. *Arquivos Brasileiros de Endocrinologia & Metabologia*, SciELO Brasil, v. 44, p. 81–90, 2000. Citado na página 70.

MANN, G. S.; MCCALLUM, A. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of machine learning research*, v. 11, n. 2, 2010. Citado na página 30.

MARQUES, R. L.; DUTRA, I. Redes bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações. *Coppe Sistemas–Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil*, 2002. Citado na página 72.

MARTINS, A. B.; TUFIK, S.; MOURA, S. M. G. P. T. Síndrome da apnéia-hipopnéia obstrutiva do sono. fisiopatologia. *Jornal Brasileiro de Pneumologia*, SciELO Brasil, v. 33, p. 93–100, 2007. Citado 2 vezes nas páginas 42 e 70.

MASSAROLI, L. C. et al. Qualidade de vida e o imc alto como fator de risco para doenças cardiovasculares: revisão sistemática. *Revista da Universidade Vale do Rio Verde*, v. 16, n. 1, 2018. Citado na página 70.

MCGOVERN, A. et al. Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, American Meteorological Society, v. 100, n. 11, p. 2175–2199, 2019. Citado na página 40.

MCLACHLAN, S. et al. Bayesian networks in healthcare: Distribution by medical condition. *Artificial Intelligence in Medicine*, Elsevier, v. 107, p. 101912, 2020. Citado na página 72.

MICHNIEWICZ, E. et al. Patients with atrial fibrillation and coronary artery disease–double trouble. *Advances in medical sciences*, Elsevier, v. 63, n. 1, p. 30–35, 2018. Citado na página 68.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. Citado 3 vezes nas páginas 10, 38 e 39.

MILLER, S.; GUINNESS, J.; ZAMANIAN, A. Name tagging with word clusters and discriminative training. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. [S.l.: s.n.], 2004. p. 337–342. Citado na página 30.

MIYASAKA, Y. et al. Secular trends in incidence of atrial fibrillation in olmsted county, minnesota, 1980 to 2000, and implications on the projections for future prevalence. *Circulation*, Am Heart Assoc, v. 114, n. 2, p. 119–125, 2006. Citado na página 20.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, Manole Ltda, v. 1, n. 1, p. 32, 2003. Citado na página 28.

MONTAVON, G.; SAMEK, W.; MÜLLER, K.-R. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, Elsevier, v. 73, p. 1–15, 2018. Citado na página 40.

MOVAHED MEHRTASH HASHEMZADEH, M. M. J. M.-R. Diabetes mellitus is a strong, independent risk for atrial fibrillation and flutter in addition to other cardiovascular disease. *International Journal of Cardiology*, v. 105, n. 3, p. 315–318, 2005. ISSN 0167-5273. Citado na página 16.

NAZARI, S. et al. A fuzzy inference-fuzzy analytic hierarchy process-based clinical decision support system for diagnosis of heart diseases. *Expert Systems with Applications*, Elsevier, v. 95, p. 261–271, 2018. Citado na página 18.

NIEMANN, H.; MOEHRLE, M. G.; FRISCHKORN, J. Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application. *Technological Forecasting and Social Change*, Elsevier, v. 115, p. 210–220, 2017. Citado na página 72.

OKADA, S.; OHZEKI, M.; TAGUCHI, S. Efficient partition of integer optimization problems with one-hot encoding. *Scientific reports*, Nature Publishing Group, v. 9, n. 1, p. 1–12, 2019. Citado na página 35.

OMS. 2022. Acessado em 01/04/2022. Disponível em: <https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1>. Citado na página 16.

OPAS/OMS. 2022. Acessado em 01/04/2022. Disponível em: <https://www.paho.org/pt/topicos/doencas-cardiovasculares>. Citado na página 16.

ORNELAS, C. et al. Relação entre doenças pulmonares obstrutivas e síndrome de apneia obstrutiva do sono. *Rev Port Imunoalergologia*, v. 27, n. 2, p. 115–25, 2019. Citado na página 70.

PICKHARDT, P. J. et al. Automated ct biomarkers for opportunistic prediction of future cardiovascular events and mortality in an asymptomatic screening population: a retrospective cohort study. *The Lancet Digital Health*, Elsevier, 2020. Citado na página 21.

PLATANIOS, E. A. et al. Learning from imperfect annotations. *arXiv*, p. arXiv–2004, 2020. Citado 2 vezes nas páginas 31 e 33.

RADFORD, A. et al. Language models are unsupervised multitask learners. *OpenAI blog*, v. 1, n. 8, p. 9, 2019. Citado na página 39.

RAHAMAN, A. et al. Developing iot based smart health monitoring systems: A review. *Rev. d'Intelligence Artif.*, v. 33, n. 6, p. 435–440, 2019. Citado na página 75.

RATNER, A. et al. Snorkel: Rapid training data creation with weak supervision. In: NIH PUBLIC ACCESS. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*. [S.l.], 2019. v. 11, n. 3, p. 269. Citado 7 vezes nas páginas 10, 30, 31, 32, 33, 34 e 67.

RATNER, A. et al. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, Springer, v. 29, n. 2, p. 709–730, 2020. Citado 2 vezes nas páginas 31 e 34.

RATNER, A. et al. Training complex models with multi-task weak supervision. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2019. v. 33, p. 4763–4771. Citado 7 vezes nas páginas 10, 31, 33, 34, 35, 44 e 45.

RATNER, A. J. et al. Data programming: Creating large training sets, quickly. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2016. p. 3567–3575. Citado 3 vezes nas páginas 33, 34 e 43.

REN, Y. et al. A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records. *BMC medical informatics and decision making*, Springer, v. 19, n. 2, p. 131–138, 2019. Citado na página 17.

RILOFF, E.; SHEPHERD, J. A corpus-based bootstrapping algorithm for semi-automated semantic lexicon construction. *Natural Language Engineering*, Citeseer, v. 5, n. 2, p. 147–156, 1999. Citado na página 30.

ROBERTO, J.; BRITO, J.; ROGÉRIO, R. Consenso brasileiro de doença pulmonar obstrutiva crônica (dpoc). *J Pneumol*, v. 26, n. Supl 1, p. 1, 2000. Citado na página 70.

RUSH, B.; CELI, L. A.; STONE, D. J. Applying machine learning to continuously monitored physiological data. *Journal of clinical monitoring and computing*, Springer, v. 33, n. 5, p. 887–893, 2019. Citado na página 18.

RUSH, B.; CELI, L. A.; STONE, D. J. Applying machine learning to continuously monitored physiological data. *Journal of clinical monitoring and computing*, Springer, v. 33, n. 5, p. 887–893, 2019. Citado na página 20.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence: a modern approach. Malaysia*. [S.l.]: Pearson Education Limited, 2016. Citado 3 vezes nas páginas 28, 29 e 30.

SALIBA, W. et al. Usefulness of chads2 and cha2ds2-vasc scores in the prediction of new-onset atrial fibrillation: a population-based study. *The American journal of medicine*, Elsevier, v. 129, n. 8, p. 843–849, 2016. Citado 2 vezes nas páginas 17 e 21.

SARZYNSKA-WAWER, J. et al. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, Elsevier, v. 304, p. 114135, 2021. Citado na página 39.

SETHY, A.; RAMABHADRAN, B. Bag-of-word normalized n-gram models. In: *Ninth Annual Conference of the International Speech Communication Association*. [S.l.: s.n.], 2008. Citado na página 37.

SHAPLEY, L. S. A value for n-person games. *Contributions to the Theory of Games*, v. 2, n. 28, p. 307–317, 1953. Citado 2 vezes nas páginas 40 e 41.

SHICKEL, B. et al. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, IEEE, v. 22, n. 5, p. 1589–1604, 2017. Citado 2 vezes nas páginas 18 e 22.

SHICKEL, B. et al. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, IEEE, v. 22, n. 5, p. 1589–1604, 2017. Citado na página 22.

SHORTLIFFE, E. H.; SEPÚLVEDA, M. J. Clinical decision support in the era of artificial intelligence. *Jama*, American Medical Association, v. 320, n. 21, p. 2199–2200, 2018. Citado na página 18.

SOARES, J. C. T. *Lei de patentes, marcas e direitos conexos: Lei 9,279, 14.05. 1996*. [S.l.]: Editora Revista dos Tribunais, 1997. Citado na página 26.

SOLIMENE, M. C.; RAMIRES, J. A. F. Indicações de cinecoronariografia na doença arterial coronária. *Revista da Associação Médica Brasileira*, SciELO Brasil, v. 49, p. 203–209, 2003. Citado na página 42.

STEELE, A. J. et al. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PloS one*, Public Library of Science San Francisco, CA USA, v. 13, n. 8, p. e0202344, 2018. Citado na página 68.

STEVENSON, I. H. et al. Prevalence of sleep disordered breathing in paroxysmal and persistent atrial fibrillation patients with normal left ventricular function. *European heart journal*, Oxford University Press, v. 29, n. 13, p. 1662–1669, 2008. Citado na página 68.

STIPP, M. A. C. et al. O consumo do álcool e as doenças cardiovasculares: uma análise sob o olhar da enfermagem. *Escola Anna Nery*, SciELO Brasil, v. 11, p. 581–585, 2007. Citado na página 70.

SZEGEDY, C. et al. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 2818–2826. Citado na página 17.

SüTő, J.; ONIGA, S.; ORHA, I. Microcontroller based health monitoring system. In: *2013 IEEE 19th International Symposium for Design and Technology in Electronic Packaging (SIITME)*. [S.l.: s.n.], 2013. p. 227–230. Citado na página 75.

TAYEFI, M. et al. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 13, n. 6, p. e1549, 2021. Citado na página 17.

TISON, G. H. et al. Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA cardiology*, American Medical Association, v. 3, n. 5, p. 409–416, 2018. Citado na página 20.

TIWARI, P. et al. Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation. *JAMA network open*, American Medical Association, v. 3, n. 1, p. e1919396–e1919396, 2020. Citado na página 17.

TOUTANOVA, K. et al. Representing text for joint embedding of text and knowledge bases. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. [S.l.: s.n.], 2015. p. 1499–1509. Citado 2 vezes nas páginas 18 e 35.

TRSTENJAK, B.; MIKAC, S.; DONKO, D. Knn with tf-idf based framework for text categorization. *Procedia Engineering*, Elsevier, v. 69, p. 1356–1364, 2014. Citado na página 38.

TUNG, P.; ANTER, E. Atrial fibrillation and sleep apnea: Considerations for a dual epidemic. *Journal of atrial fibrillation*, CardioFront, LLC, v. 8, n. 6, 2016. Citado na página 20.

VISCONTI, M.; WEIS. *The Valuation of Digital Intangibles*. [S.l.]: Springer, 2020. Citado na página 27.

WANG, J. et al. Classification of imbalanced data by using the smote algorithm and locally linear embedding. v. 3, 2006. Citado na página 49.

WANG, K. et al. Clinical and laboratory predictors of in-hospital mortality in patients with coronavirus disease-2019: A cohort study in wuhan, china. *Clinical infectious diseases*, Oxford University Press US, v. 71, n. 16, p. 2079–2088, 2020. Citado na página 17.

WANG, P.; LI, Y.; REDDY, C. K. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 6, p. 1–36, 2019. Citado 3 vezes nas páginas 73, 74 e 75.

WU, Y. et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. Citado na página 39.

XIAO, H. et al. Ssp: semantic space projection for knowledge graph embedding with text descriptions. In: *Thirty-First AAAI conference on artificial intelligence*. [S.l.: s.n.], 2017. Citado 3 vezes nas páginas 18, 35 e 36.

YADAV, S. S.; JADHAV, S. M. Detection of common risk factors for diagnosis of cardiac arrhythmia using machine learning algorithm. *Expert Systems with Applications*, Elsevier, v. 163, p. 113807, 2021. Citado na página 70.

ZAPPONE, A. et al. Model-aided wireless artificial intelligence: Embedding expert knowledge in deep neural networks for wireless system optimization. *IEEE Vehicular Technology Magazine*, IEEE, v. 14, n. 3, p. 60–69, 2019. Citado 2 vezes nas páginas 18 e 36.

ZHANG, X. S. et al. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. [S.l.: s.n.], 2019. p. 2487–2495. Citado na página 17.

ZHONG, J.; GAO, C.; YI, X. Categorization of patient disease into icd-10 with nlp and svm for chinese electronic health record analysis. In: *Proceedings of the 2018 International Conference on Artificial Intelligence and Pattern Recognition*. [S.l.: s.n.], 2018. p. 101–106. Citado 2 vezes nas páginas 44 e 45.