Universidade Fernando Pessoa
Faculdade de Ciência e Tecnologia

Pós doutoramento

PROPOSTA DE UM SISTEMA INTELIGENTE PARA PREDIÇÃO DO RISCO DE DOENÇAS CRÔNICAS NÃO TRANSMISSÍVEIS

Plataforma inteligente de predição de risco para detectar precocemente predisposições de um indivíduo desenvolver uma ou múltiplas doenças crônicas, usando Inteligência Artificial (IA)

Oberdan Santos da Costa

Relatório apresentado à Universidade Fernando Pessoa como requisito para o cumprimento do programa de pós doutoramento em Sistemas Inteligentes aplicado a área da saúde sob a supervisão do Prof. Dr. Luis Borges Gouveia.

Universidade Fernando Pessoa
2023

PROPOSTA DE UM SISTEMA INTELIGENTE PARA PREDIÇÃO DO RISCO DE DOENÇAS CRÔNICAS NÃO TRANSMISSÍVEIS

Oberdan Santos da Costa, sob supervisão do Prof. Doutor Luís Borges Gouveia

RESUMO

O presente trabalho, propõe desenvolver e validar um modelo de risco de condições de saúde para prever a probabilidade de um indivíduo desenvolver uma ou múltiplas Doenças Crônicas Não Transmissíveis (DCNT), antes que elas se manifestem. Inúmeras pessoas numa escala global correm o risco de desenvolver DCNT. Em concreto, os principais grupos de DCNT, incluem: doenças cardiovasculares, câncer, respiratórias crônicas e diabetes. Os grupos de doenças crônicas com características multifatoriais e não-infecciosas em sua origem têm sido motivo de crescente preocupação da sociedade e governos de todo o mundo, por colocar as pessoas em maior risco de complicações, chegando até mesmo a óbito, e colocando os sistemas de saúde em crise sistémica. A crise provocada pela COVID-19 acelerou o setor da saúde não só no sentido de repensar e reorientar a prestação de cuidados, mas também priorizar a integração da predição, prevenção e gestão de DCNT na população. Essa crise, mostrou também que os modelos preditivos e de prevenção e gestão de doenças crônicas em uso são insuficientes nas respostas as situações de saúde de condições crônicas. Uma abordagem multifatorial com fatores de riscos desencadeadores e verticalizados: Fatores de Risco Modificáveis (FRM) e Fatores de Risco Não Modificáveis (FRNM) é preferível para reverter essas situações de saúde e fornecer respostas específicas da probabilidade de um indivíduo desenvolver determinadas doenças crônicas. Nesse contexto, esse trabalho tem como objetivo principal "D*esenvolver uma Plataforma Inteligente de Predição do Risco de Doenças Crônicas (PIPRDC) com o propósito de apoiar os profissionais da saúde".* O PIPRDC tem como fundo pesquisas na literatura, consultas a médicos especialistas e um modelo Preditivo de Risco de DCNT. Utilizamos classificadores de modelos multi-label ET, RF e DT para prever dez tipos de DCNT simultaneamente. Entre os modelos do experimento, o MLC RF apresentou o melhor desempenho de precisão e F1-*score* com 96,16% e 90,48%, respectivamente.

Palavras-chave: Doenças crônicas, Predição, Prevenção, Covid-19, Risco, Modelo.

PROPOSAL OF AN INTELLIGENT SYSTEM FOR PREDICTING THE RISK OF NON-COMMUNICABLE CHRONIC DISEASES

Oberdan Santos da Costa, under the supervision of Prof. Luís Borges Gouveia

ABSTRACT

The present work proposes the development and validation of a health conditions risk model to predict the probability of a specific individual to develop one or multiple Chronic Non-Communicable Diseases (CNCD), before they manifest themselves. Countless people on a global scale are at risk of developing CNCD. Specifically, the main groups of CNCD include: cardiovascular diseases, cancer, chronic respiratory diseases and diabetes. The groups of chronic diseases with multifactorial and non-infectious characteristics in their origin have been a reason for growing concern of society and governments around the world, as they put people at greater risk of complications, even leading to death, and putting systems at risk. Both related with health and economic systemic crisis. The crisis caused by COVID-19 has accelerated the health sector not only to rethink and reorient the provision of care, but also to prioritize the integration of prediction, prevention and management of CNCD in the population. This crisis also showed that the predictive models and the prevention and management of chronic diseases in use are insufficient in responding to health situations of chronic conditions. A multifactorial approach with triggering and verticalized risk factors (Modifiable Risk Factors (FRM) and Non-Modifiable Risk Factors (FRNM) is preferable to revert these health situations and provide specific responses to the probability of an individual developing certain chronic diseases. As such, this work has as its main objective to “Developing an Intelligent Platform for Predicting the Risk of Chronic Diseases (PIPRDC) with the purpose of supporting health professionals.” The PIPRDC is based on research in the literature, consultations with medical specialists and a model CNCD Risk Predictive. We used ET, RF and DT multi-label model classifiers to predict up to nine types of CNCD simultaneously. Among the experimental models, the MLC RF showed the best performance of accuracy and F1-*score* with 96.16% and 90.48%, respectively.

Keywords: Chronic diseases, Prediction, Prevention, Covid-19, Risk, Model.

DEDICATÓRIA

Á minha maravilhosa esposa Selma Santos por seu apoio inabalável a todos instantes dessa e outras jornadas. E para minhas filhas Ana Luiza e Mariana, pela força e motivação.

AGRADECIMENTOS

Agradeço a Deus pela saúde, sabedoria e direções em todos os momentos da minha vida.

Ao meu supervisor Prof. Doutor Luís Borges Gouveia por compartilhar seu conhecimento e apoio em todos os instantes de construção deste trabalho.

A direção e equipe do Hospital-Escola da Fundação Fernando Pessoa, na pessoa do Dr. Bruno Soares pela disponibilidade dos dados, após a devida autorização da comissão de ética.

Aos familiares próximos por acreditarem em mim e no meu trabalho.

Tabela de conteúdo

1 INTRODUÇÃO

A crise provocada pela COVID-19 acelerou o setor da saúde não só para repensar e reorientar a prestação de cuidados, mas também para priorizar a predição, prevenção e gestão dos principais grupos de DCNT. Esses grupos de doenças com características multifatoriais e não-infecciosas em sua origem têm sido motivo de crescente preocupação da sociedade e governos de todo o mundo, por colocar as pessoas em maior risco de complicações, afetando a produtividade no trabalho, os custos de saúde, a renda familiar, e provocando desigualdade das condições de saúde entre a população.

DCNT são a principal causa de morte em todo mundo, em parte, deve à falta de ferramentas preditivas, práticas e de precisão. O *NCD Countdown 2030,* destaca que DCNT são as principais causa de morte e problemas de saúde, e são responsáveis por sete de dez mortes em todo o mundo. Dados da Organização mundial de Saúde (OMS) apontam que DCNT são responsáveis por mais de 70% de todas as mortes no mundo – o equivalente a 41 milhões de pessoas. A alta prevalência de doenças crônicas é um problema de saúde mundial, com 6,7 mil milhões de pessoas vivendo com doenças não transmissíveis, resultando em anos de vida perdidos substanciais (Naghavi et al, 2016). Pessoas nessas condições tem impacto profundo nos custos da saúde. De acordo com a (Organização Pan-Americana da Saúde, 2016), a epidemia de DCNT provocará um custo equivalente a US$ 21,3 trilhões em perdas econômicas nos países de renda baixa e média nas próximas duas décadas, valor próximo da soma dos produtos internos brutos (PIB) desses países em 2013 (US$ 24,5 trilhões). Isso, certamente, também deve ser decorrente da baixa taxa de detecção das predisposições de um indivíduo desenvolver determinada DCNT. Um estudo publicado no Pan American Journal of Public Health aponta que o só o custo da hipertensão, diabetes e obesidades chegou a R$ 3,45 bilhões em 2018 no sistema público de saúde brasileiro, elevando a sua carga financeira na economia. De acordo com (Rahimloo e Jafarian, 2016) predizer com mais precisão a condição dos pacientes é de extrema importância. Na busca de soluções para esse problema (Virani et al.,

2020) observa que a maioria das condições crônicas pode ser evitada por meio da implantação de uma plataforma de prevenção e gestão de doenças crônicas não transmissíveis. Essas plataformas, em geral têm como fundo um modelo, que são divididos em dois grupos, um para predição e outro para prevenção e gestão de DCNT destinados a identificar indivíduos com risco de desenvolver doenças crônicas. O grupo 1, compreende calculadora/instrumentos/modelos estatísticos validados e individualizados de predição de risco. Eles foram desenvolvidos para fornecer predição de risco individualizado de um indivíduo do sexo masculino/feminino desenvolver determinada doença crônica. Por característica, eles tendem a se concentrar em um amplo conjunto de fatores que desencadeiam estas doenças, que são denominados de fatores de risco. Esses modelos são: modelo Bach, LCDRAT, Framingham (ERF), QRISK-2 (ERQ), Reynolds*,* índice ADO, índice BODE, AUSDRISK, QDScore, FINDRISC e Cambridge Risk Score, KORA basic e DESIR equação clínica. O grupo 2, relaciona os principais modelos de prevenção e gestão de doenças crônicas, incluindo: CCM Expandido de Barr et al (2003), Ontario's Modelo de prevenção e gestão de doenças crônicas (CDPM) Framework (2007), modelo de atenção às condições crônicas (MACC) de (Mendes,2011). Esses modelos têm em comum as suas ampliações com base no modelo de atenção crônica *(Chronic Care Model)* – CCM e foram desenvolvidos como resposta às situações de saúde de alta prevalência de condições crônicas. Eles tornaram-se base para a reorganização dos sistemas de saúde, privilegiando a atenção primária, quanto ao planejamento e articulação das ações de combate a doenças crônicas. De modo geral, esses modelos direcionam esforços preventivos para doenças em seus estágios iniciais. Embora os modelos dos grupos 1 e 2 sejam robustos, eles são limitados nas suas aplicações, pois não fornecem respostas especificas para detectar predisposições de um indivíduo desenvolver múltiplas DCNT de forma integrada.

Agora, imagine-se um indivíduo ir a um posto de atenção primária de saúde, ter uma abordagem qualificada, personalizada, passar por uma triagem e depois de responder ao médico algumas questões pontuais, conhece os seus riscos de desenvolver DCNT, tem opções de tratamento mediante nível de urgência, procura uma especialista, se necessário e começar a fazer escolhas de estilos de vida mais saudáveis. Isso seria valioso para todos os indivíduos, médicos, gestores, sociedade e governos. O facto é que os sistemas atuais não foram projetados para uma obtenção

de diferentes informações, tão pouco têm capacidade de resposta para traçar perfis e detectar predisposições de um indivíduo desenvolver determinadas doenças crônicas, o que dificulta uma orientação médica mais precisa junto ao paciente. Neste contexto, até onde sabemos, não há uma estrutura abrangente na literatura que compile e harmonize dados de DCNT, desde a compreensão de domínios até a implantação do modelo. Portanto, há uma necessidade latente de modelos integrados e práticos para detecção das predisposições de um indivíduo desenvolver determinadas DCNT, que podem ser usados na atenção primária de saúde e em outros serviços de saúde.

A disposição de um modelo baseado em dados – forma hormônica de fácil interpretação é fundamental para auxiliar os médicos da atenção primária de saúde na avaliação da probabilidade de DCNT, bem como melhorar a precisão do diagnóstico antes de decidir sobre um procedimento invasivo, entre outros. Além disso, o modelo funcionará como um farol para governos, seguradoras e farmacêuticas nos seus planos preventivos e estratégias de saúde. Assim, prever DCNT usando um modelo preditivo de risco com características abrangentes é essencial e devem ser tratadas prontamente para evitar maiores complicações ou morte em todo o mundo.

Nesse contexto, este trabalho tem como objetivo principal "*desenvolver uma Plataforma Inteligente de Predição do Risco de Doenças Crônicas (PIPRDC) com o propósito de apoiar os profissionais da saúde na previsão precisa de uma pessoa desenvolver uma ou múltiplas DCNT, antes que elas se manifestem*". Os principais componentes do PIPRDC com seus respectivos módulos construtivos são: (C1) Grupos de DCNT, nomeadamente (G_DCNT), que diz respeito as doenças crônicas não transmissíveis e (C2) FRNM e FRM, base do Modelo do Preditivo de Risco (MPR), que combina uma matriz de dados, contribuindo para impulsionar a relações significativas da produção e transformação de dados em conhecimento derivando-os para predição de DCNT.

O estudo é relevante e tem o potencial de impactar positivamente no campo econômico – produtividade no trabalho e redução dos custos em assistência médica e no campo social – proporcionar benefícios de sobrevivência em tempo e igualdade das condições de saúde entre a população.

Espera-se que, dos resultados deste trabalho, se mostre um bom desempenho para prever com precisão DCNT, bem como permita que, as pessoas conheçam os seus riscos de desenvolver DCNT e façam escolhas de estilos de vida mais saudáveis; forneça uma alternativa de experiência de assistência médica integrada, ágil e personalizada na Atenção Primária de Saúde (APS). Esse trabalho, atende ao 3º item de compromisso da carta de Ottawa 1986 e ao 3º item dos Objetivos de Desenvolvimento Sustentável – ODS; e fornece respostas especificas e precisas da predição de DCNT para proporcionará benefícios de sobrevivência e ajudar a reduzir a alta taxa de mortalidade por DCNT.

1.1. Objetivo de Pesquisa

O objetivo principal da pesquisa é o desenvolvimento de uma Plataforma Inteligente de Predição do Risco de Doenças Crônicas (PIPRDC) com o propósito de apoiar os profissionais da saúde na previsão precisa de uma pessoa desenvolver uma ou múltiplas DCNT, ainda antes que elas se manifestem.

1.1.1 Objetivos específicos

Em complemento, para a pesquisa, foram considerados os seguintes objetivos específicos:

1. Examinar na literatura os principais modelos de prevenção e predição de doenças crônicas;
2. Consultar médicos especialistas;
3. Construir um modelo Preditivo de Risco de DCNT;
4. Desenvolvimento da PIPRDC para validar e testar a modelo Preditivo de Risco de DCNT.

2 FUNDAMENTAÇÃO TEÓRICA

Estudos empíricos existentes como os de (El_Jerjawi e Abu-Naser, 2018; Singh, Leavline e Baig, 2017; LIU et al., 2019; Deberneh e Kim, 2021; Tigga e Garg, 2020; Chun et al., 2021; Yang et al., 2020; Bharti et al., 2021; Akella e Akella, 2021; Aleksandrova et al., 2021; Syed et al., 2019; Wells et al., 2014 ;Pashayan et al., 2020; Aladwani et al., 2019; Vikas e Kaur, 2021; Matheson et al., 2018; Chaurasia et al., 2018; Albright et al., 2015; Barber et al., 2018) sobre fatores preditivos envolvendo DCNT, concentram-se principalmente nos fatores que têm influência significativa em alguma das doenças de um dos quatro grupos de DCNT, incluindo Doenças Cardiovasculares, Câncer, Doenças Respiratórias crônicas e Diabetes. Essas pesquisas, em geral focam-se na predição de único tipo de DCNT e são base de contribuições importantes para a seleção de atributos preditivos relacionados com arcabouços teóricos para a construção do modelo proposto neste trabalho. A maioria dos modelos foram desenvolvidos a partir de adaptações dos registros de dados dos pacientes compostos por grupos de variáveis independentes para prever uma variável dependente (desfecho).

Apesar de haver estudos empíricos que relatam razões pelas quais uma pessoa pode ser acometida de algum tipo de DCNT, ainda são poucos os estudos com foco na predição de mais de um tipo de DCNT. Por não contar com um número satisfatório de estudos, percebe-se ainda que muitos dos modelos existentes não foram validados.

1.2 Modelos empíricos de predição de DCNT e limitações

O modelo proposto por Chun et al. (2021) combina duas abordagens, sendo uma delas aprendizagem de máquina para prever o risco de acidente vascular cerebral em adultos Chineses. Eles usaram um sistema eletrônico administrado pelo entrevistador (questionário) para coletar dados sobre fatores sociodemográficos; fatores de estilo de vida (por exemplo, tabagismo, álcool, hábitos alimentares); histórico médico e medicação atual; atividade física e medidas físicas, incluindo altura, peso, circunferência do quadril e da cintura, bioimpedância, pressão arterial e

frequência cardíaca. Acrescentam ainda que a *Idade Avançada* associada a história prévia de DAC, hipertensão e tabagismo são presumidas pelos autores como atributos significativos na predição de risco de acidente vascular cerebral. Entretanto, o modelo contém algumas limitações, entre elas, que as equações de risco utilizadas não foram projetadas para implementação imediata na prática clínica. Além do mais, torna-se necessário um maior investimento financeiro para suporte do esforço de validação do modelo. O custo final para operacionalização pode tornar-se inviável devido a empregabilidade de recursos na implementação.

Usando um sistema de prontuário eletrônico (Yang et al., 2020) extraíram diversos atributos para construção de modelos. Após análise de regressão logística, eles constataram que 30 atributos estavam relacionados a DCV. Idade avançada; renda; tabagismo; consumo excessivo de álcool; obesidade; cintura grande; colesterol anormal; HDL-C baixo; FPG anormal e baixa capacidade de ação foram presumidas pelos autores como características significativas na predição de risco de doenças cardiovasculares. Quatro limitações estão presentes nesse estudo, são elas: Quantidade de atributos que se torna um obstáculo na implementação; dados insuficientes para treinar o modelo; falta de validação externa; e, por fim, o conjunto de dados utilizado está desequilibrado.

O modelo proposto por Hussan et al. (2022), usando fatores derivados de registros eletrônicos de saúde (EHR), utiliza 25 atributos e aprendizado de máquina para prever o risco de Câncer Colorretal (CCR). A renda por CEP, indicação da colonoscopia, sangramento gastrointestinal e quartis do índice de massa corporal foram presumidas como atributos de alto risco. As principais limitações nesse trabalho são a fonte de dados e dados desequilibrados.

Usando um conjunto de dados de diagnóstico de câncer de mama da Universidade de Wisconsin-Madison, Naji et al. (2021), selecionaram 11 atributos para construção de um modelo a partir da aplicação de cinco algoritmos de aprendizado de máquina para prever o risco de câncer de mama. As principais limitações do trabalho são: números restritos de instancias e testes em uma única fonte de dados.

O modelo proposto por Oyewo et al. (2020), usou um conjunto de dados obtido em *http://github.com/selva86/datasets/masters/prostate.csv*, extraíram 9 atributos

preditivos e combinaram com algoritmos de aprendizagem de máquina para prever o risco câncer de próstata. As principais limitações do trabalho são: número limitados de instancias e teste em uma única fonte de dados.

Usando um conjunto de dados criado pelo usuário *sta427ceyin* no site *data world*, a partir de 15 atributos, Nasser. (2019) desenvolveu um modelo para prever o risco de câncer de pulmão. O autor presumiu que fatores como idade, gênero e tosse associados a sintomas como dedos amarelos, ansiedade, doença crônica, fadiga, alergia, entrou outros, são significativos na predição de risco de câncer de pulmão. Algumas limitações desse trabalho incluem a não validação do modelo e testes em uma única fonte de dados.

O modelo proposto por Spathis e Vlamos (2019), usou dados de 132 pacientes de uma clínica no subúrbio de Thessaloniki-Grécia, foi combinado com técnicas de aprendizagem de máquina para prever o risco de asma e DPOC. Observaram que o tabagismo, o volume expiratório forçado e idade do grupo de pacientes estudados foram identificados como atributos significativos na predição de risco de asma. Por outro lado, o fluxo expiratório máximo, volume expiratório forçado, fumar e idade do grupo foram presumidas como atributos significativos na predição de risco de DPOC. As limitações incluem a não validação do modelo, testes em uma única fonte de dados e número limitados de instancias.

Usando um conjunto de dados do Registro Eletrônico de Saúde desidentificado de Optum (LI et al., 2021), extraíram 10 atributos para construir o modelo de previsão do risco de diabetes tipo 1. Fator HbA1c mais alta, glicose e hospitalização foram presumidos como atributos significativos na predição de risco de diabetes tipo 1. As limitações do modelo incluem dados incompletos e não validação do modelo.

O modelo proposto por Rani (2020) valeu-se de um conjunto de dados de diabetes disponibilizado em *https://www.kaggle.com/johndasilva/diabetes* para extrair 8 atributos considerados significativos e aplicou vários algoritmos de aprendizagem de máquina para prever o risco de diabetes tipo 2. Glicose, índice de massa corporal, idade, histórico familiar de diabetes e pressão arterial foram presumidas como atributos significativos na predição de risco de diabetes tipo 2. Algumas limitações do

modelo incluem números limitados de instancias, não validação do modelo e testes com uma única fonte de dados.

Entre os modelos empíricos analisados, a exceção do modelo proposto por Spathis e Vlamos (2019), é possível observar que a concentração de esforços está na previsão de uma única DCNT, entretanto assemelham-se pela limitação da falta de validação.

Esse conjunto de modelos fornecem uma estrutura abrangente sobre o qual se desenvolveu o modelo de predição do risco de DCNT.

1.3 Mineração de dados: Classificadores Multi-Label e Métricas de Avaliação

Em geral, duas técnicas de ML são as mais utilizadas, a saber: Supervisionada e Não Supervisionada. Neste estudo, adotaremos a técnica supervisionada com ênfase na Multi-Label Classification (MLC) e nas suas métricas de avaliação. A técnica supervisionada é parte da estrutura de sistemas classificadores de aprendizado que são baseados em regras (Urbanowicz e Moore, 2009). Essa técnica desenvolve aproximações ótimas localmente, como estimativas de classificação precisas ou aproximações de função.

Na técnica de MLC, os dados podem pertencer a mais de um rótulo simultaneamente. Assim, ao fazer previsões, uma determinada entrada pode pertencer a mais de um rótulo. Uma tarefa MLC começa a partir de um espaço de atributos d-dimensional $X \subseteq \mathbb{R}^{d}$ e um conjunto de rótulos Y = $y^1, y^2, \ldots, yq$, onde q>1. Seguindo o formato tradicional de classificação, na MLC, um modelo aprende a partir de um conjunto de dados de N instâncias D = {(XI, Yi), i = 1,...,N}. Para a i-enésima instância, xi ∈ X é seu vetor de atributos (d-dimensional) e Yi ⊆ Y é seu rótulo definido. De acordo com García et al. (2021), se yj ∈ Yi, ou seja, se a instância xi tiver associado o rótulo yj, yj é dito ser relevante para xi. Senão, yj é dito ser irrelevante para xi, ∀i = 1, 2,...,N, j = 1,2,...,q. Assim, os MLC têm se mostrado uma grande promessa para aprender uma função a partir de um conjunto de instâncias em várias aplicações, incluindo categorização de texto, classificação de imagens, recuperação de informações e estão se expandindo para outros campos, incluindo diagnóstico médico, bioinformática, etc.

A classificação de rótulos estuda o problema de aprender um mapeamento de instâncias para classificações em um conjunto predefinido de rótulos (Furnkranz et al., 2008). Nos últimos anos, a popularidade do MLC vem aumentando devido à sua capacidade de resolver uma variedade de problemas com base em uma vetorização e em tempo real. No seu estudo, Kassim et al (2021) destacam três abordagens gerais que são usadas para lidar com problemas de MLC, incluindo Métodos de Transformação de Problemas (MTPs), Métodos de Adaptação de Algoritmos (MAAs) e Métodos Ensemble (EMs). Os MTPs transformam um conjunto de dados de vários rótulos em um conjunto de dados de rótulo único usando diferentes métodos de transformação, como Least Frequent Label (LFL), Most Frequent Label (MFL) ou escolhendo qualquer rótulo aleatoriamente (Lee et al., 2016). Ao contrário dos MTPs, os MAAs lidam com o problema de aprendizado multi-label adaptando alguns algoritmos de ML diretamente para o cenário de classificação multi-rótulo. Os EMs requerem um classificador base do método de adaptação do algoritmo ou método de transformação do problema e parâmetros relevantes do método (Kassim et al., 2021).

Neste estudo, métodos *ensemble* desenvolvidos em cima das técnicas de adaptação de algoritmos, incluindo Classificadores *Random Forest* (RF), *Extra Trees* (ET) e Decisão *Tree* (DT) serão a nossa base de trabalho. Eles fazem um voto ponderado das suas previsões e em geral são usados para aumentar o desempenho preditivo e resultados de alta precisão. Utilizamos versões especializadas dessas fórmulas padrão para prever os rótulos de cada classe com um algoritmo de classificação separado.

RF é um método *ensemble* que se baseia na construção de vários classificadores independentes de árvores de decisão em diferentes subconjuntos do conjunto de dados. Ele considera a combinação (geralmente a média) da saída de cada classificador independente para melhorar o desempenho na produção de previsões gerais (Kouchaki et al., 2020). O uso de vários rótulos (ou seja, todas as DCNT simultaneamente, em vez de considerar cada uma independentemente) pode reduzir o tempo de treinamento, pois apenas um modelo é aprendido, e o desempenho preditivo pode ser aumentado (Evgeniou e Pontil, 2004) devido à correlação de aprendizado entre as entradas e o múltiplas saídas. O modelo RF pode ser estendido para aprender e prever vários medicamentos simultaneamente, considerando uma pontuação conjunta índice de Gini (Equação 1) em todos os medicamentos

considerados (Faddoul et al., 2012). Para eles, especificamente em cada árvore de decisão, para cada par (f, x) de uma característica f (mutação) e um valor x (isolado) com um rótulo y (fenótipo de resistência) no nó (t):

$$\text{Gini index, } GI_j\,(t, f, x) = \sum_{y \in Y} GIy\,(t, f, x) \qquad (1)$$

Onde Y é o número de rótulos e GI_j e GI_y são os índices de Gini conjuntos e por rótulo, respectivamente. O objetivo é minimizar a Equação (1) e, portanto, (f, x) é selecionado para separar melhor (definido por um índice de Gini de junta inferior) os dados em cada nó na árvore.

As ET são uma variante de uma árvore de decisão aleatória em várias subseções do conjunto de dados e calcula sua média para melhorar a precisão e o controle da previsão excesso de convite. Uma abordagem de aprendizado do conjunto conhecida como classificador de árvore altamente aleatório combina os resultados de classificação de várias árvores de decisão não conectadas reunidas em uma “floresta” para obtenção de um resultado. Quando aplicada, a primeira amostra de treinamento é usada para construir cada árvore de decisão na floresta. Esse processo permite que a árvore de decisão decida qual recurso usar para particionar os dados com base em um critério matemático após receber uma amostra aleatória de k recursos do conjunto de recursos em cada nó de teste (geralmente, em um nó Gini índice). A previsão final é estabelecida usando a votação por maioria por conta de classificação.

O funcionamento do ET é diferente de outros métodos de *ensemble* baseados em árvore de decisão. Ele divide o nó aleatoriamente escolhendo pontos de corte, o que diminuirá a variância melhor do que outras estratégias de randomização (Juan et al., 2021). Baseado na divisão aleatória, o tempo de execução do ET é mais rápido. Devido à eficiência computacional, o algoritmo *Extratrees* tem aplicações massivas para classificação e regressão (Ampomah et al., 2020).

As DT vêm sendo constantemente usadas em pesquisas operacionais, particularmente em análise de decisão para identificar estratégias com maior probabilidade de atingir um objetivo. Esse é um típico MLC, que é representado como

um vetor x de d valores de atributo, desenhado para um domínio de entrada, dado um conjunto finito de rótulos predefinidos. Um dos pontos fortes desse algoritmo na solução de problemas multi-label estar na modificação da fórmula (Equação 2) para o cálculo de entropia. Segundo Nareshpalsingh e Modi (2017), a entropia modificada soma todas as entropias para cada rótulo individual.

$$\text{Entropia (D)} = \sum_{i=1}^{q} -p_j \, log_2 \, p_j \; - \; (1-p_j) \, log_2 \, (1-p_j) \qquad (2)$$

A propriedade chave do ML-DT é sua eficiência computacional:

Onde D é o conjunto de instâncias no conjunto de dados e pj é uma fração de instâncias em D que pertence ao rótulo j.

Trata-se de um modelo de decisões e suas possíveis consequências, incluindo resultados de eventos aleatórios, custos de recursos e utilidade (Nijil e Mahalekshmi, 2018).

Em classificadores *multi-label*, os dados podem pertencer a mais de um rótulo simultaneamente. Assim, as previsões para cada instância são um conjunto de rótulos, e a avaliação de desempenho dos classificadores pode ser calculada com base na pontuação média de uma métrica de avaliação ou comparando diretamente as pontuações de cada classe. Na literatura, existem várias métricas para avaliação de modelos de classificação *multi-label*. Em geral, essas métricas de avaliação são divididas em dois grupos, sendo elas: métricas baseadas em exemplos e métricas baseadas em rótulos. Para avaliar o desempenho de classificadores *multi-label* é essencial considerar medidas de avaliação múltiplas e contrastantes devido aos graus adicionais de liberdade que o *multi-label* configuração introduz (Madjarov et al., 2012). Em nossos experimentos, usamos medidas de avaliação baseadas em exemplos que têm sido sugeridos por estudos anteriores (Elkafrawy et al., 2015) e são definidos na Tabela 1, tais como: perda de Hamming, acurácia, precisão, recall e F1-*score*.

Tabela 1 – Métricas de avaliação

Métricas de avaliação para MLC		
Métricas	Base exemplo	Descrição
Perda de Hamming	Perda_hamming(h) $=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{Q}\lvert h(x_i)\Delta y_i\rvert$	Mede o quão bem o classificador prevê cada um dos rótulos, calculando a média das amostras e depois dos rótulos
Acurácia	Acurácia(h) $=\frac{1}{N}\sum_{i=1}^{N}\lvert\frac{h(x_i)\cap y_i}{h(x_i)\cup y_i}\rvert$	Performance geral do modelo. Definido como a proporção de previsões corretas
Precisão	Precisao(h) $=\frac{1}{N}\sum_{i=1}^{N}\frac{\lvert h(x_i)\cap y_i\rvert}{\lvert y_i\rvert}$	proporção de previsões corretas entre todas as previsões para uma determinada classe
Recall	Recall(h) $=\frac{1}{N}\sum_{i=1}^{N}\frac{\lvert h(x_i)\cap y_i\rvert}{\lvert h(x_i)\rvert}$	Também chamado de sensibilidade. proporção de exemplos de uma classe específica que foram previstos pelo modelo como pertencentes a essa classe.
F1-*score*	F1$=\frac{1}{N}\sum_{i=1}^{N}\frac{2x\lvert h(x_i)\cap y_i\rvert}{\lvert h(x_i)\rvert+\lvert y_i\rvert}$	Mede uma média ponderada de precisão e recall, onde ambas têm o mesmo impacto na pontuação.

Fonte: Elaborado pelos autores

Os métodos médias micro *average*, macro *average*, *weighted* e *sample average*, usados neste trabalho são parâmetros utilizados pela métrica precisão, recall e F1-*score*. Em geral esses parâmetros são utilizados nas classificações *multi-classe* ou *multi-label*. De acordo com Li et al. (2017), Macro de precisão significa uma concordância média por classe dos rótulos de classe de dados com os de um classificador fornecido. Eles destacam ainda que a macro de *recall* tem uma eficácia média por classe de um classificador para identificar rótulos de classe e que a macro F1-*score* fornece os relacionamentos entre os rótulos positivos e os fornecidos por um classificador com base na média por classe.

3 CONSTRUÇÃO DO MODELO PREDITIVO DE RISCO (MPR)

O MPR é suportado por um fundo da literatura, consultas a médicos especialistas e ampliação das várias características dos modelos empíricos aqui apresentados. A sua estrutura é formada por três conjuntos de elementos abrangentes que são teorizados conforme ordenação (ver Figura 1). O primeiro conjunto de elementos está focado nos fatores preditores não modificáveis. O segundo conjunto de elementos centra-se nos fatores preditores modificáveis. Por fim, o terceiro conjunto de elementos apresenta os resultados das probabilidades das DCNT. Teoriza-se que à medida que os fatores preditores modificáveis estão equilibrados positivamente, sugere que a pessoa se encontra em condições saudáveis de saúde. Contrariamente, ou seja, identificando alterações negativas, isto é, desequilíbrio em algum dos fatores não modificáveis conjuntamente com algum dos fatores modificáveis, ocorrerá uma instabilidade, que pode resultar na probabilidade, ou não, de uma pessoa desenvolver uma ou múltiplas doenças crônicas.

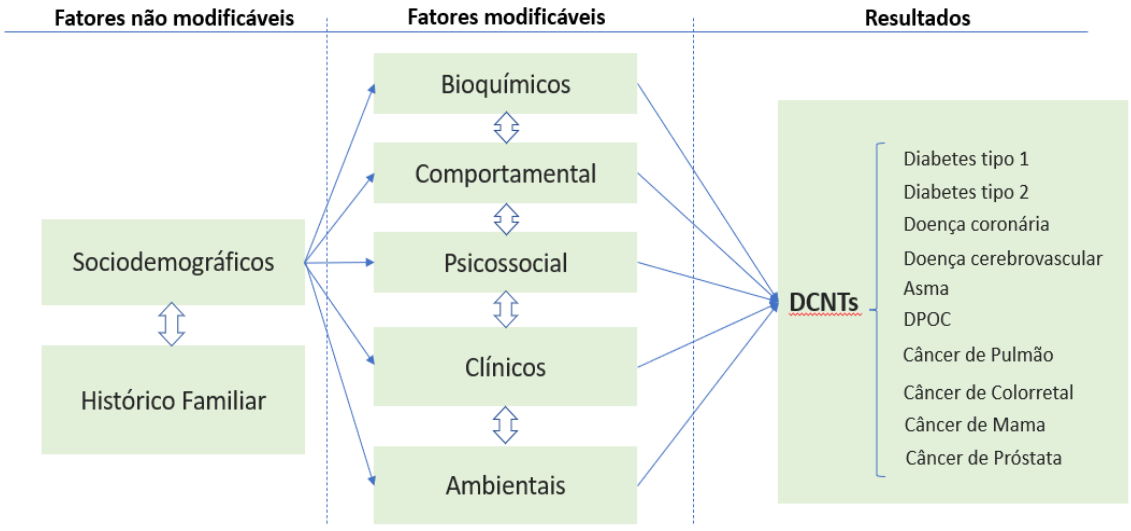


Figura 1 – Modelo Preditivo de Risco
Fonte: Elaborado pelos autores

Os fatores preditores não modificáveis do modelo tem em sua estrutura um conjunto formado pelos fatores sociodemográficos e histórico familiar. Eles foram tomados dos estudos de: (Deberneh e Kim, 2021; Matheson et al., 2018; Queiroz et al., 2021; Akella Akella, 2020; Syed et al., 2019; Wells et al., 2014; Chaurasia et al.,

2018; Albright et al., 2015; Barber et al., 2018; Pasquelli et al., 2020). Fatores preditores não modificáveis são caracterizados por não serem controlados. Alguns exemplos são: idade, sexo, história familiar e raça ou etnia. Em geral, esses fatores são significativos para prever a maioria das DCNT.

Os fatores preditores modificáveis do modelo apresenta em sua estrutura uma composição formada por fatores bioquímicos, comportamentais, psicossocial, clínicos e ambientais. a) Fatores bioquímicos, incluindo pressão arterial, glicemia, hemoglobina, colesterol, triglicerídeos, PSA etc., quando isolados ou combinados e desequilibrados desempenham papéis-chave em influenciar a ocorrência de diabetes, doenças cardiovasculares, câncer de colorretal e próstata (Liu et al., 2019; Yang et al., 2020; Theerthagiri, 2021; Deberneh and Kim, 2021; Geetha et al., 2021). Nesse sentido (Benjamin et al., 2019) corroboram destacando que muitos fatores podem causar cardiopatia, como mudanças dinâmicas no estilo de vida, tabagismo, hábitos alimentares, falta de atividade física, obesidade, diabetes e fatores bioquímicos como da pressão arterial ou glicemia; B) Fatores comportamentais como baixo consumo de frutas/verdura (vegetais), tabagismo, consumo de bebidas alcoólicas, inatividade física etc., quando isolados ou combinados e desequilibrados desempenham papéis-chave em influenciar a ocorrência de diabetes, doenças respiratórias, cardiovasculares, câncer de colorretal, mama, próstata e de pulmão (Chuarasia et al., 2018; Matheson et al. 2018; Ahmad e Mayya, 2020; Aleksandrova et al., 2021; Chun et al., 2021; Deberneh e Kim, 2021). Esses fatores afetam fortemente a sobrevida. Para a (WHO, 2005) o somatório de comportamentos de risco está associado com a diminuição da expectativa de vida; c) Fator psicossocial como estresse (stresse), quando isolado ou combinado, desempenha papéis-chave em influenciar a ocorrência de diabetes tipo 2, doenças respiratórias e cardiovasculares (Toskala e Kennedy, 2015; Akella e Akella, 2020; Tigga e Garg, 2020). Nesse sentido (Santos et al., 2021) relatam que fatores psicológicos, incluindo estresse no trabalho ou na vida familiar coletivamente, podem desempenhar um papel no desenvolvimento de doenças cardiovasculares; d) Fatores clínicos, incluindo tosse afetada, respiração afetada, dor no peito, fadiga, idade tardia da primeira gravidez, alteração da densidade mamaria etc., quando isolados ou combinados desempenham papéis-chave em influenciar a ocorrência de doenças respiratórias, cardiovasculares, câncer de pulmão e de mama (Mccoy et al., 2006; Giardiello et al., 2019; Vikas e Kaur,2021). Nesse sentido (Al-Hajj

et al., 2003) corroboram destacando que a idade tardia da primeira gravidez, a menopausa e a idade precoce da menarca (primeira menstruação ocorre antes dos oito anos de idade) estão ligadas a um aumento considerável no desenvolvimento do câncer de mama; e) Fatores ambientais como poluição no seu local de residência, poluição no seu local de trabalho e riscos ocupacionais, quando isolados ou combinados, desempenham papéis-chave em influenciar a ocorrência de doenças respiratórias e câncer de pulmão (Matheson et al., 2018; Vikas e Kaur, 2021). Estimativas globais sugerem que a poluição ambiental externa (*outdoors*) cause 1,15 milhões de óbitos em todo o mundo (correspondendo a cerca de 2% do total de óbitos) e seja responsável por 8,75 milhões de anos vividos a menos ou com incapacidade (WHO, 2009). Já a poluição no interior dos domicílios cause aproximadamente 2 milhões de óbitos prematuros e 41 milhões de anos vividos a menos ou com incapacidade (Oberg et al.,2011).

4 PLATAFORMA INTELIGENTE PROPOSTA

A Figura 2 apresenta um esquema associado com a plataforma inteligente desenvolvida, associada com o experimento realizado.

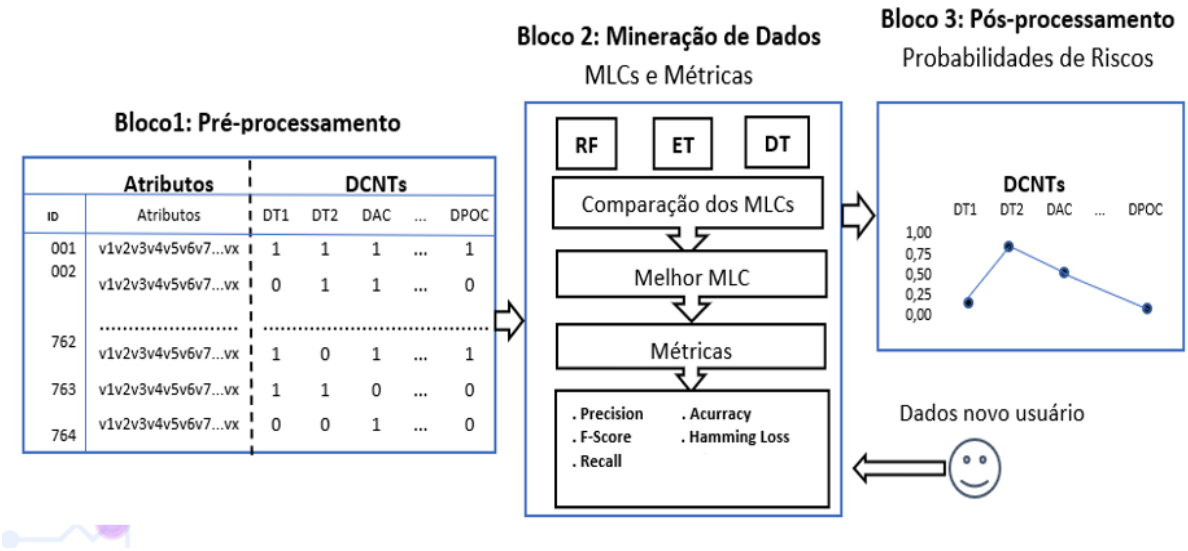


Figura 2 – Diagrama modular do PIPRDC
Fonte: Elaborada pelos autores

Pelo experimento apresentado na Figura 2, relatamos que o Bloco 1, fornece uma relação estruturada tipo matriz entre fatores com seus respectivos atributos e DCNT. O total de 38 atributos utilizados nessa matriz são derivados de um Modelo de Preditivo de Risco (MPR) que consiste em sete fatores preditores, incluindo sociodemográficos e histórico familiar bioquímicos, comportamentais, psicossocial, clínicos e ambientais.

No bloco 2, utilizamos Aprendizagem *Multi-Label* (AML), ou seja, vários *Multi-Label Classification* (MLC), incluindo *Random Forest* (RF), ExtraTreesClassifier (ET) e *Decision Tree* (DT). A AML além de inovadora, fornece uma solução potencial associada para tais desafios. Ela é uma técnica de classificação importante se cada amostra no conjunto de dados estiver devidamente associada a vários rótulos (por exemplo, idade, histórico familiar, álcool, etc.; a várias DCNT) e se houver correlações entre os rótulos. Em vez de considerar um atributo a cada DCNT individualmente, a

técnica *multi-label* aprende um único modelo para todas as DCNT e faz uma previsão no nível da amostra apresentada. Este método está mais próximo da realidade clínica, onde em geral os fatores de risco associados a DCNT não são tipicamente independentes uns dos outros. A co-ocorrência de atributos do MPR é especialmente comum em DCNT, uma vez que os esquemas padrão exigem que sejam usados em conjunto. Após comparação dos MLC, o melhor é aplicado e checado as medidas de desempenho de generalização do classificador. Por fim, o Bloco 3, trata da avaliação de padrões dos dados, ou seja, consistência útil e apresentação do conhecimento descobertos para um indivíduo por cada vez.

5 METODOLOGIA

O objetivo principal deste trabalho foi desenvolver uma Plataforma Inteligente de Predição do Risco de Doenças Crônicas (PIPRDC) com o propósito de apoiar os profissionais da saúde na previsão precisa de uma pessoa desenvolver uma ou múltiplas DCNT, antes que elas se manifestem.

Do ponto vista dos seus objetivos, trata-se de um trabalho com múltiplos objetivos complementares entre si e que levaram à definição de uma estrutura geral de pesquisa organizada em quatro seções, a saber: (1) examinar na literatura os principais modelos de prevenção e predição de doenças crônicas; (2) consulta a médicos especialistas; (3) construção do Modelo Preditivo de Risco de DCNT; e (4) desenvolvimento da PIPRDC para Validação e teste do modelo preditivo de risco.

5.1 Examinar na literatura os principais modelos de prevenção e predição de doenças crônicas

Nesta seção fizemos uma revisão abrangente da literatura de trabalhos relacionados ao uso de *Machine Learning* (ML) para predição DCNT. O resultado direcionou a obtenção de linhas de base para identificar lacunas existentes e auxiliou na proposta de uma solução.

5.2 Consulta a médicos especialistas

Aproximadamente 2000 médicos especialistas de todo o Brasil e Portugal foram consultados sobre um tipo de DCNT dentro de sua especialidade por meio de uma pesquisa encaminhada via e-mail. Os vários grupos de médicos especialidades consultadas, incluem cardiologistas, endocrinologistas, pneumologistas, oncologistas e mastologista. Os especialistas de suas respectivas áreas responderam à pesquisa. Juntos, eles, somam 365 anos de experiência no campo da medicina.

O foco da pesquisa junto aos especialistas foi identificar fatores e variáveis preditivas significativas de risco das dez principais DCNT (Diabetes tipo 1 e 2,

Doenças cardiovasculares (DAC e AVC), Doenças respiratórias (Asma e DPOC) e Câncer de colorretal, mama, próstata e de pulmão) em adultos e idosos do sexo masculino e feminino. A estrutura da pesquisa é formada por fatores de riscos modificáveis e não modificáveis, e tem como fundo estudos nacionais e internacionais.

5.3 Construção do Modelo Preditivo de Risco de DCNT

O MPR é suportado por um fundo da literatura, consultas a médicos especialistas e ampliação das várias características dos modelos empíricos aqui apresentados. A sua estrutura é formada por três conjuntos de elementos abrangentes que são teorizados conforme ordenação no item 3 deste trabalho.

5.4 Desenvolvimento da PIPRDC para validação e teste do modelo

Para o desenvolvimento da PIPRDC usamos como fundo o *framework* básico *Knowledge Discovery in Database* (KDD) e o adaptamos de (Fayyad et al., 1996), nomeadamente KDD_AZ. Ele consiste na combinação de ponta a ponta de métodos e ferramentas estatísticas, inteligência artificial, banco de dados e visualização para encontrar padrões válidos e úteis que gerem conhecimento. A sequência de três blocos do processo KDD_AZ compreende: pré-processamento, mineração de dados e pós-processamento, cada uma com as suas respectivas tarefas e fases de operação.

Bloco 1 – Pré-processamento: Compreende cinco tarefas, incluindo obtenção de dados, seleção de dados, preparação de dados, exploração de dados e transformação de dados.

Fonte de dados

Utilizamos características demográficas, dados clínicos, laboratoriais, etc., obtidos de dados de registro eletrônico, de janeiro de 2015 a junho de 2022, para pacientes admitidos em consulta e internações no Hospital-Escola da Universidade Fernando Pessoa (HE-UFP). Este estudo foi aprovado pela Comissão de Ética para Saúde do HE-UFP. Um total de 852.542 linhas de dados não estruturados foram fornecidas com os mais variados pacientes e tipos de doenças. Considerando o foco

para DCNT, foram elegíveis para este estudo 892 pacientes. Essa redução de elegíveis deu-se em função do total de 38 recursos/atributos nos dados brutos do registro eletrônico de saúde. Além disso, para recursos com poucos valores faltantes no conjunto de dados em alguns pacientes, adotamos o valor médio e modo, bem como técnicas anteriores e próximos para preencher os indicadores contínuos e discretos, respectivamente. As variáveis demográficas dos pacientes com as dez DCNT estão resumidas na Tabela 2.

Tabela 2 – Apresentação de características demográficas

	Distribuição de pacientes com complicações por DCNT									
	Diab_1	Diab_2	AVC	DAC	Asma	DPOC	ca_pulmão	CCR	ca_mama	ca_prostata
Total	12	690	118	100	116	124	23	20	36	44
Idade (anos)										
18-24	1	7	0	0	11	1	0	0	0	0
25-34	1	10	1	0	8	3	0	0	0	0
35-44	0	24	1	1	16	1	0	0	1	0
45-54	1	54	4	4	19	3	0	2	4	3
55-64	4	74	14	6	22	5	1	4	3	8
≥ 65	5	521	98	89	40	111	22	14	28	33
Sexo										
Masculino	6	308	60	67	39	70	11	13	0	44
Feminino	6	382	58	33	77	54	12	7	36	0
IMC										
IMC categoria										
≤ 18.4	5	7	0	1	1	0	0	0	0	0
18.5 – 24.9	3	67	7	14	4	10	0	0	1	13
25 – 29.9	3	310	69	42	72	74	20	11	13	12
30 - 34.9	1	195	28	26	24	26	2	9	20	15
≥ 35	0	111	14	17	12	14	1	0	2	4

Diab_1: Diabetes tipo 1; Diab_2: Diabetes tipo 2; AVC: Acidente Vascular Cerebral; DAC: Doença Arterial Coronariana; DPOC: Doença Pulmonar Obstrutiva Crônica; ca_pulmão: Câncer de pulmão; CCR: Câncer Colorretal; ca_mama: Câncer de Mama; ca_prostata: Câncer de Próstata; IMC: Indice de Massa Corporal.

Como pode ser visto na Figura 2, há um desequilíbrio de classe no conjunto de dados para alguns rótulos. Quando a observação em uma classe é maior do que em outras classes, existe um desequilíbrio de classe. Isso ocorre em função de algumas doenças que não têm a mesma frequência de registro, ou seja, algumas doenças ocorrem mais que outras. Temos duas classes em cada rótulo, sendo pacientes com e sem DCNT. Portanto, para aliviar o nível de desequilíbrio de classe para cada rótulo minoritário, ou seja, menos frequente, adotaremos métodos *Ensemble*, pois consta de

um conjunto de algoritmos mais robustos e complexos pautado em uma hierarquia de perguntas. Forçando assim, ambas as classes a serem endereçadas.

Seleção de recursos

O processo de seleção de recursos é uma etapa relevante, pois dependendo de como executado pode-se ter implicações na eficiência do algoritmo, tempo computacional e complexidade. Em se tratando de aprendizado de máquina e na precisão do algoritmo, essa é uma etapa crítica e altamente dependente desse processo. A estrutura proposta da PIPRDC contém 38 recursos significativos nativo do MPR conforme mostrado no item 3. A Tabela 3 apresenta um resumo dos recursos. O uso do MPR visa contribui para redução de *overfitting* e uma melhor precisão dos resultados de predição.

Tabela 3 – Apresentação dos recursos do MPR

Recursos do MPR		
1.idade	14.habitos alimentares	27.psa total
2.sexo	15.atividade física	28.pressão arterial sistólica
3.raça	16.tabagismo	29.pressão arterial diastólica
4.menarca precoce	17.alcool	30.poluição residencial
5.idade primeiro parto	18.IMC	31. poluição trabalho
6.duração amamentação	19.círculo da cintura	32.respiração afetada
7.número de gestação	20.espessura da pele	33.tosse afetada
8.idade menor pausa	21.dor no peito	34.alergia
9.reposição hormonal	22.estresse	35.sibilização
10.histórico familiar de diabetes	23.hemoglobina glicada	36. fadiga
11.historico familiar de DCV	24.glicose	37.dificuldade de engolir
12.historico familiar de infecções respiratórias	25.colesterol total	38.alteração densidade da mama
13.historico familiar de câncer	26.triglicerideos	

Bloco 2 – *Data Mining* (DM) é o núcleo do modelo KDD_AZ e consiste em um processo contínuo em que algoritmos inteligentes são aplicados de acordo com os requisitos e particularidade de cada técnica de ML para identificar padrões e conhecimento. Neste estudo adotamos os classificadores *Random Forest* (RF), *Extra Trees* (ET) e Decisão *Tree* (DT) como base de trabalho. Para tanto, utilizou-se *o* programa/linguagem de programação Python versão 3.9 conjuntamente com os

aplicativos Pandas versão *1.5.2*, Numpy versão 1.24.0 e a biblioteca *sklearn.esemble e* MS Excel para demonstração dos dados.

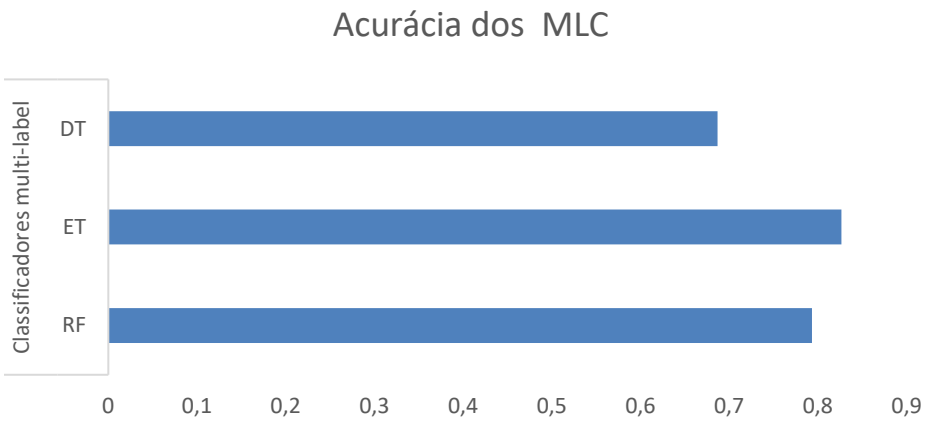
Bloco 3 – Pós-processamento: Trata os resultados da avaliação e apresentação da descoberta.

6 RESULTADOS

Para realizar o objetivo do estudo adotamos três classificadores *multi-label* e cinco métricas de desempenho de generalização dos classificadores. Usamos as classes 0 e 1 para representar duas categorias diferentes em cada um dos dez rótulos: (0) pacientes sem DCNT e (1) pacientes com DCNT. Todo o conjunto de dados foi dividido aleatoriamente em dois subconjuntos: conjunto de dados de treinamento (80%) e conjunto de dados teste (20%). Após comparação de MLC, adotaremos o modelo base com melhor desempenho como modelo principal para conduzir a classificação de tarefa.

No primeiro passo, realizamos experimentos comparativos para avaliar a proporção de previsões corretas introduzida pelos modelos MLC. Os resultados das previsões corretas é mostrado na Gráfico 1. Esse resultado usa a métrica de avaliação de desempenho acurácia do modelo. Conforme mostrado na Gráfico 1, entre os três modelos MLC, os classificadores *Extra Trees* (ET) e *Random Forest* (RF) apresentaram desempenho de 82,68% e 79,33% respectivamente.

Gráfico 1 – Precisão de previsões corretas dos MLC



Fonte: Elaborado pelos autores

Diante do primeiro resultado dos MLC, a acurácia do modelo MLC ET foi de 82,68% com uma contagem de previsões corretas de 148, mostrando assim ser o melhor modelo para essa métrica. Esse, é um classificador robusto, tem um tempo de

execução mais rápido e divide o nó aleatoriamente escolhendo pontos de corte, o que diminuirá a variância melhor do que outras estratégias de randomização (Juan et al., 2021). Destacamos ainda que os modelos MLC RF e DT apresentaram desempenho de avaliação de 79,33 e 68,71%, bem com contagem de previsões corretas de 142 e 123, respectivamente.

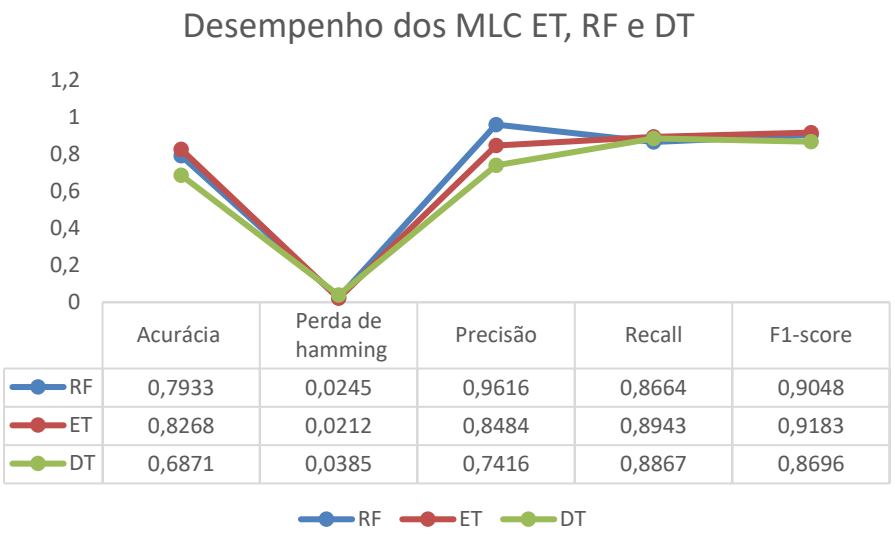
Buscando melhorar ainda mais o desempenho de generalização dos modelos, comparamos os seus desempenhos usando além acurácia, métricas perda de *hamming*, precisão, *recall* e F1-*score*, conforme mostrado na Gráfico 2.

Após experimento com as novas métricas observamos que embora os três MLC apresentem bons resultados, nós encontramos nos MLC ET e RF desempenho consistente para as métricas acurácia, métricas perda de *hamming*, precisão, *recall* e F1-*score*. Conforme mostrado no Gráfico 2, o MLC ET apresentou o melhor desempenho para as métricas Acurácia, *Recall* e F1-*score* e menor perda de *hamming* para previsão de DCNT. Além disso, a métrica *Recall* de desempenho desse modelo é melhor do que outros modelos, representando uma maior sensibilidade para detectar amostras de pacientes com DCNT.

No entanto, chama atenção o resultado do MLC RF. Ele tem desempenho semelhante ao MLC ET, superou significativamente na métrica de precisão com resultado igual a 96,16% e obtém também a segunda menor perda *hamming*, que foi de 2,45%. Isso nos permiti dizer que o número de observações que MLC RF previu como sendo de uma classe, realmente são.

Em resumo, os resultados demonstram que embora o MLC ET esteja com bons resultados, no geral o MLC RF demonstrou melhor desempenho de previsão e F1-*score* para DCNT.

Gráfico 2 – Desempenho dos MLC ET, RF e DT



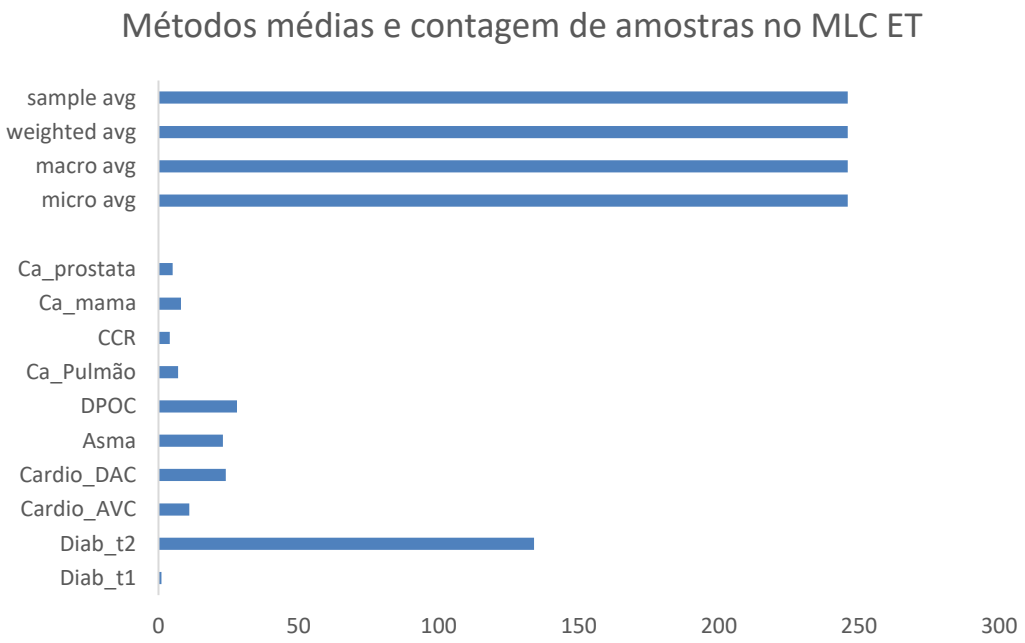
	Acurácia	Perda de hamming	Precisão	Recall	F1-score
RF	0,7933	0,0245	0,9616	0,8664	0,9048
ET	0,8268	0,0212	0,8484	0,8943	0,9183
DT	0,6871	0,0385	0,7416	0,8867	0,8696

Fonte: Elaborado pelos autores

Além de avaliar o desempenho de generalização dos MLC ET, RF e DT de forma global, avaliamos o desempenho de cada um dos MLC por DCNT. O nosso experimento nessa direção foi iniciado com o MLC ET.

O Gráfico 3 apresenta a contagem de amostras ou suporte de cada classe no conjunto de dados real usados pelo MLC ET. Há dois grupos no gráfico, um referente aos métodos de médias, na qual totaliza entrada de dados de 245 para cada uma das médias, incluindo *sample avg*, *weighted avg*, *macro avg* e *micro avg*. O outro grupo refere-se ao total de dados de entrada para cada uma das dez DCNT, incluindo diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata. Diabetes tipo 2 tem a maior quantidade de dados 134 observações, enquanto que diabetes tipo 1 tem a menor quantidade de dados 1 observação utilizada pelo MLC ET.

Gráfico 3 – Métodos de médias e Contagem de amostra no MLC ET



Fonte: Elaborado pelos autores

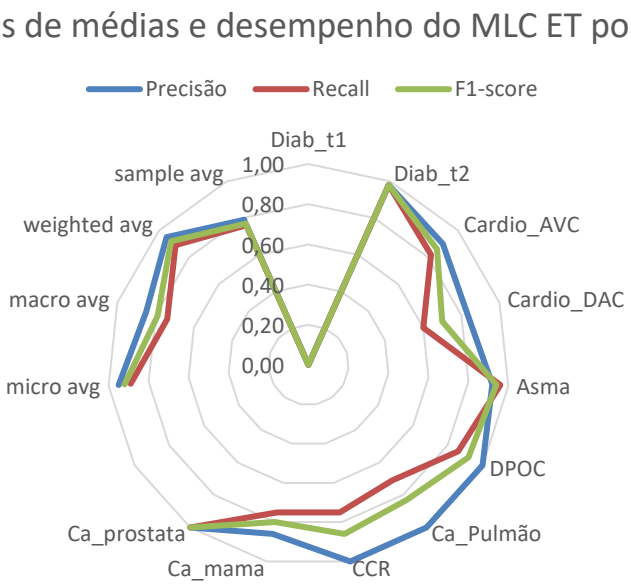
O Gráfico 4 apresenta a pontuação dos métodos de médias *sample avg*, *weighted avg*, *macro avg* e *micro avg*, e avaliação de desempenho do MLC ET utilizando as métricas Precisão, *Recall* e F1-*score* para cada uma das dez DCNT. A Tabela 2 e Gráfico 3 mostram que o conjunto de dados está desequilibrado.

Assim, a proporção de correspondências corretas (conhecida como precisão) seria ineficaz na avaliação do desempenho dos modelos MLC ET, RF e DT. Diante desse cenário optamos por usar método de médias *macro avg* combinado com as métricas Precisão, *Recall* e F1-*score* para cada uma das dez DCNT. Entre os quatro métodos de médias a macro avg destaca-se por tratar todas as classes igualmente importantes, independentemente dos seus valores de amostra, ou seja, de suporte**.**

Conforme mostrado no Gráfico 4, a pontuação de avaliação apresentada no MLC ET para *macro avg* de precisão, *Recall* e F1-*score* foi de 0,85, 0,74 e 0,79, respectivamente. Em *macro avg* de precisão para previsão de DCNT o desempenho para diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata foi de 0,00, 0,98, 0,90, 0,83, 0,92, 1,00, 1,00, 1,00, 0,86, 1,00, respectivamente. Na *macro avg* de *Recall*

para previsão de DCNT o desempenho para diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata foi de 0,00, 0,98, 0,82, 0,60, 0,96, 0,86, 0,71, 0,75, 0,75, 1,00, respectivamente. Para *macro avg* de F1-*score* para previsão de DCNT o desempenho para diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata foi de 0,00, 0,98, 0,86, 0,70, 0,94, 0,92, 0,83, 0,86, 0,80, 1,00, respectivamente.

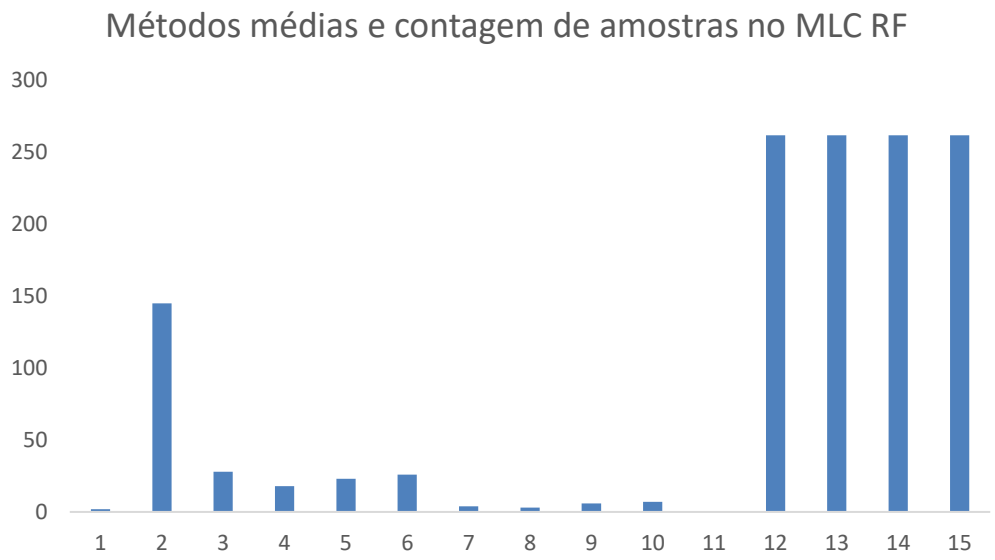
Gráfico 4 – Métodos de médias e desempenho do MLC ET por DCNT



Fonte: Elaborado pelos autores

O Gráfico 5 apresenta a contagem de amostras ou suporte de cada classe no conjunto de dados real usados pelo MLC RF. Há dois grupos presentes neste gráfico, um referente aos métodos de médias, na qual totaliza entrada de dados de 262 para cada uma das médias, incluindo *sample avg*, *weighted avg*, *macro avg* e *micro avg*, o outro grupo se refere ao total de dados de entrada para cada uma das dez DCNT, incluindo diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata. Diabetes tipo 2 tem a maior quantidade de dados 145 observações, enquanto diabetes tipo 1 tem a menor quantidade de dados 2 observações utilizada pelo MLC RF.

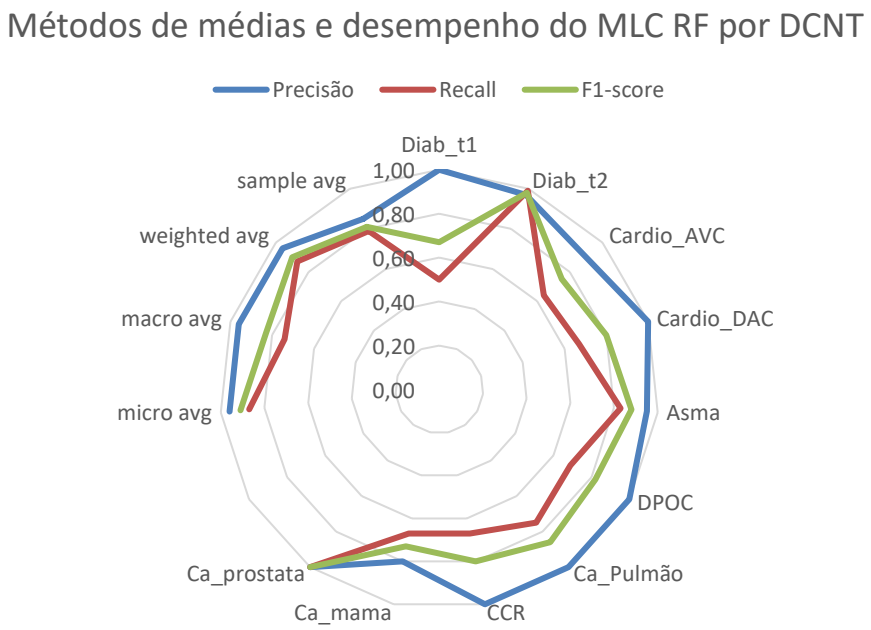
Gráfico 5 – Métodos de médias e Contagem de amostra no MLC RF



Fonte: Elaborado pelos autores

Conforme mostrado no Gráfico 6, a pontuação de avaliação apresentada no MLC RF para *macro avg* de precisão, *Recall* e F1-*score* foi de 0,96, 0,74 e 0,83, respectivamente. Em *macro avg* de precisão para previsão de DCNT o desempenho para diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata foi de 1,00, 0,97, 0,90, 1,00, 0,95, 1,00, 1,00, 1,00, 0,80, 1,00, respectivamente. Na *macro avg* de *Recall* para previsão de DCNT o desempenho para diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata foi de 0,50, 0,99, 0,64, 0,67, 0,83, 0,69, 0,75, 0,67, 0,67, 1,00, respectivamente. Para *macro avg* de F1-*score* para previsão de DCNT o desempenho para diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata foi de 0,67, 0,98, 0,75, 0,80, 0,88, 0,82, 0,86, 0,80, 0,73, 1,00, respectivamente.

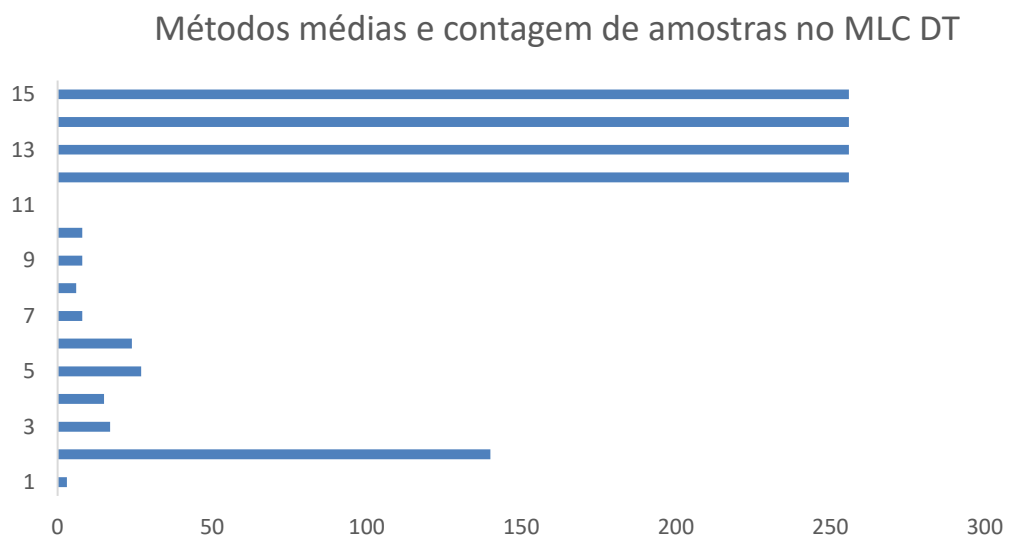
Gráfico 6 – Métodos de médias e desempenho do MLC RF por DCNT



Fonte: Elaborado pelos autores

O Gráfico 7 apresenta a contagem de amostras ou suporte de cada classe no conjunto de dados real usados pelo MLC DT. Há dois grupos presentes neste gráfico, um referente aos métodos de médias, na qual totaliza entrada de dados de 256 para cada uma das médias, incluindo *sample avg*, *weighted avg*, *macro avg* e *micro avg*, o outro grupo se refere ao total de dados de entrada para cada uma das dez DCNT, incluindo diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata. Diabetes tipo 2 tem a maior quantidade de dados 145 observações, enquanto que diabetes tipo 1 tem a menor quantidade de dados 2 observações utilizada pelo MLC DT.

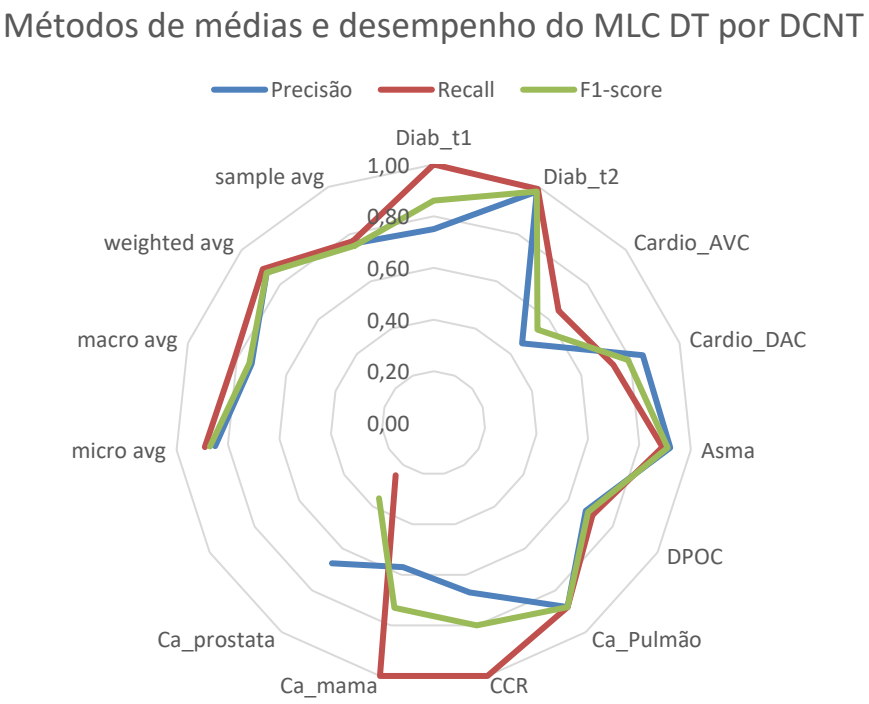
Gráfico 7 – Métodos de médias e Contagem de amostra no MLC DT



Fonte: Elaborado pelos autores

Com base no Gráfico 8, a pontuação de avaliação apresentada no MLC DT para *macro avg* de precisão, *Recall* e F1-*score* foi de 0,74, 0,81 e 0,75, respectivamente. Em *macro avg* de precisão para previsão de DCNT o desempenho para diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata foi de 0,75, 0,98, 0,46, 1,85, 0,92, 0,68, 0,88, 0,67, 0,57, 0,67, respectivamente. Na *macro avg* de *Recall* para previsão de DCNT o desempenho para diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata foi de 1,00, 0,99, 0,65, 0,73, 0,89, 0,71, 0,88, 1,00, 1,00, 0,25, respectivamente. Para *macro avg* de F1-*score* para previsão de DCNT o desempenho para diabetes tipo 1, diabetes tipo 2, cardio_AVC, cardio_DAC, Asma, DPOC, Ca_pulmão, Câncer de colorretal (CCR), Ca_mama e Ca_prostata foi de 0,86, 0,98, 0,54, 0,79, 0,91, 0,69, 0,88, 0,80, 0,73, 0,36, respectivamente.

Gráfico 8 – Métodos de médias e desempenho do MLC DT por DCNT



Fonte: Elaborado pelos autores

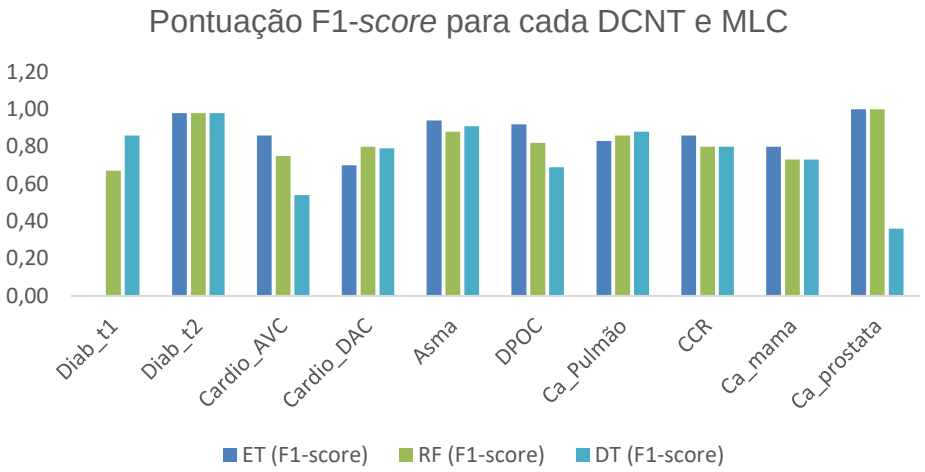
7. DISCUSSÃO

Em nosso estudo, comparamos três modelos de MLC (ET, RF e DT) para previsões dos dez tipos de DCNT simultaneamente e avaliamos o desempenho com cinco métricas diferentes, incluindo acurácia, métricas perdas de *hamming*, precisão, *recall* e F1-*score*. Nossos resultados clarificam que o modelo MLC RF supera os outros métodos MLC e pode prever efetivamente DCNT.

Os modelos de classificação multi-label ET, RF e DT têm uma ampla gama de aplicações, por exemplo, para classificações de emoções, fatores de virulência, diabetes, resistência fenotípica de sequências genômicas inteiras, doenças cardíacas. Por exemplo, Zhou *et al.* (2021) relataram nos seus estudos, que o MLC RF teve o melhor desempenho no diagnóstico de quatro complicações diabéticas simultaneamente. MLC RF também foram usados para prever resistência fenotípica de sequências genômicas inteiras (Kouchaki et al., 2020). Aqui, primeiramente e especificamente aplicamos e comparamos os modelos MLC ET, RF e DT para prever os dez tipos de DCNT simultaneamente. Os nossos resultados apontam que modelo MLC RF é consistente e obteve o melhor desempenho de precisão 96,16% entre os demais modelos de MLC.

O desempenho dos modelos de MLC em cada DCNT é diferente. Embora todos os modelos de MLC tenham obtido bom desempenho na predição de cada uma das DCNT. O modelo MLC RF reafirma a sua superação tomando como base a avaliação de desempenho por cada uma das DCNT. Para tal, usamos a métrica F1-*score* e obtivemos resultado de 0,83. Conforme mostra o Gráfico 9, o desempenho comparativo de pontuação mais baixa no MLC ET foi para a DCNT diabetes tipo 1, no entanto ele foi quase predominantemente mais alto em outras DCNT. A baixa quantidade de observações para diabetes tipo 1 no conjunto de dados utilizado em nosso experimento levou a esse resultado. Essa é uma das limitações de nosso trabalho.

Gráfico 9 – Pontuação F1-*score* para cada DCNT e MLC



Fonte: Elaborado pelos autores

CONCLUSÃO

Os algoritmos de aprendizado de máquina, especificamente os classificadores *multi-label* têm tido uma ampla gama de aplicações. Na área da saúde, esses algoritmos têm tido eficácia comprovada na mineração de dados para fornecer ajuda aos médicos nos diagnósticos e nas suas tomadas de decisões.

Dados da Organização mundial de Saúde (OMS) apontam que Doenças Crônicas Não Transmissíveis (DCNT) são responsáveis por mais de 70% de todas as mortes no mundo. Elas fazem parte de quatro grupos de doenças incluindo: doenças cardiovasculares, câncer, respiratórias crônicas e diabetes, que têm sido motivo de crescente preocupação da sociedade e governos de todo o mundo, por colocar as pessoas em maior risco de complicações, chegando até mesmo a óbito, e colocando os sistemas de saúde em crise sistémica.

Nossa abordagem em resposta ao problema das DCNT foi desenvolver uma plataforma inteligente para predição do risco de doenças crônicas, antes que elas se manifestem. Para tal, utilizamos modelos MLC ET, RF e DT para prever até dez tipos de DCNT simultaneamente. Entre os modelos do experimento, o MLC RF alcançou o melhor desempenho de precisão e F1-*score* com 96,16% e 90,48%, respectivamente.

Considerando o rápido crescimento de DCNT e os seus impactos para a sociedade e governos de todo o mundo. Nossos resultados sugerem que o MLC RF é um classificador promissor para predição de DCNT, que pode ser usado como uma abordagem de referência pelas equipes médicas para melhorar os diagnósticos precoce e tratamentos dos pacientes na atenção primária de saúde e, assim, contribuir para reduzir risco de complicações e mortes por DCNT, e proporcionar benefícios de sobrevida. Assim, dentro do escopo de contribuições, a plataforma é também um forte candidato para apoiar o 3º item dos Objetivos de Desenvolvimento Sustentável – ODS e fornecer respostas especificas e precisas da predição de DCNT.

Temos duas limitações, uma diz respeito a quantidade de dados e outra se referente a prevalência da faixa etária de pessoas com mais 45 anos no conjunto de dados. Os trabalhos futuros, além de direcionados a hospedagem da plataforma na

nuvem e disponibilizar para parceiros, apresentaremos os pesos de fatores de maior risco conjuntamente com as probabilidades de desenvolver DCNT.

REFERÊNCIAS

Ahmad, A.S. & Mayya, A.M. A new tool to predict lung cancer based on risk factors. Heliyon. 2020; 6(2): e03402. Published 2020 Feb 26. https://linkinghub.elsevier.com/retrieve/pii/S2405844020302474 .Consultado em 6 de abril de 2023

Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J., & Clarke, M. F. (2003). Prospective identification of tumorigenic breast cancer cells. *Proceedings of the National Academy of Sciences*, *100*(7), 3983-3988.

Akella, A. & Akella, S. Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution. Future Sci OA. 2021 Mar 29;7(6):FSO698. doi: 10.2144/fsoa-2020-0206. PMID: 34046201; PMCID: PMC8147740.

Aladwani, S., Alosaini, M. E., Althunayan, S. A., Alrowaidan, A. K., Al-Abrah, S. M., Alhawas, F. A. A., Aloyayri, M. A. & Almulhem, A. A. (2019). A survey to assess osteoporosis knowledge of the general population of Riyadh, Saudi Arabia. International Journal of Pharmaceutical Research & Allied Sciences, 8(4), 174-179

Albright, R., J. Benthuysen, N. Cantin, K. Caldeira & K. Anthony (2015), Coral reef metabolism and carbon chemistry dynamics of a coral reef flat, Geophysical Research Letters, 42(10), 3980-3988

Aleksandrova, K., Koelman, L., & Rodrigues, C. E. (2021). Dietary patterns and biomarkers of oxidative stress and inflammation: A systematic review of observational and intervention studies. Redox Biology, xxxx, 101869. *https://doi.org/10.1016/j.redox.2021.101869.* Consultado em 10 de abril de 2023

Ampomah, E. K., Qin Z. & Nyame, G. Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement. Information. 2020; 11(6):332.

Barber, K. E., Pogue, J. M., Warnock, H. D., Bonomo, R. A., & Kaye, K. S. (2018). Ceftazidime/avibactam versus standard-of-care agents against carbapenem-resistant Enterobacteriaceae harbouring bla KPC in a one-compartment pharmacokinetic/pharmacodynamic model. *Journal of Antimicrobial Chemotherapy*, *73*(9), 2405-2410.

Barr, V.J., Robinson, S., Marin-Link, B., Underhill, L., Dotts, A., Ravensdale, D. & Salivaras, S. The expanded Chronic Care Model: an integration of concepts and strategies from population health promotion and the Chronic Care Model. Hosp Q. 2003;7(1):73-82. DOI: *10.12927/hcq.2003.16763*. PMID: 14674182.

Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C.W., Carson, A.P et al. Heart Disease and Stroke Statistics—2019 Update: A Report from the American Heart Association. Circulation. 2019.

Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S. & Singh, P. (2021). Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Computational Intelligence and Neuroscience, 2021. *https://doi.org/10.1155/2021/8387680.* Consultado em 8 de abril de 2023

Chaurasia, V., Pal, S. & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms and Computational Technology, 12(2), 119–126. DOI: 10.1177/174830181875622 5

Chun, M., Clarke, R., Cairns, B. J., Clifton, D., Bennett, D., Chen, Y. et al. The China Kadoorie Biobank Collaborative Group, Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. Journal of the American Medical Informatics Association, 2021 Aug. 28(8):1719-1727. DOI: 10.1093/jamia/ocab068

Deberneh, H. M., & Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. International Journal of Environmental Research and Public Health, 18(6), 3317. Advance online publication. PMID:33806973. DOI:10.3390/ijerph18063317

El_Jerjawi N. S. & Abu-Naser S. S. (2018). Diabetes prediction using artificial neural network. *Int. J. Adv. Sci. Technol.* 121:54–64.

Elkafrawy, P., Mausad. A. & Esmail, H. Experimental comparison of methods for multi-label classifcation in diferent application domains. Int J Comput Appl. 2015;114:1

Evgeniou, T. & Pontil, M. (2004). Regularized multi-task learning," Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Seattle, WA: ACM), 109–117. DOI: 10.1145/1014052.1014067

Faddoul, J. B., Chidlovskii, B., Gilleron, R. & Torre, F. (2012). Learning multiple tasks with boosted decision trees. In Machine Learning and Knowledge Discovery in Databases, volume 7523, pages 681–696. Springer Berlin Heidelberg.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17[3], 1-6.

Fundação Nacional de Saúde. OMS lista as 10 principais ameaças para a saúde em 2019. Disponível em: *https://portalfns-antigo.saude.gov.br/ultimas-noticias/2375-oms-lista-as-10-principais-ameacas-para-a-saude-em-2019.* Consultado em 5 de abril de 2023

Furnkranz, J., Hullermeier, E., Mencía, E. & Brinker, K. 2008. Multilabel classification via calibrated label ranking. Machine Learning 73(2):133–153.

García, S. M., Castellano, J. G., Ruiz, C. J. M. & Abellán, J. (2021, August). Using credal C4. 5 for calibrated label ranking in multi-label classification. In *International Symposium on Imprecise Probability: Theories and Applications* (pp. 220-228). PMLR.

Geetha, P. S., Kanchana, S., Pasupathi, E., Murugan, M. & Rohini, C. (2021). A review on putative mechanism of action of nootropic herb Bacopa monnieri. Pharma Innovation. 10 (5), 672-681.

Giardiello, S., Gerbino, M., Pagano, L., Errard, J., Gruppuso, A., Ishino, H., Lattanzi, M., Natoli, P., Patanchon, G., Piacentini, F. & Pisano, G. (2021). Detailed study of HWP non-idealities and their impact on future measurements of CMB polarization anisotropies from space. Astronomy & Astrophysics.

Madjarov, G., Kocev, D., Gjorgjevikj, D. & Dzeroski, S. An extensive experimental comparicion of methods for multi-label learning. Pattern Recognition, 45(9):3084–3104, 2012.

Hussan, H., Zhao, J., Badu-Tawiah, A. K., Stanich, P., Tabung, F., Gray, D. et al. Utility of machine learning in developing a predictive model for early-age-onset colorectal neoplasia using electronic health records. PLoS ONE 2022;17(3):e0265209. DOI: *10.1371/journal. pone.0265209*.

Juan, Y., Dai, Y., Yang, Y. & Zhang, J. (2021). Accelerating materials discovery using machine learning. *Journal of Materials Science & Technology*, *79*, 178-190.

Kassim, B., Mohan, S. & Muneer, K. A. Modified ML-kNN and rank SVM for multi-label pattern classification, Journal of Physics: Conference Series, vol. 1921, Article ID012027, 2021. *(2) (PDF) Lightweight Multireceptive Field CNN for 12-Lead ECG Signal Classification*. Available from: *https://www.researchgate.net/publication/362560268_Lightweight_Multireceptive_Field_CNN_for_12-Lead_ECG_Signal_Classification*. Consultado em 2 de março de 2023

Kouchaki, S., Yang, Y., Lachapelle, A., Walker T. M. & Walker, A. S. CRyPTIC Consortium, Peto TEA, Crook DW and Clifton DA (2020) Multi-Label Random Forest Model for Tuberculosis Drug Resistance Classification and Mutation Ranking. Front. Microbiol. 11:667. DOI: 10.3389/fmicb.2020.00667

Lee, J., Kim, H., Kim, N. et al. An approach for multi-label classification by directed acyclic graph with label correlation maximization. Inform Sciences 2016; 351: 101–114.

Li, R., Liu, W., Lin, Y., Zhao, H. & Zhang, C. (2017). An ensemble multilabel classification for disease risk prediction. *Journal of healthcare engineering*, *2017*.

Liu, L. W., Xing, Q. Q., Zhao, X., Tan, M., Lu, Y., Dong, Y. M. et al. (2019). Proteomic Analysis Provides Insights into the Therapeutic Effect of GU-BEN-FANG-XIAO Decoction on a Persistent Asthmatic Mouse Model. Front. Pharmacol. 10, 441. DOI:10.3389/fphar.2019.00441

Madjarov, G., Kocev, D., Gjorgjevikj, D., Džeroski, S. (2012) An extensive experimental comparison of methods for multi-label learning Pattern Recognition 45:3084-3104

Mendes, E. V. As redes de atenção à saúde. Brasília: Organização Pan-Americana da Saúde; 2011.

Matheson, F. I., Sztainert, T., Lakman, Y., Steele, S. J., Ziegler, C. P. & Ferentzy, P. (2018). Prevention and treatment of problem gambling among older adults: A scoping review. Journal of Gambling Issues: Special Issue, (39), 6–66. DOI: 10.4309/jgi.2018.39.2

McCoy, S. J. B. R., Beal J. M., Shipman, S. B. M., Payton, M. E. & Watson, G. H. Risck factors for postpartum depression: A retrospective investigation of at 4- Weeks postnatal and a review of the Literature. JAOA v. 106, n.4 April, 2006. 193-8.

Naghavi, M., Abajobir, A. A., Abbafati, C. et al; Colaboradores das Causas de Morte do GBD 2016. Global, regional e mortalidade nacional específico idade-sexo por 264 causas de morte, 1980-2016: uma análise sistemática para o Global Burden of Disease Study 2016. *Lancet*. 2017; 390 (10100): 1151-1210. DOI: *10.1016 / S0140-6736 (17) 32152-9*

Naji, M. A., El-Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., Debauche, O. et al. Machine Learning Algorithms for Breast Cancer Prediction And Diagnosis. Procedia Computer Science 2021; 191:487-492. ISSN 1877-0509. DOI: *10.1016/j.procs.2021.07.062*.

Nijil, R. N. & Mahalekshmi, T. Multilabel Classification of Membrane Protein in Human by Decision Tree (DT) Approach. Biomed Pharmacol J 2018; 11(1).

Nareshpalsingh, J. & Modi, H. 2017. Multi-label Classification Methods: A Comparative study. International Research Journal of Engineering and Technology (IRJET), (Volume 4), Issue 12, 2017

Nasser, I. Lung Cancer Detection Using Artificial Neural Network (2019). International Journal of Engineering and Information Systems (IJEAIS), 2019 Mar;3(3):17-23.

Öberg, K. I., Qi, C., Fogel, J. K. J. et al. 2011. The Astrophysical Journal, 734, 98

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. Fatores de risco para doenças crônicas não transmissíveis nas Américas: Considerações sobre o fortalecimento da capacidade regulatória. Documento de Referência Técnica REGULA. Washington, DC ; OPAS, 2016.

Ontario Ministry of Health and Long-Term Care. Preventing and Managing Chronic Disease: Ontario's Framework. 2007.

Oyewo, A. O. & Boyinbode, O. K. Prediction of Prostate Cancer using Ensemble of Machine Learning Techniques. International Journal of Advanced Computer Science and Applications (IJACSA) 2020;11(3). DOI: 10.14569/IJACSA.2020.0110318.

Pashayan, N., Antoniou, A. C., Lee, A., Wolfson, M., Chiquette, J., Eloy, L., Eisen, A. et al. Should Age-Dependent Absolute Risk Thresholds Be Used for Risk Stratification in Risk-Stratified Breast Cancer Screening? Journal of Personalized Medicine, 2021, 11 (9) *https://doi.org/10.3390/jpm11090916.* Consultado em 10 de abril de 2023

Pasquali, R., Casanueva, F., Haluzik, M. et al. European Society of Endocrinology Clinical Practice Guideline: Endocrine work-up in obesity. Eur J Endocrinol. 2020 Jan;182(1):G1- G32

Queiroz, A. M., Sousa, A. R., Moreira, W. C., Nóbrega, M. D. P. S. S., Santos, M. B., Barbossa, L. J. H., Rezio, L. A., Zerbetto, S. R., Marcheti, P. M., Nasi, C. & Oliveira, E. O 'NOVO' da COVID-19: impactos na saúde mental de profissionais de enfermagem?. 2021. Acta Paul Enferm., 34, eAPE02523.

Rahimloo, P. & Jafarian, A. (2016) Prediction of diabetes by using artificial neural network logistic regression statistical model and combination of them. Bull Soc Sci Liege 85

Rani, K. J. Diabetes Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science. Engineering and Information Technology 2020 Jul-Agu; 6(4);294-305. DOI: *10.32628/CSEIT206463*

Santos, G. M. R. F., Silva, M. E., & Belmonte, B. R. (2021). COVID-19: Emergency remote teaching and university professors' mental health. Revista Brasileira de Saúde Materno Infantil, 21(sup.1): s237-s243. https://doi. org/10.1590/1806-9304202100
Singh, D. A. A. G., Leavline, E. J., & Baig, B. S. (2017). Diabetes Prediction Using Medical Data. Journal of Computational Intelligence in Bioinformatics, 10(1), 1-8.

Spathis, D. and Vlamos, P. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. Health Informatics J. 2019 Sep; 25(3):811-827. DOI: *10.1177/1460458217723169*. Epub 2017 Aug 18. PMID: 28820010.

Syed, S. A., Ali, S., Abdulaziz, A. A., Ibrahim, H. B. & Bard, N. A. (2019). Uropathogens and their antimicrobial resistance patterns: Relationship with urinary tract infection. International Journal of Health Sciences. (Quassim), 13(2):48-55

Theerthagiri, P. (2021). Probable Forecasting of Epidemic COVID-19 in Using COCUDE Model. EAI Endorsed Transactions on Pervasive Health and Technology V.7, n.26. DOI: 10.4108/eai.3-2-2021.168601

Toskala, E. & Kennedy, D. W. Asthma risck factors. International forum of allergy & rhinology, Hoboken, v.5 n.1 p. 11-16, set. 2015.

Tigga, N. P. & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. Proc. Comp. Sci. 167, 706–716. DOI: 10.1016/j.procs.2020.03.336

Vikas, P. K. & Kaur, P. (2021). Lung cancer detection using chi-square feature selection and support vector machine algorithm. *International Journal of Advanced Trends in Computer Science and Engineering*.

Urbanowicz, R. J., & Moore, J. H. (2009). Learning classifier systems: a complete introduction, review, and roadmap. *Journal of Artificial Evolution and Applications*, *2009*.

Virani, S. S., Alonso, A., Benjamin, E. J, Bittencourt, M. S., Callaway, C.W., Carson, A.P. et al. Heart disease and stroke statistics-2020 update: A report from the American Heart Association. Circulation. 2020; 141: e139-96. DOI: *10.1161/CIR.0000000000000757*

Wells, Y., Bhar, S., Kinsella, G., Kowalski, C., Merkes, M., Patchett, A., Salzmann, B., Teshuva, K. & van Holsteyn, J. What works to promote emotional wellbeing in older people: A guide for aged care staff working in community or residential care settings. 2014. Melbourne: Beyond Blue. *https://www.beyondblue.org.au/about-us/about-our-work/older-adults-program/what-works-to-promoteemotional-wellbeing-in-older-people*. Consultado em 9 de março de 2023

World Health Organization. Preventing chronic diseases: a vital investment. Geneva: World Health Organization. 2005.

World Health Organization. WHO guidelines on hand hygiene in health care. In *WHO guidelines on hand hygiene in health care*. 2009. (pp. 270-270).

World Health Organization. Health statistics and information systems: disease burden and mortality estimates. Geneva: WHO; 2016.

Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X. et al. Study of cardiovascular disease prediction model based on random forest in eastern China. Sci Rep. 2020 Mar; 10(1):5245. DOI: *10.1038/s41598-020-62133-5*. PMID: 32251324; PMCID: PMC7090086

Zhou, L., Zheng, X., Yang, D. *et al.* Aplicação de modelos de classificação multirrótulo para o diagnóstico de complicações diabéticas. *BMC Med Informa Decis Mak* 21, 182 (2021). DOI: 10.1186/s12911-021-01525-7

Apêndice 1 – Apresentação do trabalho

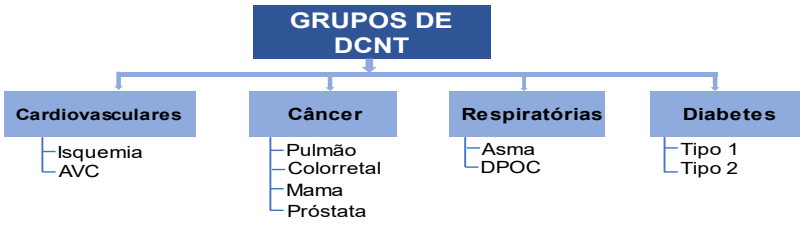
Ampliação e aprimoramento dos serviços da Atenção Primária (AP) nos hospitais : Predição de Doenças Crônicas Não Transmissíveis.

Porto - Portugal 2023

1

Ameaça global à saúde

Dados da OMS revelam que 74% das Mortes e incapacidades são causadas pelos quatro principais grupos de Doenças Crônicas Não Transmissíveis (DCNT) em todo o mundo.



2

DCNT é uma pauta de preocupações crescentes de...

- . Governos de todo mundo
- . Operadoras de planos de saúde
- . Hospitais
- . Sociedade em geral

3

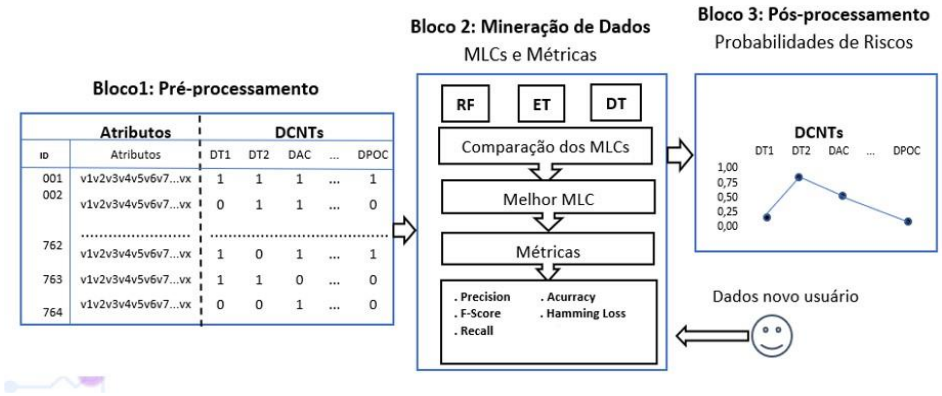
Soluções em curso

- ✓ Abordagem reativa de prevenção e controle de DCNT A doença já estar estabelecida.
- ✓ Predição de DCNT de forma isolada ou com no máximo duas DCNT do mesmo grupo. (Ex: diabetes Tipo 1 e 2)

4

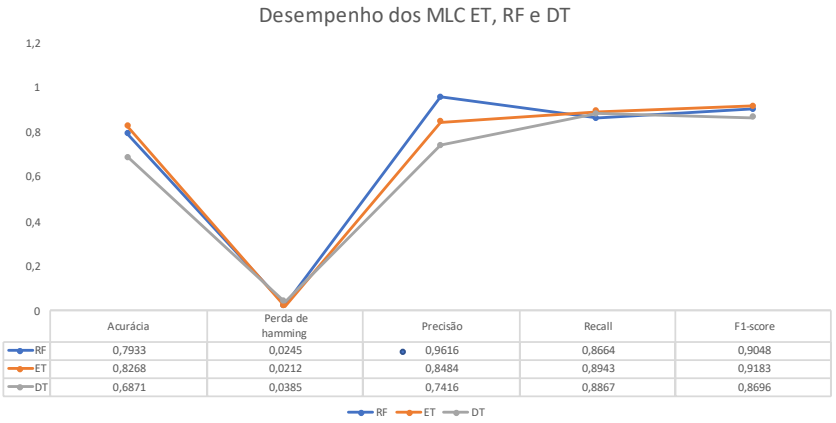
Incorporação tecnológica na AP– Oportunidade

. **Plataforma Inteligente de Predição do Risco de Doenças Crônicas**



5

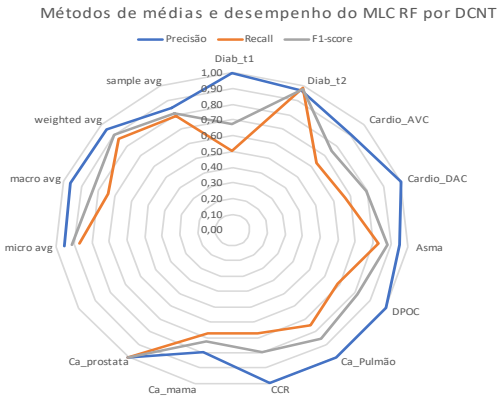
Desempenho dos algoritmos com uso de dados do HE-FFP



	Acurácia	Perda de hamming	Precisão	Recall	F1-score
RF	0,7933	0,0245	0,9616	0,8664	0,9048
ET	0,8268	0,0212	0,8484	0,8943	0,9183
DT	0,6871	0,0385	0,7416	0,8867	0,8696

6

Algoritmo de melhor desempenho por DCNT, usando dados do HE-FFP



7

Benefícios verticais da PIPRDC

Sociedade

. Acesso a serviços de saúde integrado, ágil e personalizado.

. Conhecer seus riscos de desenvolver umaou múltiplas DCNT

Hospitais

. Diagnóstico inteligente para prever DCNT

. Crescimento da base de usuários de saúde : E stimulo a realização de consultas e exames

. Redução de custos

. Ganho de eficiência : atendimento padronizado a pacientes com risco de DCNT

. Redução da sobrecarga no atendimento de urgência e emergência e custos social geral

Operadoras de PS

. Redução de custos

. Oferta de planos de saúde diferenciados com menor risco : planos customizados

. Crescimento da base de usuários de saúde

Governos

. Um sistema de saúde mais inclusivo

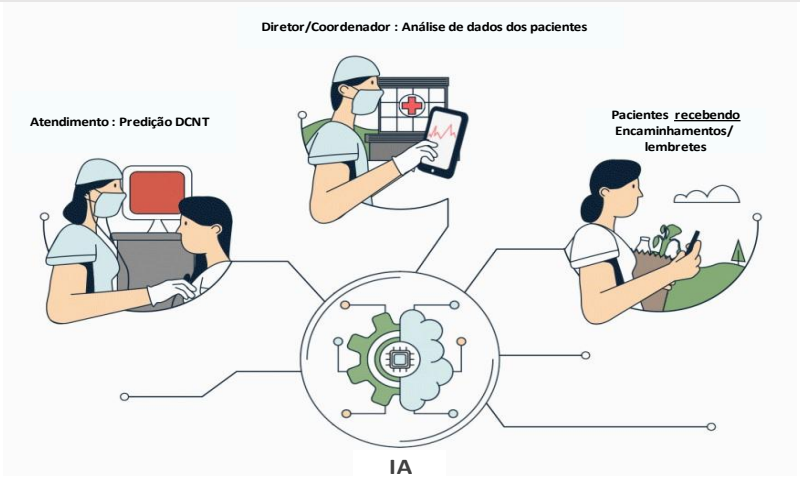
. Redução da taxa de incidência e mortalidade por DCNT

. Melhorar a Vigilância Epidemiológica

. Apoio as politicas e estratégias de promoção da saúde

8

Visão prática da aplicação



Fonte: Adaptado de autor desconhecido.

Buscamos parceria inicialmente para:

- ✓ Execução conjunta da ampliação de prestação de serviços do hospital por meio da plataforma inteligente de predição de risco de DCNT para atendimento as necessidades de saúde.
- ✓ Otimização de recursos : Padronização de atendimento de pessoas com risco de DCNT
- ✓ Apoio na emissão de boletins para uma avaliação abrangente do atendimento no hospital em relação a doenças crônicas e seus fatores de risco. Visão futura: Portugal mais saudável.
- ✓ Disponibilidade de dados para o apoio aos projetos e ações estratégicas (ex: campanhas) do hospital nas políticas de promoção, prevenção e controle de doenças crônicas não transmissíveis.
- ✓ Participações em chamadas de propostas (Ex: Aliança Global para Doenças Crônicas (GACD) e edições de premiações no campo da saúde: Ex: 12ª Edição do Prémio Saúde Sustentável (categoria: inovação).

10

Passos futuros

- Numa segunda fase, já com a expertise adquirida, a plataforma poderá ser utilizada também para o monitoramento, avaliação e aprendizagem ativa de ações efetuadas pelos médicos juntos aos pacientes identificados e tratados com DCNT que lograram êxito.
- Permite que os médicos prescrevam novas opções de tratamento para DCNT com base na aprendizagem ativa da plataforma.
- “Compartilhamento de conhecimento e descobertas”

11

Apêndice 2 – Certificados de participação em congressos


SISTEMAS INTELIGENTES PARA A SAÚDE: DESAFIOS DA ÉTICA E GOVERNANÇA
CBIS 22
XIX Congresso Brasileiro de Informática em Saúde
SBIS
CERTIFICADO
A Sociedade Brasileira de Informática em Saúde - SBIS certifica que o trabalho
Uma proposta para um Sistema Inteligente de Previsão do Risco de Doenças Crônicas
dos autores Oberdan Costa, Luís Gouveia e apresentado por Oberdan Santos da Costa em sessão poster no XIX Congresso Brasileiro de Informática em Saúde - CBIS22, realizado de 29/11 a 2/12/22 em Campinas SP - Brasil.
LUIS GUSTAVO GASPARINI KIATAKE
Presidente SBIS
OSMEIRE CHAMELETTE SANZOVO
Presidente CBIS22
JULIANO GASPAR
Presidente Comissão Científica CBIS22


SISTEMAS INTELIGENTES PARA A SAÚDE: DESAFIOS DA ÉTICA E GOVERNANÇA
CBIS 22
XIX Congresso Brasileiro de Informática em Saúde
SBIS
CERTIFICADO
A Sociedade Brasileira de Informática em Saúde - SBIS certifica que
Oberdan Costa
participou da Oficina de Revisão das competências em Informática em Saúde realizada durante o XIX Congresso Brasileiro de Informática em Saúde - CBIS22, realizado de 29/11 a 2/12/22 em Campinas SP - Brasil.
LUIS GUSTAVO GASPARINI KIATAKE
Presidente SBIS
OSMEIRE CHAMELETTE SANZOVO
Presidente CBIS22
JULIANO GASPAR
Presidente Comissão Científica CBIS22

CERTIFICADO

Certificamos que:

A Seven Publicações LTDA. em parceria com Home Publishing Brazil, declara que o artigo **"ABORDAGEM PROATIVA NA ATENÇÃO PRIMÁRIA À SAÚDE: UM MODELO DE REFERÊNCIA PARA PREDIÇÃO DO RISCO DE DOENÇAS CRÔNICAS"**, foi apresentado no II Seven International Congress of Health nos dias 24 e 25 de abril de 2023, com carga horária de 36 horas.

Autores:
Oberdan santos da costa e Luiz Borges Gouveia

Por fim, declaro os termos da seguinte declaração
São José dos Pinhais, 27 de abril de 2023, Brasil.

II SEVEN INTERNATIONAL CONGRESS OF HEALTH

Nathan Albano Valente
EDITOR-CHEFE

HOME PUBLISHING BRAZIL

Fernanda Chaves Aloisio
PRESIDENTE DA COMISSÃO DO EVENTO

Apêndice 3 – Trabalhos apresentados, publicados e em andamento para publicação

Costa, O. e Gouveia, L. (2022). Uma proposta para um Sistema Inteligente de Previsão do Risco de Doenças Crônicas. Estudos de demonstração, relatos de experiências e revisões de literatura. XIX Congresso Brasileiro de Informática em Saúde (CBIS-2022). Disponível em: https://www.researchgate.net/publication/366841322_Uma_proposta_para_um_Sistema_Inteligente_de_Previsao_do_Risco_de_Doencas_Cronicas

Costa OS da, Gouveia LB. Mortalidade pelos principais grupos de doenças crônicas não transmissíveis em 25 municípios do Maranhão, Brasil. RECIMA21 - Revista Científica Multidisciplinar - ISSN 2675-6218. Recima21combr [Internet]. 2023 Feb 26 [cited 2023 Feb 23]; Disponível em: https://recima21.com.br/index.php/recima21/article/view/2724

Costa, O., & Borges Gouveia, L. (2023). FATORES DE RISCOS E PROTEÇÃO ASSOCIADOS À PREVALÊNCIA DE HIPERTENSÃO, DIABETES E OBESIDADE NA POPULAÇÃO ADULTA LUDOVICENSE. *RECISATEC - REVISTA CIENTÍFICA SAÚDE E TECNOLOGIA - ISSN 2763-8405*, *3*(4), e34277. https://doi.org/10.53612/recisatec.v3i4.277

Costa, O. e Gouveia, L. (2023). Abordagem proativa na Atenção Primária à Saúde um modelo de referência para predição de doenças crônicas. II Seven International Congress of Health. CIÊNCIAS DA SAÚDE E SUAS DESCOBERTAS CIENTÍFICAS - V1. e-book. Seven Publicações. ISBN: 978 65 84976 36 8.

Em andamento...

Costa OS da, Gouveia LB. (2023). Predição de Doenças Crônicas Não Transmissiveis (DCNT) na atenção primária de saúde, usando inteligência artificial.