PLATAFORMA INTELIGENTE, DE PREDIÇÃO DO RISCO DE DOENÇAS CRÔNICAS NÃO TRANSMISSÍVEIS, DE APOIO À DECISÃO CLÍNICA NA ATENÇÃO PRIMÁRIA DE SAÚDE, USANDO INTELIGÊNCIA ARTIFICIAL

OBERDAN COSTA*
LUÍS BORGES GOUVEIA**

Resumo: *A combinação de envelhecimento da população, escassez de profissionais da saúde, aumento da carga de Doenças Crônicas Não Transmissíveis (DCNT) e restrições de recursos são preocupações crescentes para a sociedade e governos em todo o mundo. Particularmente, quatro grupos de DCNT, incluindo doenças cardiovasculares, câncer, doenças respiratórias e diabetes, vêm afetando o ecossistema de saúde de várias maneiras, incluindo aumento da pressão no atendimento de urgência e emergência, custos com internações e maior exposição de pacientes a infecções hospitalares. Essas doenças colocam as pessoas em maior risco de complicações, invalidez e morte, afetando a produtividade do trabalho, os custos de saúde e causando desigualdade nas condições de saúde entre a população. Como solução propõe-se a plataforma inteligente de predição do risco de doenças crônicas não transmissíveis, fundamentada numa abordagem proativa multifatorial e inteligência artificial, para fornecer respostas específicas à probabilidade de um indivíduo desenvolver DCNT, antes que ocorram.*

Palavras-chave: *Atenção primária; Decisão clínica; Doenças crônicas; Inteligência artificial; Saúde.*

Abstract: *The combination of an aging population, shortage of health professionals, the increased burden of Chronic Non-Communicable Diseases (CNCDs) and resource constraints are growing concerns for society and governments around the world. Particularly, four groups of CNCDs, including cardiovascular diseases, cancer, respiratory diseases, and diabetes, have been affecting the healthcare ecosystem in several ways, including increased pressure on urgent and emergency care, hospitalisation costs and greater exposure of patients to hospital-acquired infections. These diseases put people at greater risk of complications, disability, and death, affecting work productivity, healthcare costs and causing inequality in health conditions among the population. As a solution to mitigate this problem, an intelligent platform for predicting the risk of chronic non-communicable diseases is proposed, based on a proactive multifactorial approach and artificial intelligence, which provides specific answers to the probability of an individual developing CNCDs, before they occur.*

Keywords: *Primary care; Clinical decision making; Chronical diseases; Artificial intelligence, Health.*

* Universidade Fernando Pessoa. Email: oberdan.costa@ufp.edu.pt. ORCID: https://orcid.org/0000-0002-2448-5247.
** Universidade Fernando Pessoa; CITCEM (UIDB/04059/2020; DOI: https://doi.org/10.54499/UIDB/04059/2020). Email: lmbg@ufp.edu.pt. ORCID: https://orcid.org/0000-0002-2079-3234.

INTRODUÇÃO

Dados da World Health Organization (2022) mostram que as Doenças Crônicas Não Transmissíveis (DCNT) levam a óbito cerca de 41 milhões de indivíduos a cada ano, sendo responsáveis por 74% das mortes no mundo. No Brasil, segundo Bernal et al. (2019), elas correspondem a 75% das causas de mortes. Isso certamente 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 só o custo da hipertensão, diabetes e obesidade chegou a R$ 3,45 bilhões em 2018 no sistema público de saúde brasileiro, elevando a carga financeira na economia. De acordo com Rahimloo e Jafarian (2016), predizer com mais precisão a condição dos pacientes é de extrema importância.

No Brasil, estado do Maranhão, em 2022, aproximadamente 500 mil internamentos foram realizados nos hospitais das 19 regiões de saúde com um custo total de R$452 296 740,29 (Brasil. Ministério da Saúde 2022). Deste valor, quase 40% foram gastos com pacientes portadores de doenças evitáveis DCNT. Ainda em 2022, a cada 1 hora, 14 (quatorze) maranhenses foram internados com problemas de uma ou mais DCNT e 13 820 pessoas foram a óbito. Do total de óbitos, 52,78%, ou seja, 7294 foram ocasionadas por DCNT. Isso é uma dimensão clara de que as DCNT acarretam um custo económico elevado tanto para o sistema de saúde como para a sociedade, impactando negativamente sobre o desenvolvimento do Estado.

Em resposta a essas preocupações, especificamente as DCNT, desenvolvemos a Plataforma Inteligente de Predição do Risco de Doenças Crônicas (PIPRDC), cujo objetivo é apoiar à decisão clínica dos profissionais da saúde na atenção primária à saúde, usando Inteligência Artificial (mais concretamente, à Aprendizagem de Máquina). A predição, prevenção e o tratamento de doenças crônicas são prioridades de saúde pública e privada. Embora existam algumas aplicações de apoio a combate a DCNT, elas são limitadas em suas aplicações, pois não fornecem respostas específicas para detectar predisposições de uma pessoa para desenvolver múltiplas DCNT de forma integrada.

Dispor de uma ferramenta para auxiliar os médicos e gestores da saúde, no enfrentamento das DCNT, antes que elas ocorram é um grande passo para cuidar e salvar vidas da população.

A estrutura da plataforma é composta por quatro módulos, incluindo: predição do risco de DCNT, encaminhamento ao especialista, gestão inteligente e aprendizagem contínua, conforme Figura 1, diagrama modular da PIPRDC.

Embora os módulos da plataforma sejam integrados, cada módulo tem seu ponto alto no enfrentamento de DCNT:

- o Módulo M1, além de auxiliar os médicos no diagnóstico precoce de DCNT, contribui para políticas e estratégias de promoção da saúde, reduzindo complicações, invalidez e taxas de mortalidades;

- o Módulo M2, além de apoiar as equipes das UBS (Unidades Básicas de Saúde) com encaminhamento assertivos para outros pontos de atenção (ex.: especialista), melhora o ganho de eficiência no atendimento ao paciente;
- o Módulo M3 possibilita visualizar informações sobre encaminhamentos médicos, situação de saúde, doenças crônicas e fatores de risco prevalentes da população, auxiliando as tomadas de decisões dos gestores, coordenadores de saúde, etc.;
- o Módulo M4 aprimora abordagens terapêuticas promissoras no combate a DCNT, compartilhando conhecimento de resolutividades (decisões médicas e/ou clínicas) e fornecendo recomendações as equipes de saúde.

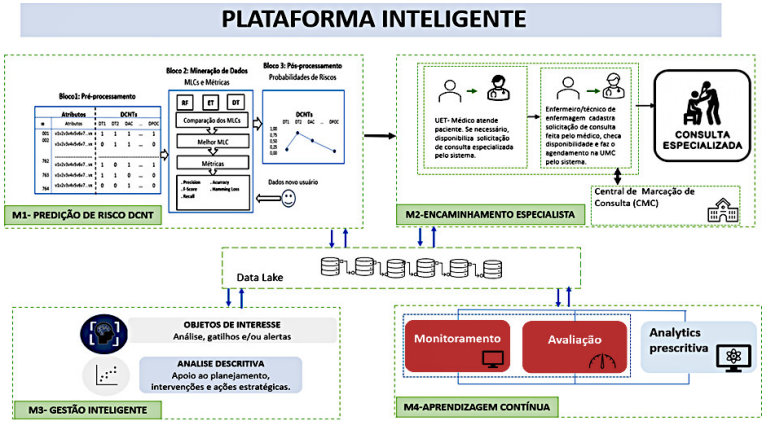


Fig. 1. Diagrama modular da PIPRDC
Fonte: Elaborado pelos autores

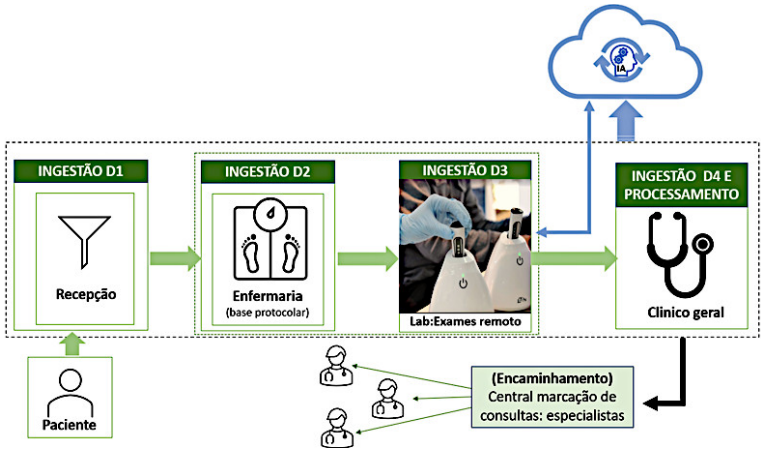


Fig. 2. Fluxo de atendimento ao paciente via plataforma: predição de risco DCNT
Fonte: Elaborado pelos autores

A plataforma ajuda a simplificar o atendimento do paciente, mudando o paradigma da abordagem reativa de prevenção e controle para uma abordagem proativa centrada no paciente, ou seja, de prever um evento, antes que ele ocorra. Do ponto de vista prático

da plataforma, o processo é iniciado quando o paciente se dirige ao atendimento ágil, proativo e personalizado da atenção primária à saúde, passa por uma triagem e depois de responder ao médico algumas perguntas (anamnese), conhece os seus riscos de desenvolver DCNT, é direcionado para um médico especialista, tem opções de tratamento mediante nível de urgência, se necessário, e começar a fazer escolhas de estilos de vida mais saudáveis.

Esse processo de fluxo do uso da plataforma envolvendo os atores recepção, enfermaria, laboratório remoto, clínico geral e especialista é mostrado na Figura 2.

Nesse contexto, o restante do texto está organizado da seguinte forma: a Seção 1 apresenta a metodologia; o *framework* é proposto na Seção 2; e, por fim, a Seção 3 apresenta a conclusão e trabalhos futuros.

1. METODOLOGIA

A metodologia aqui empregada tem como fundo uma abordagem proativa multifatorial e inteligência artificial com foco especificamente para o módulo predição de risco de DCNT. Os demais módulos são apresentados conceitualmente. Os recursos e bibliotecas da linguagem Python versão 3.9 foram essenciais para o desenvolvimento e apresentação de resultados do trabalho.

O módulo predição de risco de DCNT da plataforma foi estruturado com base em três esforços, incluindo análise dos principais modelos de prevenção e predição de doenças crônicas na literatura, consulta a médicos especialidades no Brasil e em Portugal, e construção do modelo preditivo do risco de DCNT.

1.1. Análise dos principais modelos de prevenção e predição de doenças crônicas na literatura

Uma revisão abrangente da literatura de artigos relacionados ao uso de modelos de aprendizado de máquina (ML) para predição de DNT. O resultado nos direcionou à obtenção de linhas de base para identificar as lacunas existentes e nos auxiliou na proposição da referida solução. A Tabela 1 apresenta um resumo dos trabalhos relevantes para o domínio em discussão em cada um dos quatro Grupos de Doenças Crônicas (GDC) com seus respectivos resultados, incluindo: 1 – Cardiovascular; 2 – Câncer; 3 – Respiratório e 4 – Diabetes.

1.2. Consulta a médicos especialidades no Brasil e em Portugal

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 *email*, mas menos de 10% responderam à pesquisa. Os vários grupos de médicos especialidades consultados 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.

Tabela 1. Trabalhos relevantes por grupo de doenças crônicas

						Precisão %	
GDC	**Autor**	**Ano**	**Dataset**	**Recursos**	**Classificadores**	**Homem**	Mulher
1	Chun et al.	2021	*Interviewer-adm electronic*	10	*Gradient boosted trees* (XGBoots)	83,30	83,60
	Yang et al.	2020	*electronic health record* – Zhejiang	30	*Random Forest*	78,70	
	Singh et al.	2018	*California Irvine Repository* (UCI)	11	*Logistic regression*	87,10	
	Sharma et al.	2017	Cleveland	14	*Decision tree*	93,20	
2	Hussan et al.	2022	*electronic health record (EHR)*	25	*Gradient Boosting*	86,00	
	Naji et al.	2021	*Breast Cancer Wisconsin*	11	*SVM*	97,20	
	Oyewo et al.	2020	*Github*	9	*Ensemble*	99,06	
	Nasser	2019	*site data world*	15	*ANN*	96,67	
3	Spathis e Vlamos	2019	*Clinical patients* Thes, Grécia	20	*Random Forest* (Asma-Copd)	80,30	
				20		97,70	
4	Li et al.	2021	EHR Optum	10	XGBoost	80,00	
	Rani	2020	Kaggle	8	*Decision Tree*	99,00	

Fonte: Elaborado pelos autores

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 significativos de riscos modificáveis e não modificáveis.

1.3. Construção do modelo preditivo do risco de DCNT

O *framework* básico *Knowledge Discovery in Database* (KDD) adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996) foi tomado para convergir o esforço dessa construção. 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 alinha o processo KDD, que compreende: pré-processamento, mineração de dados (classificadores *multi-label* e métricas de avaliação) e pós-processamento, cada uma com suas respectivas tarefas e fases de operação, conforme Figura 3.

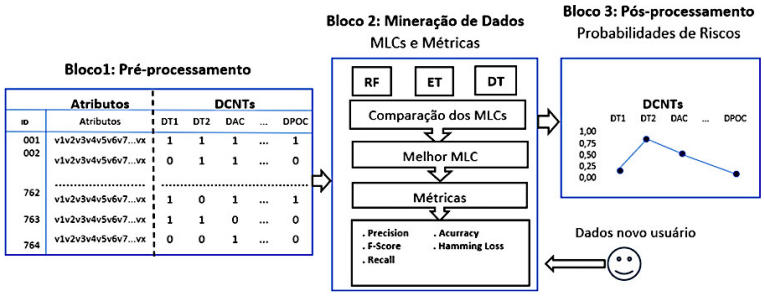


Fig. 3. Módulo predição de risco de DCNT
Fonte: Elaborado pelos autores

Bloco 1 – Pré-processamento: trata-se de cinco módulos e tarefas. A primeira, recuperação de dados, faz a importação de dados localmente, remotamente ou diretamente nos *sites*. A segunda, seleção de dados, define o tamanho da amostra, seleciona atributos relevantes e cria um banco de dados. A terceira, preparação de dados, limpa, integra e constrói dados. A quarta, exploração de dados, explora e verifica a qualidade dos dados brutos. A quinta, transformação de dados, normaliza os dados, seleciona subconjuntos de atributos, discretiza e generaliza.

Bloco 2 – *Data Mining* (DM) é o núcleo do modelo KDD e consiste em um processo contínuo no qual algoritmos inteligentes são aplicados de acordo com os requisitos e particularidades de cada técnica de aprendizagem de máquina (ML – *Machine Learning*) para identificar padrões e conhecimentos. Em geral, duas técnicas de ML são as mais utilizadas, a saber: Supervisionado e Não Supervisionado. Neste estudo adotaremos a técnica supervisionada com ênfase na Classificação Multirrótulos (MLC) e suas métricas de avaliação.

1.3.1. Classificação Multirrótulo (MLC)

Na técnica 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. Os MLC têm se mostrado muito promissores para aprender uma função a partir de um conjunto de instâncias em diversas aplicações, incluindo categorização de texto, classificação de imagens, recuperação de informações e têm se expandido para outros campos, incluindo diagnósticos médicos, bioinformática, etc.

A classificação de rótulos estuda o problema de aprender um mapeamento de instâncias para classificações em um conjunto predefinido de rótulos (Fürnkranz et al. 2008). Segundo Sharma e Mehrotra (2018), nos últimos anos, a popularidade do MLC aumentou devido à sua capacidade de resolver problemas em tempo real. Em seu estudo, Kassim, Mohan e Muneer (2021) destacam três abordagens gerais que estão sendo usadas para lidar com problemas de MLC, incluindo Métodos de Transformação de Problemas (PTM), Métodos de Adaptação de Algoritmos (AAM) e Métodos de Conjunto (EM). Os PTM transformam um conjunto de dados multirrótulo em um conjunto de dados de

rótulo único usando diferentes métodos de transformação, como rótulo menos frequente (LFL), rótulo mais frequente (MFL) ou escolhendo qualquer rótulo aleatoriamente (Lee et al. 2016). Ao contrário dos PTM, os AAM lidam com o problema de aprendizagem multirrótulo adaptando alguns algoritmos de ML diretamente para o cenário de classificação multirrótulo. Os EM requerem um classificador-base do método de adaptação do algoritmo ou método de transformação do problema e parâmetros relevantes do método (Kassim, Mohan e Muneer 2021).

Neste trabalho são considerados Métodos *Ensemble*, tomando os classificadores--base dos métodos de adaptação de algoritmos, incluindo *Random Forest* (RF), *XGBoost* e *Support Vector Machine* (SVM). Esses quatro classificadores básicos são então combinados com várias técnicas de classificação multirrótulos de métodos de transformação de problemas, incluindo *Classifer Chains* (CC), *Ensemble Classifer Chains* (ECC), *Label Power Set* (LP) e *Calibrated Label Ranking* (CLR).

RF é um método conjunto baseado na construção de vários classificadores de árvore de decisão independentes em diferentes subconjuntos do conjunto de dados. 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). XGBoost é um classificador multirrótulo. De acordo com Chen e Guestrin (2016), o XGBoost é um sistema escalável de aprendizado de máquina para cultivo de árvores, oferecendo cultivo paralelo de árvores. SVM é um tipo de classificador que calcula classificações construindo hiperplanos em um espaço multidimensional que separa amostras de diferentes classes. Sua extensão *Ranking-SVM* transforma o SVM, que é uma abordagem tradicional de rótulo único de alta eficiência, em um método usado diretamente para classificação multirrótulo (Gerevini et al. 2018). CC é o transformador de relevância binária aprimorado ao construir uma cadeia condicionada bayesiana. Semelhante à relevância binária, a cadeia classificadora trata cada rótulo como um classificador separado, mas não independente (Zhang et al. 2022). O ECC contém vários CC com ordens diferentes e pode ser aplicado para estudar dependências entre rótulos (Read et al. 2021). A abordagem LP transforma um problema multirrótulo em um problema multiclasse de rótulo único, que é treinado em todas as combinações únicas de rótulos encontradas nos dados de treinamento (Costa Júnior et al. 2017). O CLR explora as correlações entre pares de rótulos e transforma a tarefa MLC em um problema de classificação de rótulos, onde a pontuação de cada rótulo é determinada por meio de comparações entre esse rótulo e o rótulo restante (Moral-García e Abellán 2021).

1.3.2. Métricas de avaliação – Modelos MLC

As métricas de avaliação dos classificadores multirrótulos são uma derivação e adaptação do desempenho das diversas medidas de classificação multiclasse. Em classificadores multirrótulos, 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. Este trabalho emprega diversas métricas para avaliar o desempenho de classificadores multirrótulos, incluindo precisão, acurácia, pontuação F, *recall*, perda de *Hamming*, perda 0/1 e similaridade de Jaccard.

Etapa 3 – Pós-processamento: Trata-se de dois módulos. O sétimo, interpretação e avaliação, consiste em verificar a qualidade dos padrões que representam o conhecimento com base em medidas interessantes. Os resultados da avaliação dão o direcionamento para alterar ou não as medidas em prol de melhores resultados. O oitavo módulo, conhecimento, diz respeito à utilização e *feedback* dos padrões e resultados de descoberta obtidos no processo de mineração de dados (*Data Mining*). Isto determina a eficácia da ferramenta preditiva em favor das DCNT.

2. RESULTADOS DA PIPRDC

Nosso ponto de partida para implementação da Plataforma Inteligente de Predição do Risco de Doenças Crônicas (PIPRDC) foi a aprovação do experimento pela comissão de ética para saúde do HE-UFP (Hospital Escola da Universidade Fernando Pessoa). Após aprovação, a direção do HE-UFP nos forneceu um conjunto de dados com 852 542 linhas de dados não estruturados. Esses dados foram objeto de um pré-processamento e enquadramento no Modelo Preditivo de Risco de DCNT composto por fatores preditivos não modificáveis e modificáveis, incluindo: sociodemográfico, histórico-familiar, bioquímico, comportamental, psicossocial, clínico e ambiental. Esse conjunto de dados nutriu 38 atributos que serviram de base de treino e teste para a PIPRDC. Considerando o foco da plataforma para DCNT, somente 892 pacientes foram elegíveis para os testes.

Para realizar o objetivo do estudo adotámos três classificadores *multi-label* e cinco métricas de desempenho de generalização dos classificadores. Usámos 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, realizámos experimentos comparativos para avaliar a proporção de previsões corretas introduzida pelos modelos MLC. Os resultados das previsões corretas são mostrados na Figura 4. Esse resultado usa a métrica de avaliação de desempenho acurácia do modelo. Conforme mostrado na Figura 4, entre os três modelos MLC, os classificadores *Extra Trees* (ET) e *Random Forest* (RF) apresentaram desempenho de 82,68% e 79,33% respectivamente.

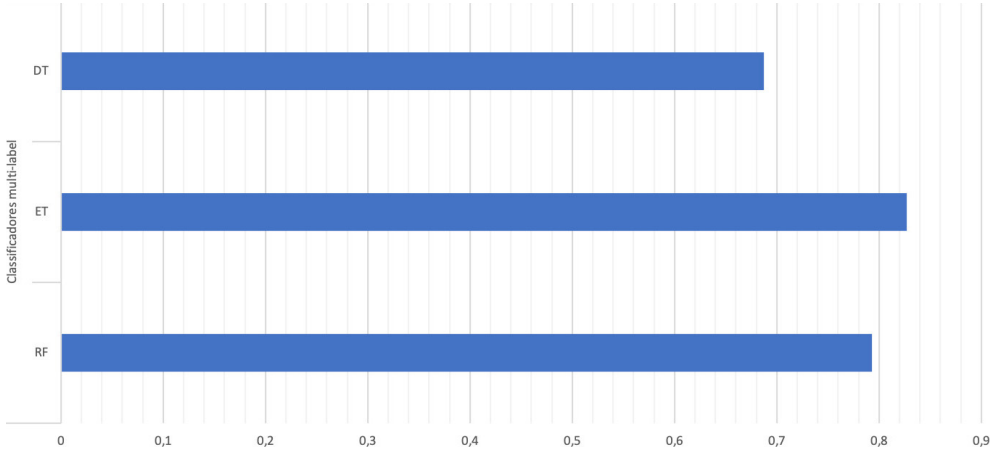
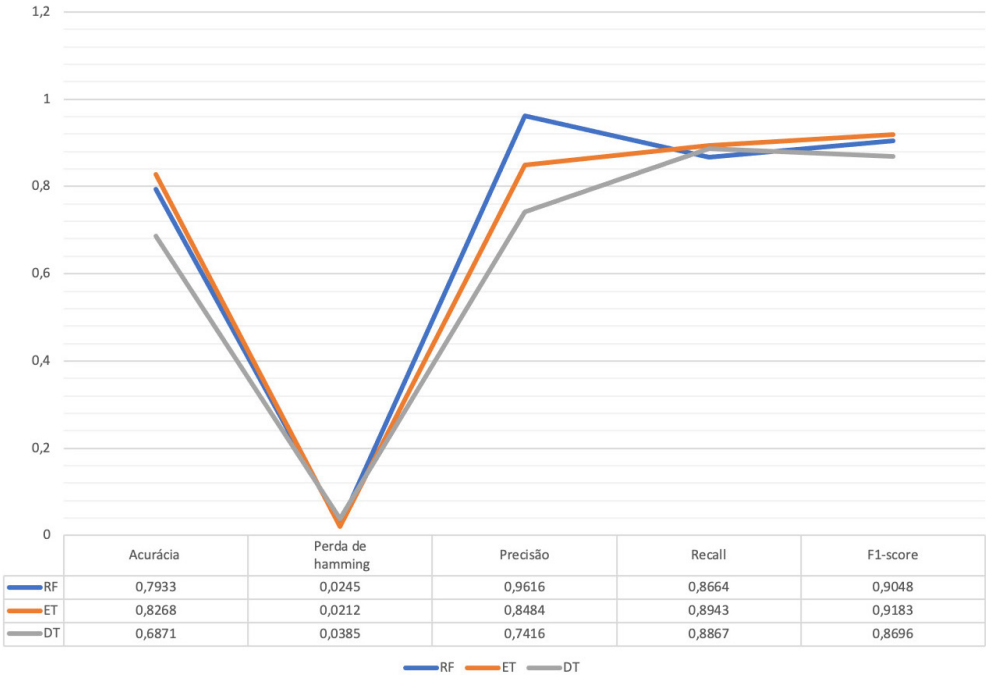


Fig. 4. Precisão de previsões corretas dos MLC
Fonte: Elaborado pelos autores

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-Cruz, Fast e Sonke 2021). Destacamos ainda que os modelos MLC RF e DT apresentaram desempenho de avaliação de 79,33 e 68,71%, bem como contagem de previsões corretas de 142 e 123, respectivamente.

Buscando melhorar ainda mais o desempenho de generalização dos modelos, comparámos seus desempenhos usando além de acurácia, métricas perdas de *hamming*, precisão, *recall* e *F1-score*, conforme mostrado na Figura 5. Em seus resultados, a PIPRDC usa a métrica Precisão, na qual refere-se a quão bem o algoritmo foi treinado para prever as classes corretas de DCNT e o fez. A Figura 5 apresenta o desempenho do MLC ET, RF e DT com base nas cinco métricas. O MLC ET apresentou melhor desempenho para as métricas *Accuracy*, *Recall* e *F1-score* e menor perda de *hamming* para predição de DCNT. Porém, o MLC RF com pontuações próximas ao MLC ET, superou-o significativamente na métrica de precisão com resultado igual a 96,16% e obteve a segunda menor perda de *hamming*, que foi de 2,45%.

A Figura 6 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 Figura 6 mostra 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 optámos por usar método de médias *macro avg*



	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

Fig. 5. Desempenho dos MLC ET, RF e DT
Fonte: Elaborado pelos autores

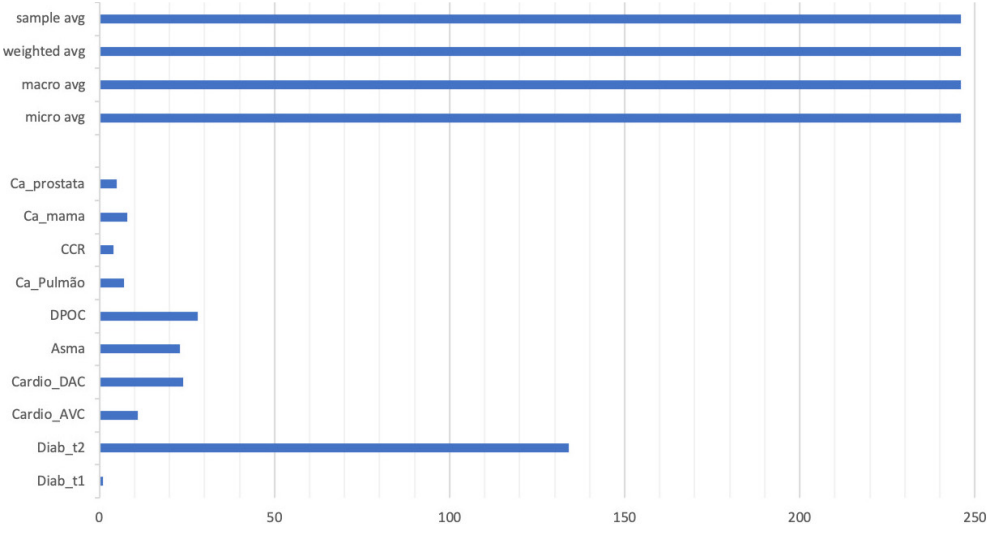


Fig. 6. Métodos de médias e contagem de amostra no MLC ET
Fonte: Elaborado pelos autores

combinado com as métricas Precisão, *Recall* e *F1-score* para cada uma das dez DCNT. Entre os quatro métodos de médias a *macro avg* se destaca por tratar todas as classes igualmente importantes, independentemente de seus valores de amostra, ou seja, de suporte.

Conforme mostrado na Figura 7, 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.

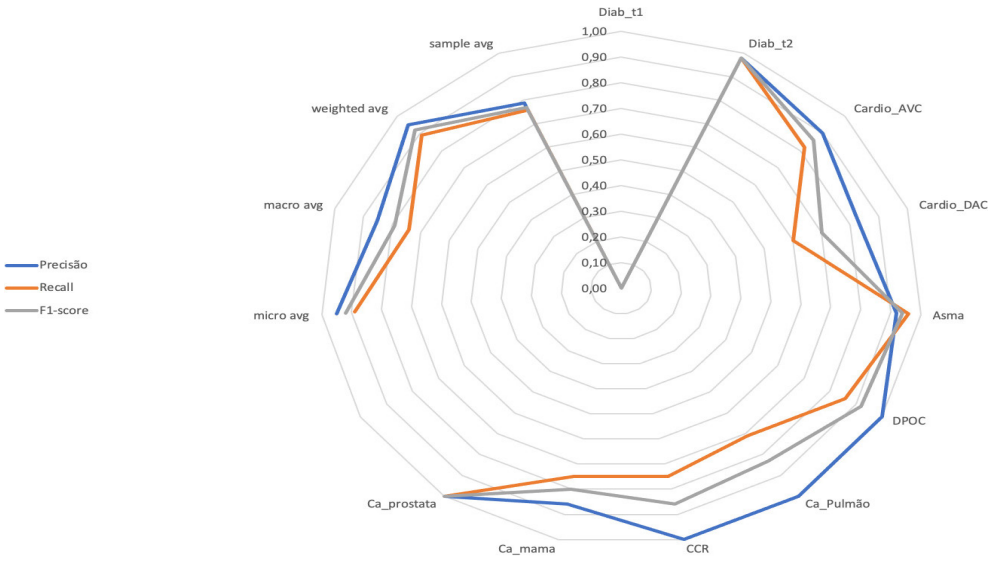


Fig. 7. Métodos de médias e desempenho do MLC ET por DCNT
Fonte: Elaborado pelos autores

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 decisão. 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 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 seus impactos para a sociedade e governos de todo o mundo, esse estudo é um forte candidato a capacidade de resposta aos problemas de DCNT. Além de auxiliar os profissionais da saúde na detecção precoce de DCNT, reduzindo o risco de complicações, invalidez e morte e custo de saúde na atenção primária de saúde, ele pode contribuir fortemente para o 3.º item dos Objetivos de Desenvolvimento Sustentável-ODS e fornecer respostas específicas e precisas da predição de DCNT e que proporcionará benefícios de sobrevivência e apoio na redução da alta taxa de mortalidade por DCNT. Os trabalhos futuros compreendem hospedagem da plataforma na nuvem e disponibilizar para parceiros.

REFERÊNCIAS

BERNAL, R. T. I., et al., 2019. Indicadores de doenças crônicas não transmissíveis em mulheres com idade reprodutiva, beneficiárias e não beneficiárias do Programa Bolsa Família. *Revista Brasileira de Epidemiologia* [Em linha]. **22**(Supl. 2) [consult. 2023-09-20]. Disponível em: https://doi.org/10.1590/1980-549720190012.supl.2.

BRASIL. Ministério da Saúde, 2022. *DATASUS. Tabnet* [Em linha]. Brasília, DF: Ministério da Saúde [consult. 2023-09-20]. Disponível em: https://datasus.saude.gov.br/informacoes-de-saude-tabnet.

CHEN, T. Q., e C. GUESTRIN, 2016. Xgboost: A Scalable Tree Boosting System. Em: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge, Discovery and Data Mining, 13-17 August 2016, San Francisco* [Em linha]. Nova Iorque: Association for Computing Machinery, pp. 785-794 [consult. 2023-09-20]. Disponível em: https://doi.org/10.1145/2939672.2939785.

CHUN, M., et al., 2021. Utility of single versus sequential measurements of risk factors for prediction of stroke in Chinese adults. *Scientific Reports* [Em linha]. **11** [consult. 2023-09-20]. Disponível em: https://doi.org/10.1038/s41598-021-95244-8.

COSTA JÚNIOR, J. D., et al., 2017. Label Powerset for Multi-Label Data Streams Classification with Concept Drift. Em: *5th Symposium on knowledge Discovery, mining and learning (KDMile), 2-4 October 2017, Uberlândia, MG, Brazil* [Em linha] [consult. 2023-09-20]. Disponível em: https://www.researchgate.net/publication/331047823_Label_Powerset_for_Multi-label_Data_Streams_Classification_with_Concept_Drift.

FAYYAD, U., G.PIATETSKY-SHAPIRO, e P. SMYTH, 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*. **17**(3), 1-6.

FÜRNKRANZ, J., et al., 2008. Multilabel classification via calibrated label ranking. *Machine Learning*. **73**(2), 133-153.

GEREVINI, G., et al., 2018. Slugging Attenuation Using Nonlinear Model Predictive Control in Offshore Oil Production. *Journal of Petroleum Science and Engineering* [Em linha]. **165**, 187-198 [consult. 2023-09-20]. Disponível em: https://doi.org/10.1016/j.petrol.2018.01.054.

HUSSAN, H., et al., 2022. Utility of machine learning in developing a predictive model for early-age-onset colorectal neoplasia using electronic health records. *PLOS ONE* [Em linha]. **17**(3) [consult. 2023-09-20]. Disponível em: https://doi.org/10.1371/journal.pone.0265209.

JUAN-CRUZ, C., M. F. FAST, e J.-J. SONKE, 2021. A Multivariable study of deformable image registration evaluation metrics in 4DCT of thoracic cancer patients. *Physics in Medicine & Biology* [Em linha]. **66**(3) [consult. 2023-09-20]. Disponível em: https://doi.org/10.1088/1361-6560/abcd18.

KASSIM, B., S. MOHAN, e K. A. MUNEER, 2021. Modified ML-kNN and rank SVM for multi-label pattern classification. *Journal of Physics: Conference Series* [Em linha]. **1921**(1) [consult. 2023-08-31]. Disponível em: https://iopscience.iop.org/article/10.1088/1742-6596/1921/1/012027.

KOUCHAKI, S., et al., 2020. Multi-Label Random Forest Model for Tuberculosis Drug Resistance Classification and Mutation Ranking. *Frontiers in Microbiology* [Em linha]. **11** [consult. 2023-09-20]. Disponível em: https://doi.org/10.3389/fmicb.2020.00667.

LEE, J., et al., 2016. An Approach for multi-label classification by directed acyclic graph with label correlation maximization. *Information Sciences* [Em linha]. **351**, 101-114 [consult. 2023-09-20]. Disponível em: https://doi.org/10.1016/j.ins.2016.02.037.

LI, L., et al., 2021. Performance assessment of different machine learning approaches in predicting diabetic ketoacidosis in adults with type 1 diabetes using electronic health records data. *Pharmacoepidemiology & Drug Safety* [Em linha]. May, **30**(5), 610-618 [consult. 2023-09-20]. Disponível em: https://doi.org/10.1002/pds.5199.

MORAL-GARCÍA, S., e J. ABELLÁN, 2021. Required mathematical properties and behaviors of uncertainty measures on belief intervals. *International Journal of Intelligent Systems* [Em linha]. **36**(1) [consult. 2023-09-20]. Disponível em: https://doi.org/10.1002/int.22432.

NAJI, M. A., et al., 2021. Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis. *Procedia Computer Science* [Em linha]. **191**, 487-492 [consult. 2023-09-20]. Disponível em: https://doi.org/10.1016/j.procs.2021.07.062.

NASSER, I., 2019. Lung Cancer Detection Using Artificial Neural Network. *International Journal of Engineering and Information Systems (IJEAIS)* [Em linha]. Mar., **3**(3), 17-23 [consult. 2023-09-20]. Disponível em: https://ssrn.com/abstract=3700556.

OYEWO, O. A., e O. K. BOYINBODE, 2020. Prediction of Prostate Cancer using Ensemble of Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications (IJACSA)* [Em linha]. **11**(3) [consult. 2023-09-20]. Disponível em: https://doi.org/10.14569/IJACSA.2020.0110318.

RAHIMLOO, P., e A. JAFARIAN, 2016. Prediction of diabetes by using artificial neural network logistic regression statistical model and combination of them. *Bulletin de la Société Royale des Sciences de Liège* [Em linha]. **85** [consult. 2023-09-20]. Disponível em: https://doi.org/10.25518/0037-9565.5938.

RANI, K. J., 2020. *Diabetes Prediction Using Machine Learning* [Em linha] [consult. 2023-09-20]. Disponível em: https://doi.org/10.32628/CSEIT206463.

READ, J., et al., 2021. Classifier chains: A review and perspectives. *Journal of Artificial Intelligence Research*. **70**, 683-718.

SHARMA, S., e D. MEHROTRA, 2018. Comparative Analysis of Multi-Label Classification Algorithms. Em: *First International Conference on Secure Cyber Computing and Communication (ICSCCC), 15-17 December 2018.*

SHARMA, T., S. VERMA, e S. KAVITA, 2017. Prediction of heart disease using Cleveland dataset: A machine learning approach. *International Journal of Recent Research Aspects*. **4**(3), 17-21.

SINGH D. A. A. G., E. J. LEAVLINE, e B. S. BAIG, 2017. Diabetes prediction using medical data. *Journal of Computational Intelligence in Bioinformatics*. **10**(1), 1-8.

SPATHIS, D., e P. VLAMOS, 2019. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health Informatics Journal* [Em linha]. Sept., **25**(3), 811-827 [consult. 2023-09-20]. Disponível em: https://doi.org/10.1177/1460458217723169.

WORLD HEALTH ORGANIZATION, 2022. *Noncommunicable diseases* [Em linha]. [Genebra]: World Health Organization [consult. 2023-09-20]. Disponível em: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases#:~:text=Key%20facts,%2D%20and%20middle%2Dincome%20countries.

YANG, X., et al., 2020. Prevalence of high-risk coronary plaques in patients with and without metabolic syndrome and the relationship with prognosis. *BMC Cardiovascular Disorders*. **20**(1), 73.

ZHANG, C., et al., 2022. Synthetic biology in chimeric antigen receptor T (CAR T) cell engineering. *ACS Synthetic Biology* [Em linha]. **11**(1), 1-15 [consult. 2023-09-20]. Disponível em: https://doi.org/10.1021/acssynbio.1c00256.