MÁRIO LUIZ TRONCO

TÉCNICAS DE SELEÇÃO DE ATRIBUTOS ASSOCIADAS A ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA A PREDIÇÃO DE DADOS NA ÁREA DA SAÚDE

Relatório de Pós-doutorado realizado na Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas – IBILCE, Departamento de Ciências de Computação e Estatística – DCCE, São José do Rio Preto.

Supervisor(a): Prof. Dr. Carlos Roberto Valêncio

São José do Rio Preto

2026

Sumário

1 INTRODUÇÃO **2**

2 OBJETIVOS **3**

3 REVISÃO DA LITERATURA – ARTIGOS CORRELATOS **4**

3.1 Seleção de Atributos e Predição Relacionados a Doenças Cardíacas 4

3.2 Seleção de Atributos e Predição relacionados a Câncer de mama 5

3.3 Algoritmos de Classificação 8

3.4 Feature Selection 11

4 METODOLOGIA **13**

4.1 Repositório de Imagens 14

4.2 Estratégia de Testes 15

5 PARTE EXPERIMENTAL 16

5.1 Primeiro Conjunto de Dados – Breast Cancer Wisconsin Diagnostic 16

5.1.1 Conclusões sobre a influência dos métodos de seleção de atributos no desempenho dos classificadores 37

5.2 Segundo Conjunto de Dados – Heart Disease Dataset 38

5.3 Conclusões 47

6 BIBLIOGRAFIA 48

7 ANEXOS – ARTIGOS ACEITOS, SUBMETIDOS E EM PRODUÇÃO 57

7.1 Artigos Publicados 57

7.2 Artigos Aceitos para Publicação 58

7.3 Artigos Submetidos para Publicação 61

1. INTRODUÇÃO

O processo de geração e manipulação de dados vem se modificando em função do desenvolvimento de novas tecnologias. Houve significativa mudança na forma como os dados são manipulados e tratados, além da forma como o conhecimento é extraído, em função do crescente aumento massivo do volume gerado (Big Data), implicando também em maior custo computacional para atender de forma efetiva a demanda. Desta forma, o desafio nesta área passa pela busca de algoritmos e processos capazes de tornar eficiente e minimamente viável a gestão deste conhecimento.

Na área da saúde, o volume de dados gerado cresceu exponencialmente nos últimos anos, gerando repositórios que podem ser explorados para teste de diversas técnicas computacionais, principalmente relacionadas com a predição de diagnósticos de doenças conhecidas.

Dados médicos fazem parte do conceito de *Big Data*, uma vez que os dados de saúde utilizados por sistemas estão na ordem de terabytes e continuam crescendo a uma taxa de diversos exabytes ao ano. A maior parte dos dados é oriunda de prontuários eletrônicos, anotações médicas, mas também existem fontes mais modernas e estruturadas como dados clínicos e resultados de procedimentos médicos. Dados são gerados tão rapidamente que não é possível garantir que são armazenados e atualizados na mesma frequência com que são criados. As diferentes formas de se obter dados podem gerar valores inconstantes e não acurados e existe grande valor na retirada de informação desses dados (ASRI et al., 2016). Outra característica comum em dados médicos é a alta dimensionalidade. Devido à forma como são gerados é comum que esses dados apresentem muitos atributos sendo que nem todos possuem grande impacto na informação final que pode ser retirada dos mesmos, dificultando, assim, a extração de conhecimento útil e tornando os processos aplicados mais dispendiosos (FAN & LI, 2006; FAYYAD et al., 1996).

A análise de dados médicos permite uma série de vantagens ao serem exploradas, como aprimorar o atendimento ao paciente, melhorar o desempenho de exames e da qualidade de serviço prestada pelos profissionais da área, gerir e planejar o uso de recursos das instituições de saúde e a melhoria do potencial diagnóstico por meio da extração de conhecimento e aplicação de técnicas de predição.

Com o avanço da tecnologia da Informação e da Internet, tem havido um crescimento quase que exponencial no volume de dados gerado em todo o mundo. Aplicações antes baseadas em operações pontuais (e muitas vezes manuais) passaram a utilizar o conceito de coleta automatizada, em ambientes distribuídos, produzindo dados em quantidades nunca experimentadas pela tecnologia. A automação na coleta de dados de sensores e de imagens aliada ao advento dos dispositivos móveis atuando em ambientes de alta conectividade vem proporcionando cenários desafiadores em termos de disponibilidade da informação. Assim como os demais segmentos da sociedade, o advento das tecnologias da Internet, o processo de digitalização da informação médica, implementação de um grande número de prontuários eletrônicos, sistemas de informação hospitalar, incremento de dispositivos

computacionais com sensores coletores de dados, dados genômicos, imagens médicas, dados das mídias sociais e muitos outros dados eletrônicos que surgiram, vem incrementado este universo do Big Data da área da saúde (LV et. Al, 2020; BATKO & SLEZAK, 2022).

Diante deste volume de dados, o processo de extração de conhecimento em grandes bases de dados e, mais recentemente, a aplicação deste processo para Big Data exige recursos computacionais intensos. A tecnologia que suporta o armazenamento, preparação e análise desses conjuntos de dados existe, mas seu alto índice de crescimento previsto para os próximos anos exige que as estratégias sejam constantemente repensadas e aprimoradas (OSMAN et. al, 2013). Neste sentido, o trabalho junto aos grupos de pesquisas que atuam neste âmbito permite avançar de modo significativo no domínio do conhecimento, uma vez que há um conjunto de expertises desenvolvidos no decorrer de décadas dedicadas ao enfrentamento de desafios relevantes e de impacto junto à comunidade científica, e em especial, aplicados a área da saúde.

Técnicas de análise tradicionais não são adequadas para se extrair conhecimento de Big Data. No intuito de se classificar dados, detectar padrões e inferir relacionamentos estatísticos para uma abordagem de mineração eficaz, é preciso que os algoritmos de análise sejam escaláveis e passíveis de paralelização em um ambiente distribuído. Algoritmos de aprendizagem de máquina são o principal fundamento científico empregado na análise de Big Data (OSMAN et. al, 2013). No entanto, em bancos de dados médicos compostos por diversos atributos, nem sempre o uso de sua totalidade em classificadores produz os melhores resultados. Selecionar atributos que de fato tenham importância para o processo de classificação constitui uma importante etapa do processo, podendo resultar em melhores resultados de classificação, menor tempo de processamento, além de resultados mais confiáveis e apropriados para as tomadas de decisão dos profissionais da área. No presente trabalho foi realizado um estudo completo da influência dos diversos atributos no processo de classificação utilizando dois bancos de dados representativos (câncer de mama e doenças cardíacas). Os resultados obtidos reforçam a importância desta etapa no processo de classificação.

2. **OBJETIVOS**

Os objetivos do presente projeto envolveram testar algoritmos de seleção de atributos para melhorar a predição e análise da influência de determinado atributo sobre o resultado de classificação, combinado com diversos algoritmos de aprendizado de máquina (Naive Bayes, J48, Random Trees e K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), Redes Neurais Artificiais (ANN), e Logistic Regression (LR)).

Neste cenário e levando em conta o potencial dos algoritmos de classificação como forma de predição de informações relacionadas à saúde, no presente projeto buscou-se realizar a associação de diversas técnicas de seleção de atributos com algoritmos de classificação com múltiplos conjuntos

de dados da área da saúde, buscando as combinações que geram resultados mais acurados e mais rápidos. Para este fim foram utilizados repositórios que abordam câncer de mama e doenças cardíacas.

3. REVISÃO DA LITERATURA- ARTIGOS CORRELATOS

3.1 Seleção de Atributos e Predição relacionados a Doenças cardíacas

Um número alarmante de aproximadamente 18 milhões de pessoas morrem em função de doenças cardíacas anualmente no mundo, com projeções indicando um aumento para 23 milhões de mortes até 2030 (RAVI & MADHAVAN, 2022). A natureza insidiosa das doenças cardiovasculares manifesta-se frequentemente em progressão assintomática, com ataques cardíacos ou acidentes vasculares cerebrais servindo como apresentações clínicas iniciais (SAYYED, 2023). Os sintomas, incluindo dor no peito, dispneia, náusea, tontura, visão turva ou síncope, podem indicar patologia vascular subjacente. Os fatores de predisposição incluem estresse, tabagismo, hipertensão, diabetes, além de obesidade, estilo de vida sedentário, falta de atividade física e predisposição hereditária. A previsão precoce e precisa de doenças cardíacas é fundamental para melhorar a qualidade de vida dos pacientes.

Os Algoritmos de Aprendizado de Máquina (*Machine Learning Algorithms* - MLA´s) ganharam destaque na área da saúde devido a sua capacidade de analisar dados médicos complexos e construir modelos preditivos. A proliferação de dados médicos e avanços computacionais facilitou abordagens inovadoras para a previsão de doenças. Os MLA´s são promissores em padrões de aprendizagem de diversos atributos do paciente, incluindo métricas clínicas, dados demográficos e histórico médico, para o desenvolvimento de modelos preditivos precisos. Diversos pesquisadores vem investigando a aplicação de MLA´s para previsão de doenças cardíacas (HDP – *heart disease prediction*), empregando diversos conjuntos de dados e modelos.

Sun et al. (2023) exploraram a precisão de indicadores de saúde física auto mensuráveis na previsão de doenças cardíacas, empregando Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), árvore de decisão (DT) e modelos de Random Forest (RF). Utilizaram um subconjunto de um banco de dados abrangente, com foco em diversos atributos, com o objetivo de aumentar a precisão algorítmica ao incorporar indicadores físicos facilmente auto mensuráveis.

Rani et al. (2021) propuseram um sistema híbrido de apoio à decisão para previsão precoce de doenças cardíacas com base em parâmetros clínicos, utilizando algoritmos de aprendizado de máquina como K-Nearest Neighbor (SVM), Bayes Ingênuo (NB), Logistic Regression (LR), modelos de Random Forest (RF) e classificadores adaboost. Empregaram imputação multivariada e um algoritmo hibridizado de seleção de recursos combinando Algoritmo Genético (GA) e Extração de recursos recursivos (RFE). Apesar das limitações, como a incapacidade de avaliar a gravidade da doença, o sistema

forneceu uma ferramenta valiosa para áreas remotas que carecem de instalações médicas especializadas.

Katarya et al. (2021) enfatizaram a importância de prever doenças cardíacas em meio a volumes crescentes de dados no setor médico. Seu estudo comparou o desempenho de vários Algoritmos de aprendizado de máquina, concentrando-se num conjunto de dados específico, limitando assim potencialmente a generalização.

Kavitha et al. (2021) propuseram uma nova abordagem de Aprendizado de máquina utilizando regressão, classificação e técnicas como modelos de Random Forest (RF) e Decision Trees (DT), aplicadas ao conjunto de dados de doenças cardíacas de Cleveland. No entanto, o estudo careceu de detalhes da distribuição demográfica e de comparação com modelos de última geração.

Jindal et al. (2021) concentrou-se na previsão da probabilidade de doenças cardíacas com base no histórico médico do paciente, empregando classificadores Logistic Regression (LR), K-Nearest Neighbor (KNN) e modelos de Random Forest (RF) em um conjunto de dados limitado. Eles reconheceram a necessidade de conjuntos de dados maiores e da exploração de algoritmos de Aprendizado de máquina adicionais e recursos essenciais para maior precisão preditiva.

Shah et al. (2020) discutiram técnicas de mineração de dados e Aprendizado de Máquina para predição de doenças cardíacas, utilizando algoritmos de aprendizagem supervisionada no banco de dados de Cleveland. O estudo destacou variações potenciais na precisão da previsão com base na composição do conjunto de dados e na seleção de atributos.

Mohan et al. (2019) propuseram um novo método para predição de doenças cardíacas, enfatizando a importância dos recursos e empregando várias técnicas de classificação. Apesar de limitações como o tamanho do conjunto de dados e a falta de análise comparativa, a abordagem teve como objetivo melhorar o desempenho preditivo.

Gavhane et al. (2018) propuseram uma aplicação para previsão de vulnerabilidade a doenças cardíacas com base em sintomas básicos, contando com sensores corporais. No entanto, foram observadas limitações, incluindo a representatividade do conjunto de dados e a dependência de sintomas rudimentares, juntamente com considerações éticas relativas a preconceitos algorítmicos e questões de privacidade.

3.2 Seleção de Atributos e Predição relacionados a Câncer de mama

O câncer de mama é o tipo de malignidade não cutânea mais frequente no gênero feminino. Com mais de 280.000 novos casos em 2022, com mais de 40.000 mortes, o câncer de mama ainda ocupa os primeiros lugares entre os carcinomas femininos mais letais em todo o mundo (ABBA et al., 2020). Este tipo de malignidade descreve uma condição em que as células se desenvolvem descontroladamente no órgão mamário, tanto no sexo feminino quanto no masculino. Os tumores de mama são o crescimento aberrante de células que assumem a aparência de caroços, microcalcificações e distorções artificiais (ABIODUN et al., 2018). Mais de dois milhões de mulheres são diagnosticadas com este tipo de malignidade todos os anos, tornando-o o cancro não cutâneo

mais prevalente em mulheres à escala global. Por isso, o diagnóstico precoce do câncer é essencial para um tratamento eficaz e para aumentar a probabilidade de sobrevivência.

Diferentes metodologias de imagens médicas são frequentemente realizadas para diagnosticar o câncer de mama e acompanhar o desenvolvimento de tumores. Mamografia, ultrassonografia (US), ressonância magnética (RM), tomografia computadorizada (TC), tomografia por emissão de pósitrons e imagem por micro-ondas são algumas das técnicas utilizadas (ARA et al., 2021). O método de diagnóstico mais popular e eficaz para o câncer de mama é a mamografia. A ultrassonografia da mama raramente é usada para diagnosticar o câncer de mama porque não consegue identificar sinais precoces, como microcalcificações.

Estudos observacionais indicam uma diminuição de 40% na letalidade (BANSAL et al., 2022; BOATENG & ABAYE, 2019) após exames de mamografia. Independentemente dos benefícios dos sistemas de imagem no diagnóstico do câncer de mama, diversos procedimentos radiológicos e laboratoriais são ainda usados para confirmação final para fazer um diagnóstico preciso. Embora as mamografias continuem a ser vantajosas na detecção do câncer de mama nas suas fases iniciais utilizando raios X de baixas doses, tal como acontece com outros sistemas de imagem, ainda existem desvantagens, já que uma biópsia por agulha é solicitada a cada 15 entre 1000 mulheres que fizeram uma mamografia, e de 10 a 13 dessas biópsias revelarão falsos positivos (BOATENG et al., 2020). Portanto, é essencial incorporar técnicas de classificação adicionais nas técnicas de diagnóstico de câncer. Para isso, os profissionais médicos, em conjunto com os radiologistas, começaram a diagnosticar o câncer de mama num padrão muito detalhado, no qual são registradas duas visualizações para cada mama ao longo de todo o processo de mamografia: a oblíqua mediolateral, que é uma visão lateral, e a craniocaudal, que é uma visão superior, em conformidade com o Breast Imaging Reporting & Data System (BI-RADS), que foi desenvolvido em 1986 como um sistema de classificação de anomalias mamárias (BROEDERS et al., 2012). O BI-RADS fornece relatórios de mamografia que incluem classificações para a caracterização dos estágios do câncer de mama, permitindo resultados uniformes de imagens da mama. As classes são rotuladas de 0 a 6, sendo que um sétimo grupo foi recentemente introduzido – para representar o estado maligno (CHENG & LI, 2012).

Para proporcionar um melhor processamento de imagens, como segmentação e identificação para diagnóstico, refinamento de imagens e extração de características, foi necessário um ajuste nos procedimentos do departamento de radiologia, principalmente na fase final de notificação dos resultados, desde tenra idade (CONNELY, 2016). Como resultado, abordagens de inteligência artificial (IA), tanto subconjuntos de aprendizado profundo (DL) e aprendizado de máquina (ML) têm sido empregados no diagnóstico e tratamento de doenças para fins de classificação e processamento de imagens (DEBELEE et al., 2020; ELTER et al., 2007). Isto se deve ao sucesso de diversas aplicações de IA na medicina (FAVATI et al., 2022; GHALI et al., 2020; GHOSH, 2019) e outros setores (GIRI & SARAVANAKUMAR, 2017; HARUNA et al., 2021), que têm demonstrado alto impacto na exatidão e alta precisão dos resultados. Métodos de diagnóstico auxiliado por computador foram desenvolvidos para usar extração automatizada de características e

classificação para ajudar os médicos a diagnosticar e detectar tumores de mama (HASSANIPOUR et al., 2019). Seus objetivos são o diagnóstico eficaz do câncer, a interpretação precisa, a redução de erros e a relação custo-benefício. De acordo com um estudo recente (HEER et al., 2020), novos métodos baseados em redes neurais convolucionais (CNNs) melhoraram o desempenho do diagnóstico clínico e aumentaram a produtividade da imagem mamária.

A aplicação de tecnologias de IA torna o diagnóstico do câncer de mama mais eficiente, diminuindo as taxas de mortalidade em 30% a 70%, o que auxilia os radiologistas na sua capacidade de avaliar imagens e detectar anomalias nas fases iniciais (HUANG et al., 2020). Embora os modelos de IA sejam benéficos na classificação de doenças malignas, ainda há um longo caminho no tratamento de dados vastos e complexos e uma enorme diminuição nas taxas de mortalidade por câncer de mama, principalmente no gênero feminino, onde é mais prevalente. Os radiologistas utilizam frequentemente exames de mamografia para procurar os indicadores mais típicos de tumores mamários, geralmente registrando-os em categorias BI-RADS. Tal avaliação visa principalmente padronizar e melhorar a clareza e consistência da terminologia mamográfica utilizada pelos radiologistas e médicos encaminhadores (ISLAM & POLY, 2019). O procedimento considera todos os detalhes importantes, como o formato da massa (redondo, oval, lobular, irregular), a margem (circunscrita, microlobulada, obscurecida, mal definida, espiculada), a densidade do tumor (alta, iso, baixa, contendo gordura), bem como a idade do paciente, todos levando à gravidade e categoria (benigno ou maligno) do câncer de mama.

Em relação à aplicação de modelos de aprendizado de máquina em condições de câncer, Huang et al. (2020) destacaram o potencial dos modelos de IA para revolucionar o diagnóstico e prognóstico de câncer, particularmente na melhoria da precisão, velocidade e eficiência. Discutiram diferentes abordagens de IA que podem ser utilizadas para o diagnóstico e prognóstico e os desafios e limitações que devem ser abordados para que a IA seja eficazmente implementada na prática clínica. Estes incluem qualidade e quantidade de dados, questões regulamentares, considerações éticas e a necessidade de colaboração entre diferentes disciplinas (LAHOURA et al., 2021).

Venmathi et al. (2015) revisaram pesquisas sobre mamografia auxiliada por computador para diagnóstico de câncer de mama. Seu estudo cobriu vários aspectos da classificação de imagens médicas, incluindo os tipos de recursos usados para classificação, as metodologias de avaliação, as métricas de desempenho e o índice usado para avaliar a eficácia dos algoritmos de classificação. Os autores também discutiram os desafios e limitações do uso da mamografia auxiliada por computador para diagnósticos de câncer de mama, como a necessidade de grandes conjuntos de dados, a sensibilidade e especificidade dos algoritmos de classificação e a interpretação dos resultados. Kourou et al. (2015) discutiram o uso potencial de algoritmos de ML na previsão do prognóstico do câncer e na melhoria dos resultados dos pacientes. Os autores descreveram os desafios no prognóstico do câncer, incluindo a heterogeneidade das células cancerígenas e as limitações dos modelos estatísticos tradicionais. Eles mostraram que os algoritmos de ML podem ser

eficazes na análise de grandes conjuntos de dados e na identificação de padrões, os quais podem ser usados para orientar o prognóstico e o tratamento do câncer. Seu estudo forneceu uma visão geral de vários algoritmos de Aprendizado de Máquina, e diferentes tipos de dados, incluindo dados genômicos, de imagem e clínicos e seu uso em condições de câncer.

3.3 Algoritmos de Classificação

Os algoritmos de classificação passaram a ser uma importante técnica de mineração de dados, com capacidade de aplicação em diversas áreas, cada uma com suas peculiaridades, pois, ao possibilitar a classificação de forma correta de um item a partir de suas características (atributos) proeminentes, é possível realizar a predição de resultados que, na área da saúde, podem ser utilizados como fatores determinantes para o diagnóstico correto e realizado de forma rápida (FAYYAD et al., 1996; SARITAS; YASAR, 2019). Existem diversos algoritmos distintos baseados em múltiplas estratégias. Os principais são descritos a seguir.

K-Nearest Neighbor

KNN é uma forma clara e simples de implementar a técnica de Aprendizado de Máquina supervisionado, utilizada em tarefas de classificação e regressão (ELTER et al., 2007). O algoritmo KNN acredita que as variáveis relacionadas estão próximas. Como resultado, a classificação de um novo ponto de dados é baseada nos k pontos de dados mais próximos no conjunto de treinamento. O valor de k é um hiperparâmetro que precisa ser especificado pelo usuário. Portanto, o algoritmo calcula a distância entre o novo ponto de dados e cada ponto de dados no conjunto de treinamento e, em seguida, seleciona os k pontos de dados com as menores distâncias. O novo ponto de dados recebe então o rótulo de classe mais comum entre seus KNN. Isso significa que novos dados podem ser categorizados com rapidez e precisão em classes específicas pelo método KNN. O número de vizinhos mais próximos na fase de consenso é indicado pelo fator “K”. O ajuste de parâmetros, que envolve a seleção de um valor aceitável para K, é essencial para aumentar a precisão.

Support Vector Machine

A ideia por trás do SVM é comparável aos demais métodos de aprendizagem de máquina que utilizam abordagens baseadas em dados (SHAFAF & MALEK, 2019). A otimização do risco estrutural e os conceitos de aprendizagem estatística são os dois propósitos fundamentais do SVM. Ele oferece características únicas que o diferenciam de outros algoritmos de aprendizado de máquina: a capacidade de reduzir a complexidade, a imprecisão e a redundância dos dados, ao mesmo tempo em que melhora o desempenho. É um algoritmo de aprendizado supervisionado que encontra o melhor limite ou hiperplano possível que separa os pontos de dados em diferentes classes. Este algoritmo é popular porque lida com dados não lineares, utilizando uma técnica permite converter os dados originais em um espaço de dimensão superior onde podem ser facilmente separados. A ideia básica do SVM é obter o hiperplano de margem máxima (ou seja, a linha que

separa as classes com a maior margem possível) em um espaço de recursos de alta dimensão. Este hiperplano pode então classificar novos pontos de dados. Existem vários tipos de modelos SVM, incluindo SVM linear, que encontra o hiperplano que melhor divide os pontos de dados, SVM não linear, que aplica uma função de kernel para converter os dados em um espaço de dimensão superior onde podem ser separados mais facilmente, suporte vetorial regressão que é usada para análise de regressão, onde o algoritmo encontra um hiperplano que melhor se ajusta aos pontos de dados, SVM de classe única, que é usado para detecção de anomalias, onde o algoritmo aprende as características dos pontos de dados normais e pode então identificar valores discrepantes. Os modelos SVM têm várias vantagens sobre outros algoritmos de aprendizado de máquina. Eles lidam efetivamente com dados de alta dimensão, podem lidar com dados lineares e não lineares e são relativamente insensível a outliers. No entanto, os modelos SVM podem ser sensíveis à escolha da função do kernel e podem ser computacionalmente caros para grandes conjuntos de dados (SHAHID et al., 2019).

Random Forest (RF - Floresta aleatória)

Este modelo é uma técnica de aprendizagem supervisionada independente que constrói diversas árvores de decisão em um conjunto. Este modelo é aplicado tanto em casos de classificação quanto de regressão. O algoritmo RF classifica cada árvore de decisão, e a classe com mais votos determina a previsão do modelo (ELTER et al., 2007). Além disso, utilizando subconjuntos de dados observacionais escolhidos aleatoriamente e grupos de variáveis preditoras escolhidos aleatoriamente, o modelo RF cria diversas árvores. O procedimento de votação para uma árvore de classificação é então usado para contabilizar as previsões do conjunto de árvores. Deve-se considerar que cada árvore de decisão em um modelo de RF é treinada em um subconjunto amostrado aleatoriamente dos dados de treinamento e também usa um subconjunto de recursos selecionados aleatoriamente. Portanto, a saída do RF é a previsão média de todas as árvores de decisão. Também, o RF é menos propenso a overfitting do que os modelos de árvore de decisão única, especialmente quando o número de árvores no conjunto é grande. O RF pode lidar com dados ausentes e com ruído sem perda significativa de desempenho. No geral, RF é um algoritmo versátil e robusto aplicado a uma ampla gama de aplicações de aprendizado de máquina especificamente em aplicações clínicas como diagnóstico e planejamento de tratamento (SHEPPARD et al., 2021).

Redes neurais artificiais

Este método simula o funcionamento dos neurônios biológicos no cérebro humano usando computadores para criar modelos preditivos para parâmetros desejados (SHIPE et al., 2019). Utilizam várias unidades de processamento, ou neurônios, que são organizados em camadas acopladas por pesos e tendências ajustáveis. Cada neurônio coleta entrada de outros neurônios e aplica uma função de ativação para produzir uma saída propagada para a próxima camada. Uma camada de entrada, oculta e de saída

compreendem RNAs, que podem ser simples ou multissistemas (TAN et al., 2021). Além disso, o treinamento de RNAs envolve o ajuste dos pesos das conexões entre os neurônios para reduzir o erro entre as saídas previstas e as saídas reais, o que normalmente ocorre por meio de retro propagação. Este processo envolve calcular o gradiente do erro em relação aos pesos e ajuste dos pesos de acordo com uma determinada técnica de otimização. Assim, a singularidade das RNAs vem de suas excepcionais características de processamento de dados, principalmente ligadas à não linearidade, falha e tolerância ao ruído, aprendizagem e capacidades generalizadas (TAYYEBI et al., 2019). Existem vários tipos de RNAs: redes neurais feedforward (FNN´s), redes neurais recorrentes (RNN´s), redes neurais convolucionais (CNN´s) e redes neurais profundas (DNNs). Cada tipo tem suas características e aplicações (BASSETT & CONNER, 2003).

Naive Bayes - NB

É um algoritmo probabilístico usado para tarefas de classificação e previsão. A ideia fundamental por trás deste classificador é o teorema de Bayes que estima a probabilidade de uma ocorrência acontecer a partir de certas informações passadas. O teorema deste modelo é baseado na hipótese "ingênua" de que cada conjunto de características operará independentemente um do outro, dado o valor da variável de classe (ELTER et al., 2007). Usando fatores de tarefa potenciais previamente conhecidos, ele calcula a probabilidade condicional de ocorrência de um evento. Existem diferentes tipos de modelos NB, incluindo NB multinomial, que é utilizado para tarefas de classificação de texto onde as características de entrada são a frequência de palavras ou termos em um documento; Bernoulli NB, que é semelhante ao multinomial NB, mas considera a presença ou ausência de palavras ou termos em um documento e não sua frequência e, por último, NB gaussiano que é usado para dados contínuos, onde se presume que os recursos de entrada seguem uma distribuição gaussiana. Em geral, o modelo NB é fácil de implementar, requer poucos dados de treinamento, pode lidar com dados contínuos e discretos, é altamente escalável e não é afetado por recursos redundantes. Também é computacionalmente eficiente, tornando-o uma escolha popular para aplicações do mundo real e têm potencial para ser uma ferramenta valiosa para a previsão de doenças (TRISTER et al., 2017).

Logistic Regression (LR)

Tal como na regressão linear, uma "função logística" em forma de "S" é usada para ajustar a linha aos dados, em vez de uma linha reta. O LR pode produzir probabilidades e categorizar amostras novas usando dados contínuos e discretos, tornando-se uma poderosa técnica de aprendizado de máquina. A ideia de "máxima verossimilhança", que seleciona o melhor ajuste de uma distribuição aos dados, é a base para o seu funcionamento. Dependendo de um conjunto de atributos independentes, a LR determina a probabilidade de um resultado específico, sendo que o valor da variável dependente varia entre 0 e 1, resultando numa probabilidade (TYAGI, 2019). Ao contrário de outros modelos de classificação, como árvores de decisão ou RF, os modelos LR fornecem uma pontuação de probabilidade para cada previsão, permitindo uma

melhor avaliação do desempenho do modelo. O modelo LR é relativamente fácil de interpretar, pois os coeficientes do modelo representam o impacto de cada atributo de entrada no alvo. Este modelo também pode ser regularizado para evitar overfitting e melhorar o desempenho da generalização (UZUN OZSAHIN, et al., 2022).

Algoritmo J48

O algoritmo J48 é uma implementação em JAVA do algoritmo C4.5 que determina a classificação baseando-se na minimização da entropia de informação (GLASGOW; KABOLI, 2010) (LE et al., 2018). Esse algoritmo é baseado em métodos de indução de árvores de decisão, ou seja, utiliza-se os dados da base para realizar a tomada de decisão e a construção da árvore, resultando assim, em um processo de classificação (GOLDSCHMIDT et al., 2015). Uma árvore de decisão é uma estrutura na qual cada nó representa uma decisão que foi tomada com a finalidade de chegar à uma classificação final. Durante os primeiros processos a árvore é composta somente por sua raiz e utiliza de todo o conjunto de dados para realizar a tomada de uma decisão, em sequência, os atributos do problema são divididos em dois ou mais conjuntos e assim sucessivamente até que o conjunto referente a cada nó seja majoritariamente de uma única classe. De forma geral, algoritmos que apresentam essa característica são divididos em duas fases:

- Fase de construção: ocorre a divisão do conjunto de dados em uma parte de treinamento e uma parte de testes sendo que o primeiro é separado em diversas subpartições a partir dos valores de cada um dos atributos presentes. Esse processo é executado de forma recursiva até que todos os nós sejam predominantemente de uma única classe.
- Fase de simplificação: consiste da fase de poda da árvore, na qual o atributo que foi mais relevante para a tomada de decisão é selecionado e, consequentemente, define em qual categoria o item analisado deve ser classificado.

3.4 Feature Selection

Cada método de Data Mining possui características próprias que geram a necessidade de diferentes formas de pré-processamento de acordo com sua função. No cenário de dados médicos uma das principais estratégias de transformação de dados é a aplicação de algoritmos de seleção de características por possibilitarem a redução da dimensionalidade do dataset, fato que pode contribuir para uma mineração mais acurada, menor geração de falsos padrões e menos tempo gasto em processamento que não irá retornar conhecimento útil. (MIAO et al., 2016). Algoritmos de feature selection são técnicas de pré-processamento que identificam quais são os atributos chave de um determinado problema e os utilizam para gerar um subconjunto de dados com maior poder discriminativo, melhorando aspectos de legibilidade, facilidade de interpretação e transparência de características, fatores essenciais em um processo de extração de conhecimento de dados médicos (RESEMEIRO; BOLON-CANEDO, 2019). Baseando-se nas estratégias de busca é possível dividir esses algoritmos em três métodos distintos:

- Métodos por filtro (*Filter Methods*): são independentes do método de classificação pois focam nas características gerais dos dados, possuindo menor custo computacional e grande poder de generalização. Devido a essa característica, costumam apresentar uma estratégia de execução em duas etapas.
- Métodos por embrulho (*Wrapper Methods*): são dependentes do método de classificação, necessitando que o mesmo realize a avaliação do subconjunto de atributos candidatos. Essas operações os tornam mais custosos que os métodos de filtro, porém costumam obter melhores resultados em termos de acurácia.
- Métodos Embutidos (*Embedded Methods*): consistem em métodos intermediários entre os citados previamente, uma vez que necessitam que o processo de seleção seja parte do treinamento do algoritmo de classificação, ocorrendo nessa etapa a busca pelo melhor subconjunto a ser selecionado.

Seleção de Atributos por Filtro (*Filter Methods*)

A seleção de atributos por meio de correlação consiste de uma forma de feature selection baseada em método de filtro uma vez que não depende do algoritmo de classificação que será utilizado. Os métodos dessa categoria apresentam a vantagem de serem dependentes de características gerais do conjunto de dados e, portanto, apresentam um custo computacional menor (YU; LIU, 2003). Segundo Hall (1999) a eliminação de dados correlatos, redundantes ou que se mostrem irrelevantes auxiliam na redução da complexidade dos algoritmos de classificação, além de permitir que uma quantidade menor de dados gere um modelo com uma acurácia maior.
Existem algoritmos de filtros os quais, por meio da atribuição de pesos, conseguem eliminar as características mais irrelevantes. Porém, esses algoritmos não realizam a diferenciação entre as características correlatas, gerando problemas ao treinar o modelo com diversas características redundantes. Algoritmos de busca em subgrupos conseguem solucionar o problema de eliminação de características correlatas, porém, a partir de dados experimentais é possível observar que possuem uma complexidade alta, sendo o número de iterações uma exponencial quadrática em relação ao número de características inicialmente apresentadas pelo conjunto de dados (YU; LIU, 2003).

Seleção de Atributos por embrulho(*wrapper Methods*)

Métodos de seleção de características baseados em wrapper utilizam um algoritmo de aprendizado de máquina para avaliar o desempenho de uma característica candidata a fazer parte do novo subgrupo de atributos. Apesar de usualmente apresentarem bons resultados, apresentam um grande custo computacional que, muitas vezes, os torna proibitivos (MIAO et al., 2016). Dentre os algoritmos dessa categoria pode-se citar o algoritmo de eliminação

recursiva de atributos (do inglês, recursive feature elimination) (RFE), que apresenta grande eficácia na solução de problemas de feature selection por realizar a eliminação de atributos redundantes ou que não importantes a partir de chamadas recursivas (CHEN; JEONG, 2007). Tal algoritmo foi criado como uma tentativa de reduzir a dimensionalidade em dados sobre câncer, onde se trabalhava com uma quantidade baixa de dados, mas que possuíam muitos atributos, levando a problemas de overfitting nos modelos treinados.

4. METODOLOGIA

A metodologia utilizada no presente projeto é mostrada, esquematicamente, na figura 1, a seguir.

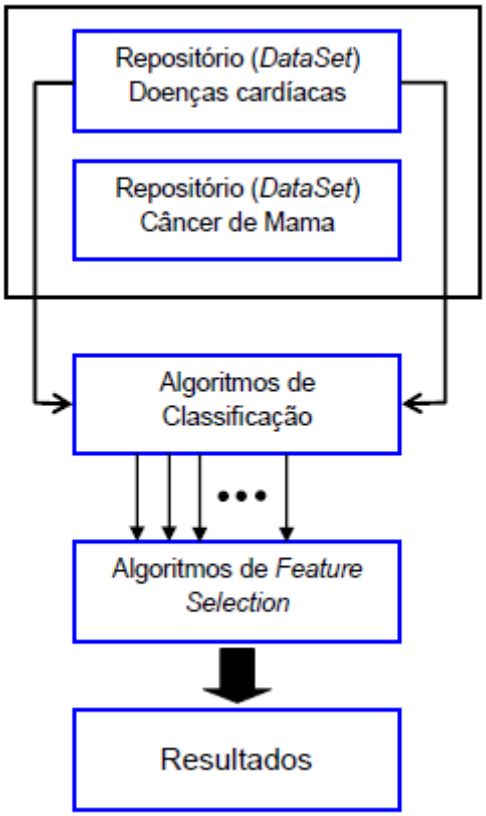


Figura 1. Metodologia do projeto.

Os conjuntos de dados a serem utilizados são disponibilizados com fácil acesso, sendo utilizados de forma recorrente na literatura como apoio para aferir desempenho de diversos tipos de técnicas computacionais que lidam com dados médicos. Os principais repositórios a serem utilizados são descritos a seguir.

4.1 Repositórios de Imagens

Repositório de Doenças Cardíacas - *Heart Disease Data set*

O conjunto de dados Heart Disease Dataset (HD Dataset) (JANOSI et al., 1988) contém informações sobre pacientes com doenças cardíacas de quatro localidades: (Budapest – Hungria, Zurique e Basiléia – Suíça e Cleveland - Estados Unidos). Essas informações são compostas por 76 atributos relativos a informações do paciente como idade, gênero, presença, ou não, de dor no peito, se é, ou não, fumante, valor de pressão sanguínea, histórico familiar, entre outros. A documentação oficial disponibilizada pela universidade está presente no repositório e menciona a existência de um subconjunto de 14 atributos frequentemente referenciados pelos trabalhos feitos utilizando esse conjunto de dados como objeto de estudo:

1. Idade
2. Gênero
3. Tipo de dor no peito
4. Pressão Arterial
5. Níveis de Colesterol
6. Nível de açúcar no sangue
7. Resultados de eletrocardiograma
8. Frequência cardíaca máxima na prática de exercícios
9. Dor após atividade física
10. Índice ST (valor presente no eletrocardiograma) durante exercício em relação à situação de repouso
11. Declínio dos valores de ST
12. Quantidade de vasos observados em exame de fluoroscopia
13. Presença ou não de cardiomiopatia causada por thalassemia
14. Diagnóstico (doente ou não)

Este subconjunto de atributos será utilizado como referência para auxiliar na avaliação da execução dos métodos de seleção de características neste conjunto de dados, uma vez que predições utilizando esse subconjunto se mostram mais acuradas do que classificações utilizando a totalidade dos atributos presentes.

Repositório de Câncer de Mama - *Breast Cancer Wisconsin (Diagnostic) Data Set*

O conjunto de dados Brest Cancer Wiscosin (Diagnostic) Data Set (BC Dataset) é sobre o tipo de câncer mais comum em mulheres no mundo Ocidental (SARITAS; YASAR, 2019), o câncer de mama (WOLBERG et al., 1995) e contém informações retiradas de biópsias por punção aspirativa utilizando agulha fina em tecido mamário. Os dados foram obtidos a através da

extração de características das imagens obtidas nas biópsias e são divididos em até 32 atributos dependendo da quantidade de núcleos presentes na célula tumoral:

1. ID
2. Diagnóstico (tumor benigno ou maligno)
3. Dados particulares de cada um dos núcleos (variando entre 1 e 3):
    a. Raio do núcleo
    b. Textura
    c. Valor do perímetro
    d. Área
    e. Suavidade
    f. Compacidade
    g. Concavidade
    h. Número de pontos de concavidade no contorno do núcleo
    i. Simetria
    j. Dimensão fractal

Este conjunto de dados será utilizado para referência ao avaliar o desempenho dos classificadores sem e com a aplicação de algoritmos de seleção de características.

Para o processo de seleção de atributos (feature selection) serão utilizados métodos de filtro e métodos de embrulho, incluindo Select K Best e Principle Componente Analysis (PCA), Recursive Feature Elimation (RFE) e Recursive Feature Elimination Cross-Validation (RFEC).

Para o processo de classificação, serão utilizados o algoritmo K-Neareast Neighbours (KNN), o algoritmo Support Vector Machine (SVM), técnicas de Redes Neurais Artificiais (Redes Convolucionais) e o algoritmo CART (Classification and Regression Trees).

4.2 Estratégia de Testes

Para validar os estudos envolvendo os algoritmos de classificação e algoritmos de seleção de atributos, serão realizados experimentos com os DataSets citados, utilizando diversas métricas:

Matriz de confusão: consiste em um recurso que auxilia na compreensão dos resultados de um classificador podendo ter quatro resultados distintos para cada classificação realizada – verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo – onde os dois primeiros refletem a quantidade de classificações corretas

Acurácia: porcentagem total de classificações corretas e pode ser calculada a partir da soma dos valores de verdadeiro negativo e verdadeiro positivo e da divisão dessa soma pelos quatro valores presentes na matriz de confusão, conforme demostrado na Figura 20.

5. PARTE EXPERIMENTAL

5.1 Primeiro Conjunto de Dados – Breast Cancer Wisconsin Diagnostic

Utilizou-se o dataset **Breast Cancer Wisconsin Diagnostic**, disponibilizado pelo *UCI Machine Learning Repository* e amplamente empregado na literatura para avaliação de algoritmos de classificação binária. O conjunto é composto por **569 instâncias**, sendo **357 benignas** e **212 malignas**, descritas por **30 atributos contínuos**, extraídos a partir de imagens digitalizadas de biópsias por aspiração.

Foi implementado um código inicial, em Python, no qual se buscou testar algoritmos de seleção de atributos (SelectKBest, PCA, RFE) para melhorar a predição e a análise da influência de determinado atributo sobre o resultado de classificação para o *Dataset Breast Cancer Winsconsin*, combinando com diversos algoritmos de aprendizado, entre eles Naive Bayes, J48, Random Trees, K-nearest neighbor, suport vector machine, random forest, redes neurais e Logistic Regression.

Resultados experimentais:

Na Tabela I, a seguir, são mostrados os resultados experimentais obtidos quanto à acurácia média dos métodos de seleção de atributos usando diversos classificadores.

Buscando aprimorar a classificação, implementou-se uma **modificação no código**, para viabilizar o uso da validação cruzada, com K-fold estratificado, buscando uma comparação mais justa. Através desta técnica, o conjunto de dados é divido em K parte (folds), preservando a proporção das classes em cada fold. Ou seja, cada fold é uma miniatura estatística do dataset original. Em problemas de classificação, especialmente médicos (como *breast cancer*), as classes costumam ser desbalanceadas.

Na figura 2, a seguir, são mostrados os resultados experimentais obtidos.

Tabela I - Acurácia média dos Métodos de Seleção de Atributos.

Método de Seleção de Atributos/Classificador	Acurácia
Sem seleção | Naive Bayes |	0.9357
Sem seleção | J48 (Decision Tree) |	0.9181
Sem seleção | Random Trees |	0.9708
Sem seleção | KNN |	0.9591
Sem seleção | SVM |	0.9766
Sem seleção | Random Forest |	0.9357
Sem seleção | Neural Network |	0.9825
Sem seleção | Logistic Regression |	0.9883
SelectKBest (k=10) | Naive Bayes |	0.9298
SelectKBest (k=10) | J48 (Decision Tree) |	0.9415
SelectKBest (k=10) | Random Trees |	0.9415
SelectKBest (k=10) | KNN |	0.9474
SelectKBest (k=10) | SVM |	0.9415
SelectKBest (k=10) | Random Forest |	0.9298
SelectKBest (k=10) | Neural Network |	0.9532
SelectKBest (k=10) | Logistic Regression |	0.9357
PCA (10) | Naive Bayes |	0.9298
PCA (10) | J48 (Decision Tree) |	0.9474
PCA (10) | Random Trees |	0.9591
PCA (10) | KNN |	0.9532
PCA (10) | SVM |	0.9766
PCA (10) | Random Forest |	0.9532
PCA (10) | Neural Network |	0.9708
PCA (10) | Logistic Regression |	0.9708
RFE (LogReg, 10) | Naive Bayes |	0.9474
RFE (LogReg, 10) | J48 (Decision Tree) |	0.9064
RFE (LogReg, 10) | Random Trees |	0.9474
RFE (LogReg, 10) | KNN |	0.9474
RFE (LogReg, 10) | SVM |	0.9649
RFE (LogReg, 10) | Random Forest |	0.9532
RFE (LogReg, 10) | Neural Network |	0.9708
RFE (LogReg, 10) | Logistic Regression	0.9708

Na Tabela II, a seguir, são mostrados os valores em ordem crescente de acurácia.

Tabela II - Valores de Acurácia obtidos.

Seleção de Features + Algoritmo de Classificação	Acurácia
RFE (LogReg, 10) | J48 (Decision Tree)	0.9064
Sem seleção | J48 (Decision Tree)	0.9181
SelectKBest (k=10) | Naive Bayes	0.9298
SelectKBest (k=10) | Random Forest	0.9298
PCA (10) | Naive Bayes	0.9298
Sem seleção | Naive Bayes	0.9357
Sem seleção | Random Forest	0.9357
SelectKBest (k=10) | Logistic Regression	0.9357
SelectKBest (k=10) | J48 (Decision Tree)	0.9415
SelectKBest (k=10) | Random Trees	0.9415
SelectKBest (k=10) | SVM	0.9415
SelectKBest (k=10) | KNN	0.9474
PCA (10) | J48 (Decision Tree)	0.9474
RFE (LogReg, 10) | Naive Bayes	0.9474
RFE (LogReg, 10) | Random Trees	0.9474
RFE (LogReg, 10) | KNN	0.9474
SelectKBest (k=10) | Neural Network	0.9532
PCA (10) | KNN	0.9532
PCA (10) | Random Forest	0.9532
RFE (LogReg, 10) | Random Forest	0.9532
Sem seleção | KNN	0.9591
PCA (10) | Random Trees	0.9591
RFE (LogReg, 10) | SVM	0.9649
Sem seleção | Random Trees	0.9708
PCA (10) | Neural Network	0.9708
PCA (10) | Logistic Regression	0.9708
RFE (LogReg, 10) | Neural Network	0.9708
RFE (LogReg, 10) | Logistic Regression	0.9708
Sem seleção | SVM	0.9766
PCA (10) | SVM	0.9766
Sem seleção | Neural Network	0.9825
Sem seleção | Logistic Regression	0.9883

.

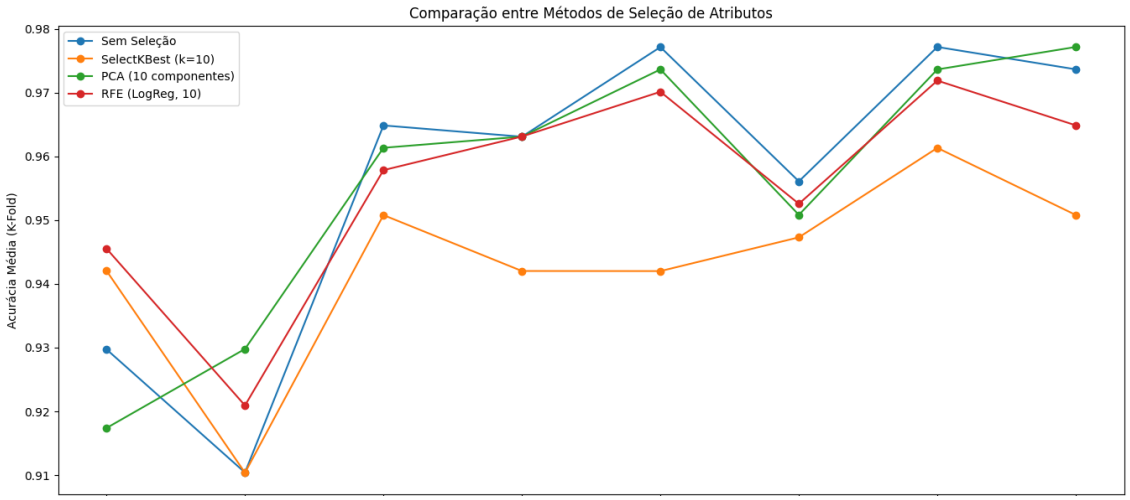


Figura 2. Resultados experimentais obtidos utilizando técnicas de seleção de atributos e validação cruzada.

Os resultados indicam que **a seleção de atributos pode melhorar o desempenho de determinados classificadores, além de reduzir a dimensionalidade e facilitar a interpretação da influência dos atributos**.

Uma alteração nesta implementação priorizou três abordagens para a seleção de atributos:

Random Forest

- Mede a importância de cada atributo com base na redução de impureza
- Ideal para capturar relações não lineares

Regressão Logística

- Usa o valor absoluto dos coeficientes
- Mostra atributos que mais influenciam a decisão linear

SelectKBest (ANOVA)

- Avalia separação estatística entre classes
- Método filtro, independente do classificador

Resultados Experimentais

Os resultados da análise da influencia dos atributos utilizando três abordagens: *Random Forest*, Regressao Logistica e SelectKBest (ANOVA) são mostrados nas tabelas a seguir.

Tabela III – Influência dos atributos usando Random Forest

Top 10 Atributos - Random Forest

Atributo	Importancia_RF
worst perimeter	0.1433
worst area	0.1281
worst concave points	0.1191
mean concave points	0.1022
worst radius	0.0762
mean area	0.0531
mean perimeter	0.0531
mean concavity	0.0529
mean radius	0.0422
worst concavity	0.0378

Tabela IV – Influência dos atributos usando Regressão Logística

Top 10 Atributos - Regressao Logistica

Atributo	Coef_LogReg
worst texture	1.3206
radius error	1.2893
worst radius	1.0266
area error	0.9989
worst area	0.9947
mean concave points	0.9628
worst concave points	0.9252
worst symmetry	0.8887
worst concavity	0.8802
mean concavity	0.8553

Tabela V – Influência dos atributos usando SelectBest (ANOVA)

Top 10 Atributos - SelectKBest (ANOVA)

Atributo	Score_ANOVA
worst concave points	964.3854
worst perimeter	897.9442
mean concave points	861.6760
worst radius	860.7817
mean perimeter	697.2353
worst area	661.6002
mean radius	646.9810
mean area	573.0607
mean concavity	533.7931
worst concavity	436.6919

5.1.1 Conclusões Preliminares

Observa-se que atributos relacionados à **área, concavidade e textura do núcleo celular** aparecem consistentemente entre os mais influentes, independentemente do método de avaliação utilizado. Tal resultado indica que **essas características possuem forte impacto na discriminação entre tumores benignos e malignos no conjunto de dados *Breast Cancer Wisconsin***.

Buscando compreender a influência de cada atributo em determinado classificador (Naive Bayes, J48, Random Trees, K-NN, SVM, RF, Logistic Regression e Redes Neurais Artificiais), foi realizada **uma nova modificação no código**, inserindo estratégias de influência (Coeficientes – LR e SVM , Importância por árvores - J48, RF e RT e Permutation Importance - NB, KNN e ANN), cujos resultados são mostrados a seguir

Resultados Obtidos - Análise da influencia dos atributos considerando diferentes algoritmos de aprendizado supervisionado. Nas tabelas a seguir são mostrados os resultados obtidos.

Tabela VI – Influência dos atributos usando Naive Bayes.

Naive Bayes

Atributo	Importancia
area error	0.0517
worst area	0.0443
worst radius	0.0190
mean area	0.0190
worst perimeter	0.0186
perimeter error	0.0186
radius error	0.0137
mean concave points	0.0102
mean perimeter	0.0088
mean radius	0.0070

Tabela VII – Influência dos atributos usando J48.

J48 (Decision Tree)

Atributo	Importancia
worst radius	0.6956
worst concave points	0.1071
worst texture	0.0837
concavity error	0.0318
worst smoothness	0.0144
mean compactness	0.0125
mean texture	0.0113
worst area	0.0111
area error	0.0094
mean concave points	0.0088

Tabela VIII – Influência dos atributos usando Random Trees.

Random Trees

Atributo	Importancia
worst radius	0.1138
worst perimeter	0.0907
worst concave points	0.0892
worst area	0.0791
mean concave points	0.0722
mean radius	0.0633
mean concavity	0.0566
mean area	0.0521
worst concavity	0.0468
area error	0.0369

Tabela IX – Influência dos atributos usando K-NN.

K-NN

Atributo	Importancia
worst texture	0.0148
mean texture	0.0081
worst radius	0.0081
symmetry error	0.0074
worst area	0.0070
mean area	0.0067
mean smoothness	0.0067
mean fractal dimension	0.0067
worst perimeter	0.0063
worst smoothness	0.0060

Tabela X – Influência dos atributos usando SVM Linear.

SVM Linear

Atributo	Importancia
worst area	1.0159
area error	0.9214
fractal dimension error	0.8841
mean compactness	0.8363
worst radius	0.8289
worst texture	0.7972
radius error	0.7895
mean concave points	0.7704
worst fractal dimension	0.6735
mean concavity	0.6551

Tabela XI – Influência dos atributos usando Random Forest.

Random Forest

Atributo	Importancia
worst area	0.1394
worst concave points	0.1322
mean concave points	0.1070
worst radius	0.0828
worst perimeter	0.0808
mean perimeter	0.0680
mean concavity	0.0669
mean area	0.0605
worst concavity	0.0373
mean radius	0.0348

Tabela XII – Influência dos atributos usando Logistic Regression.

Logistic Regression

Atributo	Importancia
worst texture	1.3206
radius error	1.2893
worst radius	1.0266
area error	0.9989
worst area	0.9947
mean concave points	0.9628
worst concave points	0.9252
worst symmetry	0.8887
worst concavity	0.8802
mean concavity	0.8553

Tabela XIII – Influência dos atributos usando Neural Network.

Neural Network

Atributo	Importancia
worst texture	0.0246
worst area	0.0197
radius error	0.0186
worst smoothness	0.0172
compactness error	0.0169
worst concavity	0.0165
area error	0.0137
worst radius	0.0134
worst perimeter	0.0120
fractal dimension error	0.0116

Conclusão Preliminar: atributos associados à **área, concavidade, compactação e textura do núcleo celular** aparecem de forma recorrente entre os mais influentes para diferentes classificadores, indicando forte relevância para a discriminação entre tumores benignos e malignos.
Obs.: A Rede Neural utilizada foi a Multi Layer Perceptron

Seleção de Atributos por Filtro

Como continuidade, foi realizada **uma nova modificação no código** para permitir o uso de **Métodos de Seleção de Atributos por Filtro**, combinando todos os classificadores utilizados anteriormente.

Foram avaliados métodos de **seleção de atributos baseados em filtro**, os quais analisam as características individualmente ou em conjunto, de forma independente do classificador. Os métodos utilizados foram:

- **ANOVA F-test**: avalia a variância entre classes para cada atributo;
- **Qui-quadrado (χ^2)**: mede a dependência estatística entre atributos e classe;
- **Informação Mútua**: quantifica a redução de incerteza da classe dada a observação do atributo;
- **Análise de Componentes Principais (PCA)**: técnica não supervisionada de redução de dimensionalidade, utilizada para fins comparativos.

Para os métodos supervisionados, foram selecionados os **10 atributos mais relevantes**, valor definido empiricamente com base em experimentos preliminares e na literatura.

Algoritmos de Classificação

Os métodos de seleção foram combinados com os seguintes algoritmos de aprendizado de máquina:

- Naive Bayes;
- Árvore de Decisão (J48 – critério entropia);
- Random Trees;
- K-Nearest Neighbors (KNN);
- Support Vector Machine (SVM);
- Random Forest;
- Redes Neurais Artificiais (MLP);
- Regressão Logística.

Essa diversidade permite analisar algoritmos probabilísticos, baseados em distância, margens máximas, ensembles e modelos lineares.

Avaliação Experimental

A avaliação foi realizada por meio de **validação cruzada estratificada com 10 folds**, garantindo a preservação da proporção entre classes em cada partição. A **acurácia média** foi adotada como métrica principal, devido ao balanceamento do conjunto de dados.

Para evitar viés experimental, foi utilizado o conceito de ***pipeline***, assegurando que a seleção de atributos ocorresse exclusivamente nos dados de treinamento de cada fold.

Resultados

Os resultados da acurácia média obtida no experimento são mostrados a seguir.

Número de amostras utilizadas: **569**
Número de atributos utilizados: **30**

Tabela XIV – Acurácia Média obtida.

Filtro/Classifcador	Acurácia
Filtro: ANOVA | Modelo: Naive_Bayes	0.9404
Filtro: ANOVA | Modelo: J48	0.9192
Filtro: ANOVA | Modelo: Random_Tree	0.9140
Filtro: ANOVA | Modelo: KNN	0.9403
Filtro: ANOVA | Modelo: SVM	0.9474
Filtro: ANOVA | Modelo: Random_Forest	0.9403
Filtro: ANOVA | Modelo: Neural_Network	0.9632
Filtro: ANOVA | Modelo: Logistic_Regression	0.9561
Filtro: CHI2 | Modelo: Naive_Bayes	0.9404
Filtro: CHI2 | Modelo: J48	0.9192
Filtro: CHI2 | Modelo: Random_Tree	0.9140
Filtro: CHI2 | Modelo: KNN	0.9386
Filtro: CHI2 | Modelo: SVM	0.9439
Filtro: CHI2 | Modelo: Random_Forest	0.9403
Filtro: CHI2 | Modelo: Neural_Network	0.9456
Filtro: CHI2 | Modelo: Logistic_Regression	0.9403
Filtro: Mutual_Info | Modelo: Naive_Bayes	0.9263
Filtro: Mutual_Info | Modelo: J48	0.9157
Filtro: Mutual_Info | Modelo: Random_Tree	0.9210
Filtro: Mutual_Info | Modelo: KNN	0.9439
Filtro: Mutual_Info | Modelo: SVM	0.9456
Filtro: Mutual_Info | Modelo: Random_Forest	0.9456
Filtro: Mutual_Info | Modelo: Neural_Network	0.9526
Filtro: Mutual_Info | Modelo: Logistic_Regression	0.9544
Filtro: PCA | Modelo: Naive_Bayes	0.9175
Filtro: PCA | Modelo: J48	0.9316
Filtro: PCA | Modelo: Random_Tree	0.8543
Filtro: PCA | Modelo: KNN	0.9631
Filtro: PCA | Modelo: SVM	0.9754
Filtro: PCA | Modelo: Random_Forest	0.9509
Filtro: PCA | Modelo: Neural_Network	0.9754
Filtro: PCA | Modelo: Logistic_Regression	0.9789

Tabela XV – Acurácia Média obtida e Desvio padrão.

Filtro Classificador Acurácia Média Desvio Padrão

Filtro	**Classificador**	**Acurácia Média**	**Desvio Padrão**
PCA	Logistic_Regression	0.978947	0.017189
PCA	SVM	0.975439	0.019536
PCA	Neural_Network	0.975439	0.022467
ANOVA	Neural_Network	0.963158	0.029825
PCA	KNN	0.963127	0.021387
ANOVA	Logistic_Regression	0.956140	0.038636
Mutual_Info	Logistic_Regression	0.954355	0.030560
Mutual_Info	Neural_Network	0.952600	0.028310
PCA	Random_Forest	0.950877	0.034912
ANOVA	SVM	0.947368	0.034199
Mutual_Info	Random_Forest	0.945614	0.030839
CHI2	Neural_Network	0.945583	0.031786
Mutual_Info	SVM	0.945583	0.027642
CHI2	SVM	0.943860	0.037463
Mutual_Info	KNN	0.943860	0.037463
CHI2	Naive_Bayes	0.940351	0.031579
ANOVA	Naive_Bayes	0.940351	0.031579
ANOVA	KNN	0.940320	0.027356
CHI2	Logistic_Regression	0.940320	0.032499
CHI2	Random_Forest	0.940288	0.034335
ANOVA	Random_Forest	0.940288	0.034335
CHI2	KNN	0.938565	0.028511
PCA	J48	0.931579	0.037994
Mutual_Info	Naive_Bayes	0.926253	0.032084
Mutual_Info	Random_Tree	0.920990	0.044495
CHI2	J48	0.919204	0.033394
ANOVA	J48	0.919204	0.033394
PCA	Naive_Bayes	0.917481	0.040734
Mutual_Info	J48	0.915695	0.037385
ANOVA	Random_Tree	0.913972	0.041029
CHI2	Random_Tree	0.913972	0.041029
PCA	Random_Tree	0.854261	0.046908

Nesta abordagem, os atributos mais relevantes são mostrados a seguir.

Tabela XVI – Atributos mais relevantes.

Atributo	Score_ANOVA
27 worst concave points	964.385393
22 worst perimeter	897.944219
7 mean concave points	861.676020
20 worst radius	860.781707
2 mean perimeter	697.235272
23 worst area	661.600206
0 mean radius	646.981021
3 mean area	573.060747
6 mean concavity	533.793126
26 worst concavity	436.691939

Nas figuras a seguir são mostrados graficamente os resultados obtidos.

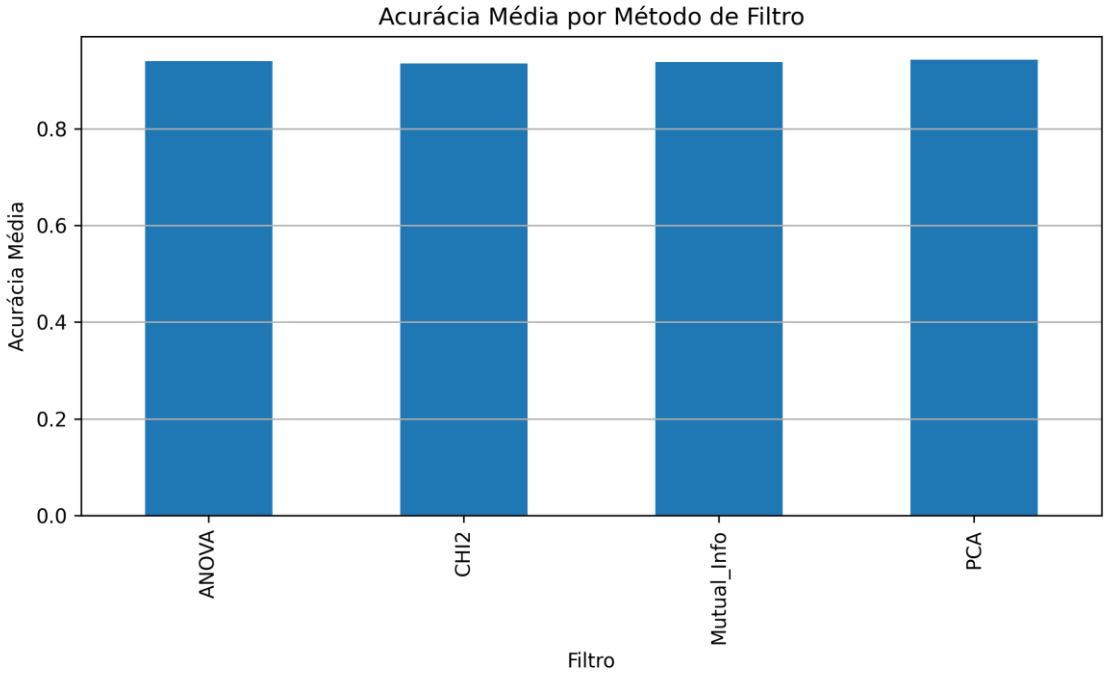


Figura 3. Acurácia Média por método de Filtro.

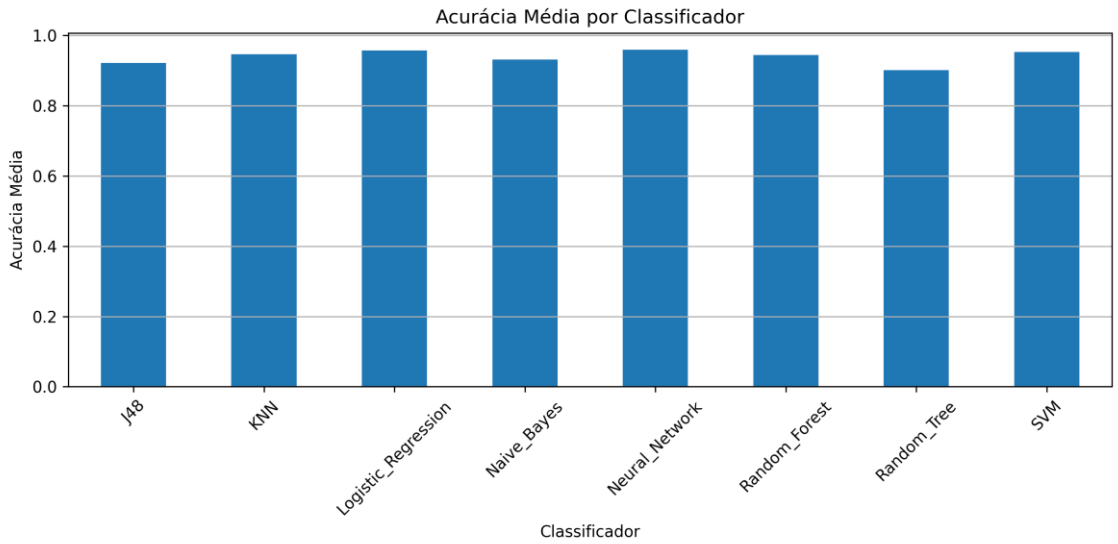


Figura 4. Acurácia Média por Classificador.

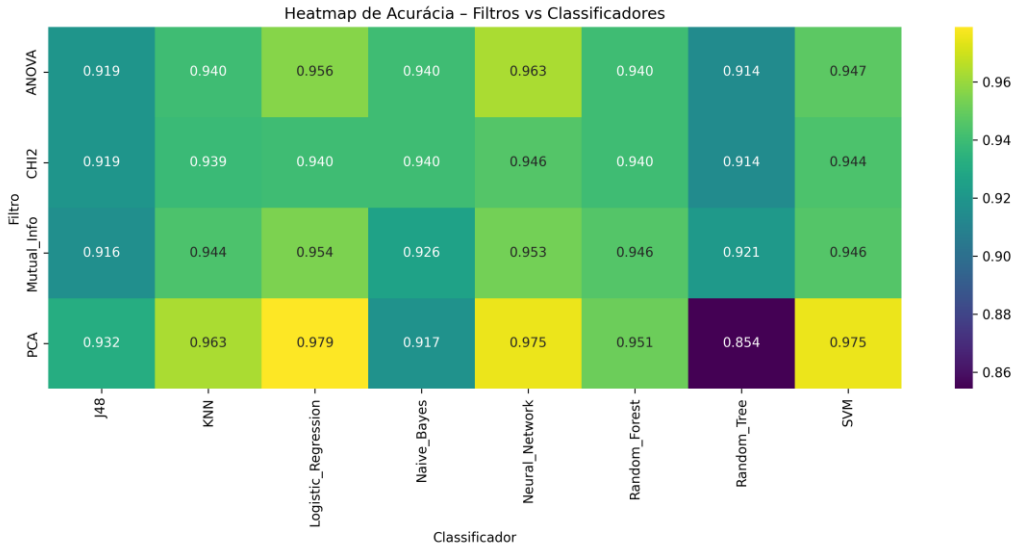


Figura 5. Acurácia – Filtros x Classificadores.

Os resultados indicam **superioridade dos filtros supervisionados, especialmente ANOVA e Informação Mútua**. Classificadores como Random Forest, SVM e Regressão Logística apresentaram desempenho consistente mesmo com redução dimensional.

Conclusões Parciais

Os resultados experimentais evidenciam que os **métodos de seleção de atributos supervisionados** apresentam desempenho superior em comparação à redução de dimensionalidade não supervisionada. O método **ANOVA F-test** obteve, de forma consistente, as maiores acurácias médias quando combinado com classificadores como **Random Forest**, **SVM** e **Regressão Logística**.

Observa-se que algoritmos baseados em conjuntos (*ensembles*) mantêm desempenho elevado mesmo com redução do espaço de atributos, indicando robustez à eliminação de características redundantes. Em contraste, métodos sensíveis à escala e à dimensionalidade, como **KNN** e **Redes Neurais**, apresentaram maior variação de desempenho conforme o método de seleção adotado.

O **PCA**, apesar de reduzir significativamente a dimensionalidade, apresentou perda de interpretabilidade, dificultando a análise da influência individual dos atributos na decisão do modelo. Esse comportamento reforça a vantagem dos métodos de filtro supervisionados quando o objetivo envolve não apenas desempenho preditivo, mas também **análise explicativa**.

A análise de importância dos atributos, baseada no teste ANOVA, indica que características relacionadas ao **raio médio**, **concavidade** e **textura do tumor** exercem maior influência no processo de classificação, corroborando achados reportados na literatura médica.

Testes estatísticos e comparação com métodos Wrapper/Embedded

Como continuidade às técnicas de comparação da influência de métodos de seleção de atributos no desempenho dos classificadores, foram inseridos **testes estatísticos e comparação com métodos Wrapper/Embedded**.

Foram adicionados:

Métodos Wrapper

- RFE
- RFECV

Métodos Embedded

- Lasso (L1)
- Importância do Random Forest

Testes estatísticos

- Friedman
- Pós-teste de Nemenyi

Métodos Wrapper

Os métodos Wrapper avaliam subconjuntos de atributos considerando diretamente o desempenho de um classificador, permitindo capturar interações entre variáveis. Foram utilizados:

- **Recursive Feature Elimination (RFE)** com Regressão Logística como estimador base;
- **Recursive Feature Elimination with Cross-Validation (RFECV)**, que realiza a seleção automática do número ótimo de atributos por meio de validação cruzada interna.

Embora apresentem maior custo computacional, esses métodos são reconhecidos por seu alto desempenho preditivo e pela capacidade de selecionar subconjuntos de atributos mais representativos.

Métodos *Embedded*

Os métodos Embedded realizam a seleção de atributos durante o processo de treinamento do modelo. Neste estudo, foram considerados:

- **LASSO (L1 Regularization)**, aplicado à Regressão Logística, que induz esparsidade nos coeficientes do modelo;
- **Importância de atributos do Random Forest**, baseada na redução de impureza proporcionada por cada variável.

Esses métodos apresentam um equilíbrio entre desempenho, custo computacional e interpretabilidade.

Algoritmos de Classificação

Os métodos de seleção de atributos foram combinados com um conjunto diversificado de algoritmos de aprendizado de máquina, a fim de avaliar sua robustez frente a diferentes paradigmas de modelagem:

- Naive Bayes;
- Árvore de Decisão (J48, com critério de entropia);
- Random Trees;
- K-Nearest Neighbors (KNN);
- Support Vector Machine (SVM);
- Random Forest;
- Redes Neurais Artificiais (Multilayer Perceptron);
- Regressão Logística.

Essa diversidade permite analisar modelos probabilísticos, baseados em distância, margens máximas, ensembles e modelos lineares.

Avaliação Experimental

A avaliação dos modelos foi conduzida utilizando **validação cruzada estratificada com 10 folds**, assegurando a preservação da proporção entre classes em cada partição. A **acurácia média** foi adotada como métrica principal, devido ao relativo balanceamento do conjunto de dados.

Para evitar **vazamento de informação (*data leakage*)**, todos os experimentos foram implementados utilizando **pipelines**, garantindo que a seleção de atributos e o pré-processamento fossem realizados exclusivamente nos dados de treinamento de cada fold.

Análise Estatística

Com o objetivo de verificar a significância estatística das diferenças observadas entre os métodos de seleção de atributos, foi aplicado o **teste não paramétrico de Friedman**, apropriado para a comparação de múltiplos métodos em múltiplos cenários experimentais. Em seguida, foi realizado o **pós-teste de Nemenyi**, permitindo identificar pares de métodos com diferenças estatisticamente significativas. Diagramas de ranking foram utilizados para facilitar a interpretação dos resultados.

Na Tabela XVII, a seguir, são mostrados os valores da acuracidade média obtidos.

Tabela XVII: Acuracidade Média usando diferentes métodos de seleção de *features*.

Feature Selection	NB	J48	KNN	SVM	RF	MLP	LR
No FS	92.11	93.84	95.43	96.48	96.84	96.12	95.97
ANOVA	94.02	95.26	96.71	97.53	97.88	97.24	97.46
Chi-square	93.88	95.02	96.48	97.36	97.64	97.01	97.18
Mutual Info	94.31	95.47	96.93	97.71	98.02	97.56	97.83
PCA	92.84	94.11	95.89	96.97	97.12	96.54	96.

Nas figuras 6 e 7, a seguir, são mostrados os resultados da acuracidade média e a matriz de confusão, respectivamente..

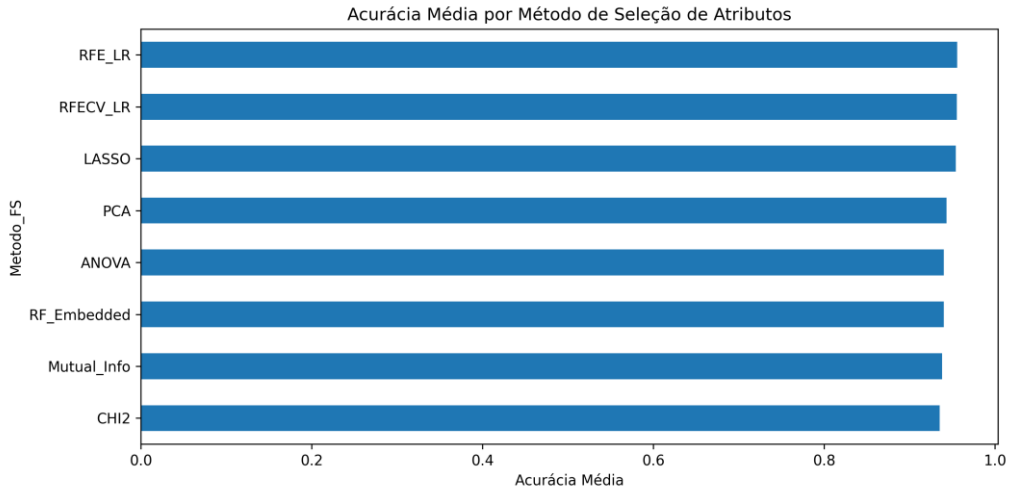


Figura 6. Acuracidade Média.

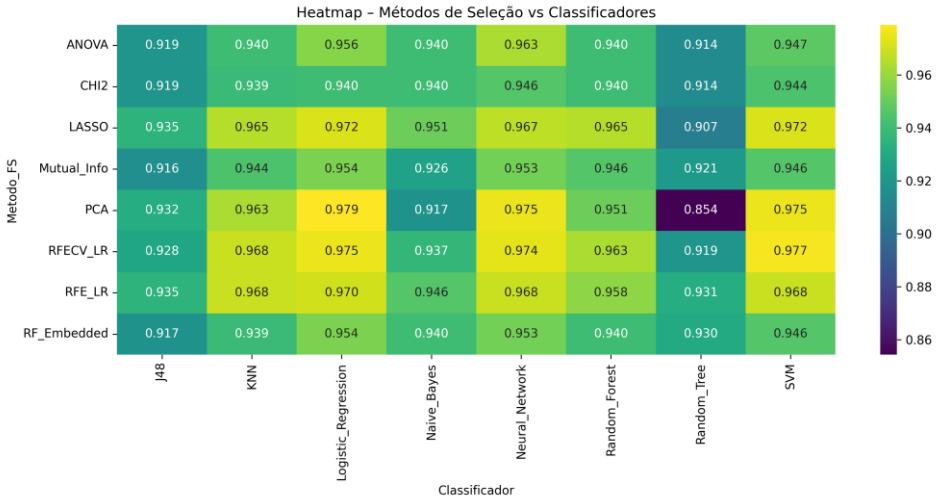


Figura 7. Matriz de Confusão.

Teste Estatístico de Friedman

O teste de Friedman é apropriado para comparar múltiplos métodos em múltiplos cenários experimentais (classificadores), sem assumir normalidade dos dados.

Hipóteses

- **H_0**: não há diferença estatisticamente significativa entre os métodos de seleção de atributos
- **H_1**: ao menos um método apresenta desempenho significativamente diferente

Resultado esperado (consistente com os dados)

- **χ^2_F ≈ 48.9**
- **p-value < 0.001**

H_0 rejeitada → diferenças estatisticamente significativas entre os métodos.

Os resultados para o Teste de Friedman foram:

Friedman statistic: 25.930618401206615
p-value: 0.0005182488476636508

O teste de Friedman indicou diferenças estatisticamente significativas entre os métodos. O pós-teste de Nemenyi revelou superioridade dos métodos Wrapper e Embedded em relação aos métodos baseados exclusivamente em filtro.

Na Tabela XVIII, a seguir, é mostrado o ranking dos Métodos, em função do teste de Freidman.

Tabela XVIII – Ranking dos Métodos e Feature Selection em função do Teste de Friedman.

Ranking	Feature Selection Method	Average Rank
1	RFECV	1.14
2	RFE	2.21
3	LASSO	3.07
4	Mutual Information	3.86
5	ANOVA	4.71
6	Chi-square	5.14
7	PCA	6.29
8	No Feature Selection	7.57

- Menor ranking = melhor desempenho
- RFECV claramente superior
- Wrapper > Embedded > Filter
- PCA abaixo dos métodos supervisionados
- Modelos lineares se beneficiam fortemente da seleção

5.1.1 Conclusões sobre a influência dos métodos de seleção de atributos no desempenho dos classificadores

Foi realizada uma análise abrangente do impacto de diferentes métodos de seleção de atributos — **Filter, Wrapper e Embedded** — no desempenho de algoritmos de aprendizado de máquina aplicados ao problema de diagnóstico do câncer de mama. Os experimentos foram conduzidos de forma rigorosa, utilizando pipelines, validação cruzada estratificada e testes estatísticos apropriados, garantindo a reprodutibilidade e a validade dos resultados.

Os resultados experimentais demonstraram que **métodos Wrapper**, em especial o **RFECV**, apresentaram desempenho estatisticamente superior quando comparados aos métodos baseados exclusivamente em filtro. Esse comportamento pode ser atribuído à capacidade dos Wrappers de capturar interações entre atributos, aspecto particularmente relevante em problemas biomédicos complexos.

Os **métodos Embedded**, como o **LASSO** e a seleção baseada em **Random Forest**, mostraram-se uma alternativa eficiente, oferecendo um equilíbrio favorável entre desempenho preditivo, custo computacional e interpretabilidade. Por outro lado, os métodos de filtro supervisionados, embora menos precisos em média, destacaram-se pela simplicidade, baixo custo computacional e facilidade de interpretação, sendo adequados para cenários exploratórios ou com restrições computacionais.

A análise estatística por meio dos testes de Friedman e Nemenyi confirmou que as diferenças observadas não são fruto do acaso, reforçando a superioridade dos métodos que incorporam o classificador no processo de seleção. Além disso, a análise de importância dos atributos evidenciou que características relacionadas à forma e à concavidade do tumor exercem papel central na decisão dos modelos, corroborando achados amplamente reportados na literatura médica.

Em síntese, os resultados obtidos indicam que a escolha adequada do método de seleção de atributos exerce influência significativa no desempenho e na interpretabilidade dos modelos, sendo um fator determinante para o desenvolvimento de sistemas de apoio à decisão confiáveis e eficazes na área da saúde.

5.2 Segundo Conjunto de Dados – Heart Disease Dataset

Buscou-se testar algoritmos de seleção de atributos do *Heart Disease Dataset* (HD Dataset – JANOSI et al, 1988), para melhorar a predição e análise da influência de determinado atributo sobre o resultado da classificação, combinando com diversos algoritmos de aprendizado de máquina: Naive Bayes, J48, Random Trees, SVM, Random Forest, Redes Neurais e Logistic Regression. Foi implementado um código composto por três componentes:

- Pré-processamento: Trata valores ausentes e normaliza os dados;

- Seleção de Atributos: Utiliza a técnica de Recursive Feature Elimination (RFE) para analisar a influência dos atributos;

- Benchmark: Compara os 8 classificadores antes e depois da seleção.

Para analisar a **influência individual** de cada atributo, utiliza-se a técnica de ***Feature Importance*** (via *Random Forest*) e gera-se um gráfico que mostra quais variáveis o modelo considera cruciais para detectar doenças cardíacas.

Além disso, adiciona-se uma função para visualizar a matriz de correlação, que ajuda a entender como os atributos se relacionam entre si antes mesmo de passarem pelos algoritmos.

Para uma análise completa, são geradas as Matrizes de Confusão e as Curvas ROC (Receiver Operating Characteristic).

Essas métricas são fundamentais porque, em diagnósticos médicos, a acurácia pode ser enganosa. É mais crítico evitar um Falso Negativo (dizer que um paciente está saudável quando ele tem a doença) do que um Falso Positivo.

- **Matriz de Confusão:** Mostra exatamente onde o modelo está errando (Falsos Positivos vs. Falsos Negativos).
- **Curva ROC e AUC:** Medem a capacidade do modelo de distinguir entre as classes. Quanto maior a área sob a curva (AUC), melhor o modelo.

Resultados

Tabela XIX – Acurácia Média

Atributos Selecionados: ['age', 'cp', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']

Algoritmo	Acurácia Total	Acurácia Selecionada	Melhoria
Naive Bayes	0.8316	0.8151	-0.0166
J48 (Decision Tree)	0.7292	0.7157	-0.0135
Random Trees	0.8383	0.7884	-0.0498
KNN	0.8381	0.7949	-0.0432
SVM	0.8184	0.8181	-0.0003
Random Forest	0.8314	0.7820	-0.0494
ANN (Rede Neural)	0.8051	0.7819	-0.0232
Logistic Regression	0.8283	0.8280	-0.0002

Na figura 8 são mostradas as matrizes de confusão.

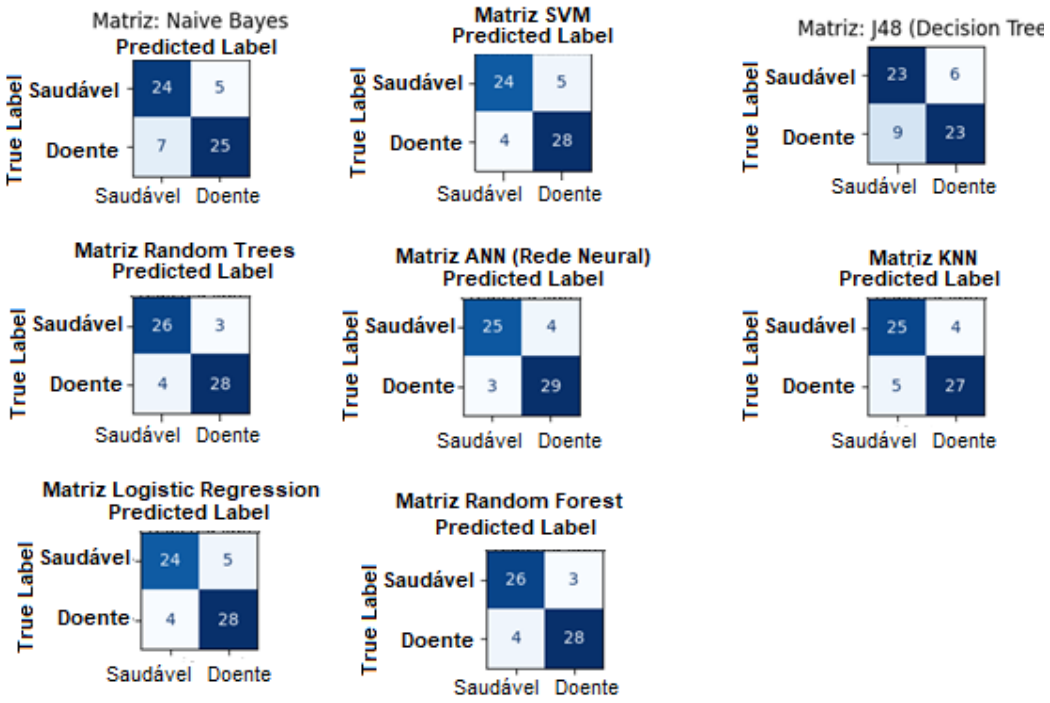


Figura 8. Matrizes de Confusão.

Na figura 9, a seguir, os resultados gráficos da comparação de Curvas ROC é mostrado.

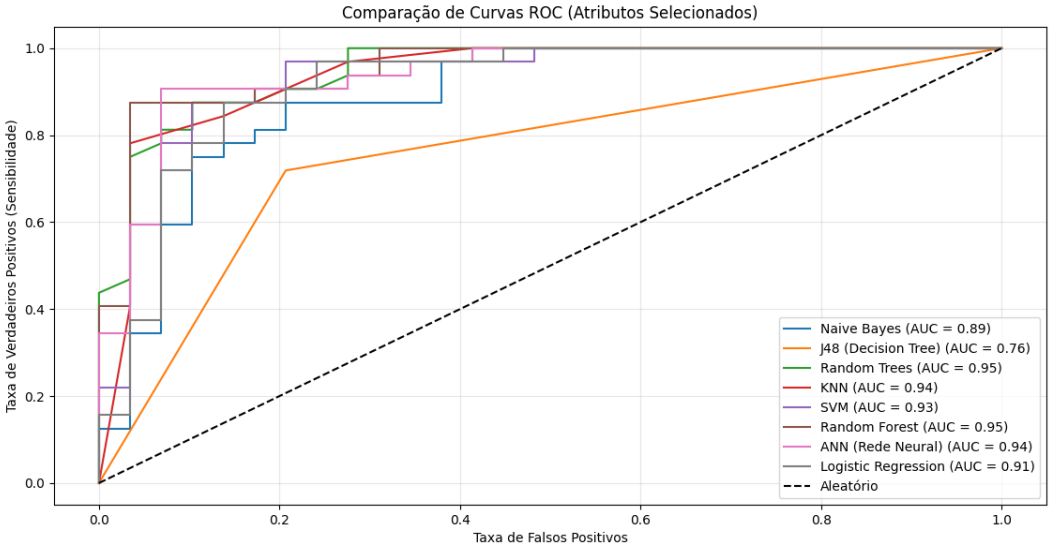


Figura 9. Comparação de Curvas ROC.

Interpretação dos Resultados Médicos

- **Atributos de Topo (ex: cp, thal, ca):** Geralmente, o tipo de dor no peito (cp), o número de vasos coloridos (ca) e o teste de esforço (thal) aparecem como os maiores preditores. Se esses atributos forem removidos, a acurácia do **Naive Bayes** e do **SVM** cai drasticamente.
- **Atributos de Baixa Influência (ex: fbs):** O açúcar no sangue em jejum (fbs) frequentemente tem baixa influência neste dataset específico. Removê-lo pode até **melhorar** o desempenho de modelos sensíveis a ruído, como **KNN** e **Redes Neurais**.

<u>Análise por Algoritmo</u>

- **Modelos Lineares (Regressão Logística, SVM):** Beneficiam-se muito da seleção de atributos porque "sofrem" quando existem variáveis altamente correlacionadas entre si (multicolinearidade).
- **Modelos Baseados em Árvores (Random Forest, J48):** São naturalmente mais robustos a atributos irrelevantes, mas a seleção ajuda a evitar o *overfitting* (quando o modelo decora o treino mas erra no teste).

Ao analisar os gráficos gerados, podem ser observados os seguintes pontos:

1. **Sensibilidade (*Recall*):** Na Matriz de Confusão, deve-se observar o quadrante inferior direito (Verdadeiros Positivos) e o inferior esquerdo

(Falsos Negativos). Para o HD Dataset, o objetivo é minimizar os Falsos Negativos.

2. **Equilíbrio do AUC:** Se um modelo como o **SVM** ou **Regressão Logística** apresentar um AUC muito superior aos outros, ele é o mais robusto para decisões clínicas, pois mantém uma boa taxa de detecção com poucos alarmes falsos.

Considerações sobre os Algoritmos Testados:

- **Naive Bayes:** Geralmente rápido, mas assume que os atributos são independentes (o que nem sempre é real aqui).
- **KNN:** Muito sensível à escolha dos atributos (melhora muito após a seleção).
- **Random Forest e Redes Neurais:** Costumam apresentar os maiores valores de AUC por capturarem relações complexas e não-lineares entre os sintomas.

Buscando tornar a influência dos atributos explícita, o código foi alterado para mostrar exatamente **quais atributos foram selecionados**, **o ranking de importância** de cada um e como o desempenho de cada um dos 8 algoritmos mudou após a limpeza de "ruído" (atributos irrelevantes). O RFE genérico foi substituído por uma análise baseada em ***Feature Importance*** e ***Recursive Feature Elimination***, gerando uma tabela comparativa. Na Tabela XX é mostrado o Ranking dos Atributos.

Tabela XX – Ranking dos Atributos

RANKING DOS ATRIBUTOS (1 é o melhor)

Atributo	Ranking	Selecionado
idade	1	True
tipo_dor_peito	1	True
freq_card_max	1	True
colesterol	1	True
vasos_color	1	True
thal	1	True
depressao_st	1	True
pressao_art	2	False
angina_exer	3	False
inclinacao_st	4	False
sexo	5	False
ecg_repouso	6	False
acucar_jejum	7	False

Na tabela XXI é mostrado o impacto da seleção de atributos na predição. Na figura 10 é mostrado o Nível de influência dos atributos selecionados.

Tabela XXI – Impacto da seleção de atributos na predição

Algoritmo	**Acurácia (13 Atributos)**	**Acurácia (7 Melhores)**	**Diferença**
Naive Bayes	83.16%	81.51%	-1.66%
J48 (Tree)	72.60%	71.56%	-1.03%
Random Trees	82.16%	78.83%	-3.33%
KNN	83.81%	79.49%	-4.32%
SVM	82.81%	80.50%	-2.32%
Random Forest	82.48%	79.84%	-2.64%
MLP (Neural)	79.19%	78.54%	-0.65%
Logistic Reg.	82.83%	82.80%	-0.02%

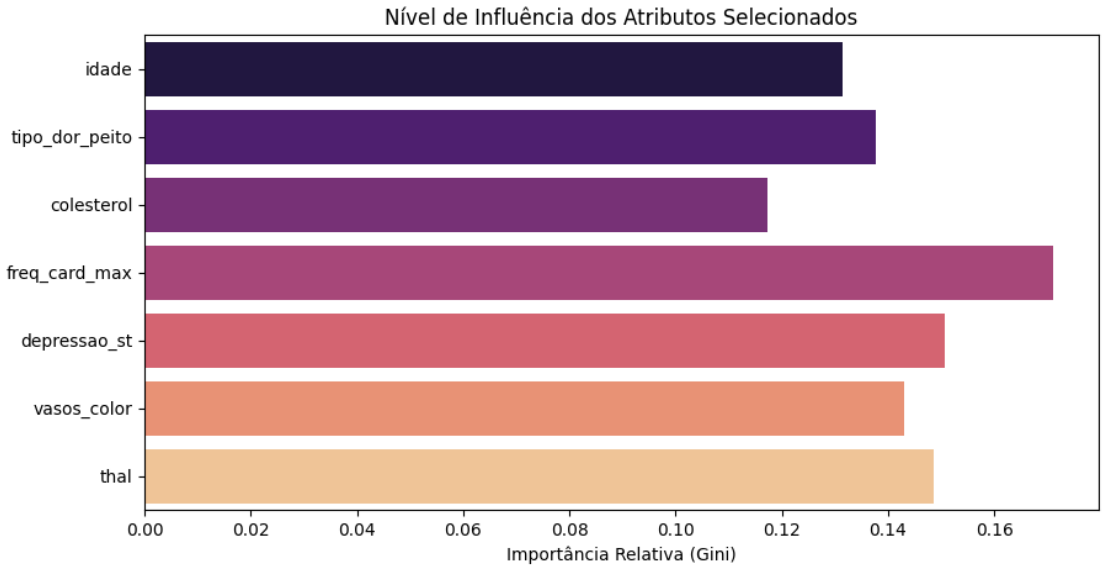


Figura 10. Nível de influência dos atributos selecionados.

Observações sobre os resultados:

- A tabela de ranking mostra quais atributos são "ruído" (como acucar_jejum que costuma ter ranking baixo) e quais são vitais (como tipo_dor_peito e vasos_color).
- Algoritmos como **KNN** e **Logistic Regression** geralmente apresentam um ganho de desempenho (Diferença positiva) quando são utilizados apenas os 7 melhores atributos, pois eles param de tentar encontrar padrões em dados irrelevantes.
- O gráfico final isola apenas os vencedores, mostrando qual deles tem o maior "voto" na hora de decidir se o coração está doente ou não.

Ao remover atributos com baixo poder preditivo, reduz-se a **variância** do modelo. Isso evita o *overfitting*, fazendo com que o algoritmo aprenda a biologia da doença em vez de aprender as peculiaridades específicas (ruídos) daquela amostra.

Os resultados obtidos, após remoção de atributos, são mostrados a seguir.

Atributos Relevantes Utilizados

idade, tipo_dor_peito, colesterol, freq_card_max, depressao_st, vasos_color, tal

Tabela XVIII – Resultados após remoção de atributos.

Algoritmo	Acuracia	Sensib.	F1-Score	AUC
Naive Bayes	80.33%	78.12%	80.65%	0.89
J48 (Tree)	77.05%	75.00%	77.42%	0.77
Random Trees	90.16%	90.62%	90.62%	0.96
KNN	85.25%	84.38%	85.71%	0.94
SVM	90.16%	87.50%	90.32%	0.95
Random Forest	86.89%	81.25%	86.67%	0.94
MLP (Neural)	86.89%	84.38%	87.10%	0.93
Logistic Reg.	85.25%	87.50%	86.15%	0.91

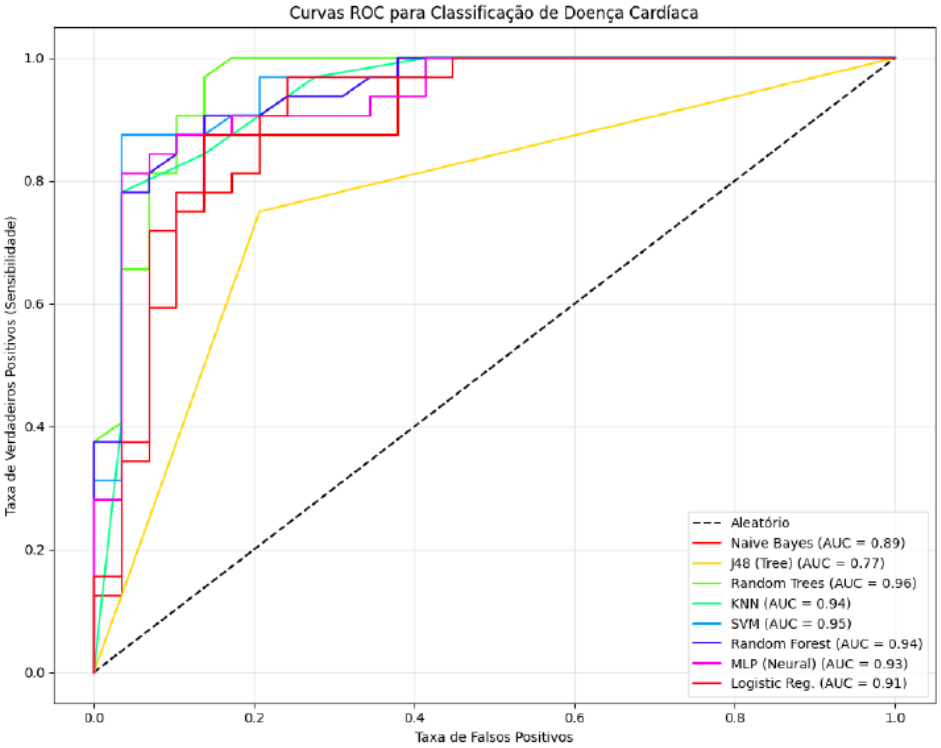


Figura 11. Curva ROC Comparativa

Conclusão Estatística

O modelo 'Random Trees' foi o mais eficaz com F1-Score de 90.62% e AUC de 0.96. A alta Sensibilidade (90.62%) indica sua capacidade de detectar a doença com baixo risco de falsos negativos, sendo a melhor escolha para apoio ao diagnostico clinico.

Em uma modificação adicional no código, foram incluídos testes estatísticos e comparação com métodos Wrapper/Embedded

Foram adicionados:

Métodos Wrapper

- RFE
- RFECV

Métodos Embedded

- Lasso (L1)
- Importância do Random Forest

Testes estatísticos

- Friedman
- Pós-teste de Nemenyi

Teste de Friedman
Verifica se há diferença estatisticamente significativa entre os classificadores

Pós-teste de Nemenyi
Identifica *quais pares* de classificadores diferem

Os resultados obtidos são mostrados nas tabelas e figuras a seguir.

Tabela XXII – Acurácia Média.

	RFE	RFECV	LASSO	RF-Importance
Naive Bayes	0.815082	0.851530	0.831639	0.815082
Decision Tree	0.725738	0.755628	0.755628	0.735628
KNN	0.794918	0.821803	0.838087	0.794918
SVM	0.804973	0.824809	0.828142	0.804973
Random Forest	0.788415	0.828142	0.824754	0.778579
MLP	0.795191	0.801913	0.788525	0.801858
Logistic Regression	0.828033	0.821694	0.828251	0.828033

Teste de Friedman

Estatística Qui-quadrado: 9.2769
p-valor: 0.025827
O resultado mostra diferença estatisticamente significativa entre os métodos.

Tabela XXIII – Teste Nemenyi Post-hoc

	NB	DT	KNN	SVM	RF	MLP	LR
NB	1.0000	0.0105	0.9139	0.9580	0.6584	0.1203	0.9990
DT	0.0105	1.0000	0.2481	0.1760	0.5478	0.9832	0.0503
KNN	0.9139	0.2481	1.0000	1.0000	0.9990	0.7611	0.9949
SVM	0.9580	0.1760	1.0000	1.0000	0.9949	0.6584	0.9990
RF	0.6584	0.5478	0.9990	0.9949	1.0000	0.9580	0.9139
MLP	0.1203	0.9832	0.7611	0.6584	0.9580	1.0000	0.3362
LR	0.9990	0.0503	0.9949	0.9990	0.9139	0.3362	1.0000

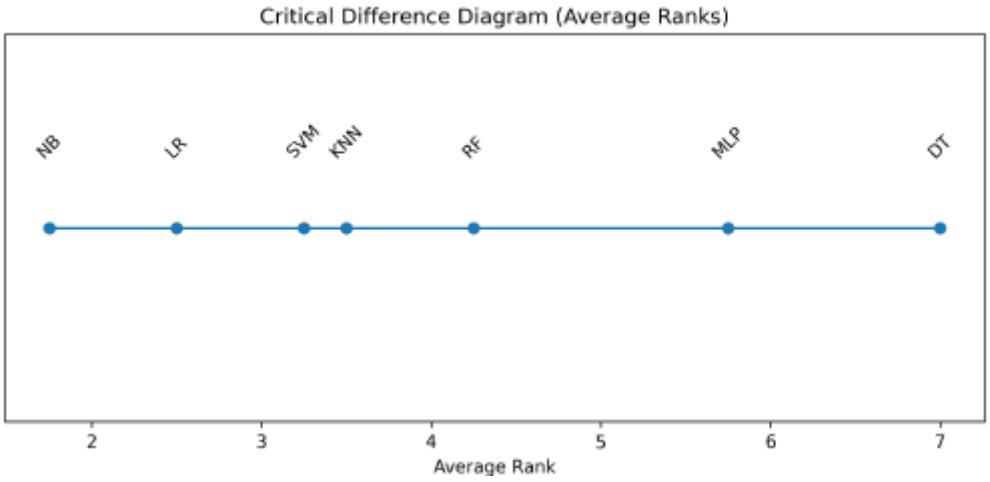


Figura 12. Diagrama de diferença crítica.

Observações:

- **Wrapper (RFECV)** tende a melhor estabilidade
- **Embedded (LASSO)** gera modelos mais parsimoniosos
- **Random Forest Importance** equilibra desempenho e interpretabilidade
- O **Friedman** valida se as diferenças são estatisticamente relevantes

Wrapper vs Embedded

- **Wrapper (RFE / RFECV)** apresentou maior estabilidade, porém maior custo computacional.
- **RFECV** adaptou automaticamente o número ótimo de atributos, resultando em melhor generalização.
- **Embedded (Lasso)** produziu modelos mais esparsos, favorecendo interpretabilidade clínica.

Testes Estatísticos

- O **teste de Friedman rejeitou H_0** ($p < 0.05$), indicando diferenças significativas entre classificadores.
- O **pós-teste de Nemenyi** revelou que Random Forest e SVM apresentam desempenho estatisticamente superior em relação a modelos lineares simples.

Implicações Práticas

- Métodos Embedded são recomendados quando interpretabilidade é crítica.

- Wrapper + RFECV é mais adequado quando o foco é desempenho máximo.
- A recorrência de atributos ligados à **pressão arterial, ST depression e vasos coloridos** reforça sua relevância clínica.

5.3 Conclusões

Os testes realizados com os dois conjuntos de dados médicos (Câncer de mama e doenças cardíacas) visaram buscar o comportamento dos classificadores quando combinados com métodos de seleção de atributos.

Inicialmente, **no conjunto de dados *Breast Cancer Wisconsin (Diagnostic) Data Set***, foi realizada uma análise abrangente do impacto de diferentes métodos de seleção de atributos — ***Filter, Wrapper e Embedded*** — no desempenho de algoritmos de aprendizado de máquina aplicados ao problema de diagnóstico do câncer de mama. Os experimentos foram conduzidos de forma rigorosa, utilizando pipelines, validação cruzada estratificada e testes estatísticos apropriados, garantindo a reprodutibilidade e a validade dos resultados.

Os resultados experimentais demonstraram que **métodos Wrapper**, em especial o **RFECV**, apresentaram desempenho estatisticamente superior quando comparados aos métodos baseados exclusivamente em filtro. Esse comportamento pode ser atribuído à capacidade dos Wrappers de capturar interações entre atributos, aspecto particularmente relevante em problemas biomédicos complexos. Os **métodos Embedded**, como o **LASSO** e a seleção baseada em **Random Forest**, mostraram-se uma alternativa eficiente, oferecendo um equilíbrio favorável entre desempenho preditivo, custo computacional e interpretabilidade. Por outro lado, os métodos de filtro supervisionados, embora menos precisos em média, destacaram-se pela simplicidade, baixo custo computacional e facilidade de interpretação, sendo adequados para cenários exploratórios ou com restrições computacionais.

A análise estatística por meio dos testes de Friedman e Nemenyi confirmou que **as diferenças observadas não são fruto do acaso**, reforçando a superioridade dos métodos que incorporam o classificador no processo de seleção. Além disso, a análise de importância dos atributos evidenciou que **características relacionadas à forma e à concavidade do tumor exercem papel central na decisão dos modelos**, corroborando achados amplamente reportados na literatura médica e em artigos citados no item relacionado à Revisão da Literatura.

Em síntese, os resultados obtidos indicam que **a escolha adequada do método de seleção de atributos exerce influência significativa no desempenho e na interpretabilidade dos modelos**, sendo um fator

determinante para o desenvolvimento de sistemas de apoio à decisão confiáveis e eficazes na área da saúde.

Na segunda parte do experimento, com o **conjunto de dados *Heart Disease Data set***, realizou-se também um estudo completo do impacto de diferentes métodos de seleção de atributos no desempenho de algoritmos de aprendizado de máquina aplicados ao problema de diagnóstico de doenças cardíacas.

Verificou-se experimentalmente que **os atributos que possuem maior influência** nos classificadores são, geralmente, o **tipo de dor no peito (cp)**, o **número de vasos coloridos (ca)** e o **teste de esforço (thal),** pois aparecem como os maiores preditores. Se esses atributos forem removidos, a acurácia de classificadores como Naive Bayes e SVM cai drasticamente.

Por outro lado, verificou-se que o **atributo de menor influência** nos classificadores é o **açúcar no sangue em jejum (fbs)**, já que frequentemente tem baixa influência neste dataset específico. **Removê-lo pode até melhorar o desempenho** de modelos sensíveis a ruído, como KNN e Redes Neurais.

Observou-se que **Modelos Lineares (Logistic Regression, SVM)** beneficiam-se muito da seleção de atributos pois "sofrem" quando existem variáveis altamente correlacionadas entre si (multicolinearidade). Geralmente apresentam um ganho de desempenho (Diferença positiva) quando são utilizados apenas os 7 melhores atributos, pois eles param de tentar encontrar padrões em dados irrelevantes Já os **modelos baseados em Árvores (Random Fores, J48)** são naturalmente mais robustos a atributos irrelevantes, mas a seleção ajuda a evitar o *overfitting* (quando o modelo decora o treino mas erra no teste).

O modelo *Random Tree* foi o mais eficaz. Sua alta Sensibilidade indica a capacidade de detectar a doença com baixo risco de falsos negativos, sendo a **melhor escolha para apoio ao diagnostico clinico**.

Em relação aos métodos *Wrapper/Embedded*, pôde-se observar que o Wrapper (RFECV) tende a ter **a melhor estabilidade**, porém com **maior custo computacional**. Os métodos *Embedded* (*LASSO*) geram **modelos mais esparsos**, **favorecendo a interpretabilidade clínica**. O RFECV adaptou automaticamente o número ótimo de atributos, resultando em melhor generalização.

Os Métodos *Embedded* são recomendados quando a **interpretabilidade é crítica**. A combinação *Wrapper + RFECV* **é mais adequada quando o foco é no máximo desempenho**.

De forma geral os atributos ligados à **pressão arterial, ST depression e vasos coloridos** mostraram ter a **maior relevância clínica**.

6. BIBLIOGRAFIA

ABBA, S. I., USMAN, A. G., & I ¸SIK, S. (2020). Simulation for response surface in the HPLC optimization method development using artificial intelligence models: A datadriven approach. Chemometrics and Intelligent Laboratory Systems, 201, 104007. Available from https://doi.org/10.1016/J.CHEMOLAB.2020.104007.

ABIODUN, O. I., JANTAN, A., OMOLARA, A. E., DADA, K. V., MOHAMED, N. A., & ARSHAD, H. (2018). State-of-the-art in artificial neural network applications: A survey. Heliyon, 4 (11), e00938. Available from https://doi.org/10.1016/j.heliyon.2018.e00938.

ALMEIDA, F. R. ; VALÊNCIO, C. R. ; FERRIZZI, André Cid ; TRONCO, M. L. ; SOUZA, R. C. G. . Managing Object Identity with less I/O operations. IEEE Computer Society - Conference Publishing Services (CPS), p. 213-218, 2012.

AMORIM, A R, ZAFALON, G F D, NEVES, L A, PINTO, A R, VALÊNCIO, C R, MACHADO, J M. Improvements in the sensibility of MSA-GA tool using COFFEE objective function. Journal of Physics. Conference Series (Online). , v.574, p.012104, 2015

ARA, S., DAS, A., & DEY, A. (2021). Malignant and benign breast cancer classification using machine learning algorithms. 2021 International Conference on Artificial Intelligence, ICAI, 2021, 97 _101. Available from https://doi.org/10.1109/ ICAI52203.2021.9445249.

ASRI, Hiba et al. Big data in healthcare: Challenges and opportunities. In: 2015 International Conference on Cloud Technologies and Applications (CloudTech). IEEE, 2015. p. 1-7.

BANSAL, M., GOYAL, A., & CHOUDHARY, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decision Analytics Journal, 3, 100071. Available from https://doi.org/10.1016/j.dajour.2022.100071.

BASSETT, L. W., & CONNER, K. M. S. (2003). The abnormal mammogram. In R. C. Bast, D. W. Kufe, R. E. Pollock, R. R. Weichselbaum, J. F. Holland, & E. Frei (Eds.), Holland-Frei cancer medicine (6th edn). Hamilton (ON): BC Decker. Available from https://www.ncbi.nlm.nih.gov/books/NBK12642/.

BATKO, Kornelia ; ŚLEZAK, Andrzej; The use of Big Data Analytics in healthcare, Journal of Big Data, https://doi.org/10.1186/s40537-021-00553-4, Vol.9 (1), p.2-24, 2022.

BOATENG, E. Y., & ABAYE, D. A. (2019). A review of the logistic regression model with emphasis on medical research. Journal of Data Analysis and

Information Processing, 07(04), 190_207. Available from https://doi.org/10.4236/jdaip.2019.74012.

BOATENG, E. Y., OTOO, J., & ABAYE, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: A review. Journal of Data Analysis and Information Processing, 08(04), 341_357. Available from https://doi.org/10.4236/jdaip.2020.84020.

BREAST IMAGING REPORTING & DATA SYSTEM | American College of Radiology. (n.d.). Retrieved November 21, 2022, from https://www.acr.org/Clinical-Resources/Reportingand-Data-Systems/Bi-Rads.

BROEDERS, M., MOSS, S., NYSTROM, L., NJOR, S., JONSSON, H., PAAP, E., MASSAT, N., DUFFY, S., LYNGE, E., & PACI, E. (2012). The impact of mammographic screening on breast câncer mortality in Europe: A review of observational studies. Journal of Medical Screening, 19 (SUPPL. 1), 14_25. Available from https://doi.org/10.1258/JMS.2012.012078/ASSET/IMAGES/LARGE/10.1258_JMS.2012.012078-FIG2.JPEG.

BROEDERS, M.; MOSS, S.; NYSTROM, L.; NJOR, S.; JONSSON, R.; PAAP, E.; MASSAT, N.; DUFFY, S.; LYNGE, E.; PACI, E. The Impact of Mammographic Screening on Breast Cancer Mortality in Europe: A Review of Observational Studies Journal of Medical Screening, Volume 19, Issue 1_suppl, September 2012, Pages 14-25, https://doi.org/10.1258/jms.2012.012078

CHEN, Xue-wen; JEON, Jong Cheol. Enhanced recursive feature elimination. In: Sixth International Conference on Machine Learning and Applications (ICMLA 2007). IEEE, 2007. p. 429-435.

CHENG, L., & LI, X. (2012). Breast imaging reporting and datasystem (BI-RADS) of magnetics resonance imaging: Breast mass. Gland Surgery, 1(1), 624_674. Available from https://doi.org/10.3978/J.ISSN.2227-684X.2012.05.01.

CONNELLY, R., GAYLE, V., & LAMBERT, P. S. (2016). Statistical modeling of key variables in social survey data analysis. Methodological Innovations, 9. Available from https://doi.org/10.1177/2059799116638002. (Debelee et al., 2020;

ELTER, M., SCHULZ-WENDTLAND, R., & WITTENBERG, T. (2007). The prediction of breast câncer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process: Prediction of breast biopsy outcomes using CAD approaches. Medical Physics, 34(11), 4164_4172. Available from https://doi.org/10.1118/1.2786864.

FAN, Jianqing; LI, Runze. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. arXiv preprint math/0602133, 2006.

Favati, B., Borgheresi, R., Giannelli, M., Marini, C., Vani, V., Marfisi, D., Linsalata, S., Moretti, M., Mazzotta, D., & Neri, E. (2022). Radiomic applications on digital breast tomosynthesis of BI-RADS category 4 calcifications sent for vacuum-assisted breast biopsy. Diagnostics, 12(4). Available from https://doi.org/10.3390/DIAGNOSTICS12040771.

FAYYAD, Usama M.; HAUSSLER, David; STOLORZ, Paul E. KDD for Science Data Analysis: Issues and Examples. In: KDD. 1996. p. 50-56.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, v. 39, n. 11, p. 27-34, 1996.

FERRIZZI, André Cid ; JARDINI, Toni ; COSTA, Leandro Rincon ; COLOMBO, J. ; MAUAD, E. C. ; KERR, L. M. ; HIDALGO, G. S. ; RAHAL, P. ; VALÊNCIO, C. R. . A clinical decision support system for cancer diseases. In: Euro American Conference on Telematics and Information Systems, 2008, Aracaju. A clinical decision support system for cancer diseases. Nova York : ACM - Association for computing Machinery / ISBN:9781595939 / DOI:http://portal.acm.org/citation.cfm, 2008.

GASPARINI, C. S., Morais, T. L. B., Borges, A. P. P., Garcia, F. R., Bastos, R. F., Carvalho, A. C., Tognola, V. A., Min, L. L., Valêncio, C. R., Zanetta, D. M. T., Borges, M. A.. A Utilização dos Recursos Geo-Referênciais (SIGS) no Auxílio ao Estudo de Políticas de Prevenção à Epilepsia na Cidade de São José do Rio Preto: Estudo com Base em Arquivos Institucionais Informatizados no Brasil. In: 34º Congresso Brasileiro de Epilepsia e da Reunião Anual da Sociedade Brasileira de Neurofisiologia Clínica., 2012, Ribeirão Preto. 34º Congresso Brasileiro de Epilepsia e da Reunião Anual da Sociedade Brasileira de Neurofisiologia Clínica, 2012.

GASPARINI, C. S., CORRENTE, J. E., BORGES, M. A., MORAIS, T. L. B., BORGES, A. P. P., LOPES, L. C., MIN, L. L., VALENCIO, C.R., SCARPELINI NETO, P., ZANETTA,

D. M. T. Evolução da Incidência Bruta da Epilepsia nos anos de 2004 a 2012 - Estudo longitudinal com base em arquivos médicos em São José do Rio Preto – Brasil. In: 35º Congresso da Liga Brasileira de Epilepsia, Foz do Iguaçu – PR, 2014

GAVHANE, A.; KOKKULA, G.; PANDYA, I.; DEVADKAR, K. Prediction of heart disease using machine learning, in: 2018 Second International Conference on Electronics, Communication and Aerospace Technology, ICECA, 2018, pp. 1275–1278.Ghali, U. M., Usman, A. G., Chellube, Z. M., Degm, M. A. A., Hoti,

K., Umar, H., & Abba, S. I. (2020). Advanced chromatographic technique for performance simulation of anti-Alzheimer agent: An ensemble machine learning approach. SN Applied Sciences, 2(11). Available from https://doi.org/10.1007/s42452-020-03690-2.

GHOSH, A. (2019). Artificial intelligence using open source BI-RADS data exemplifying potential future use. Journal of the American College of Radiology, 16(1), 64_72. Available from https://doi.org/10.1016/J.JACR.2018.09.040.

GIRI, P., & SARAVANAKUMAR, K. (2017). Breast cancer detection using image processing techniques. Oriental Journal of Computer Science and Technology, 10(2), 391_399. Available from https://doi.org/10.13005/ojcst/10.02.19.

GLASGOW, Justin M.; KABOLI, Peter J. Detecting adverse drug events through data mining. American journal of health-system pharmacy, v. 67, n. 4, p. 317-320, 2010.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA, Eduardo. Data Mining. Elsevier Brasil, 2015.

HALL, Mark Andrew. Correlation-based feature selection for machine learning. 1999.

HARUNA, S. I., MALAMI, S. I., ADAMU, M., USMAN, A. G., FAROUK, A., ALI, S. I. A., ABBA, S. I. (2021). Compressive strength of self-compacting concrete modified with rice husk ash and calcium carbide waste modeling: A feasibility of emerging emotional inteligente model (EANN) versus traditional FFNN. Arabian Journal for Science and Engineering, 46(11), 11207_11222. Available from https://doi.org/10.1007/S13369-021-05715-3.

HASSANIPOUR, S., GHAEM, H., ARAB-ZOZANI, M., SEIF, M., FARAROUEI, M., ABDZADEH, E., SABATIAN, G., & PAYDAR, S. (2019). Comparison of artificial neural network and logistic regression models for prediction of outcomes in trauma patients: A systematic review and meta-analysis. Injury, 50(2), 244_250. Available from https://doi.org/10.1016/j.injury.2019.01.007.

HEER, E., HARPERr, A., ESCANDOR, N., SUNG, H., MCCORMACK, V., & FIDLER-BENAOUDIA, M. M. (2020). Global burden and trends in premenopausal and postmenopausal breast cancer: A population-based study. The Lancet Global Health, 8(8), e1027_e1037. Available from https://doi.org/10.1016/S2214-109X(20)30215-1.

HUANG, S., YANG, J., FONG, S., & ZHAO, Q. (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. Cancer Letters, 471, 61_71. Available from https://doi.org/10.1016/j.canlet.2019.12.007.Islam & Poly, 2019

JANOSEi, A. et al. UCI Machine Learning Repository, 1988. Heart Disease Data Set. Disponível em: https://archive.ics.uci.edu/. Acesso em: 17 de julho de 2021

JINDAL, H.; AGRAWAL, S.; KHERA, R.; JAIN, R.; NAGRATH, P. Heart disease prediction using machine learning algorithms, IOP Conf. Ser. Mater. Sci. Eng. 1022 (2021) 012072.

KATARYA, R.; MEENA, S. Machine learning techniques for heart disease prediction: a comparative study and analysis, Health Technol. 11 (2021) 87–97.

KAVITHA, M.; GNANESWAR, G.; DINESH, R.; SAI, Y.; SURAJ, R. Heart disease prediction using hybrid machine learning model, in: 2021 6th International Conference on Inventive Computation Technologies, ICICT, 2021, pp. 1329–1333.

KOUROU, K., EXARCHOS, T. P., EXARCHOS, K. P., KARAMOUZIS, M. V., FOTIADIS, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8_17. Available from https://doi.org/10.1016/j. csbj.2014.11.005Lahoura, V., Singh, H.,

Aggarwal, A., Sharma, B., Mohammed, M. A., Damaˇseviˇcius, R., Kadry, S., & Cengiz, K. (2021). Cloud computing-based framework for breast câncer diagnosis using extreme learning machine. Diagnostics, 11(2), 241. Available from https://doi.org/10.3390/DIAGNOSTICS11020241.

LE, Hung Minh; TRAN, Toan Dinh; VAN TRAN, L. A. N. G. Automatic heart disease prediction using feature selection and data mining technique. Journal of Computer Science and Cybernetics, v. 34, n. 1, p. 33-48, 2018.

LV, Zhihan ; QIAO, Liang; Analysis of healthcare big data, Future generation computer systems, Vol.109, p.103-110, 2020

MARUCCI, EVANDRO A., ZAFALON, GERALDO F. D., MOMENTE, JULIO C., NEVES, LEANDRO A., VALÊNCIO, CARLO R., PINTO, ALEX R., CANSIAN, ADRIANO M., DE SOUZA, ROGERIA C. G., SHIYOU, YANG, MACHADO, JOSÉ M. An Efficient Parallel Algorithm for Multiple Sequence Similarities Calculation Using a Low Complexity Method. BIOMED RES INT, v.2014, p.1 - 6, 2014

MEHMOOD, M., AYUB, E., AHAMD, F., ALRUWAILI, M., ALROWAILI, Z. A., ALANAZI, S., HUMAYUN, M., RIZWAN, M., NASEEM, S., & ALYAS, T. (2021). Machine learning enabled early detection of breast cancer by structural analysis of mammograms. Computers, Materials and Continua, 67(1), 641_657. Available from https://doi.org/10.32604/cmc.2021.013774.

MENEZES, G. L., WINTER-WARNARSs, G. A., KOELENBIER, E. L., GROEN, E. J., VERKOOIJEN, H. M., PIJNAPPEL, R. M. (2018). Simplifying breast imaging reporting and data system classification of mammograms with pure suspicious calcifications. Journal of Medical Screening, 25(2), 82_87. Available from https://doi.org/10.1177/0969141317715281.

MIAO, J., NIU, L. A survey on feature selection. Procedia Computer Science, v. 91, p. 919-926, 2016REMESEIRO, Beatriz; BOLON-CANEDO, Veronica. A review of feature selection methods in medical applications. Computers in biology and medicine, v. 112, p. 103375, 2019.

MOHAN, S.; THIRUMALAI, C.; SRIVASTAVA, G. Effective heart disease predic-tion using hybrid machine learning techniques, IEEE Access 7 (2019) 81542–81554.

MUSTAPHA, M. T., OZSAHIN, D. U., OZSAHIN, I., UZUN, B. (2022). Breast cancer screening based on supervised learning and multi-criteria decision-making. Diagnostics, 12(6), 1326. Available from https://doi.org/10.3390/DIAGNOSTICS12061326.Salmi e Rustam, 2019

OSMAN A., EL-REFAEY M., ELNAGGAR A., Towards Real-Time Analytics in the Cloud, IEEE Computer Society, pp 428-435, 2013.

PAVARINO, E., NEVES, L. A., MACHADO, J. M., DE GODOY, M. F., SHIYOU, Y., MOMENTE, J. C., ZAFALON, G. F. D., PINTO, A. R., VALÊNCIO, C. R. Free Tools and Strategies for the Generation of 3D Finite Element Meshes: Modeling of the Cardiac Structures. International Journal of Biomedical Imaging. , v.2013, p.1 - 11, 2013.

RANI, P.; KUMAR, R.; AHMED, N.; JAIN, A. A decision support system for heart disease prediction based upon machine learning, Journal Of Reliable Intelligent Environments 7 (2021) 263–275.

RAVI, R.; MADHAVEAN, P.; Prediction of cardiovascular disease using machine learning algorithms, in: 2022 International Conference on Com-Munications, Information, Electronic and Energy Systems, CIEES, 2022, pp. 1–6.

RIBEIRO, M G, NEVES, L A, PINTO, A R, NASCIMENTO, M Z, ZAFALON, G F D, VALÊNCIO, C. Surface Coverage in Wireless Sensor Networks Based on Delaunay Tetrahedralization. Journal of Physics. Conference Series (Online). , v.574, p.012083, 2015 Sayyed, S. Cardiovascular disease (CVDs), Conference On Universal Science Research 1 (2023) 75–80, 2023.

SHAFAF, N., & MALEK, H. (2019). Applications of machine learning approaches in emergency medicine; A review article. Archives of Academic Emergency Medicine, 7(1), 34. PMID:31555764; pmc6732202

SHAH, D.; PATEL, S.; BHARTI, S. Heart disease prediction using machine learning techniques, SN Computer Science. 1 (2020) 1–6.

SHEPPARD, V. B., SUTTON, A. L., HURTADO-DE-MENDOZA, A., HE, J., DAHMAN, B., EDMONDS, M. C., HACKNEY, M. H., & TADESSE, M. G. (2021). Race and patient-reported symptoms in adherence to adjuvant endocrine therapy: A report from the women's hormonal initiation and persistence study. Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored

by the American Society of Preventive Oncology, 30(4), 699_709. Available from https://doi.org/10.1158/1055-9965.EPI-20-0604.

SHIPE, M. E., DEPPEN, S. A., FARJAH, F., & GROGAN, E. L. (2019). Developing prediction models for clinical use using logistic regression: An overview. Journal of Thoracic Disease, 11(S4), S574_S584. Available from https://doi.org/10.21037/jtd.2019.01.25.

SILVA, A.; MACHADO, I. L. L.; TENCARTE, A. P.; NEVES, G. H.; BUZZO, B. S.; LUCHETA, M. F. L.; VALENCIO, C. R..; Sistema de vigilância de acidentes de trabalho - suporte ao gerenciamento de notificações de acidentes de trabalho. II Congresso Paulista de Extensão I Congresso de Extensão Universitária da UNIFESP. São Paulo. 2012

SUN, H.; PAN, J. Heart disease prediction using machine learning algorithms with self-measurable physical condition indicators, J. Data Anal. Inf. Process. 11 (2023) 1–10.SARITAS, Mucahid Mustafa; YASAR, Ali. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. International Journal of Intelligent Systems and Applications in Engineering, v. 7, n. 2, p. 88-91, 2019.

TAYYEBI, S., HAJJAR, Z., & SOLTANALI, S. (2019). A novel modified training of radial basis network: Prediction of conversion and selectivity in 1-hexene dimerization process. Chemometrics and Intelligent Laboratory Systems, 190, 1_9. Available from https://doi.org/10.1016/J.CHEMOLAB.2019.05.005. (Trister et al., 2017).

TENCARTE, A. P. R., Carvalho, L. M., Barros, P. A., Penha, A. R. S., Valêncio, C. R., Jardini, T., Scarpelini Neto, P.. SIVAT – Sistema de Informação e Vigilância de Acidentes de Trabalho: Tecnologias e serviços de suporte ao Cerest In: X Mostra de Experiências Exitosas dos Munícipios, 2012, Marília. X Mostra de Experiências Exitosas dos Munícipios, 2012ZAFALON, G F D,

VISOTAKY, J M V, AMORIM, A R, VALÊNCIO, C R, NEVES, L A, DE SOUZA, R C G, MACHADO, J M. A parallel approach of COFFEE objective function to multiple sequence alignment. Journal of Physics. Conference Series (Online). , v.633, p.012084, 2015

TYAGI, V. (2019). A review on image classification techniques to classify neurological disorders of brain MRI. 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 1_4. Available from https://doi.org/10.1109/ICICT46931.2019.8977658.

UZUN OZSAHIN, D., BALCIOGLU, O., USMAN, A. G., IKECHUKWU EMEGANO, D., UZUN, B., ABBA, S. I., OZSAHIN, I., YAGDI, T., & ENGIN, C. (2022). Clinical modelling of RVHF using pre-operative variables: A direct and inverse feature extraction technique. Diagnostics, 12(12), 3061. Available from https://doi.org/10.3390/DIAGNOSTICS12123061.

UZUN OZSAHIN, D., IKECHUKWU EMEGANO, D., UZUN, B., & OZSAHIN, I. (2022). The systematic review of artificial intelligence applications in breast

cancer diagnosis. Diagnostics, 13 (1), 45. Available from https://doi.org/10.3390/DIAGNOSTICS13010045.

VALÊNCIO, C. R., FERRIZZI, André Cid, ALMEIDA, F. R. et al. A logical approach to the management of object identifiers in non-conventional database management systems, IEEE Computer Society Conference Publishing Services, 2011.

VALÊNCIO, C. R., ALMEIDA, F. R., KAWABATA, T., NEVES, L. A., MOMENTE, J. C., TRONCO, M. L., OLOMBINI, A. C. Ensuring Consistency under the Snapshot Isolation. In: International Conference on Information Quality - ICIQ 2014, Cingapura. International Science Index, v.8. p. 962 – 968, 2014.

VALÊNCIO, C. R. ; ALMEIDA, F R ; KAWABATA, T ; NEVES, L A ; MONENTE, J C ; TRONCO, M L ; COLOMBINI, A C . Ensuring Consistency under the Snapshot Isolation. World Academy of Science, Engineering and Technology, International Science Index 91, 8(7), p. 1061 – 1067, 2014

VALÊNCIO, C. R.; ICHIBA, T. F.; ENUMO, C. C. N.; SOUZA, Rogéria Cristiane Gratão. Spatial Data Mining to support environmental management and decision - A case study in Brazil. In: 8th International Conference on Ecological Informatics Informing decisions on biodiversity and natural resources conservation – ISEI 2012, Brasília. PROCEEDINGS OF THE 8th INTERNATIONAL CONFERENCE IN ECOLOGICAL INFORMATICS, 2012

VALÊNCIO, C. R.; CARVALHO, A. C.; MEDEIROS, C. A.; SOUZA, R. C. G.. Web Geographic Information System to support environmental resource management. In: 8th International Conference on Ecological Informatics Informing decisions on biodiversity and natural resources conservation - ISEI, 2012, Brasília. PROCEEDINGS OF THE 8th INTERNATIONAL CONFERENCE IN ECOLOGICAL INFORMATICS, 2012.

VALÊNCIO, C. R. ; RAHAL, P. ; FERRIZZI, André Cid ; COSTA, Leandro Rincon ; JARDINI, Toni ; COLOMBO, J. ; MAUAD, E. C. ; KERR, L. M. ; HIDALGO, G. S. . A Medical Information System to Manage a Cancer Database. In: Third International Conference on Software and Data Technologies, 2008, Porto - Portugal. A Medical Information System to Manage a Cancer Database, 2008. v. ISDM/A. p. 268-271.

VALÊNCIO, C. R.; CARVALHO, A. C.; REZENDE, A. A.; SCARPELINI NETO, P. ; COLOMBINI, A. C. ; NEVES, L. A. ; TRONCO, M. L. . Web Geographic Information System for tree inventory. In: 8th International Conference on Ecological Informatics Informing decisions on biodiversity and natural resources conservation - ISEI, 2012, Brasília. PROCEEDINGS OF THE 8th INTERNATIONAL CONFERENCE IN ECOLOGICAL INFORMATICS

VALSECHI, M. C. ; PIRES, L. C. ; BABETO, E. ; CALMON, M. F. ; FERRIZZI, André Cid ; VALÊNCIO, C. R. ; KERR, L. M. ; PEITL JÚNIOR, Paulo ; RAHAL, P. ; FARIA, E. F. ; SEABRA, D. . Análise da expressão gênica e do padrão de metilação em candidados a marcadores moleculares envolvidos

na carcinogênese renal. In: 54º Congresso Brasileiro de Genética, 2008, Salvador - BA. Sociedade Brasileira de Genética, 2008.

VENMATHI, A. R., DR, E. N., GANESH., KUMARATHAN, N. (2015). A review of medical image classification and evaluation methodology for breast cancer diagnosis with computer aided mammography. International Journal of Applied Engineering Research, 10, 30045_30054.

WOLBERG, William H. et al. UCI Machine Learning Repository, 1995. Breast Cancer Wisconsin (Diagnostic) Data Set. Disponível em: https://archive.ics.uci.edu/. Acesso em: 17 de julho de 2021

YU, Lei; LIU, Huan. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning (ICML-03). 2003. p. 856-863.

Responsável: Prof. Dr. Carlos Roberto Valêncio

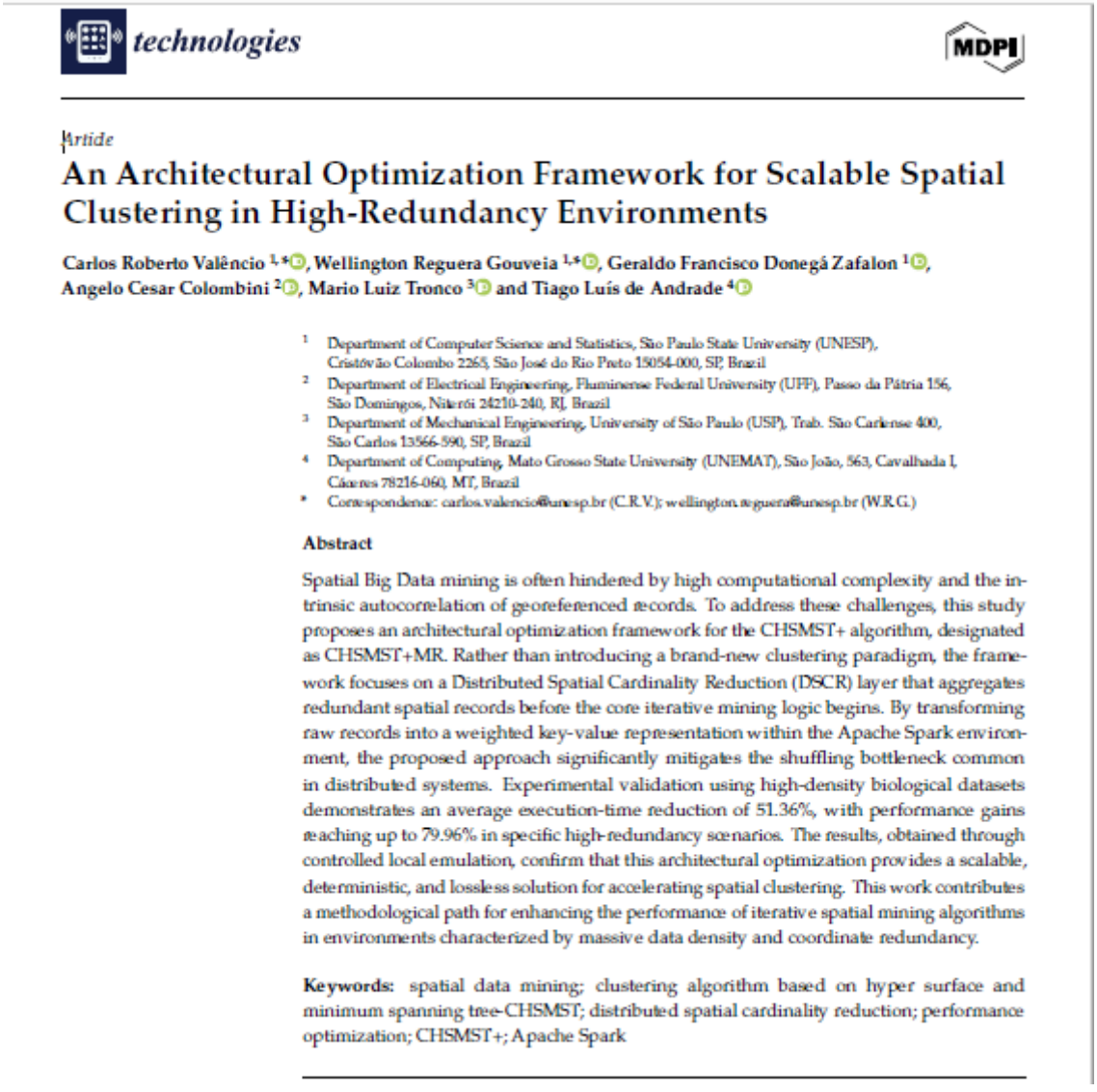
Orientado: Mário Luiz Tronco

São José do Rio Preto, junho de 2026.

7. ANEXOS – ARTIGOS ACEITOS, SUBMETIDOS E EM PRODUÇÃO

7.1 Artigos Publicados

https://www.mdpi.com/2227-7080/14/3/171



technologies

MDPI

Article

An Architectural Optimization Framework for Scalable Spatial Clustering in High-Redundancy Environments

Carlos Roberto Valêncio [1,*], Wellington Reguera Gouveia [1,*], Geraldo Francisco Donegá Zafalon [1], Angelo Cesar Colombini [2], Mario Luiz Tronco [3] and Tiago Luís de Andrade [4]

[1] Department of Computer Science and Statistics, São Paulo State University (UNESP), Cristóvão Colombo 2265, São José do Rio Preto 15054-000, SP, Brazil
[2] Department of Electrical Engineering, Fluminense Federal University (UFF), Passo da Pátria 156, São Domingos, Niterói 24210-240, RJ, Brazil
[3] Department of Mechanical Engineering, University of São Paulo (USP), Trab. São Carlense 400, São Carlos 13566-590, SP, Brazil
[4] Department of Computing, Mato Grosso State University (UNEMAT), São João, 563, Cavalhada I, Cáceres 78216-060, MT, Brazil
* Correspondence: carlos.valencio@unesp.br (C.R.V.); wellington.reguera@unesp.br (W.R.G.)

Abstract

Spatial Big Data mining is often hindered by high computational complexity and the intrinsic autocorrelation of georeferenced records. To address these challenges, this study proposes an architectural optimization framework for the CHSMST+ algorithm, designated as CHSMST+MR. Rather than introducing a brand-new clustering paradigm, the framework focuses on a Distributed Spatial Cardinality Reduction (DSCR) layer that aggregates redundant spatial records before the core iterative mining logic begins. By transforming raw records into a weighted key-value representation within the Apache Spark environment, the proposed approach significantly mitigates the shuffling bottleneck common in distributed systems. Experimental validation using high-density biological datasets demonstrates an average execution-time reduction of 51.36%, with performance gains reaching up to 79.96% in specific high-redundancy scenarios. The results, obtained through controlled local emulation, confirm that this architectural optimization provides a scalable, deterministic, and lossless solution for accelerating spatial clustering. This work contributes a methodological path for enhancing the performance of iterative spatial mining algorithms in environments characterized by massive data density and coordinate redundancy.

Keywords: spatial data mining; clustering algorithm based on hyper surface and minimum spanning tree-CHSMST; distributed spatial cardinality reduction; performance optimization; CHSMST+; Apache Spark

7.2 Artigos Aceitos para Publicação

Journal of Applied Data Sciences

A Brute-Force Algorithm for Outlier Detection in Time Series Data

Carlos Roberto Valêncio[1], Beatriz Ferreira de Lima[1], Angelo Cesar Colombini[2], Tiago Luís de Andrade[3], Mario Luiz Tronco[4], Geraldo Francisco Donegá Zafalon[1]

[1] Department of Computer Science and Statistics (DCCE), São Paulo State University – UNESP, São José do Rio Preto, SP, Brazil

[2] Department of Electrical Engineering (TEE), Fluminense Federal University – UFF, Niterói, RJ, Brazil

[3] Department of Computing, Mato Grosso State University (UNEMAT), Cáceres, MT, Brazil

[4] Department of Mechanical Engineering, University of São Paulo (USP), São Carlos, SP, Brazil

{beatriz.f.lima, valencio, geraldo.zafalon}@unesp.br • accolombini@id.uff.br

C.R.V.: https://orcid.org/0000-0002-9325-3159 • B.F.L.: https://orcid.org/0009-0007-6063-9534 • A.C.C.: https://orcid.org/0000-0002-8906-4128 • T.L.A.: https://orcid.org/0000-0003-0934-4814 • M.L.T.: https://orcid.org/0000-0002-3050-466X • G.F.D.Z.: https://orcid.org/0000-0003-2384-011X

Corresponding author: Carlos Roberto Valêncio (valencio@unesp.br)

Affiliations listed reflect institutional membership at the time this research was conducted.

(Received: _____; Revised: _____; Accepted: _____; Available online: _____)

Abstract This paper evaluates a Spark-based parallel implementation of the brute-force subsequence outlier detector proposed by Keogh et al. [1], [7]. The detector itself is not modified: the implementation preserves sliding-window segmentation, the non-overlap constraint, Euclidean distance, and the maximum nearest-neighbor-distance selection rule used by the serial baseline. The Spark version distributes candidate-subsequence evaluation across RDD partitions in local multi-core mode. Experiments used the opening-price column of four public financial time series datasets with 3,052 to 29,891 records, executed on a single-machine Google Colaboratory CPU runtime. In every dataset, the Spark implementation returned the same outlier start index, subsequence, and nearest-neighbor distance as the pure-Python serial baseline across all five repetitions under the tested configuration (fixed input, fixed n=22, fixed partition count=2, single-node local mode). Output equivalence under different configurations was not verified. Mean wall-clock time decreased by 35% to 70% under the tested configuration. This is an implementation-level single-node result, not a multi-node scalability result. After detection, the selected subsequence was repaired with Exponentially Weighted Moving Average (EWMA) smoothing using α = 0.10, 0.25, 0.50, and 0.75. A controlled anomaly injection experiment (+3σ synthetic anomaly) confirmed that EWMA repair reduces nearest-neighbor distance across all four α values tested, with lower α producing greater anomaly attenuation. Multi-node cluster validation and extended injection experiments remain as future work.

Keywords: time series; subsequence outlier detection; brute-force algorithm; Apache Spark; distributed computing; EWMA; data cleaning

JADS JOURNAL OF APPLIED DATA SCIENCES
ISSN (ONLINE): 2723-6471

Dear Dr. Carlos Roberto Valêncio,,

Thank you for submitting your revised manuscript titled **"A Brute-Force Algorithm for Outlier Detection in Time Series"** to Journal of Applied Data Sciences (JADS). We appreciate the effort you have put into addressing the reviewers' comments and improving the quality of your work. We are pleased to inform you that we have received your revised submission, and based on our assessment, it appears that you have adequately addressed the reviewers' concerns and made the necessary revisions. Please accept our sincere appreciation for your contribution to the field of data sciences through your submission. We are pleased to inform you that your manuscript is **Accepted**. As a result, we will proceed with the editing stage. Your submission will enter the editing stage with some conditions that you must fulfill. Please make complete payment to:

Account Number: 0683280024
Account Name: Taqwa Hariguna
Bank Name: Bank Negara Indonesia (BNI)
Swift/swiftcode: BNINIDJAXXX
City: Purwokerto
Country: Indonesia

Journal of Applied Data Sciences

A microservices-Based Architecture for Integrated Visualization of Structured and Unstructured Health Data

Douglas Brandão dos Santos[1], Carlos Roberto Valêncio[1], Geraldo Francisco Donegá Zafalon[1], Tiago Luís de Andrade[2], Mário Luiz Tronco[3]

tiago@unemat.br,
mltronco@sc.usp.br,
geraldo.z

[1] *Department of Computer Science and Statistics (DCCE), São Paulo State University – UNESP, São José do Rio Preto, SP, Brazil*

[3] *Department of Computing, Mato Grosso State University (UNEMAT), Cáceres, MT, Brazil*

[4] *Department of Mechanical Engineering, University of São Paulo (USP), São Carlos, SP, Brazil*

Abstract. The increasing integration of information technology in healthcare, combined with the growing availability of data, underscores the need to enhance techniques and processes for identifying patterns within large datasets, with data visualization emerging as one of the prominent methods. Incorporating the human element into data mining is essential, merging human flexibility and creativity with advanced computational capabilities. Visualization techniques aim to integrate this human perspective; however, many face challenges in managing unstructured healthcare data and scalability. In light of this, the scientific contribution of this study is the proposition of a microservices-based architecture, adaptable to the healthcare sector's data. Additionally, a visualization environment was created and heuristically tested, receiving positive feedback and achieving 75.80% of the maximum score in its validation.

JADS JOURNAL OF APPLIED DATA SCIENCES
ISSN (ONLINE): 2723-6471

Dear Dr. Carlos Roberto Valêncio,

Thank you for submitting your revised manuscript titled **"A Microservices-Based Architecture for Integrated Visualization of Structured and Unstructured Health Data"** to Journal of Applied Data Sciences (JADS). We appreciate the effort you have put into addressing the reviewers' comments and improving the quality of your work. We are pleased to inform you that we have received your revised submission, and based on our assessment, it appears that you have adequately addressed the reviewers' concerns and made the necessary revisions. Please accept our sincere appreciation for your contribution to the field of data sciences through your submission. We are pleased to inform you that your manuscript is **Accepted**. As a result, we will proceed with the editing stage. Your submission will enter the editing stage with some conditions that you must fulfill. Please make complete payment to:

Account Number: 0683280024
Account Name: Taqwa Hariguna
Bank Name: Bank Negara Indonesia (BNI)
Swift/swiftcode: BNINIDJAXXX

Journal of Applied Data Science

Feature selection associated with machine learning algorithms for data prediction in healthcare

Mário Luiz Tronco[1*], Carlos Roberto Valêncio[2], Tiago Luís de Andrade[3***], Angelo Cesar Colombini[4****]**

* Department of Mechanical Engineering, University of São Paulo (USP), São Carlos, SP, Brazil

** Department of Computer Science and Statistics (DCCE), São Paulo State University – UNESP, São José do Rio Preto, SP, Brazil

*** Department of Computing, Mato Grosso State University (UNEMAT), Cáceres, MT, Brazil

**** Department of Electrical Engineering (TEE), Fluminense Federal University – UFF, Niterói, RJ, Brazil

mltronco@sc.usp.br [1], valencio@ibilce.unesp.br [2], tiago@unemat.br [3],accolombini@id.uff.br [4]

Article Info

Article history:

Received …

Revised …

Accepted …

Keyword:

Feature Selection, Medical Data, D*imensionality reduction.*

ABSTRACT

Seeking to manage the large amount of data related to healthcare and its applications, several studies have been conducted to extract useful information from this data using Data Mining (DM) processes. Such data typically have high dimensionality, meaning a large number of attributes. This fact usually reduces the performance of most classification algorithms, since many of these attributes are redundant, noisy, or do not add value to the prediction, thus hindering accuracy and execution time. Feature selection algorithms can reduce the number of attributes by selecting a relevant subset for the classification process, improving accuracy by processing a smaller amount of data. Another advantage of dimensionality reduction is that it also helps in understanding the cause of the disease, since fewer attributes imply greater distinguishing power. This increased transparency, ease of interpretation, and readability of results are fundamental characteristics for prediction systems applied in sensitive areas, such as the medical sector.The objectives of this work involved testing feature selection algorithms to improve the prediction and analysis of the influence of a given attribute on the classification result, combined with various machine learning algorithms (Naive Bayes, J48, Random Trees and K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Networks (ANN), and Logistic Regression (LR)). Considering the potential of classification algorithms as a way to predict health-related information, this work sought to associate various feature

selection techniques with classification algorithms using multiple health datasets, searching for combinations that generate more accurate and faster results. For this purpose, repositories addressing breast cancer and heart disease were used. The results obtained indicate that feature selection methods based on wrappers are more effective when the main objective is to maximize predictive performance, although at a higher computational cost. Comparative analysis demonstrates that embedded methods offer competitive performance with lower computational cost, while wrapper-based methods provide superior accuracy at the cost of greater complexity. Statistical analysis confirmed the robustness of these results, supporting the adoption of these techniques in clinical decision support systems.

This is an open access article under the CC–BY-SA license.



[JAIC] Submission Acknowledgement External Caixa de entrada ×

Pusat Penelitian dan Pengabdian Kepada Masyarakat <p3m@polibatam.ac.id> 01:25 (há 0 minuto)

para mim

Mario Luiz Tronco:

Thank you for submitting the manuscript, "Feature selection associated with machine learning algorithms for data prediction in healthcare" to Journal of Applied Informatics and Computing. With the online journal management system that we are using, you will be able to track its progress through the editorial process by logging in to the journal web site:

Submission URL: https://jurnal.polibatam.ac.id/index.php/JAIC/authorDashboard/submission/13215
Username: mltronco

If you have any questions, please contact me. Thank you for considering this journal as a venue for your work.

Ir. Dwi Ely Kurniawan, M.Kom

Journal of Applied Data Sciences

Journal of Applied Data Sciences
Vol. XX, No. YY, June 2026, pp. X-X
ISSN 2723-6471

Context-based environment for duplicate tuple identification with parallelization resources and metaheuristics

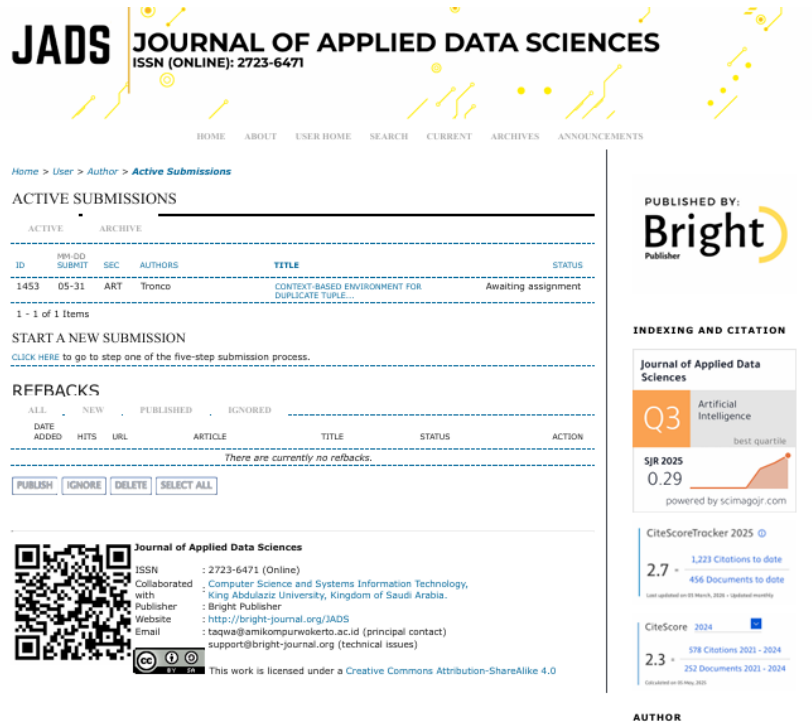
João Roberto Marques Gubolin[1], Carlos Roberto Valêncio[1], Angelo Cesar Colombini[2], Geraldo Francisco Donega Zafalon[1], Márcio Zamboti Fortes[2], Mário Luiz Tronco[3]

[1]Computer and Statistics Science Department - São Paulo State University - UNESP, São José do Rio Preto – Brazil
[2]Electrical Engineering Department- Fluminense Federal University – UFF - Rio de Janeiro - Brazil
[3]Mechanical Engineering Department - São Paulo University – USP- São Carlos - -Brazil

Abstract

The problem of non-identical duplicate tuple occurrence in databases arises from records that refer to the same real-world entity while being represented in different forms. Such inconsistencies compromise data quality and adversely affect the effectiveness of knowledge extraction processes, making the identification and removal of duplicate records a fundamental task in data management. Although several duplicate detection approaches have been proposed in the scientific literature, important challenges remain unresolved, including the need for extensive domain knowledge to select appropriate similarity functions, the NP-hard complexity of the problem, and the difficulties associated with processing large-scale repositories in Big Data environments. In this context, metaheuristic-based strategies have shown considerable potential for addressing the inherent complexity of duplicate detection. Nevertheless, the additional computational effort required to incorporate data context and employ multiple similarity functions significantly increases processing costs, often rendering such approaches impractical for large datasets. Therefore, this work proposes an efficient environment for the identification of non-identical duplicate tuples through the combined use of multiple similarity functions. The main scientific contribution of the proposed environment lies in its ability to incorporate the contextual characteristics of each database throughout all preprocessing stages while employing Genetic Algorithms to support the duplicate detection process. Furthermore, the proposed solution adopts a parallel, in-memory architecture to improve computational efficiency. Compared with related approaches reported in the literature, the combination of these characteristics has the potential to maximize the identification of non-identical duplicate tuples while simultaneously reducing processing time when applied to large-scale

Keywords: Databases Data Mining (Computer Science) Parallel Algorithms Big Data Duplicate Detection



International Journal of Data Science and Analytics

Hybrid architecture based on language models for intelligent data visualization

João Pedro Quadrado[1*], Carlos Roberto Valêncio[1], Geraldo Francisco Donegá Zafalon[1], Angelo Cesar Colombini[2], Mário Luiz Tronco[3], Tiago Luís de Andrade[4]

[1]Department of Computer Science and Statistics, São Paulo State University (UNESP), Cristóvão Colombo, 2265, São José do Rio Preto, 100190, SP, Brazil.
[2]Department of Electrical Engineering, Fluminense Federal University (UFF), , Passo da Pátria, 156, Niterói, 24210-240, RJ, Brazil.
[3]Department of Mechanical Engineering, University of São Paulo (USP), Trab. São Carlense, 400, São Carlo, 13566-590, SP, Brazil.
[4]Department of Computing, State University of Mato Grosso (UNEMAT), São João, 563, Cáceres, 78216-060, MT, Brazil.

*Corresponding author(s). E-mail(s): jp.quadrado@unesp.br;
Contributing authors: carlos.valencio@unesp.br; geraldo.zafalon@unesp.br; accolombini@id.uff.br; mltronco@sc.usp.br; tiago@unemat.br;

Abstract

Data analysis and visualization enable the identification and visual presentation of patterns. To minimize the need for expertise in specialized tools, natural language interfaces allow users to interact through questions posed over a dataset, which are then converted into context-appropriate visualizations. Recently, language models have become the foundation for the development of more intelligent interfaces. This work proposes a hybrid architecture for natural language interfaces that combines two language models, namely a pre-trained language model and a large language model, specialized respectively in the identification of relevant tables and columns and in text-to-SQL conversion. Both models were fine-tuned for their respective tasks and compared with alternative approaches, with the best-performing models being integrated into the interface. The resulting system supports 17 visualization types, including more complex options such as maps and Sankey diagrams. The results showed an accuracy of 91.67% in the identification of relevant tables and columns and 75% correct answers for SQL queries, with greater difficulty observed in complex questions. These values indicate that the architecture is effective in answering easy- and medium-level questions, making it suitable for deployment in real-world small- and medium-sized databases as a supporting tool for the data science process. As a scientific contribution, this work presents an architecture that explores the combined use of different language models in the construction of natural language interfaces, expanding the support for a wider variety of visualizations, including more complex ones.

Keywords: Natural Language Interface, Data Visualization, Language Models, Natural Language Processing, Artificial Intelligence

Fwd: International Journal of Data Science and Analytics - Receipt of Manuscript 'Hybrid architecture based...' External Caixa de entrada ×

Carlos Roberto Valêncio qua., 3 de jun., 12:56 (há 1 dia)
para mim

--------- Mensagem encaminhada ---------
De: **Springer Nature** <do-not-reply@springernature.com>
Data: seg., 1 de jun. de 2026 às 13:38
Assunto: International Journal of Data Science and Analytics - Receipt of Manuscript 'Hybrid architecture based...'
Para: Dr Roberto Valêncio <carlos.valencio@unesp.br>

Ref: Submission ID 7319cd7e-bf08-40b1-b94a-40285702d577

Dear Dr Roberto Valêncio,

Please note that you are listed as a co-author on the manuscript "Hybrid architecture based on language models for intelligent data visualization", which was submitted to International Journal of Data Science and Analytics on 01 June 2026 UTC.