Uso do KDD e Mineração de Dados para Prever Mortes por Doenças Cardíacas

Gismar P. Barbosa[1], João B. Amazonas[1], Lais F. Gregório[1], Thainnara S. Lima[1]

[1]Faculdade de Tecnologia FT – Universidade de Campinas (UNICAMP)
R. Paschoal Marmo, 1888 – Jd. Nova Itália – Limeira – SP – Brazil

gismmar_barbosa@yahoo.com.br, jamazonaz@hotmail.com,
lahgregoriiiio@gmail.com, thainnara8@gmail.com

1. Introdução

De acordo com Bourbon et al. (2016), doenças cardiovasculares (DCV) afetam o sistema circulatório, ou seja, o coração e os vasos sanguíneos (artérias, veias e vasos capilares). A maioria das DCV são causadas pelo deposito de placas de gordura e cálcio no interior das artérias, dificultando a circulação sanguínea nos órgãos, chegando a impedi-la.

As DCV's são a principal causa de morte no mundo, mais pessoas morrem anualmente por essas enfermidades do que por qualquer outra causa. Somente no ano de 2016, a OPAS/OMS (2024) estima que 17,9 milhões de pessoas morreram por doenças cardiovasculares, representando cerca de 31% de todas as mortes em nível global. A maioria das doenças cardiovasculares pode ser prevenida abordando os fatores de risco comportamentais, como uso de tabaco, dieta não saudável e obesidade, sedentarismo e uso nocivo de álcool, usando estratégias para toda a população. O controle dos fatores de risco é a melhor maneira de prevenir as DCV. Um fator de risco é uma condição que aumenta a probabilidade de sofrer uma doença cardiovascular. (Bourbon et al., 2016).

A insuficiência cardíaca é um evento comum causado por DCV's e o conjunto de dados do dataset escolhido (predição de insuficiência cardíaca) possui 12 recursos que podem ser usados para prever a mortalidade por insuficiência cardíaca. Pessoas com doenças cardiovasculares ou que apresentam alto risco cardiovascular (devido à presença de um ou mais fatores de risco) precisam de detecção e gerenciamento precoce, e um modelo de aprendizado de máquina pode ser de grande ajuda.

Tendo em vista a relevância do estudo sobre insuficiência cardíaca na sociedade atual, o grupo decidiu escolher uma base de dados referente à predição da insuficiência cardíaca, com 12 características clínicas que preveem eventos de morte. O dataset foi coletado na obra de Chicco e Jurman (2020), com o título de *Heart Failure Clinical Records (2020)* e, utilizado na sua forma integral para as análises. Esse dataset possui o acompanhamento de 299 pacientes que apresentavam algum tipo de cardiopatia.

Desta forma, o objetivo é, por meio do *Knowledge Discovery in Databases* (KDD), onde, segundo Souza (2023), utilizando as técnicas combinadas de banco de dados, estatística, aprendizado de máquina e visualização de dados, no supracitado dataset, desenvolver um modelo preditivo para identificar a ocorrência de insuficiência cardíaca em pacientes com base nos dados clínicos existentes (mortes durante período de acompanhamento).

2. Metodologia

O dados, frutos desta análise, foram retirados de maneira integra da obra *Heart Failure Clinical Records (2020)*, e, originalmente estão disponibilizados para download na url *https://archive.ics.uci.edu/static/public/519/heart+failure+clinical+records.zip*.

Para criar um ambiente controlado, o referido dataset foi baixado em nosso repositório de análise e, daqui em diante, neste documento, em todas as oportunidades que nos referirmos a ele, faremos com referência a este ambiente de análise (inclusive nos possíveis tratamentos que esse dataset possa sofrer para condução das análises).

2.1. Descrição do dataset

O dataset em questão é composto pelos seguintes metadados:

Tabela 1 - Variáveis presentes no dataset (original) utilizado para as análises

Nome da Variável	Regra	Tipo	Dado demográfico	Descrição	Unidade de Medida
age	Característica	Inteiro	Idade	idade do paciente	Anos
anaemia	Característica	Binário		diminuição de glóbulos vermelhos ou hemoglobina	
creatinine_phosphokinase	Característica	Inteiro		nível da enzima CPK no sangue	mcg/L
diabetes	Característica	Binário		se o paciente tem diabetes	
ejection_fraction	Característica	Inteiro		porcentagem de sangue saindo do coração a cada contração	%
high_blood_pressure	Característica	Binário		se o paciente tem hipertensão	
platelets	Característica	Contínuo		plaquetas no sangue	kiloplatelets/mL
serum_creatinine	Característica	Contínuo		level of serum sodium in the blood	mEq/L
sex	Característica	Binário	Sexo	mulher ou homem	
smoking	Característica	Binário		se o paciente fuma ou não	
time	Característica	Inteiro		período de acompanhamento	dias

Nome da Variável	Regra	Tipo	Dado demográfico	Descrição	Unidade de Medida
death_event	Alvo	Binário		se o paciente faleceu durante o período de acompanhamento	

Este dataset possui os registros de 299 pacientes, distribuídos entre homens e mulheres. Portanto esse é o limiar de dados que possuímos para realizar as análises.

2.2. Processo de KDD

Nesse subcapitulo, iremos descrever (em formato de lista) cada etapa realizada no processo de KDD que executamos para chegarmos aos resultados obtidos.

Em especial, por consequência dos dados obtidos do dataset já estarem rotulados, o processo de mineração de dados utilizados foi a classificação de dados que, de acordo com Marcela (2023), por meio da categorização de diferentes classes e atributos, é possível identificar padrões de comportamento e predizer ações, baseadas nestes padrões.

Eventualmente, iremos mencionar o uso de um script (SQL), que foi utilizado para tratar os dados. Esse scritpt pode ser consultado e/ou baixado na url https://github.com/gismarb/heart_attack_data_miner/blob/main/SCRIPT/script_v20240 6010943.sql . Com esse script, executamos uma boa parte dos processos de KDD.

Para realizar a mineração dos dados (e outros processos de classificação de dados), optamos pela utilização do programa Weka [WEKA], uma poderosa ferramenta de mineração de dados e aprendizado de máquina de código aberto, desenvolvida na Universidade de Waikato, na Nova Zelandia. O ambiente do Weka contém uma coleção de visualizações e ferramentas para análise e previsão de dados, com interfaces gráficas de fácil acesso, além de contar com uma comunidade robusta de usuários e desenvolvedores, bem como uma ampla documentação disponível online, o que facilita o suporte e o aprendizado da ferramenta.

E, unindo o programa Weka, com a mineração de dados por meio da classificação, utilizamos o algoritmo J48, pois, como menciona a obra de Khanna (2021), é um dos algoritmos de aprendizado de máquina mais usados para examinar os dados de forma categórica e contínua. O algoritmo C4.5 (J48) é usado principalmente entre muitos campos para classificar dados, por exemplo, interpretar os dados clínicos para o diagnóstico de doença cardíaca coronariana (como ocorre com os dados desta análise). O funcionamento do J48 envolve a construção de uma árvore de decisão de forma iterativa. Em cada etapa, o algoritmo seleciona o melhor atributo para dividir os dados com base em critérios específicos, como o ganho de informação. Essa divisão é feita de maneira recursiva, criando-nos e ramos na árvore até que todas as amostras tenham a mesma classe ou não possam ser divididas mais. A árvore resultante é utilizada para classificar novas instâncias, seguindo os ramos de acordo com os valores de seus atributos.

E, conjunto com essas ferramentas, fizemos o uso de planilhas, com o programa MS. Excel [Microsoft Excel], onde fizemos a projeção da planilhas, carregando os dados para tabulação, a cada transformação dos dados.

1. Baixamos os dados os dados originais (no formato de arquivo CSV – dataset original), que também pode ser encontrado na url https://github.com/gismarb/heart_attack_data_miner/blob/main/DATASET/CSV/heart_failure_clinical_records_dataset.csv;
2. Utilizando o Excel, fizemos o carregamento do arquivo CSV para um formato de planilha, que pode ser baixada na url https://github.com/gismarb/heart_attack_data_miner/blob/main/DOC/XLSX/heart_failure_clinical_records_dataset.xlsx ;
3. Fizemos as primeiras análises:
    a. Verificamos a estrutura dos dados, onde foi constatado que não seria necessário executar a fase de "Seleção", pois os dados constantes estavam intrinsicamente alinhados ao domínio (área da análise);
    b. Os dados estavam completos – não existiam dados faltantes (nulos);
    c. Os dados estavam disponíveis para iniciarmos uma primeira execução do classificado (fase de mineração de dados), pulando as etapas de "Classificação" e "Transformação".
4. Utilizando o arquivo CSV, sem nenhum tipo de ajuste e/ou modificação, submetemos o mesmo ao Weka, para fazer a primeira classificação. Esse procedimento não funcionou – o Weka não conseguiu reconhecer as variáveis como "*classes*", mas sim como tipo "*numeric*";
5. Utilizando o Weka, fizemos a transformação do arquivo CSV para ARFF (formato nativo do Weka), sinalizando, manualmente, quais eram as variáveis de tipo "*classe*" – o arquivo ARFF em questão pode ser consultado na url https://github.com/gismarb/heart_attack_data_miner/blob/main/DATASET/ARFF/heart_failure_clinical_records_dataset.arff . Em nosso caso, as variáveis que passaram por transformação, foram as variáveis que, em seu modelo inicial (no dataset) eram binárias – de valor "1" ou "0";
6. Posteriormente, submetemos esse arquivo ARFF ao Weka, e fizemos a classificação a primeira classificação, utilizando 70% dados para treinamento e 30% para testar o modelo, utilizando o algoritmo J48;
7. Decidimos pré-processar os dados que já tínhamos. Então, utilizando o dataset original (inicial), por meio do script (SQL), fizemos o upload desses dados para um banco de dados SQL (preenchendo uma tabela nomeada como "*dataset01*");
8. Usando o script (SQL), e aplicando tratamento de dados colunares (estrutura CASE), transformando as variáveis numéricas / binárias "1" e "0" em strings "YES" e "NO" (e, com exceção, no caso da variável "sex", "M" e "F"), visando melhorar a visualização dos dados resultantes;
9. Por meio desse tratamento de dados, enviamos esses dados tratados para outra tabela do banco de dados, a tabela "dataset02" – tendo agora um registro inicial e um tratado, armazenados em bases de dados diferentes;

10. A partir da tabela “dataset02”, baixamos os dados para um arquivo CSV, disponível na url https://github.com/gismarb/heart_attack_data_miner/blob/main/DATASET/CSV/dataset02.csv ;
11. Utilizando o Excel, fizemos o carregamento deste arquivo CSV (tratado) para um formato de planilha, visando a analisar os dados tabulares. A planilha pode ser baixada na url https://github.com/gismarb/heart_attack_data_miner/blob/main/DOC/XLSX/dataset02.xlsx ;
12. Carregamos o mesmo arquivo CSV para o Weka e fizemos sua conversão para o formato ARFF. O arquivo ARFF resultante pode ser visualizado e/ou baixado na url https://github.com/gismarb/heart_attack_data_miner/blob/main/DATASET/ARFF/dataset02.arff ;
13. Recarregamos o arquivo ARFF gerado (dataset02.arff) no Weka e fizemos a classificação a primeira classificação, utilizando 70% dados para treinamento e 30% para testar o modelo, utilizando o algoritmo J48;
14. Novamente, realizando o pré-processamento dos dados, voltamos ao banco de dados e, com a utilização do script (SQL), carregamos os dados da tabela “dataset02” para uma nova tabela “dataset03”, e aplicando tratamento de dados colunares (estrutura CASE). Nesse ponto, aplicamos a discretizarão das variáveis com dados do tipo “contínuo”. Nesse ponto, transformamos esses dados para o tipo de atributo ordinal {LOW, NORMAL, HIGH}, seguindo as regras:
    a. Levando em consideração os indicadores e/ou marcadores de saúde para as variáveis em específico, classificamos os valores, entre os atributos ordinais supracitados;
    b. Quando necessário, aplicamos conversão de tipos / dados;
    c. Fizemos a diferenciação entre o atributo sexo (em marcadores de saúdes, essa variável tem correlação direta para os marcadores);
    d. A transformação analisou o valor e classificou dentro do atributo ordinal, a depender de sua regra no domínio específico dos dados analisados;
    e. As fórmulas e transformações podem ser encontradas no script (SQL), na parte do carregamento da tabela “dataset03”.
15. A partir da tabela “dataset03”, baixamos os dados para um arquivo CSV, disponível na url https://github.com/gismarb/heart_attack_data_miner/blob/main/DATASET/CSV/dataset03.csv ;
16. Utilizando o Excel, fizemos o carregamento deste arquivo CSV (tratado) para um formato de planilha, visando a analisar os dados tabulares. A planilha pode ser baixada na url https://github.com/gismarb/heart_attack_data_miner/blob/main/DOC/XLSX/dataset03.xlsx ;

17. Carregamos o mesmo arquivo CSV para o Weka e fizemos sua conversão para o formato ARFF. O arquivo ARFF resultante pode ser visualizado e/ou baixado na url https://github.com/gismarb/heart_attack_data_miner/blob/main/DATASET/ARFF/dataset03.arff ;
18. Recarregamos o arquivo ARFF gerado (dataset03.arff) no Weka e fizemos a classificação a primeira classificação, utilizando 70% dados para treinamento e 30% para testar o modelo, utilizando o algoritmo J48;
19. Fizemos as análises finais, comparando as 3 classificações realizadas:
    a. Comparando os gráficos de correlação;
    b. Comparando os dados estatísticos resultantes da aplicação do J48;
    c. Comparando e entendendo a estrutura das árvores de decisões geradas pelo J48.

Esses procedimentos adotados, de realizar 3 rodadas de classificação com o J48, se deu pelo fato de a transformação dos dados alterar os resultados nos modelos gerados. O uso do J48, assim como o Weka, ocorreu por conta de serem as ferramentas demonstradas em aula. Por fim, o fluxo do processo, envolvendo SQL, Excel, CSV e ARFF, ocorreu por conta da visibilidade do processo e coerência com o KDD (aplicando tecnologia e técnicas multidisciplinares para a extração do conhecimento).

3. Resultados

No contexto geral, os resultados, assim como artefatos criados e/ou utilizados para a conclusão desse projeto / análise (planilhas, imagens, scripts etc.) podem ser encontrados no repositório do [GitHub], na url https://github.com/gismarb/heart_attack_data_miner/tree/main .

Adiante os, resultados serão separados em 3 seções: resultados das classificações; gráficos de correlação e árvores de decisões. Serão comparados os 3 processos de classificação para cada seção em específico. Os dados são os retirados do Weka, com a aplicação do J48, advindos dos dados oriundos do KDD (sendo o processamento do J48 e Weka, a parte da mineração dos dados).

3.1. Resultados da classificação

Tabela 2 - Comparativos das Classificações do J48

	i	ii	iii
Instâncias classificadas corretamente	83%	83%	76%
Instâncias classificadas incorretamente	17%	17%	24%

i – heart_failure_clinical_records_dataset.arff;

ii – dataset02.arff;

iii – dataset03.arff;

3.2. Gráficos de correlação

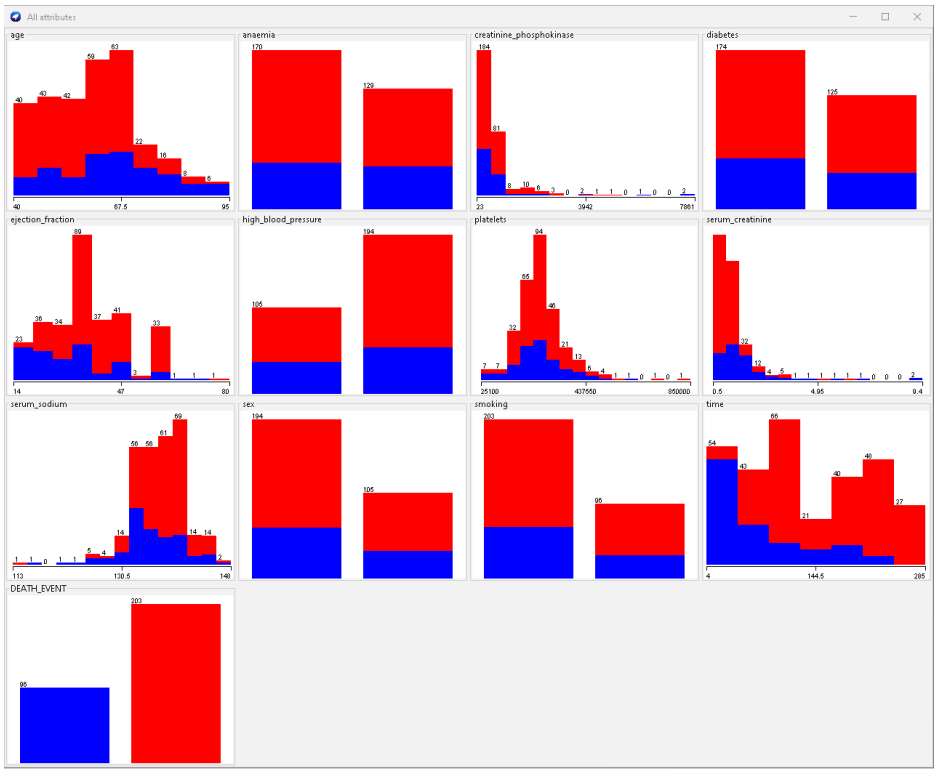

Figura 1 - Gráfico de correlação dataset02

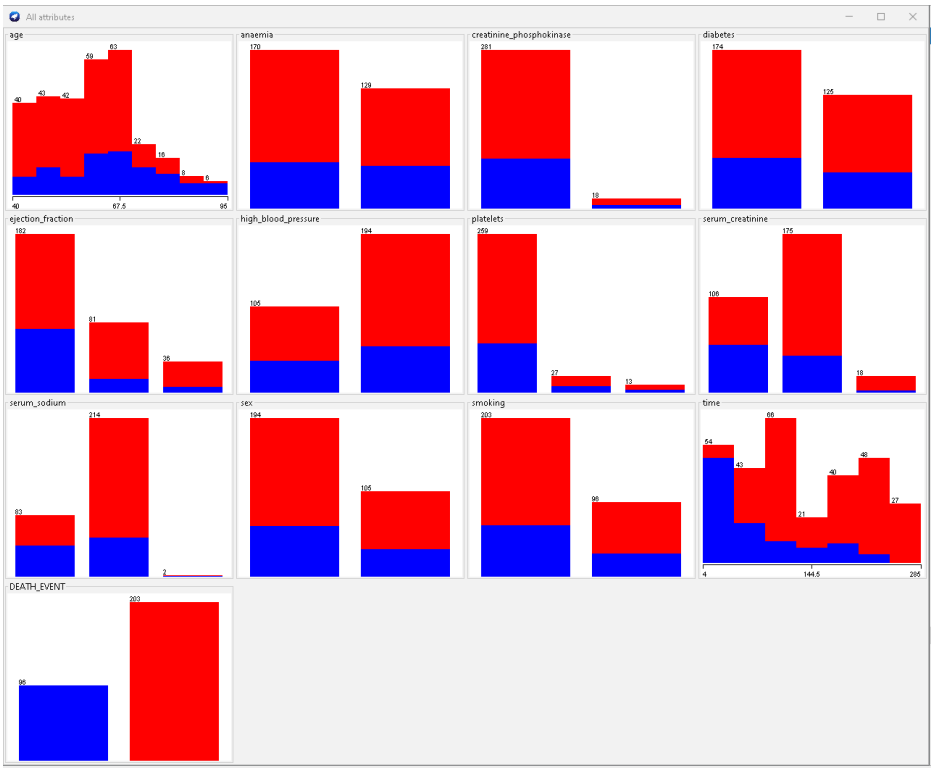

Figura 2 - Gráfico de correlação dataset03

3.3. Árvores de decisão

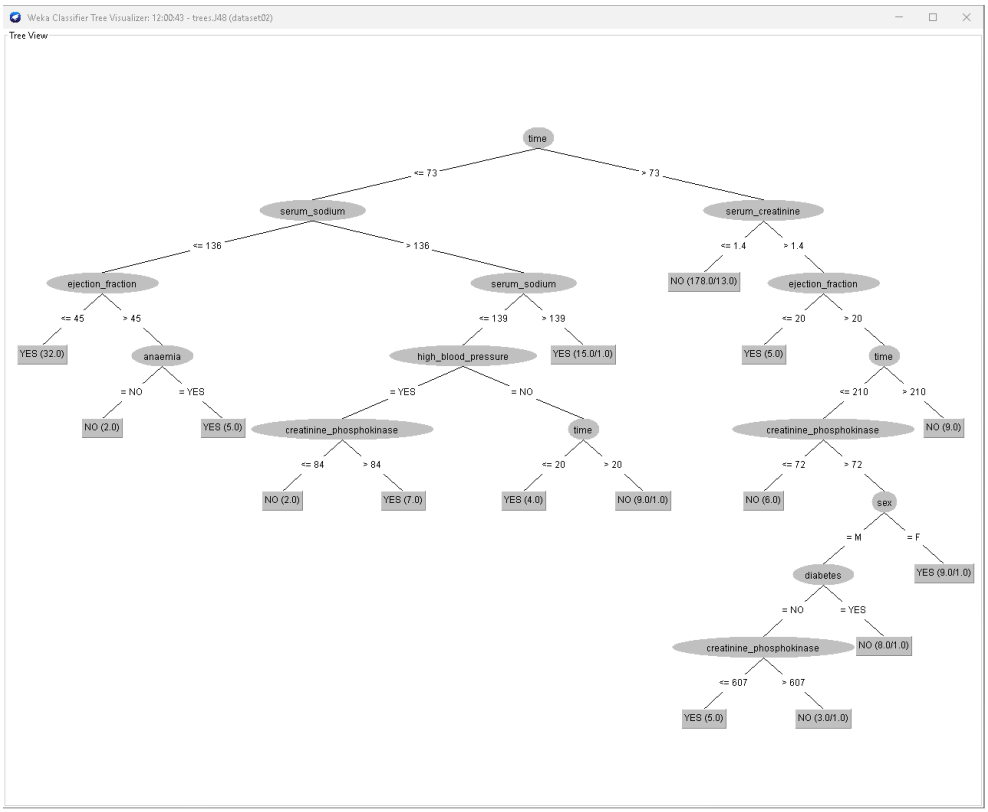

Figura 3 - Árvore J48 dataset02

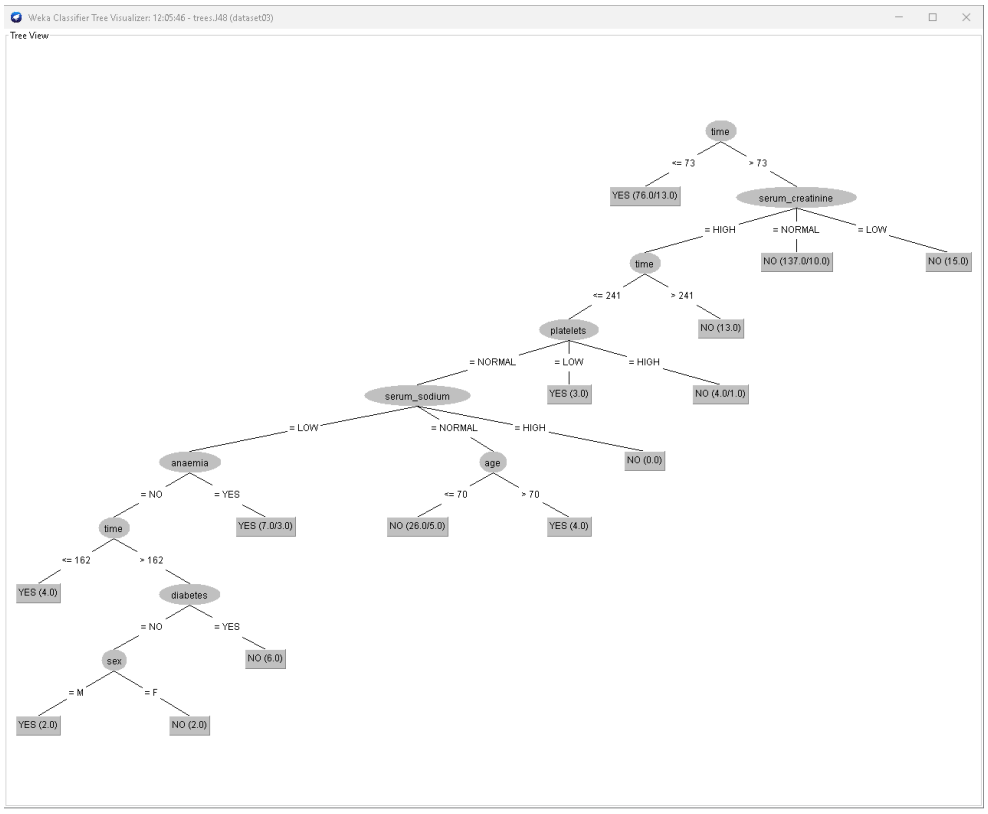

Figura 4 - Árvore J48 dataset03

4. Discussão

Após o processamento e reprocessamento dos dados, referente aos dataset utilizados neste projeto, observamos as seguintes situações:

- A variável "time", que é o "período de acompanhamento do paciente", é, sem sobra de dúvidas, o dado mais relevante para definir morte ou não morte de um paciente – ao menos é o que verificamos com a aplicação do classificado do algoritmo do J48;
- A classificação do dataset original e o que sofreu o primeiro tratamento dos dados (dataset02), demonstra uma maior taxa de consistência na classificação dos dados em relação a classificação que ocorreu no dataset03;
- A classificação do dataset original e o que sofreu o primeiro tratamento dos dados (dataset02), tem uma árvore de decisões mais balanceada – e não levou em consideração as variáveis "*platelets*" e "smoking", como sendo fatores a serem considerados para predição de morte ou não morte;
- A classificação do dataset03 (que recebeu os tratamentos de discretização), não está tão balanceada (ainda que isso possa não ser relevante) e, não levou em consideração as variáveis "*creatinine_phosphokinase*", "*ejection_fraction*", "*high_blood_pressure*" e "*smoking*", como sendo fatores a serem considerados para predição de morte ou não morte.

Em resumo, seguindo as classificações realizada (processo de mineração de dados), podemos entender que a variável "time", é o fator de maior relevância para morte, porém, da mesma forma, podemos afirmar que a estratégia de categorizar os indicadores das variáveis contínuas (ainda que realizando o mesmo processo de marcadores de exames), não seja a melhor estratégia para realizar um modelo preditivo para esse domínio. Portanto, entendemos que esses resultados não sejam conclusivos para o domínio e, talvez seja necessário um maior volume de dados para submeter ao classificador ou, até mesmo, mais variáveis a serem levadas em consideração – porém isso dependeria de auxílio de profissionais do referido domínio análise em conjunto.

5. Conclusão

Após a realização das análises, refazendo os processos de tratamento de dados, assim como submetendo ao classificador para mineração dos dados, concluímos que o processo não foi efetivo. Uma prática comum do domínio (saúde) é categorizar em classes os valores aceitáveis dos marcadores de exames. Porém ao aplicar essa mesma regra aos dados do dataset, percebemos que isso distorceu a classificação de dados. E, percebemos que fatores (de conhecimento comum) importantes para saúde cardíaca, como a pressão sanguínea alta e o uso do tabaco, foram descartados pelo classificador. Isso não está compatível com o que ocorre no dia a dia e, por isso, entendemos que, ou essa regra de processamento não possa ser aplicado neste contexto, ou sejam necessários mais dados massivos para treinar o classificado ou, por final, de fato, seja necessário manter os dados no seu formato, contínuo, pois isso, quando correlacionado, se mostra mais próximo da realidade (aumenta a precisão do classificador e fica perto do que, realmente ocorre no mundo real).

References

Bourbon, Mafalda, et al. *Doenças Cardiovasculares*. fevereiro de 2016, p. 1–24.

Chicco, Davide, e Giuseppe Jurman. *Heart Failure Clinical Records*. UCI Machine Learning Repository, 2020. *DOI.org (Datacite)*, https://doi.org/10.24432/C5Z89R.

Chicco, Davide, e Giuseppe Jurman. "Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone". *BMC Medical Informatics and Decision Making*, vol. 20, nº 1, dezembro de 2020, p. 16. *DOI.org (Crossref)*, https://doi.org/10.1186/s12911-020-1023-5.

"GitHub: Let's Build from Here". *GitHub*, 2024, https://github.com/.

Khanna, Nilima. "J48 Classification (C4.5 Algorithm) in a Nutshell". *Medium*, 18 de agosto de 2021, https://medium.com/@nilimakhanna1/j48-classification-c4-5-algorithm-in-a-nutshell-24c50d20658e.

Marcela. "Classificação de dados: Abordagens e modelos em mineração de dados". *Awari*, 31 de julho de 2023, https://awari.com.br/classificacao-de-dados-abordagens-e-modelos-em-mineracao-de-dados/.

OPAS/OMS. "Doenças cardiovasculares - OPAS/OMS | Organização Pan-Americana da Saúde". *OPAS - Organização Pan-Americana de Saúde*, 31 de janeiro de 2024, https://www.paho.org/pt/topicos/doencas-cardiovasculares.

"Software de planilha online gratuito: Excel | Microsoft 365". *Microsoft Excel*, https://www.microsoft.com/pt-br/microsoft-365/excel. Acesso em 5 de junho de 2024.

Souza, Alex. "Knowledge Discovery in Databases (KDD)". *Blog Do Zouza*, 26 de julho de 2023, https://medium.com/blog-do-zouza/knowledge-discovery-in-databases-kdd-462ea2775715.

"Weka 3 - Data Mining with Open Source Machine Learning Software in Java". *WEKA The Workbench for Machine Learning*, https://waikato.github.io/weka-site/index.html. Acesso em 31 de maio de 2024.