9º EnPE
Encontro de Ensino, Pesquisa & Extensão
Patrocínio, MG, outubro de 2022

TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL PARA O AUXÍLIO NA DETECÇÃO DE POSSIBILIDADE DE SOBREVIVÊNCIA APÓS O DESDOBRAMENTO DE UMA DOENÇA CARDÍACA

Vitória Stéfane de Souza, Danielli Araújo Lima
<vitoriasteffane5@gmail.com>,<danielli@iftm.edu.br>
Instituto Federal do Triângulo Mineiro (IFTM) Campus Patrocínio
Laboratório de Inteligência Computacional e Robótica (LICRo)
Modalidade: Pesquisa

Resumo: Doenças cardiovasculares vem proporcionando um número ocasionalmente grande de mortes no mundo inteiro, podem ser relacionadas a uma má alimentação ou podem ser hereditárias. Algumas delas são a insuficiência cardíaca, infarto agudo do miocárdio e a hipertensão. Com o passar dos anos, a tecnologia juntamente com a inteligência artificial se fez cada vez mais presente na medicina cardiovascular. Com ferramentas elaboradas para auxiliar no diagnóstico e tratamento de doenças coronárias. Será abordado no presente estudo, como o aprendizado de máquina pode ser utilizado na detecção de fatores que influenciam na probabilidade de sobrevivência de pacientes com doenças cardiovasculares, a partir de uma base de dados coletada de diferentes pacientes. Os algoritmos usados foram o Decision Tree e Support Vector Machines. Assim, foi possível perceber quais são os dados de maior relevância para o índice de sobrevivência após o desenvolvimento de alguma doença cardíaca. Os resultados apresentaram $84,512\%$ de acurácia para o algoritmo Support Vector Machines e $81,145\%$ de acurácia para a algoritmo Decision Tree.

Palavras-chave: Classificação. Aprendizado Supervisionado. Arvore de Decisão. Máquinas de Vetores Suporte. Cardiologia. Infarto.

Introdução

O coração é um dos órgãos de maior importância do corpo humano, dando suporte à vida nos seres humanos. Por ser um órgão tão importante, é necessário um grande cuidado, pois caso desenvolva uma doença cardiovascular, se não bem tratada, ou se houver uma contrariedade nos exames ou mesmo da própria pessoa, um evento de morte poderá ocorrer. Doenças cardiovasculares podem ser desenvolvidas com o tempo ou adquiridas por hereditariedade.

Muitas vezes o que acarreta o desenvolvimento de doenças cardíacas é a ma alimentação. São exemplos de algumas doenças cardíacas: hipertensão arterial - popularmente conhecida por pressão alta, ela é caracterizada por níveis elevados de pressão sanguínea nas artérias; insuficiência cardíaca - baseia-se em um distúrbio onde o coração se torna incapaz de suprir as necessidades do corpo, causando a restrição do fluxo sanguíneo, congestão do sangue nas veias

e nos pulmões; infarto agudo do miocárdio - descrito como necrose miocárdica resultante da abstração de uma artéria coronária; entre outras diversas (CHICCO; JURMAN, 2020).

Ao aplicarmos a inteligência artificial (IA) no âmbito da medicina, juntamente com a análise dos médicos, podemos proporcionar melhoras significativas na descoberta, tratamento e análise de doenças diversas. Analisando uma base de dados (BD) coletados de pacientes com insuficiência cardíaca e aplicando algoritmos de aprendizado de máquina (AM) para prever a sobrevivência de pacientes com doenças cardiovasculares, podemos identificar quais os fatores predominantes no diagnóstico, por meio da ciência de dados (CD), apoiado por especialistas.

Neste trabalho temos como objetivo aplicar os algoritmos de Decision Tree (DT) e o Support Vector Machine (SVM), considerados duas diferentes técnicas de aprendizado de máquina supervisionado, com o foco em fazer a classificação de doenças do coração em casos de sobrevivência ou morte. Neste sentido, as técnicas são comparadas utilizando-se a matriz de confusão. Ao final, os resultados são discutidos e analisados a partir dos parâmetros coletados pelos algoritmos de classificação.

Fundamentação teórica

Para compreender a aplicação as definições de doenças do coração serão apresentadas algumas dela no estudo. Neste sentido, doenças cardíacas são doenças que atacam o coração como consequência de uma alimentação ruim, ou com o tempo. Algumas delas são infarto agudo do miocárdio que é quando um coágulo bloqueia o fluxo sanguíneo para o coração entre outras como cardiopatia, hipertensão, doença cardíaca reumática. Assim, neste trabalho estaremos usando uma base de dados com população de tamanho 299, com 13 atributos, sendo eles 105 mulheres e 194 homens, com a faixa etária de 40 e não ultrapassando os 95 anos de idade. Como (CHICCO; JURMAN, 2020) sita em seu estudo, ambos possuíam disfunção sistólica ventricular esquerda e insuficiência cardíaca prévia.

Cada um desses atributos podem interferir diretamente na taxa de óbitos de pessoas com doenças do coração, que podem ser desenvolvidas com o tempo, através de uma alimentação ruim ou hábitos não saudáveis, como por exemplo fumar ou até mesmo o estresse. O atributo sexo representa a classe de homens ou mulheres. O atributo anemia representa se o paciente tem ou não anemia, e significa segundo (CHICCO; JURMAN, 2020), a pessoa é considerada anêmica se os níveis de hematócrito no sangue estiverem em níveis inferiores a 36%. Creatina fosfoquinase (CPK) é o indicador dos níveis de enzima no sangue, um alto níveis de CPK no sangue pode ser considerado um indicador de insuficiência cardíaca, pois quando ocorre a danificação muscular do tecido, a CPK flui para o sangue, como (CHICCO; JURMAN, 2020) diz. Creatina sérica é causada pela creatina quando o músculo se rompe, no entanto os médicos levam em consideração para monitorar a função renal do paciente e a fração de ejeção (FE) significa quanto sangue está sendo bombeado pelo ventrículo esquerdo a cada batimento do coração. Possuindo idade, diabetes, pressão alta, plaquetas, tabagismo, tempo e sódio sérico também como atributos.

O aprendizado de máquinas pode ser entendido como um sub-campo da ciência da computação onde o computador tem a possibilidade de aprende sem ser devidamente programado, ou seja, não havendo programação explicita, podendo aprender através de dados (ALVES; LIMA, 2018). Existem quatro tipos de aprendizado de máquinas: (1) Aprendizado Supervisionado, (2) Aprendizado semi-supervisionado, (3) Aprendizado não-supervisionado e o aprendizado por reforço. Neste trabalho focaremos no aprendizado supervisionado que é um modelo que pode aprender doravante de resultados pré definidos, fazendo uso de informações, sendo elas bem rotuladas. Podendo assim, possibilitar o treinamento do algoritmo em alguma tarefa específica.

Dentre as técnicas de aprendizado supervisionado temos: Redes Neurais Artificiais (RNA), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Gradient Boosting (GB) (LIMA; FERREIRA; SILVA, 2021). O foco deste trabalho será a comparação de resultados entre a DT e o SVM. Neste sentido, a DT é uma técnica supervisionada, onde se baseia a construção de uma árvore de decisão, onde vários pontos de decisão serão tomados, pontos esse chamados de nós da árvore, onde o resultado será seguir por uma caminho ou outro denominados ramos. O SVM criado em 1995 por Vladimir Vapnik, por sua vez, é uma técnica supervisionada que é utilizada para tarefas de regressão e comparação. Baseada na procura de uma maior margem para conseguir dividir diferentes dados. Distinguindo um hiperplano ideal para que possa realizar a separação dos dados com uma maior margem possível. Quando os dados não são separados linearmente deve-se passar para uma outra dimensão maior até que se consiga fazer uma separação de ambos. Existem diferentes ferramentas para a mineração de dados: o WEKA, a Linguagem R e o KNIME Analytics Platform que é a ferramenta usada neste trabalho. O KNIME é uma ferramenta desenvolvida pela equipe de engenheiros de software da Universidade de Konstanz liderada por Miichael Berthold, com sede em Zurich, Suíça, é uma plataforma de código aberto para empregar as técnicas de aprendizado de máquina e programação visual.

No trabalho de (AHMAD et al., 2017) é elaborado um estudo sobre a população do Paquistão com insuficiência cardíaca, estimando a taxa de sobrevivência e mortalidade. Foi utilizado 200 replicações de bootstrap, inclinação do preditor linear médio, curva de ROC, modelo de Cox e um nomograma para visualização gráfica de sobrevivência. Segundo eles, a idade, fração de ejeção, sódio, anemia, pressão arterial e creatina foram dados muito significativos para a análise. Com a curva de ROC foi possível detectar que em um tempo mais longo de acompanhamento obteve-se 81% em relação ao evento de morte, enquanto em um curto tempo é capaz de reconhecer apenas 77%.

No trabalho de (CHICCO; JURMAN, 2020) os autores abordam apenas dois parâmetros clínicos para a abordagem de sobrevivência dos pacientes com insuficiência cardíaca, utilizando a mesma base que (AHMAD et al., 2017), são eles creatina sérica e fração de ejeção, onde baseou-se a construção dos modelos de aprendizado de máquina. Foram aplicados para a conseguirem prever a sobrevivência 10 métodos diferentes de diversas áreas de aprendizado de máquina. Igualmente como (CHICCO; JURMAN, 2020) descreveu em seu estudo, também chegou a um resultado de que a fração de ejeção e a creatina sérica foram os dados mais relevante para previsão da insuficiência cardíaca. Dentre os 5 classificadores considerados pelo autor, a arvore de decisão foi quem proporcionou melhores resultados. Tendo como resultado 80% de apuração.

Metodologia

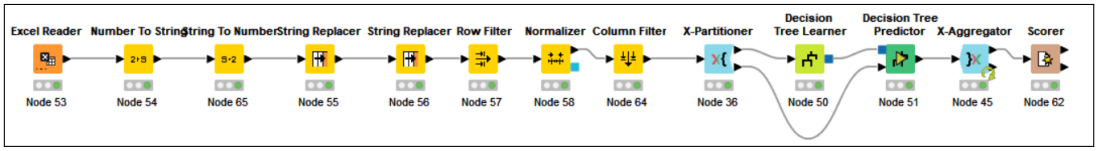
Utilizando a plataforma KNIME elaboramos 2 workflows para a classificação da base de dados de insuficiência cardíaca. Primeiramente, usamos o algoritmo DT, que baseia-se em vários pontos de decisão, datados de nós, portanto a decisão será tomada por um ou por outro existido vários caminhos a serem tomados como resultado. A partir dele, a IA poderá analisar se a pessoa sobreviveu ou não a uma doença cardiovascular. Ademais, outro algoritmo utilizado foi o SVM, baseado em vetores suporte.

Na Figura 1(a) é apresentado o workflow para a DT contando com 13 nós distintos. Temos um nó leitor da base de dados do tipo **Excel Reader**, é ele que lera os dados para a execução do fluxo. O nó **Number To String** passa uma ou várias colunas da base para o formato de string se a mesma estiver no formato de número, utilizado para passar o evento de morte que estava do tipo número para string. O nó **String To Number** tem o mesmo funcionamento de

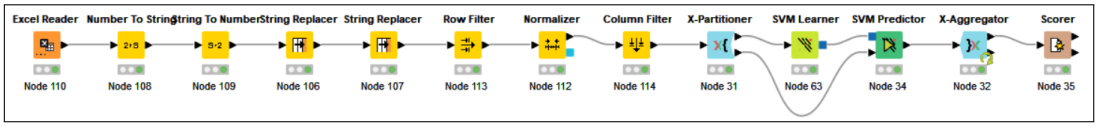
Number To String, só que passará a coluna de número para string, feito para as classes de platelets e serum creatine. Os 2 nós **String Replacer** possibilitam modificar os valores das células, onde ocorreu a modificação do evento de morte de $0 \rightarrow$ `YES` (representa os pacientes que não morreram) e $1 \rightarrow$ `NO` (pacientes que foram a óbito). O nó **Row Filter** é utilizado para filtragem de linhas, usado para determinar as idades dos pacientes que possuíam algumas não consideráveis. Foi determinado as idades de 1 a 99. O nó **Normalize** é usado para normalizar a base, pois tinham valores extremamente grandes que poderiam causar conflito no resultado final dos testes, após aplica-lo todos os valores ficaram estipulados entre $\{0, 1\}$ respectivamente.

O nó **Column Filter** possibilita a filtragem de colunas da base, podendo escolher quais serão analisadas e quais não serão para análise dos resultados. O nó **X-Partitioner** é um loop de validação cruzada (cross validation), ou seja, todos os nós que estiverem entre ele e o nó **X-Aggregator**, que é o final do loop de validação que coleta os resultados, compara as classe e gera as previsões para as linhas, executaram quantas vezes for estipulado, nesse caso foram feitas 10^2 repetições. O nó **Decision Tree Learn** induzirá a árvore de classificação, o atributo principal deverá ser nominal, fornecerá dois métodos para a divisão, o índice de Gini e a taxa de ganho e um método de poda, que se baseia no comprimento mínimo de descrição. O nó **Decision Tree Predictor** utiliza a árvore de decisão existente para descobrir o valor da classe para padrões atuais. O nó **Scorer** permite a visualização da matriz de confusão, a classificação correspondente e tem como saída a sensibilidade, especificidade, acurácia e Cohen's kappa.

Na Figura 1(b) foi apresentado o Support Vector Machine, também contando com 13 nós, possui 10 nós identicos ao de Decision Tree que são Excel Reader, Number To String, String To Number, String Replacer, Row Filter, Normalise, Colum Filter, X-Partitioner que contou também com 100 execuções, X-Aggregator e Scorer. Suas definições foram mostradas



(a) Algoritmo Decision Tree.



(b) Algoritmo Support vector machine.

Figura 1 – Workflows para a classificação binária de doenças do coração.

acima. O nó **SVM Leaner** é caracterizado por treinar uma máquina de vetor de suporte nos dados iniciais, é capaz de traçar um hiperplano para cada classe. O nó **SVM Predictor** usa o modelo que foi criado pelo SVM Learner para identificar a saída de valores específicos.

Resultados

Em algoritmos de aprendizado supervisionado, uma das melhores abordagens para verificar o sucesso do aprendizado é por meio da acurácia. A acurácia é medida por meio de quatro parâmetros: (i) Falso positivo (FP) é quando o resultado é negativo mas classifica como positivo, (ii) Falso Negativo (FN) quando o resultado é positivo mas classifica como negativo, (iii) Verdadeiro Positivo (VP) quando é realmente verdadeiro, ou seja, quantidade de mortos, e o (iv) Verdadeiro Negativo (VN) quando é realmente negativo, quantos não morreram. Ao ser feita a

contagem de todos esses termos e obter a matriz de confusão, é possível calcular métricas de avaliação da acurácia (A) para a classificação, conforme a Equação 1.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

Dessa forma, pode-se observar na Tabela 1 com base nos dados coletados, que o melhor método para analisar e prever a possibilidade de sobrevivência de pacientes com doenças cardíacas é o SVM Learner. Neste sentido o SVM proporcionou melhores resultados em sua análise

Tabela 1 – Resultados encontrados a partir dos métodos de aprendizado DT e SVM.

Métodos de Classificação	FP	FN	VP	VN	Acurácia	Erro
Decision Tree	26	172	69	30	81,145%	18,855%
SVM Learner	30	186	65	16	84,512%	15,488%

com $84,512\%$, já a DT teve uma acurácia levemente menor de $81,145\%$, superior aos 74.0% alcançados no trabalho de (CHICCO; JURMAN, 2020) para os parâmetros de "creatinine' e "ejection fraction". Neste sentido, podemos verificar que a SVM é o algoritmo mais indicado, dentre os testados para realizar a classificação da BD do coração, usada neste trabalho.

Considerações finais

O coração é um órgão de extrema importância e precisamos cuidar daquilo que nos faz vivos. Sabe-se que o estudo de parâmetros que afetam o coração pode ser muito importante para diminuir a taxa de mortalidade em casos repentinos de alguma falha. Portanto, analisando os resultados de (CHICCO; JURMAN, 2020) nós conseguimos superar os resultados apresentados pelo autor, como visto na tabela (1) com o algoritmo de aprendizado supervisionado SVM. Neste sentido, o algoritmo superou SVM o DT, também investigado no presente trabalho. Como trabalhos futuros, para esta mesma base de dados, apresentaremos os parâmetros que mais são relevantes para a predição de doenças do coração. Ademais, usaremos outros algoritmos de CD para a predição e comparação com os resultados já alcançados.

Referências

AHMAD, T. et al. Survival analysis of heart failure patients: A case study. **PloS one**, Public Library of Science San Francisco, CA USA, v. 12, n. 7, 2017. Citado na página 3.

ALVES, F. B.; LIMA, D. A. Uso de la clasificación para el análisis y la minería de datos en la herramienta de enseñanza-aprendizaje google classroom. **Nuevas Ideas en Informática Educativa**, v. 4, p. 589–594, 2018. Citado na página 2.

CHICCO, D.; JURMAN, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. **BMC medical informatics and decision making**, BioMed Central, v. 20, n. 1, 2020. Citado 3 vezes nas páginas 2, 3 e 5.

LIMA, D. A.; FERREIRA, M. E. A.; SILVA, A. F. F. Machine learning and data visualization to evaluate a robotics and programming project targeted for women. **Journal of Intelligent & Robotic Systems**, Springer, v. 103, n. 1, p. 1–20, 2021. Citado na página 3.