AVALIAÇÃO DE CONJUNTO DE DADOS DE PACIENTES CARDÍACOS ATRAVÉS DE ALGORITMOS DE REGRESSÃO e CLASSIFICAÇÃO

Mestrado: Programa de Pós Graduação em Ciência da Informação
Disciplina: Machine Learning
Professor: Gustavo Medeiros de Araujo
Aluna: Eliane Somavilla
Semestre: 2021/01
Data: 02/08/2021

Introdução

Este estudo foi focado na análise de sobrevida de pacientes com insuficiência cardíaca que foram internados no Instituto de Cardiologia e Hospital Aliado de Faisalabad-Paquistão durante Abril e Dezembro de 2015. Todos os pacientes apresentavam disfunção sistólica do ventrículo esquerdo, pertencentes às classes III e IV da classificação realizada pela New York Heart Association. Diversos algoritmos de Machine Learning capazes de analisar dados através de técnicas de regressão e classificação foram usados para prever a taxa de mortalidade de futuros pacientes com problemas similares. Foram consideradas características como idade, fração de ejeção, creatinina sérica, sódio sérico, anemia, plaquetas, creatinina fosfoquinase, pressão arterial, diabetes e tabagismo como potenciais contribuintes para a mortalidade.
Todas as características foram analisadas de forma a identificar o conjunto mínimo de informações necessárias para um diagnóstico rápido e eficiente referente à insuficiência cardíaca.

Descrição do Problema

Anualmente, as doenças cardiovasculares matam milhões de pessoas no mundo todo e se manifestam principalmente como infartos do miocárdio e insuficiência cardíaca. A insuficiência cardíaca é caracterizada pela incapacidade do coração atuar adequadamente como bomba, quer seja por déficit de contração e/ou de relaxamento, comprometendo o funcionamento do organismo. Quando não tratada adequadamente, a qualidade de vida é reduzida podendo levar o paciente a óbito.

Os principais fatores de risco cardiovasculares, segundo o Hospital Israelita Albert Einstein da cidade de São Paulo/SP -BR, estão diretamente relacionados com fatores como a hipertensão arterial, diabetes, tabagismo, dislipidemia (gorduras no sangue) e sedentarismo. A presença destes fatores isoladamente ou de forma conjunta em um organismo provocam o desenvolvimento da doença coronariana que pode levar ao infarto agudo do miocárdio ou diminuição da performance do coração.

Outras causas incluem doenças que acometem as válvulas cardíacas (degenerativas ou inflamatórias, como doenças renais e reumática), doenças congênitas, doenças genéticas, auto-imunes, inflamatórias (Ex: durante o periparto), por toxicidade (tratamento de câncer) e também infecciosas (Ex: AIDS). O abuso de bebidas alcoólicas e drogas como a cocaína também elevam o risco de desenvolvimento de doenças cardíacas que são hoje responsáveis por 31% das mortes em todo o mundo.

O diagnóstico da insuficiência cardíaca é clínico por um profissional médico especializado em cardiologia, através da história contada pelo paciente. O relato pode incluir a intolerância aos esforços, a falta de ar ao deitar e o inchaço nos membros inferiores e do abdômen. O diagnóstico normalmente também considera o exame físico de acúmulo de sangue nos pulmões e no organismo como um todo.

O exame que confirma a insuficiência cardíaca é o ecocardiograma e substâncias produzidas pelo coração insuficiente também podem auxiliar no diagnóstico, como o peptídeo natriurético tipo B, conhecido como BNP. Parte fundamental do diagnóstico é tentar encontrar a causa, uma vez que pode implicar em tratamentos específicos.

De acordo com a SCIELO-Brasil (Revista da Associação Médica Brasileira), o comprometimento sistólico do ventrículo esquerdo é responsável pela maioria dos casos de falência crônica do coração e pode ser diagnosticado ecocardiograficamente pela fração de ejeção ventricular esquerda igual ou inferior a 0,40. A disfunção diastólica é caracterizada por sintomas de insuficiência cardíaca com a fração de ejeção preservada (habitualmente > 0,45)

A insuficiência cardíaca pode se estabelecer em qualquer faixa etária e estima-se sua prevalência em 1 a 2% da população. Além disso, incide de forma progressiva com o aumento da idade, sendo que após os 70 anos, mais de 10% da população é acometida. Após os 55 anos, existe um risco de aproximadamente 30% de desenvolvimento da insuficiência cardíaca.

Visando apoiar os profissionais da saúde no diagnóstico de seus pacientes, objetiva-se com este estudo estimar as taxas de mortalidade devido à insuficiência cardíaca e comprovar sua ligação com alguns dos principais fatores de risco apontados pelo Hospital Israelita Albert Einstein da cidade de São Paulo/SP -BR e pela SCIELO-Brasil, escolhendo uma base de dados confiável para estudo e aplicação de algoritmos de Machine Learning especialistas em análise de dados por meio de técnicas de regressão e classificação.

Considerando também que a realidade de muitos hospitais e clínicas no Brasil e em diversos outros países não colabora para um diagnóstico completo e detalhado para os pacientes, neste estudo avaliamos quais são as características mínimas que um profissional de saúde deve ter acesso para realizar um diagnóstico rápido e efetivo a respeito de insuficiência cardíaca.

Metodologia

Neste estudo realizamos a análise de uma base de dados com informações reais sobre pacientes com insuficiência cardíaca através de algoritmos que viabilizam o aprendizado de máquina. Desta forma, os computadores têm a capacidade de aprender de acordo com as respostas esperadas por meio associações de diferentes dados.

Outro benefício alcançado se refere à seleção de recursos, pois, por meio de funções específicas dos algoritmos, é possível selecionar automaticamente o subconjunto de dados considerados mais relevantes para definir o resultado da análise computacional. Dessa forma, melhoramos a eficiência computacional e reduzimos o erro de generalização do modelo removendo dados ou ruídos irrelevantes.

Base de Dados

A base de dados utilizada para o estudo faz parte do repositório de arquivos do Machine Learning Repository (UCI) que mantém uma coleção de bancos de dados e geradores de dados usados pela comunidade para alimentar de algoritmos de aprendizado de máquina.O repositório foi criado em 1987. Desde então, ele tem sido amplamente utilizado por alunos, educadores e pesquisadores em todo o mundo como uma fonte primária de conjuntos de dados de aprendizado de máquina.

O conjunto de dados escolhido para o estudo contém os prontuários de 299 pacientes com insuficiência cardíaca e todos os pacientes apresentavam disfunção sistólica do ventrículo esquerdo, pertencentes às classes III e IV da classificação realizada pela New York Heart Association. A coleta dos dados foi realizada durante o período de acompanhamento no ano de 2015, especificamente entre os meses de Maio e Dezembro. Cada perfil de paciente apresenta 13 características clínicas. São elas:

1. **Idade**: idade do paciente (anos)
2. **Anemia**: diminuição dos glóbulos vermelhos ou hemoglobina (booleana)
3. **Pressão alta:** se o paciente tem hipertensão (booleana)
4. **Creatinina fosfoquinase** (CPK): nível da enzima CPK no sangue (mcg / L)
5. **Diabetes**: se o paciente tem diabetes (booleano)
6. **Fração de ejeção (FE):** porcentagem de sangue que sai do coração a cada contração (porcentagem)
7. **Plaquetas**: plaquetas no sangue (quilo placas / mL)
8. **Sexo**: mulher ou homem (binário)
9. **Creatinina sérica**: nível de creatinina sérica no sangue (mg / dL)
10. **Sódio sérico**: nível de sódio sérico no sangue (mEq / L)
11. **Tabagismo**: se o paciente fuma ou não (booleano)
12. **Tempo**: período de acompanhamento (dias)
13. **Evento de óbito**: se o paciente faleceu durante o período de acompanhamento (booleano)

A versão original do conjunto de dados foi coletada por Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab e Muhammad Ali Raza (Faculdade do Governo University, Faisalabad, Paquistão) e disponibilizado em julho de 2017.

A versão atual do conjunto de dados foi elaborada por David Chicco (Krembil Research Institute, Toronto, Canadá) e doado ao Repositório de Aprendizado de Máquina da Universidade da Califórnia sob os mesmos direitos autorais em janeiro de 2020.

Análise dos Dados

O estudo atual é baseado em 299 pacientes com insuficiência cardíaca, sendo 105 mulheres e 194 homens. Para a avaliação quantitativa da progressão da doença, os médicos utilizam normalmente a classificação funcional da New York Heart Association (NYHA), que determina quatro classes que vão desde a ausência de sintomas de atividades normais (Classe I) até um estágio em que qualquer atividade física provoca desconforto e ocorrem sintomas em repouso (Classe IV).

Todos os pacientes tinham entre 40 e 95 anos, apresentavam disfunção sistólica do ventrículo esquerdo e tinham insuficiência cardíaca prévia que os colocava nas classes III ou IV da classificação da New York Heart Association (NYHA).

O tempo de acompanhamento foi de 4 a 285 dias, com uma média de 130 dias. O tempo de acompanhamento não foi considerado como parâmetro de entrada deste estudo, pois objetivou-se focar em dados relacionados à causa da insuficiência cardíaca e não no tempo que o paciente esteve sob acompanhamento e tratamento médico. O gênero dos pacientes (sexo) também não foi considerado como parâmetro de entrada pois entende-se que é apenas uma característica utilizada categorizar os pacientes em dois grupos distintos.

A doença foi diagnosticada nos pacientes desta base de dados por laudo de ecocardiograma, exames sanguíneos ou notas escritas pelo médico. As características de Idade, sódio sérico, creatinina sérica, tabagismo, pressão arterial (PA), fração de ejeção (FE), anemia, plaquetas, creatinina fosfoquinase (CPK) e diabetes foram considerados como variáveis potenciais para explicar a mortalidade por Insuficiência Cardíaca. As plaquetas foram divididas em três níveis com base nas referências médicas por litro de sangue: valores abaixo de 150000 indicam índice baixo, entre 150000 e 450000 indicam índice normal e acima de 450000 indicam índice alto relacionado à presença de trombos ou coagulação sanguínea. A creatinina sérica maior que seu nível normal (1,5) é um indicador de disfunção renal.

Entre os 299 pacientes estudado, 96 eram fumantes, 105 apresentavam pressão alta, 67 apresentavam creatinina maior que 1,5, 125 tinham anemia, 129 tinham diabetes e 219 apresentavam fração de ejeção menor ou igual que 40, ou seja, o percentual de sangue bombeado para fora de um ventrículo esquerdo a cada batimento cardíaco era de até 40%. O total de 252 pacientes apresentaram um índice normal de plaquetas, o que significa que não havia presença de trombos ou coagulação sanguínea. O total de 96 pacientes morreram durante o período de estudo.

O sódio é um mineral que serve para o correto funcionamento dos músculos e nervos. O teste de sódio sérico é um exame de sangue que indica os níveis de sódio no sangue do paciente. Um nível

anormalmente baixo de sódio no sangue pode ser causado por insuficiência cardíaca. O sódio é considerado baixo quando inferior a 136 (mEq / L), de acordo com a SCIELO-Brasil, e nestes casos o paciente apresenta hiponatremia, que é o distúrbio hidroeletrolítico mais comum em pacientes hospitalizados. A presença de hiponatremia está associada a uma série de desfechos desfavoráveis, tais como: necessidade de internamento em unidade de terapia intensiva, hospitalização prolongada e mortalidade. Ainda não está claro na medicina se existe relação de causalidade direta ou se a hiponatremia é apenas um marcador de gravidade da doença de base. No entanto, sabe-se que o manejo inadequado de um paciente com hiponatremia pode causar graves danos neurológicos ou até mesmo a morte. No estudo, o total de 200 pacientes apresentam níveis normais de sódio do organismo e 99 apresentam níveis baixos.

Por fim, a creatinina fosfoquinase (CPK), indica o nível da enzima CPK no sangue. Quando um tecido muscular é danificado, como coração e cérebro, a CPK flui para o sangue. Portanto, altos níveis de CPK no sangue de um paciente podem indicar insuficiência cardíaca ou lesão . Os valores de referência da creatinofosfoquinase (CPK) são de 32 e 294 U/L para homens e 33 a 211 U/L para mulheres, mas podem variar dependendo do laboratório onde é realizado o exame. Considerando esta referência, 91 homens e 58 mulheres apresentam indicativo de tecido muscular danificado. De forma gráfica podemos observar os seguintes percentuais:

Qtdade de dados	299	
Característica	**Qtdade**	**Percentual**
Fumantes	96	32.11
Pressão Alta	105	35.12
Creatina maior 1.5	67	22.41
Anemia	125	41.81
Diabetes	129	43.14
FE <= 40	219	73.24
Plaquetas >400000 ou Plaquetas <150000	47	15.72
Sódio <136	99	33.11
CPK (33-211 M) (32-294 H)	150	50.17

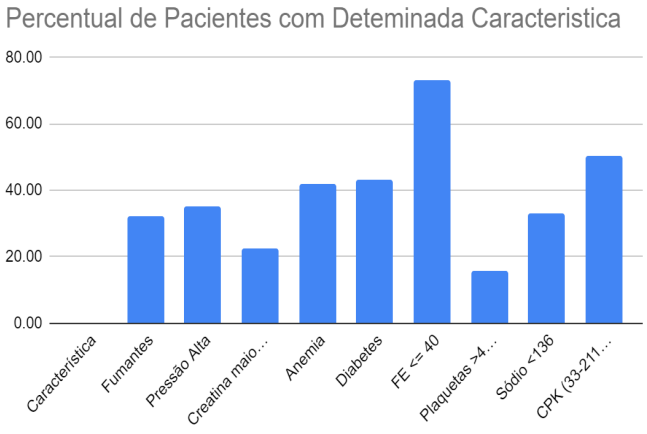

Gráfico 1 - Análise dos Dados

Para ter acesso à base de dados e a codificação do trabalho acesse: https://github.com/elianesomavilla/PCI510009---Machine-Learning-and-Deep-Learning

Machine Learning

Machine Learning, ou Aprendizado de Máquina, é uma especialidade da inteligência artificial. Por meio da aprendizagem de máquina, os computadores têm a capacidade de aprender de acordo com as respostas esperadas por meio de associações de diferentes dados. Para alcançar a aprendizagem, o computador é treinado com base em um conjunto de dados.

Neste estudo utilizamos diversos algoritmos de Machine Learning para analisar o resultado alcançado através de diferentes perspectivas. Fizemos uso da linguagem de programação Python e a principal biblioteca utilizada foi a sklearn, que oferece um conjunto de algoritmos de classificação e regressão. Antes de aplicar os algoritmos, realizamos o processamento inicial da base de dados.

Pré Processamento

Após análise inicial dos dados que compõem o dataframe, dividimos a base de dados composta por 299 registros em dois conjuntos, X e y, respectivamente. O conjunto X corresponde às variáveis independentes do dataframe, que são os dados dos exames de sangue e do ecocardiograma de cada paciente.O conjunto corresponde à variável dependente, y, nos indica se o paciente obteve óbito ou não considerando o resultado da combinação das variáveis independentes atribuídas a cada paciente.

Na sequência, usamos o recurso *SelectKBest* pertencente ao módulo *feature_selection* da biblioteca *sklearn*. Este recurso, remove todas as variáveis, exceto K características de maior pontuação, onde K é um parâmetro que indica a quantidade de variáveis que devem ser avaliadas pelo recurso utilizado.

O próximo passo foi a divisão dos conjuntos X e y no intuito de treinarmos adequadamente os algoritmos e realizar o aprendizado de máquina. O percentual de 80% dos dados foram usados para treinamento e 20% dos dados foram destinados para teste com o objetivo de verificar o comportamento de cada algoritmo ao receber dados novos e nos certificar de que o algoritmo realmente aprendeu com base nos dados usados durante o treino.

Após, utilizamos o recurso *StandardScaler* para normalizar os valores das variáveis do conjunto X que será usado para treinamento e teste. Este recurso transforma os dados de forma que a distribuição tenha um valor médio de 0 e desvio padrão de 1.

Por fim, executamos todo o processo de pré-processamento por 10 vezes e os algoritmos foram rodados com conjunto de dados cada vez menores considerando as K características mais importantes apontadas pelo *SelectKBest*. O resultado obtido em cada rodada foi registrado para comparação futura após aplicação de outros recursos similares, porém mais adequados a cada algoritmo.

Algoritmos de Regressão

A regressão é um dos métodos de previsão mais utilizados no meio estatístico. O principal objetivo é verificar como as variáveis de interesse influenciam uma variável resposta (y), além de criar um modelo matemático capaz de predizer valores de y com base em novos valores de variáveis preditoras X.

Regressão Linear Multivariada

A regressão linear multivariada é indicada quando o conjunto X contempla mais de uma variável preditora. No nosso estudo, inicialmente, são consideradas 10 variáveis no conjunto X.

Uma vez executado o pré-processamento, a eliminação recursiva de recursos (RFE) é utilizada para avaliação das melhores características especificamente para este algoritmo através da função RFE . Dado um estimador externo que atribui pesos aos recursos (por exemplo, os coeficientes de um modelo linear), o objetivo da eliminação recursiva de recursos (RFE) é selecionar variáveis considerando conjuntos cada vez menores. Primeiro, o estimador é treinado no conjunto inicial de recursos e a importância de cada recurso é obtida por meio de atributo específico qualquer (como *coef_ ou feature_importance*) . Em seguida, os recursos menos importantes são removidos do conjunto atual . Esse procedimento é repetido recursivamente no conjunto restante até que o número desejado de recursos a serem selecionados seja atingido.

Regressão Polinomial

.
A regressão polinomial faz uso de polinômios, que são expressões algébricas, usadas em casos onde as variáveis do modelo não seguem uma forma linear de dados. Neste caso, as variáveis independentes são expandidas pela quantidade de polinômios, conforme configurado. No presente estudo, os melhores resultados para este algoritmo foram observados usando o grau 3 de polinômios.

Após o pré processamento, utilizamos novamente a eliminação recursiva de recursos (RFE) para identificar a importância de cada variável do conjunto X para o modelo.

O resultado obtido foi expressado através da métrica *R2_score* que calcula o coeficiente de determinação. Esse coeficiente fornece uma métrica de quão bem as amostras futuras provavelmente

serão previstas pelo modelo. A melhor pontuação possível é 1.0 e pode ser negativa porque o modelo pode ser arbitrariamente pior. A mesma métrica foi utilizada com o algoritmo de Regressão Linear Multivariada para medir a acurácia do modelo durante a fase de teste.

No estudo todo este processo foi executado 10 vezes para cada um dos algoritmos de regressão. Cada rodada considerava apenas as melhores variáveis de acordo com o método de seleção de recursos empregado. O resultado obtido em cada rodada foi registrado para comparação futura após aplicação de outros recursos similares, porém mais adequados a cada algoritmo.

Algoritmos de Classificação

Os algoritmos de classificação tem como objetivo classificar itens ou amostras de acordo com as características observadas. Através de algoritmos que implementam este método, é possível ensinar o computador a realizar a classificação no intuito de identificar a qual categoria pertence uma determinada amostra de dados.

KNN (K-Nearest Neighbors)

O KNN (K-Nearest Neighbors) é um algoritmo de classificação e também de regressão cujo objetivo é determinar a qual grupo uma determinada amostra pertence com base na distância à K amostras vizinhas. No presente estudo, considerando todas as variáveis do conjunto independente, os melhores resultados para este algoritmo foram observados através da análise dos 6 vizinhos mais próximos, conforme tabela abaixo:

RESULTADOS DE TENTATIVAS - KNN				
3 Vizinhos	4 Vizinhos	5 Vizinhos	6 Vizinhos	7 Vizinhos
0.6	0.62	0.58	0.65	0.65

Tabela 01 - análise de número de vizinhos

Após o pré processamento, utilizamos a Seleção Sequencial Direta através da função *SequentialFeatureSelecto*r (SFS) pertencente ao módulo *feature_selection* da biblioteca *sklearn.* A SFS remove ou adiciona um recurso por vez com base no desempenho do classificador até que o subconjunto de recursos do tamanho K desejado seja alcançado.

A Seleção Sequencial Direta (SFS) difere da Eliminação de Recursive Feature (RFE) pois a RFE é computacionalmente menos complexa já que usa os coeficientes de peso do recurso (para modelos lineares) ou importância do recurso (algoritmos baseados em árvore) no intuito de eliminar recursos

recursivamente, enquanto o SFS objetiva eliminar ou adicionar recursos com base em uma métrica de desempenho de classificador / regressão.

RFC (Random Forest Classifier)

O RFC (Random Forest Classifier) é um algoritmo de classificação e também de regressão cujo objetivo é determinar a qual grupo uma determinada amostra pertence com base em árvores de decisão criadas no momento do treinamento. No presente estudo, considerando todas as variáveis do conjunto independente, os melhores resultados para este algoritmo foram observados através da criação de 7 árvores de decisão, conforme tabela abaixo:

RESULTADOS DE TENTATIVAS - RFC				
5 Árvores	6 Árvores	7 Árvores	8 Árvores	9 Árvores
0.62	0.6	0.65	0.58	0.63

Tabela 02 - análise de número de árvores de decisão

Após o pré processamento, utilizamos a função *feature_importances* disponível à todos os algoritmos ensembles, como o RFC. Este método retorna um score para cada variável. Quanto maior o score, maior é a importância dessa variável para o resultado da análise computacional.

SVM (Support Vector Machine)

O SVM (Support Vector Machine) é um algoritmo de classificação e também de regressão cujo objetivo é determinar a qual grupo uma determinada amostra pertence por meio de uma linha de separação entre duas classes. Esta linha comumente é chamada de Hiperplano.

Este algoritmo possui uma teoria um pouco mais robusta se comparada a outros algoritmos mais simples, como KNN. O SVM pode ser configurado para trabalhar com problemas lineares, usando o Kernel Linear, ou não lineares, usando o Kernel RBF (Radial Basis Function), que realiza um mapeamento para um espaço de maior dimensão.

Após o pré processamento dos dados, utilizamos a eliminação recursiva de recursos (RFE) para identificar a importância de cada variável do conjunto X para o modelo.

Nos primeiros testes realizados neste estudo, o kernel Linear atingiu 75% de acurácia considerando todo o conjunto de variáveis independentes inicial e apresentou acurácia similar conforme os recursos eram

eliminados de forma recursiva. O Kernel RBF atingiu o mesmo percentual a partir do momento em que o conjunto X passou a estar composto por 8 variáveis, conforme tabela abaixo:

RESULTADOS DE TENTATIVAS - SVM							
	4 Features	5 Features	6 Features	7 Features	8 Feacture	9 Feactures	10 Feactures
SVM - kernel RBF	0.75	0.75	0.75	0.72	0.75	0.68	0.72
SVM - kernel Linear	0.73	0.73	0.72	0.73	0.72	0.73	0.75

Tabela 03 - SVM - Kernel Linear e Kernel RBF

Optou-se por seguir o estudo com o Kernel Linear.

Logistic Regression

O Logistic Regression, ou Regressão Logística, é um algoritmo de classificação considerado o método estatístico mais utilizado para modelar variáveis categóricas. A Regressão Logística e a regressão linear são similares. Enquanto na Regressão Linear temos uma variável independente contínua, na Regressão Logística a variável resposta é binária..

Após o pré processamento dos dados, utilizamos a eliminação recursiva de recursos (RFE) para identificar a importância de cada variável do conjunto X para o modelo.

O resultado obtido para todos os algoritmos de classificação foi expressado através de uma matriz de confusão, que é uma tabela comparativa dos valores que cada algoritmo trouxe como predição em relação aos valores reais ocorridos. Ou seja, após treinar um modelo e aplicar as predições sobre o conjunto de dados separado para teste, o resultado obtido é expresso em uma coluna com as predições. As taxas computadas por uma matriz de confusão são:

- **Taxa de Verdadeiro Positivo (TVP)**: percentual que foi predito positivo sobre o total que de fato era positivo.
- **Taxa de Falso Positivo (TFP**): percentual que foi predito positivo sobre o total que de fato era negativo.
- **Taxa de Verdadeiro Negativo (TVN)**: percentual que foi predito negativo sobre o total que de fato era negativo.
- **Taxa de Falso Negativo (TFN)**: percentual que foi predito negativo sobre o total que de fato era positivo.

No estudo todo este processo foi executado 10 vezes para cada um dos algoritmos de classificação. Cada rodada considerava apenas as melhores variáveis de acordo com o método de seleção de recursos empregado. O resultado obtido em cada rodada foi registrado para comparação futura após aplicação de outros recursos similares, porém mais adequados a cada algoritmo.

Resultados e Discussão

Ao avaliar os resultados alcançados através do recurso *SelectKBest*, podemos observar que a melhor acurácia foi alcançada pelos algoritmos de classificação, considerando um conjunto mínimo de 3 a 4 variáveis para o modelo.

Dentre os 4 algoritmos avaliados nesta categoria, observamos que o KNN teve notória melhora na taxa de acurácia quando considerados apenas 3 variáveis no modelo, chegando a 77% de confiança. Os demais algoritmos de classificação apresentaram dados consistentes conforme ocorria a eliminação de variáveis do modelo, obtendo baixa variação na taxa de acurácia. Ao considerar apenas 2 variáveis, o modelo perdia eficácia para todos os algoritmos.

AVALIAÇÃO GERAL - ACURÁCIA USANDO O RECURSO "SelectBest"									
Método: método feature_selection do recurso SelectKBest da biblioteca sklearn									
Algoritmos	**10 Feactures**	**9 Feactures**	**8 Feactures**	**7 Feactures**	**6 Feactures**	**5 Feactures**	**4 Feactures**	**3 Feactures**	**2 Feactures**
KNN	0.65	0.65	0.66	0.66	0.7	0.7	0.65	0.77	0.7
SVM	0.75	0.73	0.72	0.73	0.72	0.73	0.73	0.72	0.63
Random Forest	0.65	0.73	0.68	0.66	0.7	0.7	0.66	0.72	0.66
Logistic Regression	0.72	0.73	0.72	0.72	0.7	0.72	0.73	0.73	0.65
Regressão Polinomial	0.94	0.32	0.75	0.64	0.49	0.45	0.39	0.33	0.14
Regressão Multivariada	0.15	0.16	0.17	0.17	0.17	0.17	0.17	0.16	0.06

Tabela 04 - Resultado com o recurso *SelectBest* para avaliação de variáveis

Notamos através deste recurso que os algoritmos de regressão não se mostraram boas opções para definir se um paciente virá a óbito pois, no caso da Regressão Multivariada a acurácia do modelo obteve um percentual baixo independentemente da quantidade de variáveis que compunham o modelo. O Algoritmo de Regressão Polinomial obteve acurácia superior a 90% ao considerar as 10 variáveis no modelo, todavia, ao eliminar variáveis a acurácia diminuía e não verificamos linearidade no percentual obtido como resultado. Dessa forma, considera-se que este algoritmo não apresentou consistência da mesma forma que observamos nos algoritmos de classificação. Especificamente para a Regressão Polinomial poderia ser avaliado se o modelo apresentou *overfitting*, que ocorre quando o modelo se ajusta bem ao conjunto de dados de treinamento, mas se mostra ineficaz para prever novos resultados.

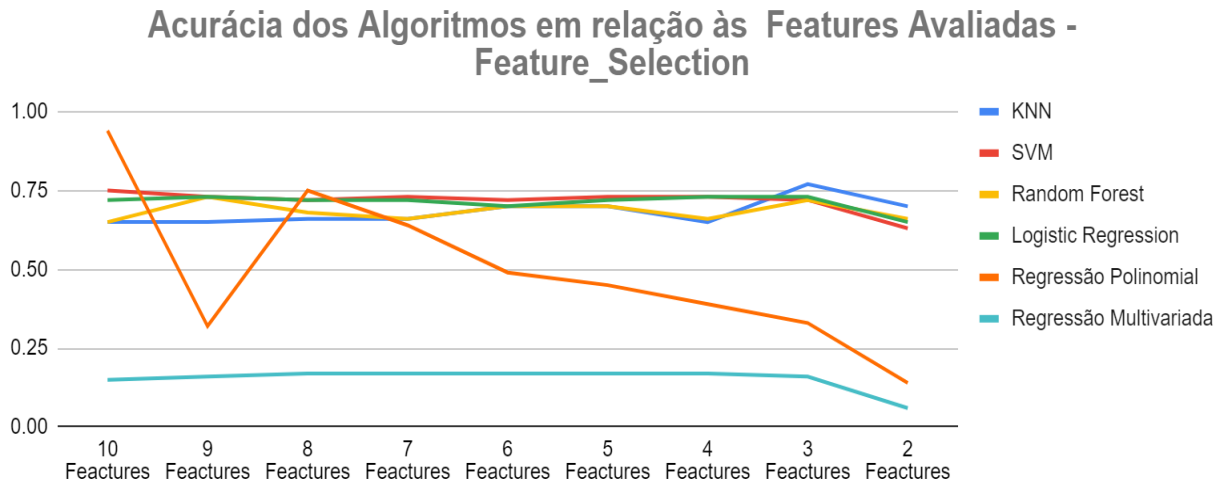

Gráfico 2 - Acurácia alcançada através do recurso *SelectBest* para avaliação de variáveis

Avaliação dos Algoritmos com Outros Recursos

Ao avaliar os resultados alcançados através de recursos específicos usados para identificar as melhores variáveis de cada algoritmo utilizado no estudo, podemos observar que a melhor acurácia foi alcançada por cada algoritmo com uma quantidade distinta de variáveis mínimas.

AVALIAÇÃO GERAL - ACURÁCIA USANDO RECURSOS DISTINTOS PARA CADA ALGORITMO									
Algoritmos	**10 Feactures**	**9 Feactures**	**8 Feactures**	**7 Feactures**	**6 Feactures**	**5 Feactures**	**4 Feactures**	**3 Feactures**	**2 Feactures**
KNN	0.65	0.62	0.65	0.65	0.65	0.65	0.65	0.68	0.7
SVM	0.75	0.75	0.75	0.73	0.75	0.72	0.72	0.72	0.63
Random Forest	0.65	0.68	0.7	0.66	0.7	0.68	0.75	0.66	0.56
Logistic Regression	0.72	0.72	0.72	0.73	0.73	0.75	0.75	0.73	0.68
Regressão Polinomial	0.94	0.72	0.72	0.69	0.57	-0.66	0.34	0.33	0.16
Regressão Multivariada	0.1564	0.1548	0.1552	0.1511	0.1655	0.1667	0.1522	0.1598	-0.0077

Tabela 05 - Resultado com diversos recursos para avaliação de variáveis

Observamos que:

- O algoritmo **KNN,** que utilizou o recurso *feature_selection* da função *SequentialFeatureSelector* da biblioteca *mlxtend*, obteve uma melhor acurácia ao ser utilizado com o mínimo de **2 variáveis**. Todavia, o algoritmo manteve baixa variação, independente de utilizar as 10 variáveis do modelo ou apenas o conjunto reduzido de 2 variáveis.
- O **SVM,** que utilizou o recurso *feature_selection* da função RFE da biblioteca *sklearn*, obteve o melhor resultado com o mínimo de **6 variáveis**. Todavia, o algoritmo manteve baixa variação, independente de utilizar as 10 variáveis do modelo ou apenas o conjunto reduzido de 3 variáveis. Sua acurácia caiu consideravelmente ao serem avaliadas apenas 2 variáveis.
- O **Random Forest Classifier,** que utilizou o recurso *feature_importances* (métodos próprio de algoritmos assembly), obteve o melhor resultado com o mínimo de **4 variáveis**. Observa- se que conforme as variáveis eram eliminadas do modelo, a acurácia aumentava, todavia, houve redução na taxa da acurácia ao avaliar o algoritmo com um conjunto reduzido de 7 ou 3 variáveis. Ao considerar um conjunto de apenas 2 variáveis a acurácia cai consideravelmente.
- O **Logistic Regression,** que utilizou o recurso *feature_selection* da função RFE da biblioteca *sklearn,* obteve uma melhor acurácia ao ser utilizado com o mínimo de **4 variáveis**. Todavia, o algoritmo manteve baixa variação, independente de utilizar as 10 variáveis do modelo ou apenas o conjunto reduzido de 3 variáveis. Sua acurácia caiu consideravelmente ao serem avaliadas apenas 2 variáveis.
- A **Regressão Polinomia**l, que utilizou o recurso *feature_selection* da função RFE da biblioteca *sklearn,* obteve o melhor resultado quando avaliado o conjunto total de 10 variáveis. Conforme as variáveis eram eliminadas do modelo a acurácia caia consideravelmente chegando a apresentar valores negativos
- A **Regressão Multivariada,** que utilizou o recurso *feature_selection* da função RFE da biblioteca *sklearn,* não apresentou acurácia superior a 16,5% em nenhum momento, chegando a apresentar valores negativos.

De forma Gráfica observamos claramente as variações:

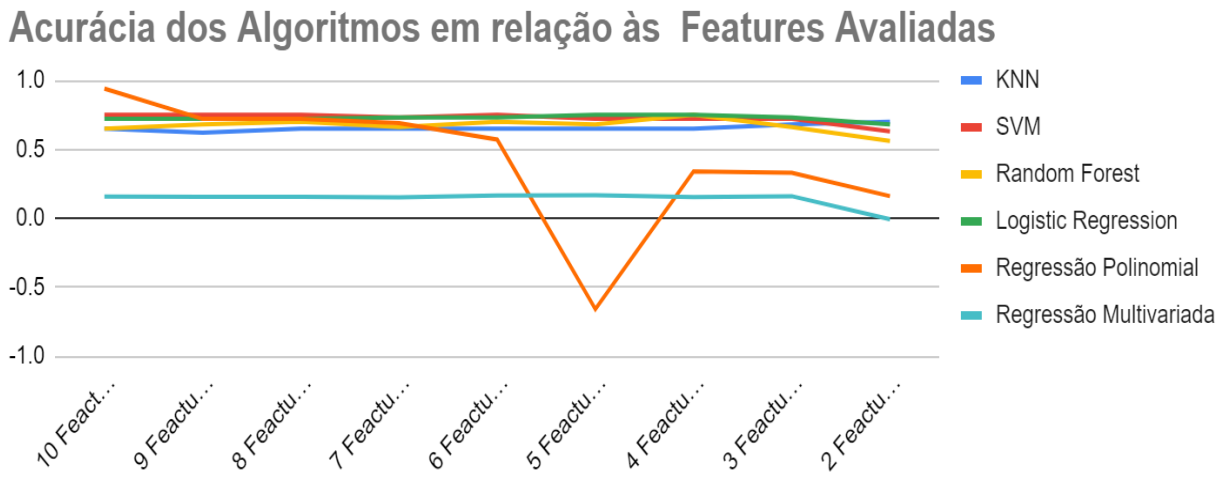

Gráfico 3 - Acurácia alcançada através de diversos recursos de avaliação de variáveis

Todos os algoritmos utilizados para classificação da amostra de dados obtiveram resultados consistentes e lineares. Além disso, confirmou-se que os algoritmos de Regressão Polinomial e Regressão Multivariada não apresentam resultados eficientes para predizer sobre o possível óbito de um paciente.

As variáveis que foram eliminadas do modelo de acordo com o recurso usado foram registradas em tabelas que podem ser visualizadas nos anexos deste estudo. De forma forma geral, a idade, a fração de ejeção (FE), que pode ser obtida através de ecocardiograma, e a creatinina sérica, que pode ser obtida através de exame sanguíneo, se mostraram as principais variáveis para prever a taxa de mortalidade através de algoritmos de classificação.

Os dados referentes à presença de anemia e diabetes, quantidade de plaquetas no sangue e se o paciente é fumante ou não foram as variáveis que menos influenciaram na maioria dos resultados obtidos. As variáveis relativas à pressão alta, sódio e creatinina fosfoquinase (CPK) se mostraram importantes contribuintes mas não determinantes sobre a possibilidade de sobrevivência de um paciente. De forma gráfica, podemos observar a importância das variáveis para o modelo através da média de vezes que cada uma compôs o conjunto reduzido de variáveis utilizadas nos algoritmos estudados.

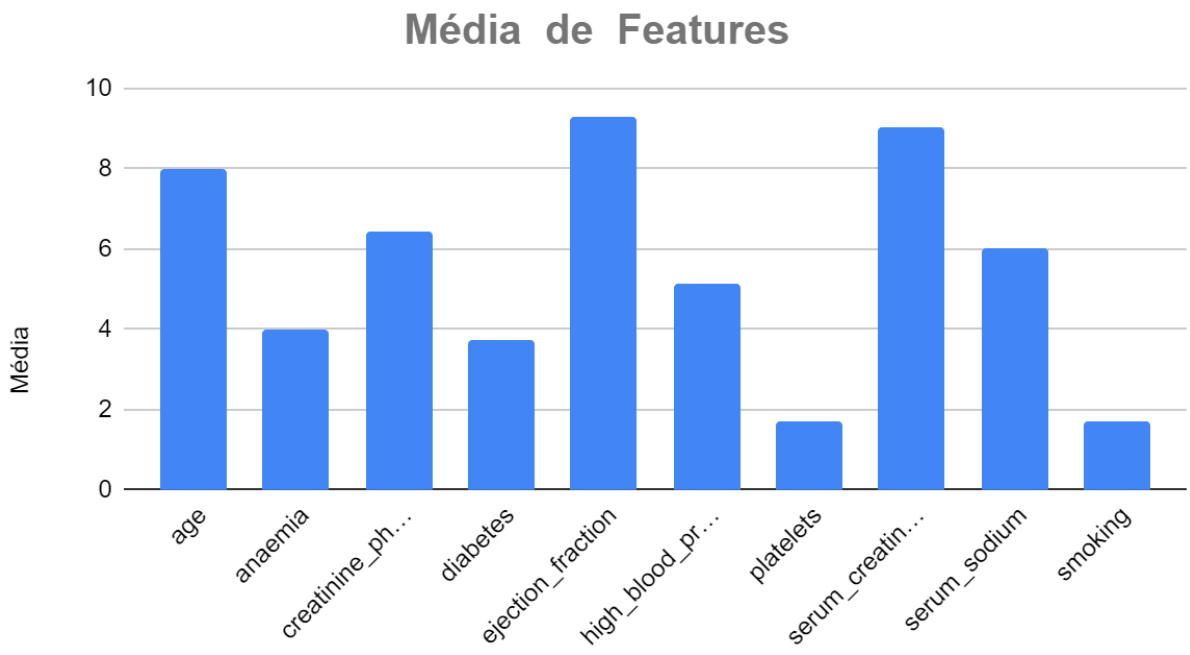

Gráfico 4 - Média de importância das variáveis no modelo

Desta forma, podemos concluir que, através de um conjunto reduzido de dados composto por idade, fração de ejeção (FE) e creatinina sérica, é possível predizer sobre a probabilidade de óbito de pacientes

considerando uma taxa de acurácia superior a 70% para todos os algoritmos de classificação utilizados neste estudo.

Considerando o cenário das clínicas e hospitais que muitas vezes não contam com todos os recursos necessários para avaliar o quadro de um paciente com insuficiência cardíaca, este estudo direciona os profissionais da saúde a priorizar os dados que apresentam maior relevância na probabilidade de sobrevivência de um paciente.

Discusão

Considerando os resultados obtidos, poderia ser desenvolvida uma pontuação para classificar o risco do paciente vir à óbito, de acordo com os dados obtidos para as variáveis estudadas. Variáveis que não tiveram grande influência no modelo, como anemia, diabetes, plaquetas e se o paciente é fumante ou não, poderiam ter um menor peso. Por outro lado, Fração de Ejeção (FE), idade e creatinina sérica teriam um peso maior. As variáveis relativas à pressão alta, sódio e creatinina fosfoquinase (CPK), poderiam ter um peso intermediário. Desta forma, os profissionais da saúde poderiam contar com uma fórmula para prever a taxa de sobrevida dos pacientes com insuficiência cardíaca.

Durante as pesquisas foram encontrados estudos sobre a influência da qualidade de vida no sucesso de tratamento para pacientes com insuficiência cardíaca. De acordo com o Hospital Israelita Albert Einstein, o sedentarismo é uma das causas de doenças cardiovasculares que podem ser prevenidas com exercícios físicos.

A qualidade de vida também é influenciada pela cultura, renda, acesso à informação e localização que cada pessoa vive, visto que estas variáveis podem limitar o acesso à atividades que ajudam na prevenção de doenças cardiovasculares. Caso houvesse uma base de dados com informações sobre a qualidade de vida dos pacientes, poderíamos analisar os resultados do estudo considerando estas realidade.

Anexo 1 - Tabela de Variáveis - SelectKBest

AVALIAÇÃO GERAL										
Método: feature_selection da função SelectKBest da biblioteca sklearn										
Qtdade de features avaliadas	**Feature 1**	**Feature 2**	**Feature 3**	**Feature 4**	**Feature 5**	**Feature 6**	**Feature 7**	**Feature 8**	**Feature 9**	**Feature 10**
10	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	smoking
9	age	anaemia	creatinine_phosphokinase		ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	smoking
8	age	anaemia	creatinine_phosphokinase		ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	
7	age	anaemia	creatinine_phosphokinase		ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
6	age	anaemia			ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
5	age				ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
4	age				ejection_fraction			serum_creatinine	serum_sodium	
3	age				ejection_fraction			serum_creatinine		
2					ejection_fraction			serum_creatinine		
1								serum_creatinine		

Anexo 2 - Tabela de Variáveis - Algoritmos de Classificação

AVALIAÇÃO - KNN										
Método: feature_selection da função SequentialFeatureSelector da biblioteca mlxtend										
Qtdade de features avaliadas	**Feature 1**	**Feature 2**	**Feature 3**	**Feature 4**	**Feature 5**	**Feature 6**	**Feature 7**	**Feature 8**	**Feature 9**	**Feature 10**
10	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	smoking
9	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	
8	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction		platelets	serum_creatinine	serum_sodium	
7	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction			serum_creatinine	serum_sodium	
6	age	anaemia		diabetes	ejection_fraction			serum_creatinine	serum_sodium	
5	age	anaemia			ejection_fraction			serum_creatinine	serum_sodium	
4		anaemia			ejection_fraction			serum_creatinine	serum_sodium	
3					ejection_fraction			serum_creatinine	serum_sodium	
2					ejection_fraction			serum_creatinine		
1								serum_creatinine		

AVALIAÇÃO - SVM										
Método: feature_selection da função RFE da biblioteca sklearn										
Qtdade de features avaliadas	**Feature 1**	**Feature 2**	**Feature 3**	**Feature 4**	**Feature 5**	**Feature 6**	**Feature 7**	**Feature 8**	**Feature 9**	**Feature 10**
10	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	smoking

9	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	smoking
8	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
7	age		creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
6	age		creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine		
5	age		creatinine_phosphokinase		ejection_fraction	high_blood_pressure		serum_creatinine		
4	age		creatinine_phosphokinase		ejection_fraction			serum_creatinine		
3	age				ejection_fraction			serum_creatinine		
2					ejection_fraction			serum_creatinine		
1								serum_creatinine		

AVALIAÇÃO - LOGISTIC REGRESSION										
Método: feature_selection da função RFE da biblioteca sklearn										
Qtdade de features avaliadas	**Feature 1**	**Feature 2**	**Feature 3**	**Feature 4**	**Feature 5**	**Feature 6**	**Feature 7**	**Feature 8**	**Feature 9**	**Feature 10**
10	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	smoking
9	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	smoking
8	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
7	age		creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
6	age		creatinine_phosphokinase		ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
5	age		creatinine_phosphokinase		ejection_fraction	high_blood_pressure		serum_creatinine		
4	age				ejection_fraction	high_blood_pressure		serum_creatinine		
3	age				ejection_fraction			serum_creatinine		

2	age				ejection_fraction					
1					ejection_fraction					

AVALIAÇÃO - RANDOM FOREST CLASSIFIER										
Método: feature_importances (métodos próprio de algoritmos assembly)										
Qtdade de features avaliadas	**Feature 1**	**Feature 2**	**Feature 3**	**Feature 4**	**Feature 5**	**Feature 6**	**Feature 7**	**Feature 8**	**Feature 9**	**Feature 10**
10	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	smoking
9	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	
8	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
7	age	anaemia	creatinine_phosphokinase		ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
6	age		creatinine_phosphokinase		ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
5	age		creatinine_phosphokinase		ejection_fraction			serum_creatinine	serum_sodium	
4	age		creatinine_phosphokinase		ejection_fraction			serum_creatinine		
3			creatinine_phosphokinase		ejection_fraction			serum_creatinine		
2			creatinine_phosphokinase							
1			creatinine_phosphokinase							

Anexo 3 - Tabela de Variáveis - Algoritmos de Regressão

AVALIAÇÃO - REGRESSÃO POLINOMIAL										
Método: feature_selection da função RFE da biblioteca sklearn										
Qtdade de features avaliadas	**Feature 1**	**Feature 2**	**Feature 3**	**Feature 4**	**Feature 5**	**Feature 6**	**Feature 7**	**Feature 8**	**Feature 9**	**Feature 10**
10	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	smoking
9	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	smoking
8	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
7	age		creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
6	age		creatinine_phosphokinase		ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
5	age		creatinine_phosphokinase		ejection_fraction			serum_creatinine	serum_sodium	
4	age		creatinine_phosphokinase		ejection_fraction			serum_creatinine		
3	age				ejection_fraction			serum_creatinine		
2	age				ejection_fraction					
1					ejection_fraction					

AVALIAÇÃO REGRESSÃO LINEAR MULTIVARIADA										
Método: feature_selection da função RFE da biblioteca sklearn										
Qtdade de features avaliadas	**Feature 1**	**Feature 2**	**Feature 3**	**Feature 4**	**Feature 5**	**Feature 6**	**Feature 7**	**Feature 8**	**Feature 9**	**Feature 10**
10	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	smoking
9	age	anaemia	creatinine_phosp	diabetes	ejection_fracti	high_blood		serum_cr	serum_s	smoking

			hokinase		on	_pressure		eatinine	odium	
8	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
7	age		creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
6	age		creatinine_phosphokinase		ejection_fraction	high_blood_pressure		serum_creatinine	serum_sodium	
5	age		creatinine_phosphokinase		ejection_fraction			serum_creatinine	serum_sodium	
4	age		creatinine_phosphokinase		ejection_fraction			serum_creatinine		
3	age				ejection_fraction			serum_creatinine		
2	age				ejection_fraction					
1					ejection_fraction					

Anexo 4 - Tabela Quantitativa de Variáveis

A tabela abaixo apresenta quantas vezes cada variável foi apresentada em cada um dos algoritmos usados no estudo de acordo com o recurso de avaliação de *features* utilizado.

Para o KNN foi utilizado o recurso *feature_selection* da função *SequentialFeatureSelector* da biblioteca *mlxtend*, Para o Random Forest foi utilizado o recurso *feature_importances* (métodos próprio de algoritmos assembly) e os demais algoritmos foram avaliados com o recurso *feature_selection* da função RFE da biblioteca *sklearn*. Todos também foram avaliados com o recurso *BestSelect* da biblioteca *sklearn*. Por fim, calculou-se a média de apresentação de cada variável no estudo.

Features	KNN	SVM	Random Forest	Logistic Regression	Regressão Multivariada	Regressão Polinomial	Todos - BestSelect	Média
age	6	8	7	9	9	9	8	8
anaemia	7	3	4	3	3	3	5	4
creatinine_phosphokinase	4	7	10	6	7	7	4	6
diabetes	5	5	3	4	4	4	1	4
ejection_fraction	9	9	8	10	10	10	9	9
high_blood_pressure	2	6	5	7	5	5	6	5
platelets	3	1	2	1	1	1	3	2
serum_creatinine	10	10	9	8	8	8	10	9
serum_sodium	8	4	6	5	6	6	7	6
smoking	1	2	1	2	2	2	2	2

Referências

https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5

https://www.einstein.br/especialidades/cardiologia/doencas-sintomas/insuficiencia-cardiaca

https://www.scielo.br/j/ramb/a/npPgV7NzQ99tWdQhvZJvCfs/?lang=pt

https://www.tuasaude.com/plaquetas/

https://minerandodados.com.br/aprenda-como-selecionar-features-para-seu-modelo-de-machine-learning/

http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/

https://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_digits.html#sphx-glr-auto-examples-feature-selection-plot-rfe-digits-py

https://machinelearningmastery.com/calculate-feature-importance-with-python/

https://www.scielo.br/j/jbn/a/ggcdv7X6mjHSyVRtcY8fTxS/?lang=pt&format=pdf

https://www.healthline.com/health/sodium-urine#purpose

https://www.tuasaude.com/exame-cpk/

https://scikit-learn.org/

https://www.cin.ufpe.br/~tg/2010-2/gmoj.pdf

https://cienciaenegocios.com/o-que-e-a-matriz-de-confusao/

https://teses.usp.br/teses/disponiveis/45/45133/tde-25062007-163150/publico/dissertacao_4.pdf

https://inferir.com.br/artigos/algoritimo-knn-para-classificacao/