UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
MESTRADO PROFISSIONAL EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

COMPARAÇÃO DE ALGORITMOS DE APRENDIZADO DE MAQUINAS PARA DESENVOLVIMENTO DE UM SISTEMA PARA TRIAGEM DE ADOLESCENTES OBESOS UTILIZANDO VARIÁVEIS CLÍNICAS

DANILO JOSÉ DOS SANTOS COSTA

SÃO LUÍS - MA
2023

UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
MESTRADO PROFISSIONAL EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

DANILO JOSÉ DOS SANTOS COSTA

COMPARAÇÃO DE ALGORITMOS DE APRENDIZADO DE MAQUINAS PARA DESENVOLVIMENTO DE UM SISTEMA PARA TRIAGEM DE ADOLESCENTES OBESOS UTILIZANDO VARIÁVEIS CLÍNICAS

Trabalho apresentado ao curso de Mestrado Profissional em Engenharia da Computação e Sistemas na Universidade Estadual do Maranhão como pré-requisito para obtenção do título de Mestre sob orientação do Prof. Ewaldo Eder Carvalho Santana.

SÃO LUÍS - MA
2023

Costa, Danilo Jose dos Santos.

Comparação de algoritmos de aprendizado de máquinas para desenvolvimento de um sistema para triagem de adolescentes obesos utilizando variáveis clínicas. / Danilo Jose dos Santos Costa. – São Luís, 2024.

59 f.

Dissertação (Mestrado Profissional em Engenharia de Computação e Sistemas) - Universidade Estadual do Maranhão, São Luís, 2023.

Orientador: Prof. Dr. Ewaldo Eder Carvalho Santana.

1. Gordura Corporal. 2. Saúde. 3. Aprendizado de máquina. 4. Adolescentes. I. Título.

CDU: 519.6:616-008.847.9

Elaborado por Francisca Elany R. Sousa Lopes - CRB 13/754

DANILO JOSÉ DOS SANTOS COSTA

COMPARAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA DESENVOLVIMENTO DE UM SISTEMA PARA TRIAGEM DE ADOLESCENTES OBESOS UTILIZANDO VARIÁVEIS CLÍNICAS

Trabalho apresentado ao curso de Mestrado Profissional em Engenharia da Computação e Sistemas na Universidade Estadual do Maranhão como pré-requisito para obtenção do título de Mestre sob orientação do Prof. Dr. Ewaldo Eder Carvalho Santana.

Aprovado em: 10 de novembro de 2023.

Documento assinado digitalmente
gov.br
CARLOS MAGNO SOUSA JUNIOR
Data: 05/02/2024 10:53:52-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Carlos Magno Sousa Júnior
Membro

Prof. Dr. Mauro Sérgio Silva Pinto
Membro

Documento assinado digitalmente
gov.br
EWALDO EDER CARVALHO SANTANA
Data: 05/02/2024 10:45:16-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Ewaldo Eder Carvalho Santana
Orientador

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus por me dar forças e sabedoria para enfrentar os desafios durante toda a trajetória do mestrado. Sua presença em minha vida foi fundamental para alcançar esta etapa.

A minha família, meu porto seguro, que me apoiou desde o início da minha jornada acadêmica, sou grato por todo amor, carinho e paciência dedicados a mim, especialmente nos momentos mais difíceis.

Ao meu orientador, agradeço por compartilhar seu conhecimento, orientação e paciência ao longo do meu trabalho de pesquisa. Seus conselhos e feedbacks foram de grande valia para o desenvolvimento deste trabalho.

Agradeço também aos amigos e colegas que acompanharam minha trajetória, me dando força e estímulo quando precisei, e a todas as pessoas que contribuíram de alguma forma para a realização deste sonho.

Por fim, expresso minha profunda gratidão à instituição de ensino pela oportunidade de realizar este mestrado e pelo suporte acadêmico que recebi. Espero ter contribuído de alguma forma para o avanço do conhecimento em minha área de pesquisa.

RESUMO

Nas últimas décadas muitos países têm entrado em processo de desenvolvimento de forma geral, devido ao mundo globalizado que exige alta dinamicidade se teve um aumento considerável no índice de má alimentação do ser humano, fazendo com que se tenha uma rápida transição nutricional e epidemiológica, resultando em inúmeros indivíduos com excesso de gordura corporal ainda na adolescência. A alta prevalência de excesso de peso na fase da adolescência tem se tornado um grande problema na saúde do ser humano em geral, a adolescência é caracterizada como a fase onde o corpo humano mais se desenvolve e está associada diretamente a uma gama de doenças que podem colocar em risco a saúde do indivíduo com excesso de peso, sendo assim o presente estudo tem o objetivo de estimar o percentual de gordura corporal em adolescentes, para se pode alcançar este objetivo foram selecionados algoritmos de aprendizado de máquina que foram comparados, para que no fim se pudesse selecionar o que tem melhor desempenho e poder entregar um resultado satisfatório de maneira geral, a base de dados utilizada para aplicação dos algoritmos é constituída por dados coletados de estudantes de escola pública de ensino na cidade de São Luís do Maranhão no ano de 2018, a base contém 772 entradas de ambos os gêneros com idade de 10 até 19 anos, com estes dados foi possível avaliar indicadores como: idade, gênero, massa corporal, estatura, circunferência da cintura, circunferência do quadril, circunferência da panturrilha e a circunferência braço, assim como o percentual de gordura corporal adquirido através da bioimpedância. Após a aplicação dos algoritmos o K-SVM teve o melhor desempenho, mostrando potencial para ser usado em uma aplicação futura.

Palavras Chave: Gordura Corporal, Saúde, Aprendizado de máquina, Adolescentes.

ABSTRACT

In recent decades, many countries have entered into a development process in general, due to the globalized world that requires high dynamics, there has been a specific increase in the rate of poor human nutrition, causing a rapid nutritional and epidemiological transition, resulting in in considerable individuals with excess body fat even in adolescence. The high prevalence of excess weight during adolescence has become a major problem in human health in general. Adolescence is characterized as a phase in which the human body develops the most and is directly associated with a range of diseases that can put the health of an overweight individual at risk, so the present study aims to estimate the percentage of body fat in adolescents. In order to achieve this objective, machine learning algorithms were selected that were compared, to Even if you could select the one that performs best and can deliver an overwhelming result in general, the database used to apply the algorithms is obtained from data collected from public school students in the city of São Luís do Maranhão. In 2018, the database contains 772 entries of both genders aged 10 to 19 years. With this data it was possible to evaluate indicators such as: age, gender, body mass, height, waist characteristics, hip characteristics, body characteristics calf and arm functionality, as well as the percentage of body fat acquired through bioimpedance. After applying the algorithms, K-SVM had the best performance, showing potential to be used in a future application.

***Keywords**: Body Fat, Health, Machine Learning, Adolescents.*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PNSN	Pesquisa Nacional de Saúde e Nutrição
IMC	Índice de massa corporal
NHANES	*Second National Health and Nutrition Examination Survey*
DEXA	*Dual X - ray absorptiometry*
KNN	*K-Nearest Neighbors*
SVM	*Support Vector Machines*
L-SVM	*Linear Support Vector Machines*
K-SVM	*Kernel Support Vector Machine*
OMS	Organização mundial de saúde
RCE	Relação cintura estatura
BIA	Análise da impedância bioelétrica (Bioimpedância)
CC	Circunferência da Cintura
CB	Circunferência do Braço
CP	Circunferência do Pescoço
Cpant	Circunferência da Panturrilha
CQ	Circunferência do Quadril

SUMÁRIO

1. INTRODUÇÃO

No ano de 2015, a população somente de adolescentes no mundo foi de 1,2 bilhões representando aproximadamente 16% da população mundial, cerca de 90% da população mundial de adolescentes vivem em países de baixa renda e pouco desenvolvidos, onde são normalmente encontradas grandes barreiras para se manter a educação alimentar plena (JÚNIOR, 2019), as diferenças que ocorrem no estado nutricional de um indivíduo podem ser decorrentes de fatores do meio ambiente, genéticos ou uma combinação de ambos, também existe uma correlação entre o excesso de peso dos pais com o excesso de peso dos filhos, mostrando características de hereditariedade em problemas de obesidade. A PNSN (Pesquisa Nacional de Saúde e Nutrição), de 1989, no Brasil, indicou que adolescentes do sexo feminino, moradoras na região Sul do País e com renda familiar per capta acima de 2,2 salários mínimos, tem o índice de massa corporal (IMC) superior ao IMC de adolescentes americanas avaliadas pelo NHANES (Second National Health and Nutrition Examination Survey) (FONSECA; SICHIER; VEIGA, 1988), mostrando que no Brasil se tem grandes números se tratando de obesidade.

O período que corresponde a adolescência vai dos 10 aos 19 anos, a adolescência é caracterizada por um período onde ocorrem uma série de mudanças no corpo do indivíduo, dentre elas podem-se destacar o crescimento, que pode acarretar um acúmulo da distribuição de gordura corporal, fato esse que pode propiciar o desenvolvimento da obesidade (JÚNIOR, 2019). Segundo (WANDERLEY; FERREIRA, 2010) obesidade é uma enfermidade causada por níveis excessivos de gordura corporal, podendo afetar a saúde do indivíduo em questão, e segundo a organização mundial de saúde (OMS) a obesidade é um mal que tem afetado uma abundância de indivíduos no mundo e está relacionada a hipertensão arterial, doenças cardiovasculares e várias outras (JÚNIOR, 2019). As diferenças sociais existentes no Brasil podem ser observadas em números relacionados à obesidade, a localização de um indivíduo pode retratar seu excesso de peso ou até mesmo a sua desnutrição, geralmente regiões com maior índice de desenvolvimento apresentam também maiores números de indivíduos com obesidade, sendo assim a obesidade é considerada uma epidemia que atinge o mundo todo não importando idade, classe, gênero, etnia ou religião. Devido ao seu longo alcance essa enfermidade se tornou

um problema de saúde pública, biologicamente a adolescência é a fase de maior velocidade quando se trata de crescimento de um indivíduo (WANDERLEY; FERREIRA, 2010), sendo assim é a fase da vida humana que merece a atenção especial quando se trata de obesidade, portanto, indivíduos que estão na adolescência são alvos de estudos que podem ajudar a combater de forma mais eficiente a obesidade.

A análise da composição corporal é um importante ponto que tem como destaque a medição do percentual de gordura corporal (REZENDE et al., 2010), em níveis elevados a gordura corporal pode resultar em resistência à insulina, trazendo grandes transtornos para a pessoa em questão, portanto, o acompanhamento deste índice é de extrema importância para todo e qualquer indivíduo, pois, um descontrole no índice de gordura corporal pode acarretar danos graves à saúde do ser humano (JÚNIOR, 2019), atualmente existem várias técnicas criadas para que possa avaliar a composição corporal dentre elas se destacam a DEXA (dual X — ray absorptiometry) e a bioimpedância elétrica, onde o DEXA é um procedimento que possui alta precisão e tem baixa exigência em termos de níveis de radiação (ZEBALLOS et al., 2021), e a bioimpedância elétrica é um exame que faz a avaliação da composição corporal utilizando correntes elétricas de baixa intensidade através do corpo (BRITTO; MESQUITA, 2008), dentre estes métodos o DEXA é o mais utilizado, porém, existem alguns aspectos negativos em sua execução, como, por exemplo, o alto custo financeiro, a necessidade de radiação e a necessidade de mão obra especializada são alguns deles (JÚNIOR, 2019).

Um indicador antropométrico de alta viabilidade devido ao seu baixo custo e simplicidade é o IMC, porém, a procedência deste método causas diversas controvérsias sobre sua eficiência, sendo que este método não diferencia tecido adiposo de massa magra, dessa forma se corre grande risco de que indivíduos com alto nível de massa magra possam ser classificados como obesos, neste contexto é de extrema importância o desenvolvimento de métodos que auxiliam na avaliação nutricional de indivíduos no estado de adolescência (JÚNIOR, 2019), como alternativa o aprendizado de máquina possui diversas técnicas que podem trazer resultados ótimos e oferecer métodos de diagnósticos mais rápidos e não invasivos.

Neste trabalho são aplicados os métodos de aprendizado de máquina: *Nayve Bayes*, KNN, regressão logística, L-SVM, arvore de decisão e K-SVM, para que se possa fazer uma comparação entre eles e encontrar o modelo que apresenta melhor resultado para que se possa medir a gordura corporal em adolescentes, visando entregar uma solução computacional que auxilie no combate a obesidade.

1.1. Objetivos

1.1.1. Objetivo Geral

Comparar diversos algoritmos de aprendizado de máquina, a fim de identificar o que tem melhor performance para estimativa de gordura corporal em adolescentes com idade entre 10 a 19 anos.

1.1.2. Objetivos Específicos

- Comparar vários algoritmos de aprendizado de máquina buscando averiguar qual tem melhor desempenho para fins de predição de índice de gordura corporal.
- Analisar o desempenho do algoritmo que tiver melhor desempenho.
- Criar uma ferramenta computacional que não recorra a procedimentos invasivos.

1.2. Metodologia

A aprendizagem de máquina é uma técnica que permite que computadores aprendam a partir de dados sem serem explicitamente programados. A utilização de algoritmos de aprendizagem de máquina tem se tornado cada vez mais popular na área da saúde, especialmente no que se refere ao estudo da gordura corporal. A gordura corporal é uma medida importante de saúde e é comumente avaliada por meio de técnicas como a bioimpedância elétrica, a densitometria e a tomografia computadorizada. No entanto, esses métodos são caros e invasivos, o que dificulta sua utilização em larga escala. A aprendizagem de máquina pode ser uma alternativa viável para avaliar a gordura corporal de forma não invasiva e de baixo custo.

Para realizar o trabalho aqui proposto, foram feitas as etapas de seleção de atributos de entrada do modelo, estudo da base de dados, definição de tecnologias, pré-processamento e avaliação do método, estas etapas estão detalhadas abaixo.

A escolha dos algoritmos depende da complexidade do problema, Por exemplo. Árvores binárias são escolhidas pela interpretabilidade; a regressão logística para problemas de classificação binária simples; o KNN se destaca em padrões complexos; enquanto o k-SVM é preferido para dados de alta dimensão e problemas complexos de classificação ou regressão não lineares.

1.2.1. Seleção de Atributos de Entrada do Modelo

Todos os atributos escolhidos para a base de dados do estudo em questão estão descritos na literatura e são usados em diversas outras aplicações, se tratando de avaliação do Estado nutricional e da saúde dos adolescentes são importantes dados como idade massa corporal estatura e gênero, estes dados são fortemente indicados pela OMS, se tratando de indicadores antropométricos a circunferência da cintura é um índice bastante usado na análise da gordura corporal e também do risco cardiovascular, as circunferências do braço quadril e panturrilha são largamente utilizadas para avaliação da perda de massa magra, portanto, todos os atributos usados neste trabalho são largamente utilizados em outros trabalhos têm diversas aplicações e trazendo resultados promissores (JÚNIOR, 2019).

1.2.2. Base de Dados

A base de dados é constituída por 772 linhas, com nove atributos, a coleta dos dados foi concretizada através de alunos da rede pública na cidade de São Luís do Maranhão, com idade de 10 a 19 anos, onde estes foram escolhidos de forma não probabilística no ano de 2018, para o cálculo amostral pela estimação de proporção que teve como base a prevalência de excesso de peso em adolescentes de 20,5%, prevalecia sugerida do desfecho de 26,9%, erro tolerável de 5% (erro tipo I) e poder do teste de 90% (erro tipo II), com adição de 10% para possíveis perdas ou recusas. Chegando a uma amostra mínima de 513 adolescentes (JÚNIOR, 2019).

No trabalho de (JÚNIOR, 2019) durante o processo de coleta de dados foram usadas algumas regras de exclusão de indivíduos que apresentavam os seguintes quadros:

- Adolescentes gestantes, amamentando ou que fazem uso de anticoncepcional;
- Uso de medicamentos que alteram o estado nutricional;
- Incapacidade física que impossibilita ou compromete as medidas antropométricas.
- Não concordância dos responsáveis ou participantes;
- Ausência na coleta de dados.

Cada medida obtida foi colhida por um único pesquisador com o mesmo instrumento calibrado, as medidas foram colhidas duas vezes para que se pudesse tirar uma média e inseri-las na base de dados, para o percentual de gordura corporal foi usada a bioimpedância tetrapolar, seguindo a tabela abaixo:

Tabela 1 - Valores de referência para classificação do %GC

Classificação/Gênero	Masc. (%)	Fem. (%)
Baixo	6 – 10	12 - 15
Ótimo	10.1 – 20	15.1 - 25
Moderadamente Alto	20.1 – 25	25.1 - 30
Alto	25.1 – 31	30.1 - 35
Muito Alto	>31	>35

Fonte: (JÚNIOR, 2019).

De acordo com a tabela utilizada no trabalho de (JÚNIOR, 2019), o ponto de corte que delimita a classificação de excesso de gordura é de 20,1 para homens e 25,1 para mulheres, portando os dados colhidos neste trabalho são:

Tabela 2 - Atributos e suas descrições

Nome do atributo	Descrição
sexo	Sexo do Individuo
peso	Peso do Indivíduo
altura	Altura do Indivíduo
cc	Circunferência da Cintura do Indivíduo
idade	Idade do Indivíduo
quadril	Medida do Quadril do Indivíduo
braço	Medida do braço do Indivíduo

pant	Medida da panturrilha do Indivíduo
classi_GC	Classe da gordura corporal do indivíduo

Fonte: Elaborada pelo autor.

Segundo (JÚNIOR, 2019), seu trabalho possui aprovação do comitê de ética em pesquisa pela universidade federal do Maranhão todos os voluntários foram incluídos somente após serem informados sobre todo o processo que para a realização do estudo logo após isso os termos de consentimento livre e esclarecido foram assinados onde os responsáveis os participantes também assinaram o termo de assentimento e o termo de consentimento livre e esclarecido onde se teve de parecer CEE: 83206118.1.0000.5087, dessa forma se garantiu a confidencialidade dos dados cedidos para o estudo em questão e também a opção de que o voluntário pode desistir participar do estudo a qualquer momento.

1.2.3. Definição de Tecnologias

Atualmente existem diversas tecnologias que tem grande capacidade no tratamento de dados e na descoberta de conhecimento destes dados, portanto, devido à natureza deste trabalho foi escolhida a linguagem de programação python na versão 3.10, devido a sua versatilidade, pequena curva de aprendizagem e grande número de bibliotecas que podem auxiliar no desenvolvimento de soluções como, por exemplo: matplotlib para geração de gráficos, pandas para manipulação de dados e scikit-learn para aprendizado de máquina.

1.2.4. Pré-Processamento e Avaliação do Método

Os atributos empregados como entradas no modelo de estimação foram cuidadosamente selecionados com base em indicadores previamente delineados na literatura, direcionados à avaliação do estado nutricional e da saúde de adolescentes. Idade, massa corporal, estatura e gênero foram escolhidos como índices recomendados pela Organização Mundial da Saúde (OMS) e foram meticulosamente mensurados em diversos estudos para a análise nutricional desta população.

Devido à base usada para este trabalho já ter sido usada em outros trabalhos, o pré-processamento foi executado anteriormente e a base não contém

irregularidades que podem comprometer o resultado de uma análise e futuro, como definição do método de avaliação foi definido que serão aplicados cinco algoritmos consolidados por ótimos resultados na literatura se tratando de aprendizado de máquina estes algoritmos são: KNN, L-SVM, árvore de decisão, K-SVM, SVM e regressão logística.

1.2.5. Analise Estatística

Para um entendimento mais aprofundado sobre a base de dados em que este trabalho se apoia foi realizada uma análise exploratória, onde o principal objetivo é entender as particularidades e características intrínsecas dos dados em questão, a tabela mostra, informações gerais sobre a base:

Tabela 3 - Dados estatísticos sobre o conjunto de dados

Estatística do banco de dados	Valor
Número de variáveis	9
Número de variáveis Categóricas	2
Número de variáveis Numéricas	7
Número de linhas	772
Dados Faltantes	0
Linhas duplicadas	0

Fonte: Elaborada pelo autor.

Outro ponto importante para se observar em uma análise exploratória é a correlação entre as variáveis pertencentes ao conjunto de dados, a correlação é uma ferramenta que tem grande importância para todas as áreas de conhecimento e não impacta somente no resultado final, mas também nas etapas que utilizam outras técnicas para análise, e se resume em verificar se existe relação entre duas ou mais variáveis, ou seja, saber se as alterações sofridas em uma das variáveis impactam em alterações na outra variável.

Tabela 4 - Correlações entre variáveis do conjunto de dados

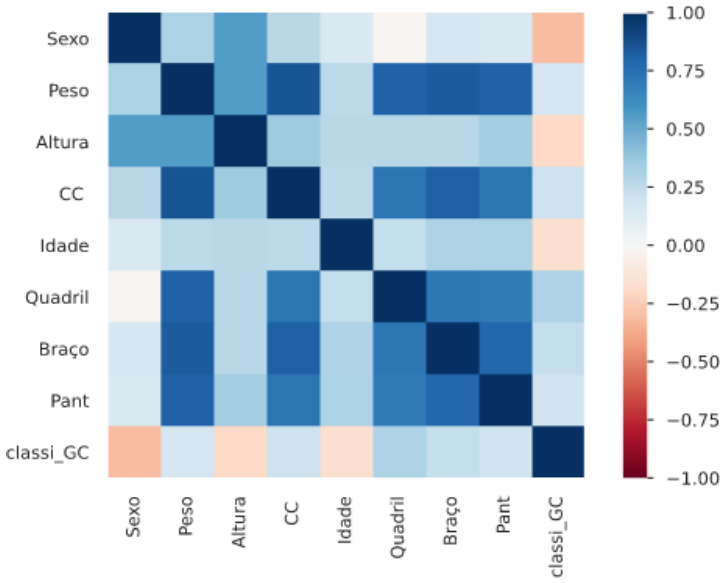
Variável	Correlação
Sexo	Alta correlação com altura.

Peso	Alta correlação com altura.
CC	Alta correlação com peso.
Quadril	Alta correlação com peso.
Braço	Alta correlação com peso.
Panturrilha	Alta correlação com peso.
Idade	Alta correlação com Quadril.

Fonte: Elaborada pelo autor.

A correlação de *Spearman* é uma técnica estatística, que se baseia em postos e exige apenas que as variáveis “X” e “Y” sejam medidas no mínimo em escala ordinal (BAUER, 2007), a figura 01 mostra o gráfico da correlação de *Spearman* entre as variáveis contidas no conjunto de dados em questão:

Figura 1 - Correlação de *Spearman*



Fonte: Elaborada pelo Autor.

Do ponto de vista estatístico existem algumas tendências que são de extrema importância para análise de uma base de dados, como, por exemplo moda, média e mediana. Estas estão descritas na tabela abaixo.

Tabela 5 - Moda, média e mediana das variáveis do conjunto de dados

Variável	Moda	Média	Mediana
Peso	62.4	57.184378	56.10
Altura	1.62	1.639355	1.63
CC	65.0	69.042293	68.00
Quadril	92.0	90.313731	90.00
Braço	25.0	24.901813	25.00
Panturrilha	34.0	33.246567	33.00
Idade	16	15.658031	16.00

Fonte: Elaborada pelo autor.

Para averiguar a normalidade dos dados foi usado o teste Kolmogorov-Smirnov, se tratando de da comparação entre os grupos de amostras independentes, no caso de distribuição normal das variáveis, foi utilizado o teste “t” de Student e o teste de Mann-Whitney U para as variáveis que não apresentarem distribuição normal

dos dados, os resultados serão considerados estatisticamente significativos para $p < 0,05$, sendo p chamado p-valor (JÚNIOR, 2019). Além disso, foram observadas outras informações gerais como que dos 772 indivíduos, 501 são do sexo masculino e 271 são do sexo feminino e se tratando da classificação sobre gordura corporal, 539 foram considerados com gordura corporal acima do ideal.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. Obesidade

A obesidade é uma enfermidade que se caracteriza pelo teor de gordura corporal, onde determinado nível pode comprometer gravemente a saúde do indivíduo em questão, trazendo prejuízos à saúde do mesmo, como, por exemplo: alterações metabólicas, dificuldades respiratórias e do aparelho locomotor, além de deixar o indivíduo propício a outros tipos de doenças, como doenças cardiovasculares, diabete e alguns tipos de câncer. A obesidade faz parte do grupo de doenças consideradas crônicas e não transmissíveis, O seu diagnóstico é feito a partir de parâmetros estipulados pela OMS, através do IMC, este por sua vez, é obtido a partir da relação entre peso corpóreo e estatura dos indivíduos em questão, através destes parâmetros são classificados os indivíduos em obesos cujo IMC se trata de valor igual ou maior que 30 kg/m^2 (WANDERLEY; FERREIRA, 2010).

A obesidade tem causas diversas sendo resultado da interação entre fatores genéticos, metabólicos, sociais, comportamentais e até mesmo culturais, em sua grande maioria os casos de obesidade tem associação direta com a ingestão calórica e o sedentarismo, para tanto o excesso de calorias se armazena como um tecido adiposo fazendo com que seja gerado um balanço energético positivo e este balanço pode ser definido como uma diferença entre a quantidade de energia adquirida e a energia gasta na realização das funções vitais e de todas as atividades do corpo humano. A obesidade pode ser responsável por diversas enfermidades que o indivíduo possa obter como, por exemplo, hipertensão arterial sistêmica, hipertrofia ventricular esquerda com ou sem influência cardíaca, doença cerebral vascular, trombose venosa profunda e várias outras (TAVARES; NUNES; SANTOS, 2010).

Segundo estudos existem diversos aspectos biológicos da obesidade que envolvem a genética e o metabolismo de um ser humano, existe também a teoria da economia energética, que vem sendo apontada como possível contribuinte para o desenvolvimento da obesidade em um ser humano. Segundo esta teoria, em situações de adversidades biológicas e também sociais onde ocorre um deficit de energia, o organismo usa uma série de mecanismos metabólicos adaptativos onde esses organismos visam promover a redução do gasto energético, caracterizando uma estratégia de sobrevivência, esta situação leva ao organismo a um novo ponto de equilíbrio, em que o gasto e a ingestão energética são diferentes do normal proporcionando uma quebra de rotina, dessa forma o aumento na ingestão de alimentos pode facilmente resultar em ganho de peso como consequência do aumento da eficiência metabólica (WANDERLEY; FERREIRA, 2010).

A morbimortalidade relacionada a obesidade tem aumentado de forma significativa principalmente quando o IMC está superior a 30 kg/m², sendo que o risco de morte do indivíduo tem um aumento significativamente e em muitos casos pode até mesmo dobrar, existem vários graus de obesidade e a obesidade de grau 3 é uma enfermidade das que mais matam em todo o mundo, somente na América Latina cerca de 200 mil pessoas morrem anualmente em decorrência das consequências relacionadas à obesidade, sendo assim diversos estudos tratam a qualidade de vida relacionado a indivíduos que possui a obesidade, com isso é fácil pressupor que a obesidade é um dos males mais avassaladores dos últimos anos (TAVARES; NUNES; SANTOS, 2010).

2.2. KNN (*K-Nearest Neighbors*)

O algoritmo KNN é uma técnica de aprendizado supervisionado do tipo *lazy*, esse algoritmo se resume em encontrar os “k” exemplos rotulados mais próximos do exemplo não classificado, ou seja, é feita uma classificação de acordo com a proximidade de outros itens já classificados, a aprendizagem supervisionada é uma área que envolve situações em que se tem os exemplos de treinamento com informações de entrada e informações de saída, ou seja, o resultado da saída é conhecido, sendo intitulado rótulo, este rótulo pode ser cedido tanto por um instrutor supervisionando o processo de aprendizagem ou também pode estar contido no

próprio exemplo como um de seus atributos, a figura 2 mostra um exemplo de estrutura para aprendizado supervisionado (JORGE, 2012).

Figura 2 - Representação de um conjunto de dados

	A_1	A_2	...	A_M	Classe (Y)
E_1	x_{11}	x_{12}	⋮	x_{1M}	y_1
E_2	x_{21}	x_{22}	⋮	x_{2M}	y_2
⋮	⋮	⋮	⋱	⋮	⋮
E_N	x_{N1}	x_{N2}	⋮	x_{NM}	y_N

Fonte: (JORGE, 2012).

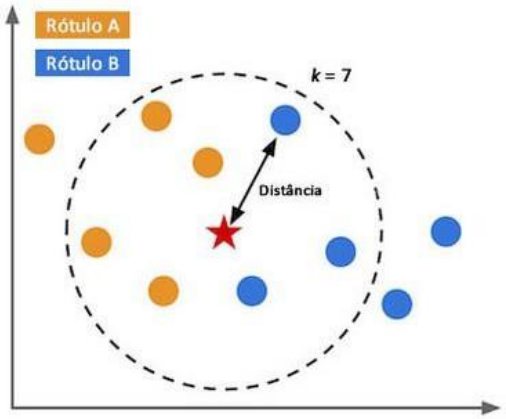
O conjunto de treinamento para um algoritmo de aprendizado supervisionado se resume em um conjunto “E” de “N” exemplos de treinamento (JORGE, 2012). O algoritmo KNN não tem um estágio de treinamento e executa a classificação fazendo o cálculo primeiro da distância entre a amostra usada como teste e as amostras de treinamento para obter seus vizinhos mais próximos, duas importantes características do algoritmo KNN são a regra de classificações e a função que a distância entre os dois pontos, a regra de classificação mostra o modo que o algoritmo vai tratar a importância de cada um dos elementos selecionados, ou seja, buscar os “K” elementos mais próximos, e a função de distância é usada para medir a distância entre dois elementos para que se possa identificar quais são os mais próximos, existem várias medidas de distância, porém, a distância euclidiana é a mais simples e mais usada dentro da literatura especialmente em casos bidimensionais, a fórmula é representada como d = a distância euclidiana e “X” e “Y” são os pontos que serão calculados, abaixo a equação da distância Euclidiana (JORGE, 2012).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad 2.1$$

Este algoritmo é muito popular sendo usado em diversas aplicações de mineração de dados e estatística e é considerado um dos principais algoritmos neste

ramo (ZHANG; ZONG; ZHU, 2018), a figura 02 mostra um exemplo de funcionamento do KNN.

Figura 3 - Exemplo de funcionamento do KNN



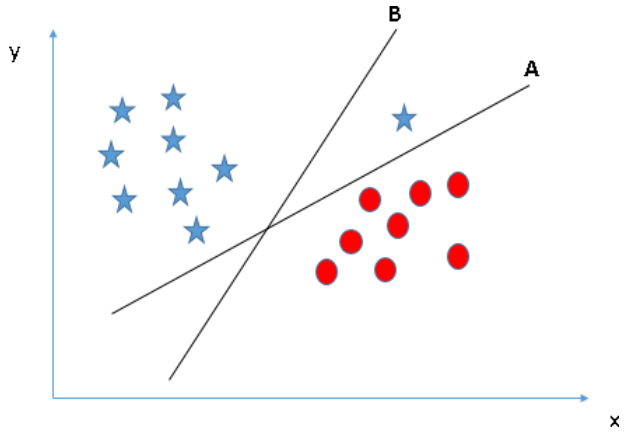
Fonte: (PACHECO, 2017).

O KNN é muito usado e tem fácil aplicação, porém, pode não ser uma boa opção quando o conjunto de dados toma, grandes proporções, pois, a classificação de uma nova instância exige cálculos de distância em relação aos vizinhos próximos, em casos com grandes quantidades de vizinhos este processo pode ser trabalhoso e fazer com que o desempenho do algoritmo, em geral, perca qualidade (FARIA, 2016).

2.3. SVM (*Support Vector Machines*)

O SVM (*Support Vector Machines*) é um algoritmo de aprendizado de máquina que pode ser usado para classificação e também para regressão, este algoritmo se baseia em planos que definem o limite de decisão, onde estes planos são responsáveis por diferenciar objetos de classes distintas, o SVM tem resultados promissores em problemas de classificação como reconhecimento de caracteres manuscritos, detecção de faces, categorização de textos e outros. Devido a sua ótima performance o algoritmo SVM é visto como um referencial dentro das pesquisas que utilizam aprendizado de máquina, para problemas que não são lineares o algoritmo utiliza funções de kernel para fazer um mapeamento do espaço de entrada que é de alta dimensão e o espaço intitulado de espaço de características onde as classes são linearmente separáveis fazendo assim com que o resultado final seja único, a figura 3 ilustra o funcionamento do algoritmo SVM (LIMA, 2014).

Figura 4 - Funcionamento do SVM

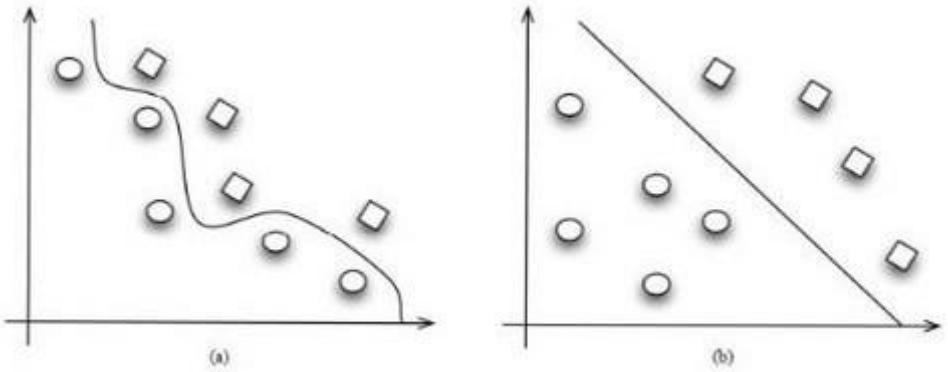


Fonte: (SOUSA, 2019).

As SVM se caracterizam por criar um hiperplano como superfície de decisão, onde a separação entre os exemplos dados seja máxima, ou seja, a mais clara possível. A margem é calculada pela distância entre o hiperplano e os vetores mais próximos a ele, estes são chamados vetores de suporte, como o espaço em questão é bidimensional, o hiperplano é como uma linha que separa linearmente os pontos azuis dos pontos vermelhos mostrados na figura 4. Uma escolha razoável é o hiperplano que representa a maior separação para minimizar erros de classificação e problemas de enviesamento do modelo (*overfitting*) e assim fazer com que o algoritmo tenha um ótimo desempenho (ARAÚJO, 2015).

A principal ideia de uma SVM é mapear o espaço original (x), em um espaço de características (f) de alta dimensão utilizando uma função de mapeamento não linear, esta técnica é considerada recente e, portantoportanto, ainda apresenta pontos que podem ser melhorados ao longo de estudos sobre a técnica, existem diferentes tipos de SVM elas podem ser usadas com diferentes tipos de kernel, devido à sua versatilidade é sensato escolher a configuração de SVM mais adequada para cada finalidade, tornando o algoritmo uma ferramenta poderosa, porém, que exige uma configuração correta para cada tipo de problema (LIMA, 2014).

Figura 5 - Ilustração de um espaço onde as classes não são linearmente separáveis, b) Ilustração de um espaço de características onde as classes são linearmente separáveis



Fonte: (LIMA, 2014)

Uma SVM usa um ou mais hiperplanos para dividir o espaço em zonas fazendo com que estas zonas contenham classes em comum, ao rotular estas zonas o sistema é capaz de identificar seus elementos através de uma amostra, então o hiperplano é usado para fazer a separação das instâncias. O SVM seleciona todas as amostras críticas e a partir disso cria uma função linear que as diferencia, O SVM tem a capacidade de fazer a separação mesmo quando elas não são linearmente separáveis, isto porque o algoritmo é capaz de projetar um problema em um espaço de alta dimensão tornando as classes em questão linearmente separáveis (LIMA, 2014).

2.3.1 L-SVM (*Linear Support Vector Machines*)

Para vetores de suporte existe uma classificação intitulada de SVC linear, o objetivo da SVC linear é com base nas características e padrões de cada classe separar as classes de forma correta, ou seja, não permitir que um objeto, seja classificado erroneamente, o classificador responsável por isso é chamado de hiperplano, em uma L-SVM o objetivo é encontrar um hiperplano onde o vetor normal “W” seja ortogonal aos vetores tangentes, para este algoritmo “n” conjuntos de dados de treinamento (x1, y1 a xin., yin) são cedidos, e é criado um classificador onde qualquer hiperplano pode ser mencionado como o conjunto de pontos “X” satisfazendo a equação (CHAPELLE; SCHOLKOP, 2001).

$$\omega . x - b = 0 \qquad 2.2$$

Onde “W” serve como o vetor normal ao hiperplano, podem haver dois tipos de margem, a margem rígida e a margem flexível, se os dados de treinamento podem ser separados de forma linear e sem qualquer erro, a margem rígida é aplicada, caso tenha algum de erro, a margem dura falha é usada (GHOSH; DASGUPTA; SWETAPADMA, 2019).

O classificador SVM linear é implementado especificamente para níveis massivos de dados e recursos, o hiperplano de decisão que é calculado é usado para classificar amostras em diferentes categorias. A seleção do fator de penalidade de erro, que expressa a tolerância ao erro, afeta significativamente a precisão da SVM linear (GHOSH; DASGUPTA; SWETAPADMA, 2019), este algoritmo é amplamente utilizado devido a seus resultados ótimos e alta capacidade de processamento de dados.

2.4. K-SVM (Kernel *Support Vector Machine*)

O algoritmo K-SVM é uma variação do SVM é usado em diversos casos na literatura, é técnica que usa um kernel para realizar a transformação de dados e a partir do resultado desta transformação encontra um limite ótimo entre as saídas possíveis, este algoritmo realiza operações complexas para que se possa descobrir uma maneira de separar o conjunto de dados com base nos rótulos ou saídas (ARAÚJO, 2015), existem diversos tipos de funções de kernel dentre elas estão: Kernel linear, Função polinomial, Função de Kernel dentre outras (SHAW; ROUTRAY, 2016), a equação que rege este algoritmo é dada por:

$$f(x) = sign\left(\sum_{i=1}^{N} y_i\, \alpha_i k\,(x, x_i) + b\right) \qquad 2.3$$

Onde k (x, xi) é o kernel que é representado pelo produto dos pontos k(x, xi) = 〈x, xi〉 onde 〈., .〉 representa o produto escalar (SHAW; ROUTRAY, 2016). O kernel tem função de treinamento e previsão de algoritmos e os parâmetros podem ser definidos para qualquer função de classe kernel assim é possível você fazer o cálculo do produto interno no espaço entre dois argumentos vetoriais, dessa forma

o K-SVM é um ótimo método de classificação não linear tendo um ótimo desempenho onde os vetores de suporte informam completamente a superfície de decisão mostrando informações no espaço do kernel (ARAÚJO, 2015).

O K-SVM é um ótimo algoritmo de classificação não linear onde os Vetores de Suporte retratam totalmente a superfície de decisão absorvendo informações através do kernel. O esquema de classificação de K-SVM trabalha com o objetivo de separar amostras pertencentes a, diferentes classes procurando o hiperplano de maior margem (SHAW; ROUTRAY, 2016), com essa metodologia o algoritmo promete resultados ótimos e, é bastante utilizado em diversos problemas no mundo real.

2.5. Regressão Logística

No ano de 1885, em um estudo onde demonstrou que a altura dos filhos não se reflete na altura dos pais, Sir Francis Galton criou o termo regressão, desde então essa técnica vem sendo aprimorada e usada em diversos tipos de problemas. Ela compreende um vasto conjunto de técnicas estatísticas que são usadas para fazer modelagem de relações entre variáveis e conseguir a partir disso predizer o valor de uma ou mais variáveis dependentes a partir de um conjunto de variáveis independentes, a correlação entre as variáveis são de extrema importância e a sua análise tem como objetivo a avaliação do grau de associação entre as variáveis em questão, ou seja, o quão forte é o relacionamento entre essas variáveis, para se quantificar a relação entre duas variáveis quantitativas se recorre ao coeficiente de correlação linear de Pearson (RODRIGUES, 2012), o coeficiente de correlação linear de Pearson entre duas variáveis “X” e “Y” é dado por:

$$Rxy = \frac{\sum_{I=1}^{n}(xi - \underline{x})(yi - \underline{y})}{\sqrt{\sum_{i=i}^{n}(xi - \underline{x})^2}\ \sqrt{\sum_{i=1}^{n}(yi - \underline{y})^2}} \qquad 2.4$$

onde:

$$\underline{x} = \sum_{i=1}^{n} \frac{xi}{n} \text{ e } \underline{y} = \sum_{i=1}^{n} \frac{y_i}{n} \qquad 2.5$$

Em outras palavras, é o quociente entre a covariância das variáveis “X” e “Y” e o produto do desvio padrão de “X” e “Y”, a partir de Rxy você pode concluir o grau de relação existente entre as variáveis “X” e “Y”. Não existe uma classificação única dentro da correlação, portanto, para se entender melhor a correlação é utilizada uma tabela que informa a partir do resultado o grau de correlação entre as variáveis (RODRIGUES, 2012).

Tabela 6 - Coeficientes e suas correlações

Coeficiente de Correlação	Correlação
R_{xy} = 1	Perfeita positiva
0,8 ≤ R_{xy} < 1	Forte positiva
0,5 ≤ R_{xy} < 0,8	Moderada positiva
0,1 ≤ R_{xy} < 0,5	Fraca positiva
0 ≤ R_{xy} < 0,1	Ínfima positiva
0	Nula
- 0,1 ≤ R_{xy} < 0	Ínfima negativa
- 0,5 ≤ R_{xy} < - 0,1	Fraca negativa
- 0,8 ≤ R_{xy} < - 0,5	Moderada negativa
-1 ≤ R_{xy} < - 0,8	Forte negativa
R_{xy} = -1	Perfeita negativa

Fonte: (RODRIGUES, 2012).

A regressão se caracteriza como uma técnica que busca modelar uma equação matemática que faz a descrição da relação entre duas variáveis, existem diversos métodos para aplicação de equações de regressão em situações envolvendo variáveis, no entanto, a análise estatística feita sobre a regressão faz somente a modelagem de qual relacionamento matemático pode existir entre as variáveis em questão, e se existe algum. A regressão linear também pode ser aplicada para fins de prevenção de valores de uma variável, é importante dizer que existem vários tipos de relações que podem assumir diversas formas dentro da aplicação da regressão linear, as equações lineares têm grande importância porque servem para fazer aproximação com relações que acontecem na vida real e também pelo fato de ter em um nível de dificuldade baixo quando se trata de implementar e interpretar seus resultados, em

outras formas de análise de regressão como a regressão múltipla e regressão corvilínica os mesmos conceitos utilizados na regressão linear simples são usados (RODRIGUES, 2012).

É importante salientar que não é em todas as situações que a equação linear vai ser o principal ponto de aproximação, em geral, é preciso desenvolver tarefas preliminares para poder saber se o modelo linear é realmente adequado ao problema em questão, o trabalho preliminar se resume em fazer o diagrama de dispersão dos dados, ou seja, verificar se os valores de “X” e de “Y” mostram uma tendência linear (SELL, 2005).

Quando a regressão é tratada somente utilizando uma equação matemática, para que se possa descrever o relacionamento entre duas variáveis onde um é dependente e a outra é independente, buscando estimar um valor para uma variável com base nesses valores conhecidos, ela é dita regressão linear simples, porém, quando a regressão trata três ou mais variáveis, com uma variável dependente e outras duas ou mais independentes buscando melhorar a capacidade de predição confrontando a regressão linear simples, ela é dita regressão linear múltipla (SELL, 2005).

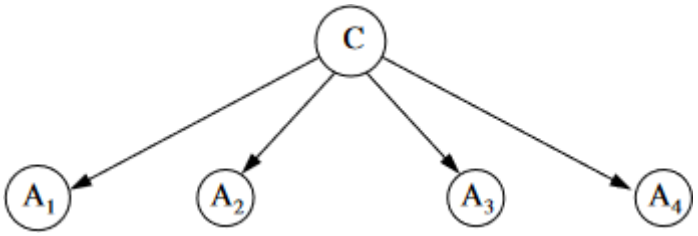
Já a regressão logística, é uma técnica que vem da área da estatística, que tem o intuito de modelar através de um conjunto de observações, a possível relação logística que existe entre uma variável resposta e um conjunto de variáveis explicativas, numéricas ou categóricas (CABRAL, 2013).

2.6. *Nayve Bayes*

O Naive Bayes é um classificador probabilístico que tem como base o teorema bayesiano que foi criado por Thomas Bayes, também é chamado de classificador ingênuo de Bayes, uma rede bayesiana é em um modelo estrutural e também um conjunto de problemas condicionais. O modelo dito estrutural é composto por um grafo direcionado em que os nós representamos atributos e os arcos representam dependências de atributos, existem dependências de atributos por probabilidades condicionais para cada um dos nós dados em seus pais. As redes de Bayes são frequentemente usadas para problemas de classificação, nos quais, um aprendiz tenta

construir um classificador a partir de um determinado conjunto de exemplos de treinamento com rótulos de classe (JIANG et al., 2007).

Figura 6 - Exemplo de *Naive Bayes*



Fonte: (JIANG et al., 2007)

O Naive Bayes é um dos algoritmos mais simples quando se trata de aprendizado de máquina, o teorema de Bayes faz uso da teoria da probabilidade onde ele mostra a relação que existe entre a probabilidade condicional e a probabilidade inversa, ou seja, a probabilidade de uma hipótese ser verdadeira onde consultada uma observação de uma evidência e essa probabilidade da evidência pela sua hipótese, este teorema representa uma das primeiras tentativas de se modelar de forma matemática uma inferência estatística (JIANG et al., 2007).

Uma rede Bayesiana é um algoritmo de aprendizado de máquina que tem a capacidade de gerar predições associadas a valores de probabilidade, uma de suas principais vantagens é que ela permite o tratamento de fenômenos associados ao tempo considerando a dependência temporal dos dados e tratando de forma rápida e eficiente séries temporais, o estudo dessas redes é complexo e exige um vasto conhecimento sobre estatística e probabilidade, no entanto, um classificador bayesiano é simples de forma conceitual e pode ser entendido sem mais problemas da mesma forma quando se trata de sua aplicação, como já dito o classificador aprende a partir dos dados cedidos para treinamento a probabilidade condicional de cada atributo cedido a ele dado o valor da classe(JIANG et al., 2007).

Para melhor entendimento do algoritmo como já dito são necessárias noções de probabilidade, a probabilidade condicional é a probabilidade que um evento aconteça dado que outro evento aconteceu, em outras palavras é a probabilidade que um evento chamado “A” aconteça depois que um evento chamado “B” aconteceu, esta

probabilidade é representada por uma barra como pode ser visto na equação (WANKE et al., 2014).

$$P\ (A \mid B) \qquad 2.6$$

Quando existem dois eventos independentes, dado um terceiro, a probabilidade de que aconteçam ao mesmo tempo, é igual à multiplicação de suas probabilidades dado o terceiro evento (WANKE et al., 2014).

$$P(A, B \mid C) = P(A|C) \cdot P(B|C) \qquad 2.7$$

A Regra de *Bayes* diz que:

$$P(A|B) = \frac{P(A, B)}{P(B)} \qquad 2.8$$

e que:

$$P(B|A) = \frac{P(A, B)}{P(A)} \qquad 2.9$$

Considerando as Eq. (3) e (4), tem-se que:

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \qquad 2.10$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \qquad 2.11$$

Para se fazer a classificação é aplicada a regra de *Bayes* para calcular a probabilidade de cada classe, dados o valor do atributo em um caso teste então se escolhe a que resulta em maior probabilidade (WANKE et al., 2014).

$$C_{casoteste} = \arg\max P(c \mid a_1, a_2, \ldots, a_n) \qquad 2.12$$

Aplicando a regra de *Bayes*:

$$C_{casoteste} = \arg\max \frac{P(a_1, a_2, \ldots, a_n|c) \cdot P(c)}{P(a_1, a_2, \ldots, a_n)} \quad 2.13$$

Então:

$$C_{casoteste} = arg\ max\ P(a_1, a_2, \ldots, a_n \mid c) \cdot P(c) \quad 2.14$$

O cálculo realizado se torna possível devido a uma suposição de que todos os atributos são independentes, dada a classe, em poucos casos a suposição é verdadeira, porém, mesmo assim o classificador tem um ótimo desempenho, dado a independência dos atributos, dada a classe.

$$C_{casoteste} = arg\ max\ P(a_1, a_2, \ldots, a_n \mid c) \cdot P(c) \quad 2.15$$

$$C_{casoteste} = arg\ max\ P(c) \cdot P(a_1|c) \cdot C\ ) \quad 2.16$$

$$C_{casoteste} = arg\ max\ P(a_1, a_2, \ldots, a_n|c) \quad 2.17$$

$$C_{casoteste} = arg\ max\ P(c) \cdot P(a_1|\ c) \cdot P(a_2|c) \cdot \ldots \cdot P(a_n|c) \quad 2.18$$

$$C_{casoteste} = arg\ max\ P(c) \cdot \prod_i P(a_i \mid c) \quad 2.19$$

O cálculo das probabilidades efetuado utilizando a contagem de exemplos contidas no conjunto de treinamento.

$$P(a_i|c) = \frac{P(a_i, c)}{P(c)} \quad 2.20$$

$$P(c) = \frac{|S_c|}{|S|} \quad 2.21$$

$$P(a_i, c) = \frac{|S_{ai \cdot c}|}{|S|} \quad 2.22$$

$$P(a_i|c) = \frac{P(a_i, c)}{P(C)} = \frac{|S_{ai \cdot c}|}{|S|} \cdot \frac{|S|}{|S_C|} = \frac{|S_{ai \cdot c}|}{|S_c|} \quad 2.23$$

Atualmente este algoritmo é usado em diversas aplicações do mundo real devido a sua grande performance e a facilidade de implementação em diversas linguagens de programação.

2.7. Composição Corporal

A partir do século XIX iniciou o interesse pela medição de quantidade dos diferentes componentes do corpo humano, no século XX esse interesse cresceu ainda mais, após vários estudos detectaram que existia a associação do excesso de gordura com diversas enfermidades, aproximadamente em 1998 a obesidade foi classificada como um fator de risco para a doença coronária pelo American Heart Association, a prevalência dessa enfermidade cresceu a cada ano que se passava inclusive nos países em desenvolvimento, até o início do século XX a análise da composição corporal era feita através da dissecação de cadáveres, técnica que é considerada até hoje a única maneira direta de medir os principais componentes do corpo humano, porém, estudiosos passaram a procurar maneiras de medir a composição corporal de forma mais rápida, eficiente e com o indivíduo ainda em vida (JÚNIOR, 2019).

A definição de composição corporal é dada pela proporção entre os componentes corporais e a massa corporal total, onde essa medida é normalmente expressa pela percentagem de massa gorda e massa magra de um indivíduo, por meio de uma avaliação corporal se tem a possibilidade de determinar os componentes do corpo humano em uma visão quantitativa, uma avaliação de gordura corporal pode trazer diversas informações de extrema importância quando se trata da saúde de um ser humano, como, por exemplo, ela pode determinar a quantidade de gordura corporal por região e a gordura total, pode identificar se os níveis de gordura do indivíduo estão altamente altos ou baixos, pode avaliar os riscos cardiometabólicos e cardiovasculares, pode monitorar a mudança corporal que tem associação com diversas doenças e também pode constatar se existe eficácia em intervenções nutricionais e exercícios físicos (JÚNIOR, 2019).

Vários estudos indicam que a gordura corporal tem relação direta com risco cardiovascular em adultos e adolescentes, assim sendo qualquer medida que possa agilizar e dar mais qualidade ao processo de medição de gordura corporal de um ser humano é de extrema importância para a saúde pública, em geral (JÚNIOR, 2019).

2.8. Técnicas de Avaliação da Composição Corporal

Uma avaliação da composição corporal informa dados sobre a quantidade dos principais componentes corporais como: músculos, ossos, gorduras, tecidos e substâncias residuais, onde a soma destes componentes, resulta no peso corporal total. Existem três técnicas de avaliação corporal, elas são divididas em três categorias: diretas, indiretas e duplamente indiretas (JÚNIOR, 2019).

No método direto existe uma separação e pesagem de cada um dos componentes corporais de forma isolada, para essa técnica é necessária a dissecação do cadáver, portanto, os procedimentos que utilizam essa técnica são extremamente raros atualmente (JUNIOR, 2019).

O método indireto é eficaz, porém, em seu uso não existe a manipulação de componentes separados, esta característica torna este método limitado pela sua aplicação e também por ter um alto custo, em geral, este método é utilizado para validação de técnicas duplamente indiretas (JUNIOR, 2019).

Além disso, este método pode causar incômodo ao paciente e os aparelhos utilizados necessitam de uma complexa e a manutenção e também de profissionais devidamente treinados, alguns exemplos de técnicas indiretas são: pesagem hidrostática, hidrometria, plestismografia e absortometria radiológica de dupla energia — DEXA (JÚNIOR, 2019).

Figura 7 - Aparelho DEXA

Fonte: (JÚNIOR, 2019).

Os métodos indiretos dão origem aos métodos duplamente indiretos, onde os métodos duplamente indiretos são validados pelos métodos indiretos, estes métodos possuem aplicação com grande facilidade e baixo custo, característica que faz com que os mesmos sejam aplicados em estudos clínicos e epidemiológicos, se destacam a bioimpedância e a antropometria (JÚNIOR, 2019).

2.9. Avaliação Antropométrica

A ciência que estuda as medidas do tamanho corporal é definida como antropometria, é um ramo das ciências biológicas que tem como foco de estudo as características mensuráveis da morfologia humana, a origem da antropometria se dá pelo tempo de antiguidade, os egípcios e os gregos estudavam a relação das diversas partes do corpo humano, o conhecimento sobre os biotipos do ser humano vem desde os tempos bíblicos onde o nome de muitas unidades de medidas são utilizadas até os dias de hoje (RODRIGUEZ-AÑEZ, 2001).

A importância das medidas antropométricas ganhou uma atenção especial na década de 40 onde aconteceu um fenômeno que se precisava de produção em massa, pois, um produto mal dimensionado poderia provocar elevação dos custos e por outro devido ao surgimento de sistemas de trabalho complexos onde a força do ser humano era a parte vital para se alcançar o sucesso a antropometria ganhou notoriedade, pois, através dela foi possível se estudar as medidas do corpo humano, uma das funcionalidades da antropometria é a avaliação do Estado nutricional de um indivíduo que se resume na utilização de processos de diagnóstico que precisam de agravos nutricionais assim se pode identificar grupos de risco (RODRIGUEZ-AÑEZ, 2001), ao se utilizar dados antropométricos é possível se quantificar e qualificar as medidas de um indivíduo assim se pode determinar os valores obtidos dentro de um intervalo que pode ser considerado normal ou não.

A sensibilidade é a capacidade de um teste detectar os indivíduos que são verdadeiramente positivos em outras palavras é a capacidade de diagnosticar corretamente os indivíduos com excesso de gordura corporal, por outro lado, a especificidade é a capacidade do teste detectar verdadeiros negativos, ou seja, os indivíduos que não possuem níveis elevados de gordura corporal, a técnica da antropometria é largamente utilizada na avaliação do Estado nutricional de indivíduos na adolescência por ser uma técnica pouco invasiva fácil de executar e relativamente barato estas características possibilitam que ela seja aplicada em grandes números de indivíduos em comparação a outros métodos ela se sobressai por estas características algumas das variáveis da composição corporal mais utilizadas em estudos são o percentual de gordura corporal a circunferência da cintura do quadril a relação cintura quadril a relação cintura estatura, dessa esta técnica é de grande importância, pois, possibilita a detecção do excesso de gordura corporal em adolescentes tendo em vista que além de ser uma técnica de fácil aplicação e que necessita de pouco recurso pode ajudar a diagnosticar o excesso de gordura ainda na adolescência fazendo com que um indivíduo possa procurar ajuda médica em uma fase inicial aumentando as chances do indivíduo para prevenir o excesso de gordura (JÚNIOR, 2019).

2.10. IMC

O índice de massa corporal de um indivíduo é utilizado para mostrar as correlações máximas entre sobrepeso e gordura corporal esta informação é obtida através da razão entre o peso corporal que é expresso em quilogramas e a estatura que é expressa em m², abaixo a equação de cálculo do IMC (DALLASTELLA, 2006).

$$IMC = \frac{Peso\ Corporal(Kg)}{Estatura^2(m)} \qquad 2.24$$

O IMC é uma técnica utilizada para análises e diagnóstico nutricional de pacientes, porém, ele possui algumas limitações, mas segue sendo utilizado pela sua facilidade de ser aplicado, pois, as informações que a técnica exige são de fácil obtenção, utilizando o resultado obtido através da fórmula citada anteriormente a OMS classifica valores conforme a tabela abaixo.

Tabela 7 - Diferentes níveis de diagnósticos por IMC

IMC	GRAU
< 18,5 kg/m²	Baixo Peso
Entre 18,5 kg/m² e 24,9kg/m²	Eutrófico
Entre 25kg/m² e 29,9 kg/m²	Sobrepeso
Entre 30kg/m² e 34,9kg/m²	Obesidade de grau l
Entre 35kg/m² e 39,9kg/m²	Obesidade de grau ll
Acima de 40kg/m²	Obesidade de grau lll

Fonte: (PEREIRA, 2022).

O IMC é um cálculo universal ele foi adotado pela organização Mundial de Saúde para que se pudesse classificar a saúde de um ser humano quando se trata do peso do mesmo, devido ao grande número de enfermidades ligadas ao sobrepeso e a obesidade de um indivíduo o IMC é muito utilizado para se aferir métricas de desnutrição e obesidade, pelo fato de não considerar a composição corporal o IMC não é recomendado para a aplicação de avaliação de saúde de atletas, pois, estes facilmente podem ter o peso mais elevado devido às atividades praticadas, deixando a desejar em algumas situações, portanto, apesar de muito o usado o cálculo do IMC não possui unanimidade em sua aplicação (PEREIRA, 2022).

2.11. Relação Cintura Estatura

A relação cintura estatura (RCE) tem base partindo do ponto de que determinada estatura exige um grau de gordura armazenada na parte superior do corpo, a técnica de RCE mostra que existe uma correlação como a gordura visceral, em outras palavras se considera uma proporção de gordura central devido à estatura do indivíduo (VIEIRA et al., 2017).

Esta técnica vem apresentando altas taxas de sucesso para identificação de riscos cardiovasculares em crianças, alguns estudos constataram que a técnica tem alta capacidade de discriminar a gordura corporal em níveis elevados em

adolescentes, portanto, é um bom indicador para o excesso de gordura corporal se tratando desta faixa etária (JÚNIOR, 2019).

A RCE pode ser um importante indicador que pode ser usado na predição de riscos metabólicos em virtude da obesidade, a técnica pode ser aplicada em adultos e crianças, um argumento bastante utilizado é que a medida isolada da Circunferência da Cintura (CC) e do IMC precisam de vários pontos de corte esses pontos dependem da etnia ou do gênero do indivíduo, essas informações podem dificultar a sua utilização (JÚNIOR, 2019).

2.12. Bioimpedância

A Bia se baseia em um modelo de condutor no formato cilíndrico, cuja área transversal e o comprimento são uniformes e homogêneos ao corpo humano, o volume cilíndrico é diretamente relacionado com a impedância total do corpo humano (EICKEMBERG et al., 2011), A análise da impedância bioelétrica tem grandes benefícios, pois, é uma técnica barata segura de fácil aplicação e portátil além de não invasiva, essas características tornam a técnica habilitada a ser usada por diversos profissionais em estudos epidemiológicos, ela pode ser definida como a capacidade do tecido biológico restringir a passagem de corrente elétrica (JÚNIOR, 2019).

Em um sistema biológico quando se trata de corrente elétrica a mesma é transmitida pelos íons diluídos nos fluidos corporais os tecidos magros são condutores com grande potencial por terem uma grande quantidade de água e eletrólitos, resultando em uma baixa resistência para a passagem de uma corrente elétrica, em contrapartida, a gordura, o osso e a pele representam componentes de baixa condutividade e, tem elevada resistência (EICKEMBERG et al., 2011), a análise da impedância bioelétrica tem grandes benefícios, pois, é uma técnica barata, segura, de fácil aplicação, portátil e não invasiva, essas características tornam a técnica habilitada a ser usada por diversos profissionais em estudos epidemiológicos, ela pode ser definida como a capacidade do tecido biológico restringir a passagem de corrente elétrica(JÚNIOR, 2019).

O método da bioimpedância é focado na diferença que ocorre na passagem de uma corrente elétrica pelos diferentes componentes do corpo humano que possuem gordura e água corporal, a impedância (Z) é considerada a medida da oposição a corrente que um circuito mostra quando se é aplicada uma tensão alternada sobre ele, esta tensão depende da resistência (R) que por sua vez é causada pela água corporal total e a reatância (Xc) que é causada pela capacitância da membrana celular(JÚNIOR, 2019), conforme mostrado na equação:

$$|Z| = \sqrt{R^2 + X_C{}^2} \qquad 2.25$$

A reatância faz a configuração da oposição quando se trata da corrente elétrica gerada através da capacitância, um capacitor é constituído por duas ou mais membranas condutoras estas membranas são separadas por um material isolante, ou com baixo poder produtivo capaz de armazenar energia elétrica, a membrana citoplasmática presente no ser humano é formada por duas camadas de um material proteico em outras palavras o material com alto poder de condutividade e uma camada de lipídios que por sua vez são isolantes, o conjunto de resistência elétrica é a distância massa corporal e estatura pode medir com confiabilidade a composição corporal, essas medidas podem sofrer ajustes dependendo do gênero, etnia, idade, peso e do quão ativo fisicamente o indivíduo é através de equações já pré-definidas (JÚNIOR, 2019).

2.13. BIA de Frequência Única

A Bia possui diferentes métodos para a realização de uma análise um desses métodos é a frequência única, onde a impedância de frequência única utiliza uma única corrente de frequência de 50kHz, onde a corrente é passada apenas no espaço extracelular em todo o corpo ou em determinadas partes do corpo, as informações geradas através deste método são colhidas com a aplicação de equações de regressão previamente derivadas utilizando os dados de referência, no geral a Bia de frequência única é uma técnica pouco custosa e de fácil aplicação, e para aplicá-la existem diversos equipamentos com eficácia já comprovada (CRISPILHO, 2019).

3. RESULTADOS

Nesta etapa serão aplicados os algoritmos escolhidos para fins de comparação, para tanto serão usadas métricas para aferir o desempenho de cada algoritmo, e ao fim poder selecionar o que melhor se adequá ao objetivo final deste trabalho. Em relação ao percentual de GC 69,7% (n = 539) da amostra apresentou gordura corporal elevada para idade. Todas as variáveis analisadas apresentaram diferença estatisticamente significativa ($p < 0,05$), com exceção da cor, quando estratificadas pelo %GC.

3.1. *Naive Bayes*

O *Naive Bayes* é um algoritmo classificador que visa rotular, instâncias descritas por um conjunto de variáveis, a tabela mostra as métricas obtidas com a aplicação do mesmo.

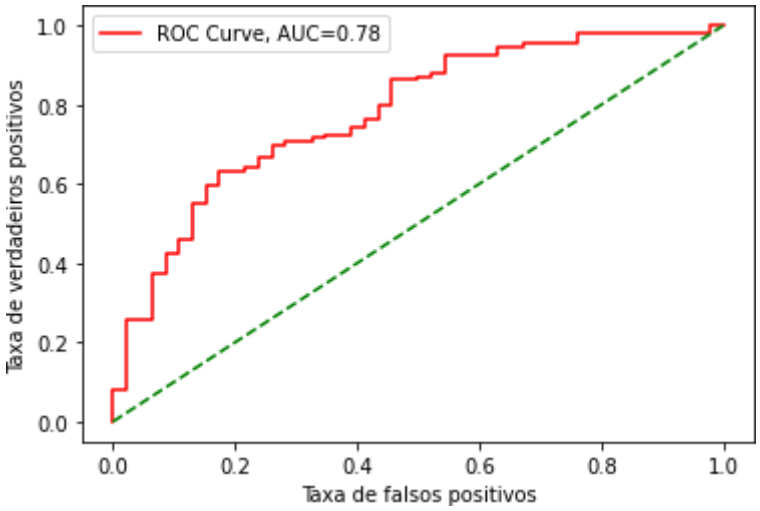
Tabela 8 - Métricas para o algoritmo *Nayve Bayes*

Métrica	Valor
Acurácia	0.6967741935483871
Sensibilidade	0.7247706422018348
Especificidade	0.6304347826086957
f1_score	0.7707317073170732
Valor Preditivo Positivo	0.8229166666666666
Valor Preditivo Negativo	0.4915254237288136
Taxa de Falso Negativo	0.27522935779816515
Prevalência	0.7032258064516129
Razão de Verossimilhança teste positivo (RV+)	1.9611440906637885
Razão de Verossimilhança teste negativo (RV−)	0.4365707054729516

Fonte: Elaborada pelo autor.

Analisando a curva ROC referente ao *Naive Bayes*, podemos observar o desempenho do algoritmo.

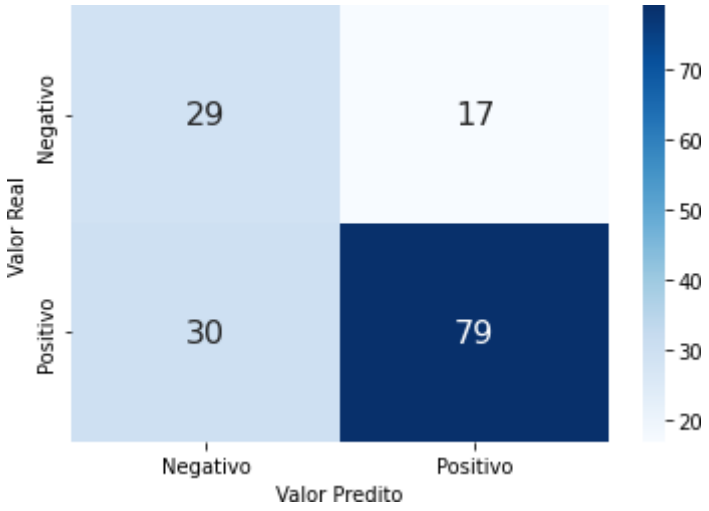
Figura 8 - Curva ROC para o algoritmo Nayve Bayes



Fonte: Elaborada pelo autor.

A curva ROC teve valor de 78% e pode se notar que de forma gráfica o desenvolvimento da mesma, demonstra uma curva com diversas oscilações, abaixo a figura 9 mostra a matriz de confusão do modelo.

Figura 9 - Matriz de confusão para o algoritmo *Nayve Bayes*



Fonte: Elaborada pelo autor.

A matriz de confusão mostra os valores que foram classificados de forma correta ou não, é possível observar que se obteve 29 verdadeiros positivos e 79 verdadeiros negativos, mostrando que o modelo é bastante promissor quando se trata de classificar indivíduos que não possuem alto percentual de gordura corporal.

3.2. *K-Nearest Neighbors* (KNN)

O KNN é um algoritmo muito conhecido no mundo da ciência de dados, tem como objetivo classificar uma nova amostra de dados considerando os vizinhos mais próximos (JIANG et al., 2007), a tabela 9 mostra as métricas do modelo.

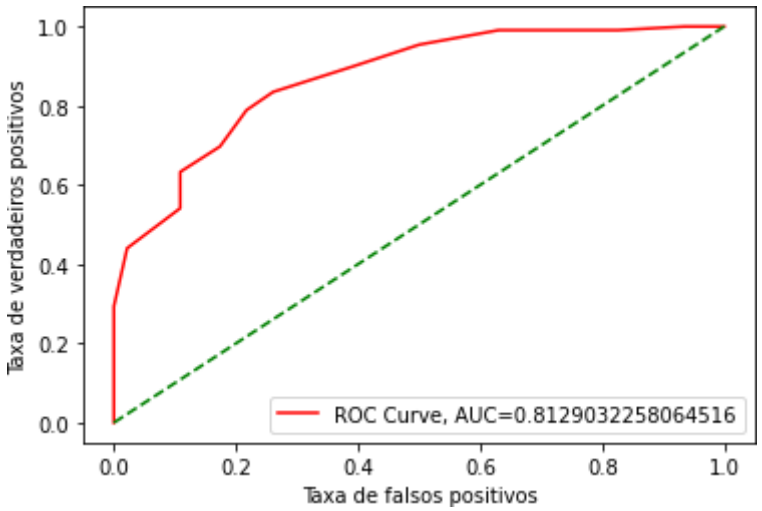
Tabela 9 - Métricas para o algoritmo KNN

Métrica	Valor
Acurácia	0.8129032258064516
Sensibilidade	0.8990825688073395
Especificidade	0.6086956521739131
f1_score	0.8711111111111111
Valor Preditivo Positivo	0.8448275862068966
Valor Preditivo Negativo	0.717948717948718
Taxa de Falso Negativo	0.10091743119266056
Prevalência	0.7032258064516129
Razão de Verossimilhança teste positivo (RV+)	2.2976554536187566
Razão de Verossimilhança teste negativo (RV−)	0.16579292267365653

Fonte: Elaborada pelo autor.

Analisando a curva ROC referente ao KNN, podemos observar o desempenho do algoritmo.

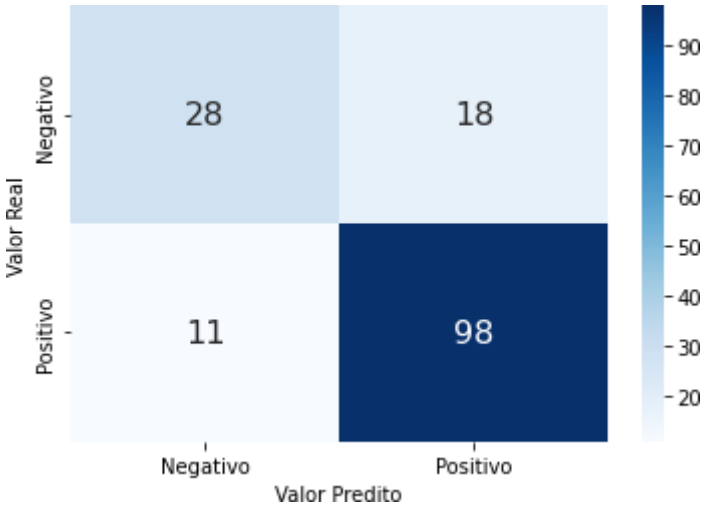
Figura 10 - Curva ROC para o algoritmo KNN



Fonte: Elaborada pelo autor.

A curva ROC teve valor de 81% e pode se notar que de forma gráfica o desenvolvimento da mesma, demonstra uma curva de comportamento variado, abaixo a figura 11 mostra a matriz de confusão do modelo.

Figura 11 - Matriz de confusão para o algoritmo KNN



Fonte: Elaborada pelo autor.

A matriz de confusão mostra os valores que foram classificados de forma correta ou não, é possível observar que se obteve 28 verdadeiros positivos e 98 verdadeiros negativos, mostrando que o modelo mostrou na maior parte do seu desempenho, maior eficiência em classificar verdadeiros negativos.

3.3. Linear SVM

LSVM é uma técnica rápida para treinar máquinas de vetores de suporte (SVMs), com base em uma abordagem iterativa simples (CHAPELLE; SCHOLKOP, 2001).

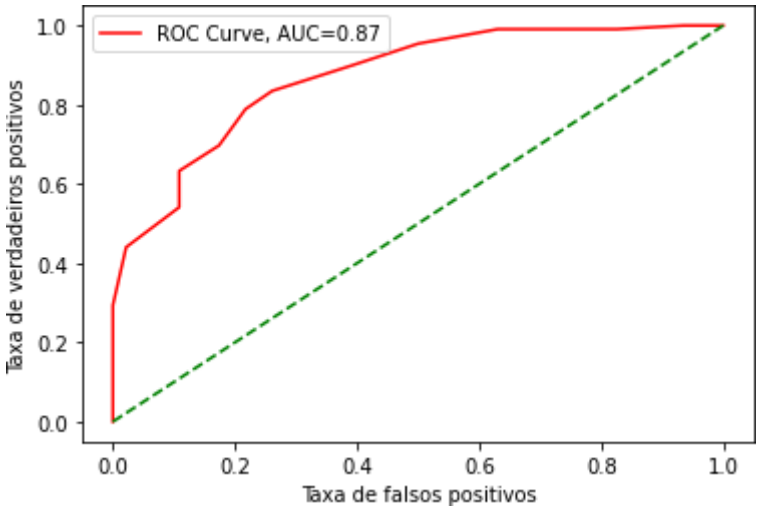
Tabela 10 - Métricas para o algoritmo LSVM

Métrica	Valor
Acurácia	0.7677419354838709
Sensibilidade	0.8807339449541285
Especificidade	0.5
f1_score	0.8421052631578947
Valor Preditivo Positivo	0.8067226890756303
Valor Preditivo Negativo	0.6388888888888888
Taxa de Falso Negativo	0.11926605504587157
Prevalência	0.7032258064516129
Razão de Verossimilhança teste positivo (RV+)	1.761467889908257
Razão de Verossimilhança teste negativo (RV−)	0.23853211009174302

Fonte: Elaborada pelo autor.

Analisando a curva ROC referente ao L-SVM, podemos observar o desempenho do algoritmo equação (CHAPELLE; SCHOLKOP, 2001).

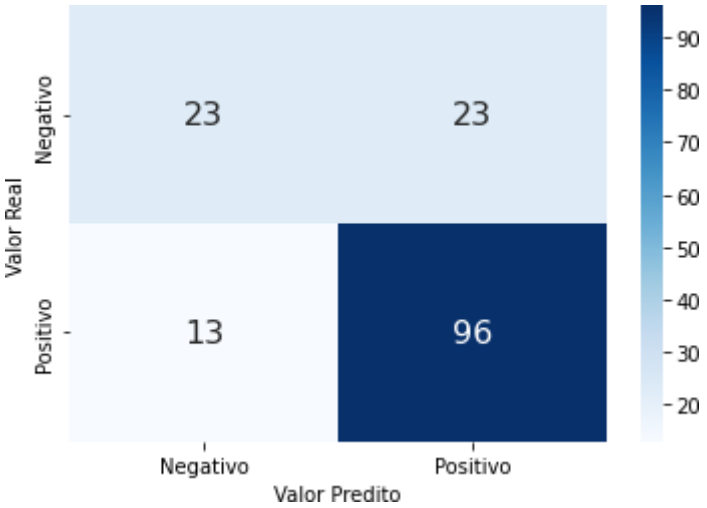
Figura 12 - Curva ROC para o algoritmo L-SVM



Fonte: Elaborada pelo autor.

A curva ROC teve valor de 87% e pode se notar que de forma gráfica o desenvolvimento da mesma, demonstra uma curva que se divide em momentos como crescente e decrescente, abaixo a figura 13 mostra a matriz de confusão do modelo.

Figura 13 - Matriz de confusão para o algoritmo LSVM



Fonte: Elaborada pelo autor.

A matriz de confusão mostra os valores que foram classificados de forma correta ou não, é possível observar que se obteve 23 verdadeiros positivos e 96 verdadeiros negativos, mostrando maior eficiência do modelo em classificar casos negativos de obesidade.

3.4. Kernel SVM

Grande parte do poder de classificação dos SVM vem da escolha do kernel, esses kernels são bastante genéricos (SHAW; ROUTRAY, 2016), fato que faz com tenham diversificadas aplicações.

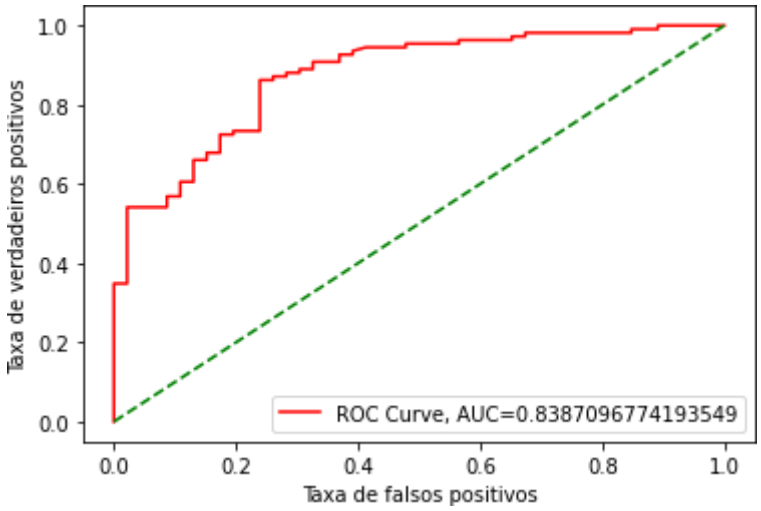
Tabela 11 - Métricas para o algoritmo KSVM

Métrica	Valor
Acurácia	0.8387096774193549
Sensibilidade	0.926605504587156
Especificidade	0.6304347826086957
f1_score	0.8898678414096917
Valor Preditivo Positivo	0.8559322033898306
Valor Preditivo Negativo	0.7837837837837838
Taxa de Falso Negativo	0.07339449541284404
Prevalência	0.7032258064516129
Razão de Verossimilhança teste positivo (RV+)	2.50728548300054
Razão de Verossimilhança teste negativo (RV−)	0.11641885479278709

Fonte: Elaborada pelo autor.

Analisando a curva ROC referente ao K-SVM, podemos observar o desempenho do algoritmo.

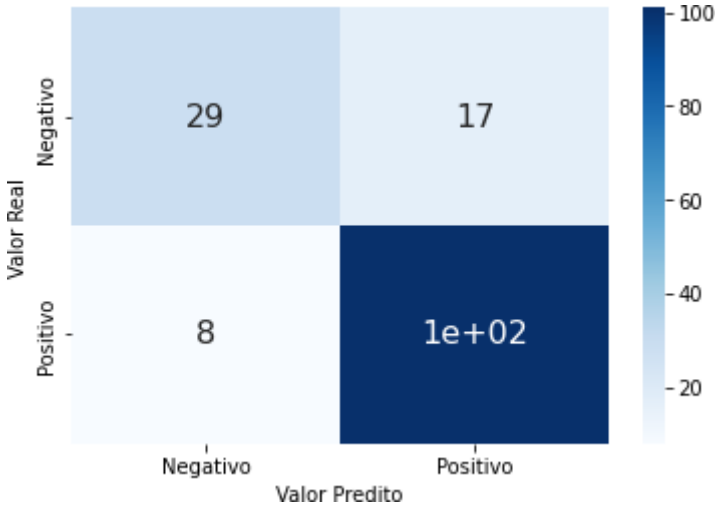
Figura 14 - Curva ROC para o algoritmo KSVM



Fonte: Elaborada pelo autor.

A curva ROC teve valor de 83% e pode se notar que de forma gráfica o desenvolvimento da mesma, demonstra uma curva com diversas mudanças de comportamento, abaixo a figura 16 mostra a matriz de confusão do modelo.

Figura 15 - Matriz de confusão para o algoritmo KSVM



Fonte: Elaborada pelo autor.

A matriz de confusão mostra os valores que foram classificados de forma correta ou não, é possível observar que se obteve 29 verdadeiros positivos e 1e+02 verdadeiros negativos.

3.5. Regressão Logística

A regressão logística se resumo em um modelo matemático que tem o objetivo de modelar usando um conjunto de observações a relação logística com um conjunto de dados (CABRAL, 2013), essa técnica é usada em grande escala no campo do aprendizado de máquinas, abaixo a tabela 11 mostra as métricas obtidas com a aplicação da técnica no conjunto de dados em questão.

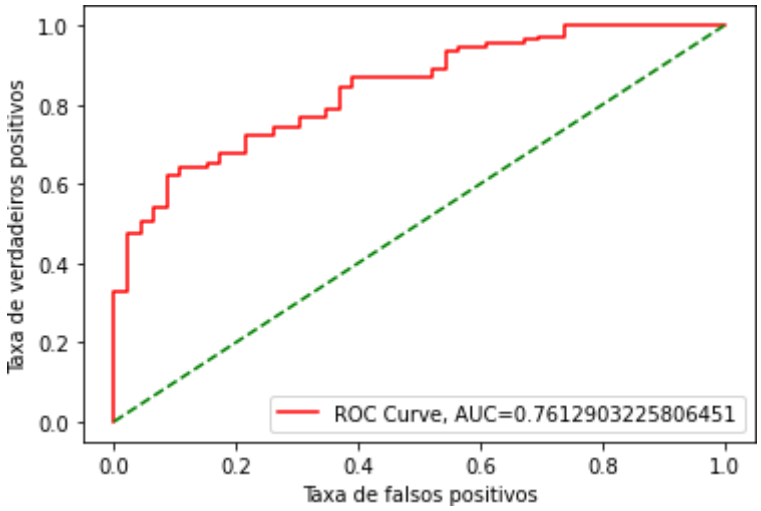
Tabela 12 - Métricas para o algoritmo de Regressão Logística

Métrica	Valor
Acurácia	0.7612903225806451
Sensibilidade	0.8715596330275229
Especificidade	0.5
f1_score	0.8370044052863436
Valor Preditivo Positivo	0.8050847457627118
Valor Preditivo Negativo	0.6216216216216216
Taxa de Falso Negativo	0.12844036697247707
Prevalência	0.7032258064516129
Razão de Verossimilhança teste positivo (RV+)	1.7431192660550459
Razão de Verossimilhança teste negativo (RV−)	0.25688073394495414

Fonte: Elaborada pelo autor.

Analisando a curva ROC referente a regressão logística, podemos observar o desempenho do algoritmo.

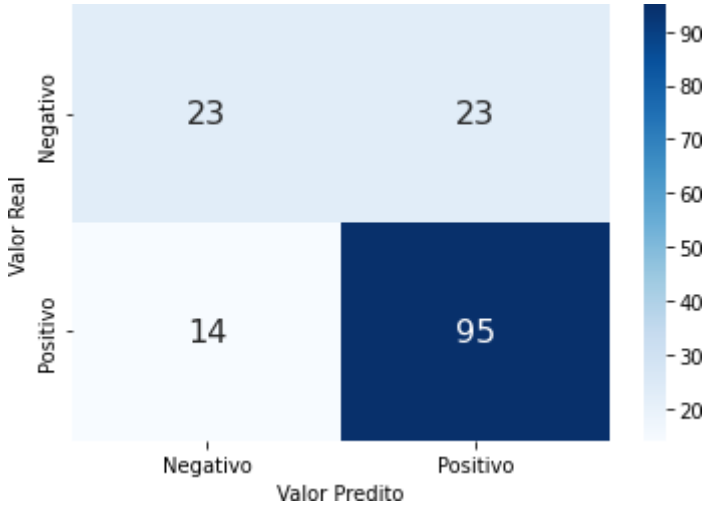
Figura 16 - Curva ROC para o algoritmo de regressão logística



Fonte: Elaborada pelo autor.

A curva ROC teve valor de 76% e pode se notar que de forma gráfica o desenvolvimento da mesma, demonstra uma curva com muitas oscilações, abaixo a figura 18 mostra a matriz de confusão do modelo.

Figura 17 - Matriz de confusão para o algoritmo de regressão logística



Fonte: Elaborada pelo autor.

A matriz de confusão mostra os valores que foram classificados de forma correta ou não, é possível observar que se obteve 23 verdadeiros positivos e 95 verdadeiros negativos, semelhante a outros já avaliados com eficiência em classificar casos de indivíduos com baixo índice de gordura corporal.

5.6 Árvore de Decisão

A árvore de decisão é um algoritmo muito usado no campo do aprendizado de máquinas, é utilizado em problemas de regressão e classificação, abaixo das métricas obtidas com a aplicação do algoritmo.

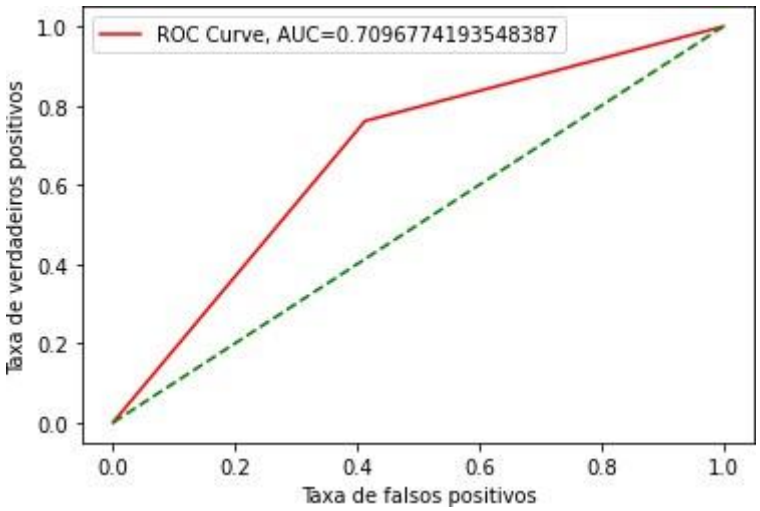
Tabela 13 - Métricas para o algoritmo de árvore de decisão

Métrica	Valor
Acurácia	0.7096774193548387
Sensibilidade	0.7614678899082569
Especificidade	0.5869565217391305
f1_score	0.7867298578199052
Valor Preditivo Positivo	0.8137254901960784
Valor Preditivo Negativo	0.5094339622641509
Taxa de Falso Negativo	0.23853211009174313
Prevalência	0.7032258064516129
Razão de Verossimilhança teste positivo (RV+)	1.8435538387252537
Razão de Verossimilhança teste negativo (RV−)	0.40638803941556234

Fonte: Elaborada pelo autor.

Analisando a curva ROC referente a arvore de decisão, podemos observar o desempenho do algoritmo.

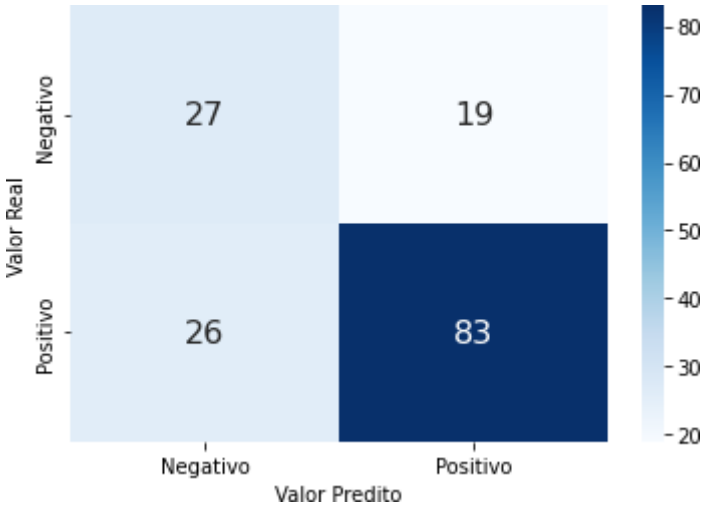
Figura 18 - Curva ROC para o algoritmo de arvore de decisão



Fonte: Elaborada pelo autor.

A curva ROC teve valor de 70% e pode se notar que de forma gráfica o desenvolvimento da mesma, demonstra uma curva com poucas oscilações, abaixo a figura mostra a matriz de confusão do modelo.

Figura 19 - Matriz de confusão para o algoritmo de arvore de decisão



Fonte: Elaborada pelo autor.

A matriz de confusão mostra os valores que foram classificados de forma correta ou não, é possível observar que se obteve 27 verdadeiros positivos e 83 verdadeiros negativos, mostrando um desempenho abaixo de outros já analisados neste trabalho.

4. COMPARAÇÃO DE METRICAS

A tabela 13 mostra as métricas extraídas de cada modelo utilizado neste trabalho.

Tabela 14 - Comparação de métricas

	NB	KNN	L-SVM	RL	D-Tree	K-SVM
Acurácia	0,7	0,82	0,77	0,76	0.71	0,84
Sensibilidade	0,72	0,9	0,88	0,87	0.76	0,93
Especificidade	0,63	0,61	0,5	0,5	0.58	0,63
f1-Score	0,77	0,87	0,84	0,84	0.79	0,89
VPP	0,82	0,84	0,81	0,8	0.81	0,86
VPN	0,49	0,72	0,64	0,62	0.51	0,78
TFN	0,28	0,1	0,12	0,13	0.24	0,07
RV+	1,96	2,23	1,76	1,74	1.84	2,51
RV-	0,44	0,17	0,24	0,26	0.41	0,12
AUC	0,78	0,87	0,87	0,84	0.67	0,87
Prevalência	0,7	0,7	0,7	0,7	0,7	0,7

Fonte: Elaborada pelo Autor

Alguns algoritmos tiveram o desempenho similar, mas fazendo uma análise das métricas é possível notar que o algoritmo teve maior desempenho em muitos quesitos é o K-SVM, portanto, ele se mostrou mais eficaz como ferramenta para auxiliar na medição de gordura corporal, este traballho teve como resultado o artigo: “Comparação de técnicas de aprendizado de máquina para previsão de índices de gordura corporal em adolescentes” públicado ao fim de todo desenvolvimento.

5. DISCURSÃO

Como já dito o excesso de peso tem relação direta com o aparecimento de diversas doenças cardiovasculares além de distúrbios metabólicos, essas comorbidades atingem boa parte da população mundial, portanto, este trabalho consiste em fazer uma comparação sobre algoritmos aplicados a uma base, as comorbidades que são consequências do excesso de peso não precisam de muito tempo para apresentar em seus sintomas sendo assim na adolescência muitas dessas doenças apresentam consequências que podem comprometer gravemente a saúde do indivíduo em questão, segundo existe uma necessidade de investigação que tenha o objetivo de realizar o diagnóstico precoce do excesso de gordura corporal, pois, uma vez que seja diminuído em um por cento a prevalência de obesidade em adolescentes isso poderia gerar uma economia de 586,3 milhões de dólares em gastos futuros com a saúde de indivíduos adultos.

Dessa forma é evidente que reduzir a obesidade traz grandes benefícios da sociedade de modo geral, este estudo analisou os resultados obtidos com a aplicação de cinco algoritmos de aprendizado de máquina estes são, esses algoritmos tiveram como atributos de entrada parâmetros clínicos com a finalidade de fazer a predição da presença de elevados níveis de gordura corporal em adolescentes em comparação também há indicadores antropométricos IMC e RCE, o objetivo principal deste trabalho é o comparativo entre os algoritmos em questão buscando encontrar o que possui melhor performance para o problema, dentre os algoritmos utilizados todos tiveram resultados satisfatórios, o fato de que este trabalho apresenta uma técnica que usa somente recursos computacionais a torna uma alternativa vantajosa tendo em vista que também foram testadas outras técnicas computacionais onde foi feito um comparativo e selecionada a que tem maior desempenho.

O IMC é um famoso indicador antropométrico largamente utilizado em estudos epidemiológicos para avaliação de excesso de gordura em indivíduos e faixa etária geral, porém, seu uso é bastante criticado pelo fato de não correlacionar a composição de massa gorda e massa magra e a distribuição da gordura corporal, isso pode acontecer como, por exemplo, como atletas que podem possuir um alto IMC devido ao seu grande aumento de massa muscular, em um estudo realizado em São Paulo

aplicado a frequentadores de uma academia, 46 adultos foram estudados, porém, vários deles foram classificados erroneamente pelo cálculo do IMC, segundo o IMC apresenta uma espécie de cidade de 92% e uma baixa sensibilidade em torno de 50%, quando se trata de detectar a obesidade com base no percentual de gordura corporal, este fator se torna preocupante, pois, o uso de um teste com baixa sensibilidade em uma prática clínica pode gerar diagnósticos e consequentemente pode trazer atraso no tratamento de comorbidades que são associadas ao acesso de cultura corporal assim como também pode prejudicar a implementação de medidas para que se possa combater a mesma.

O método proposto neste trabalho apresenta uma sensibilidade maior ao IMC e RCE, dessa forma é válida a utilização de métodos computacionais que apresentam ótimos resultados como uma alternativa a medição de gordura corporal em um indivíduo, se levando em consideração a política nacional de promoção de saúde todo o método que se mostre eficaz e que tenha baixo custo com a finalidade de resolver questões sociais de saúde deve ser levada em, pois, este método tem como objetivo prevenir situações graves e reduzir a morte prematura de indivíduos que podem vir a sofrer com as doenças em questão, é de extrema importância se ressaltar o desenvolvimento de um método para triagem da obesidade, pois, os números de indivíduos que sofrem e podem vir a sofrer com comorbidades decorrentes deste fator são alarmantes.

6. CONCLUSÕES

Conclui que este trabalho é de extrema importância, pois, mostra como a tecnologia pode ser usada nas mais diversas áreas e trazer benefícios reais para a sociedade, em geral, nesse caso específico a obesidade é um mal que atinge um número expressivo de pessoas no mundo inteiro, além de suas complicações próprias ela também abre a porta para outras diversas enfermidades, sendo assim qualquer método que possa ajudar a identificar previamente qualquer sinal de obesidade é bem-vindo, neste trabalho foi utilizado o aprendizado de máquina para que através das medidas de um indivíduo se possa calcular o índice de gordura do mesmo a proposta é de extrema importância e muito interessante para os padrões de exames

utilizados hoje em dia, pois, não é um procedimento invasivo utiliza uma tecnologia que tende a obter melhoras e assim trazer resultados cada vez mais assertivos.

O modelo baseado em teve desempenho superior aos outros e também a indicadores antropométricos que são utilizados hoje em dia em vários aspectos como, por exemplo, poder discriminatório, sensibilidade, acurácia, sendo assim uma ótima opção para auxiliar o profissional de saúde na predição de gordura corporal de um indivíduo, este modelo pode ser implementado em forma de ferramenta de forma que se torne completamente usual para qualquer indivíduo sem a necessidade de conhecimentos aprofundados em ciência de dados fazendo com que se torna uma ferramenta com alta usabilidade e que possa realmente ajudar no combate às comorbidades geradas pelo excesso de gordura e assim mostrar valor ajudando a sociedade, em geral, no combate a uma das doenças mais graves do século.

7. REFERÊNCIAS

JÚNIOR, CARLOS MAGNO SOUSA. **DESENVOLVIMENTO DE UM SISTEMA PARA TRIAGEM DE ADOLESCENTES OBESOS UTILIZANDO VARIÁVEIS CLÍNICAS**. Orientador: Ewaldo Eder Carvalho Santana. 2019. 68 f. Tese (Pós-Graduação) - UNIVERSIDADE FEDERAL DO MARANHÃO, São Luís - MA, 2019.

WANKE, Bruna dos Santos Lazéra; COSTA, Vivian Oliveira; PINA, Aloísio Carlos; FILHO, Armando Carlos de Pina. APLICAÇÃO DO CLASSIFICADOR NAIVE BAYES PARA IDENTIFICAÇÃO DE FALHAS DE UM MANIPULADOR ROBÓTICO. **ABCM Symposium Series in Mechatronics**, [*s. l.*], v. 6, p. 888-895, 2014.

WANDERLEY, Emanuela Nogueira; FERREIRA, Vanessa Alves. Obesidade: uma perspectiva plural. **Ciência & Saúde Coletiva**, Diamantina - MG, ano 2010, p. 185-194, 2010.

REZENDE, Fabiane Aparecida Canaan; ROSADO, Lina Enriqueta Frandsen Paez Lima; FRANCESCHINN, Sylvia do Carmo Castro; ROSADO, Gilberto Paixão; RIBEIRO, Rita de Cássia Lanes. 90 Rev Bras Med Esporte – Vol. 16, Mar/Abr, 2010 CLÍNICA MÉDICA DO EXERCÍCIO E DO ESPORTE Aplicabilidade do Índice de Massa Corporal na Avaliação da Gordura Corporal. **CLÍNICA MÉDICA DO EXERCÍCIO E DO ESPORTE**, Mato Grosso - BR, 2010.

ZEBALLOS, Luiza; NASCIMENTO, Érica; GRANHA, Eugênia; BERTI, Marina; BERBERT, Débora; NACIF, Marcia. AVALIAÇÃO DA COMPOSIÇÃO CORPORAL TOTAL E SEGMENTAR DE ALUNOS DO CURSO DE NUTRIÇÃO PELA DENSITOMETRIA POR DUPLA EMISSÃO DE RAIOS - X. **Revista Brasileira de Obesidade, Nutrição e Emagrecimento**, [*s. l.*], ano 2021, p. 914-919, 2021.

BRITTO, Eleonora Peixoto; MESQUITA, Evandro Tinoco. Bioimpedância Elétrica Aplicada à Insuficiência Cardíaca. **Revista SOCERJ**, [*s. l.*], p. 178-183, 2008.

PELING, International Journal of Engineering and Emerging Technology, Vol. 2, No. 1, January—June 2017 (p-issn: 2579-5988, e-issn: 2579-597X) 53 Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm Ida Bagus Adisimakrisna; ARNAWAN, I Nyoman; ARTHAWAN, I Putu Arich;

JANARDANA, IGN. Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm. **International Journal of Engineering and Emerging Technology**, [*s. l.*], ano 2017, v. 2, p. 53-57, 2017

RODRIGUES, UNIVERSIDADE DA BEIRA INTERIOR Ciências Modelo de Regressão Linear e suas Aplicações Sandra Cristina Antunes. **Modelo de Regressão Linear e suas Aplicações**. Covilhã: [*s. n.*], 2012 2012.

ZHANG, Shichao; ZONG, Ming; ZHU, Xiaofeng. Efficient kNN Classification With Different Numbers of Nearest Neighbors. **IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS**, [*s. l.*], v. 29, p. 1774-1784, 2018.

YAO, Yukai; LIU, Yang; YU, Yongqing; XU, Hong; LV, Weiming; LI, Zhao; CHEN, Xiaoyun. K-SVM: An Effective SVM Algorithm Based on K-means Clustering. **JOURNAL OF COMPUTERS**, Lanzhou,China, ano 2013, v. 8, n. 10, p. 2632-2638, 2013.

JIANG, Liangxiao; WANG, Dianhong; CAI, Zhihua; YAN, Xuesong. Survey of Improving Naive Bayes for Classification. **Lecture Notes in Computer Science**, Wuhan, China, ano 2007, v. 4632, n. 12, p. 134-145, 2007.

TAVARES, Telma Braga; NUNES, Simone Machado; SANTOS, Mariana de Oliveira. Obesidade e qualidade de vida: revisão da literatura: Obesity and quality of life: literature review. **Revista Med Minas Gerais** , [*s. l.*], ano 2010, p. 359-365, 2010.

WANDERLEY, Emanuela Nogueira; FERREIRA, Vanessa Alves. Obesidade: uma perspectiva plural: Obesity: a perspectiva plural perspective. **Ciência & Saúde Coletiva**, [*s. l.*], ano 2010, p. 185-194, 2010.

JORGE, Aikes Junior. Study of the influence of similarity measures in Time Series Prediction with the kNN-TSP algorithm. *In*: JORGE, Aikes Junior. **Estudo da influência de diversas medidas de similaridade na previsão de séries temporais utilizando o algoritmo KNN-TSP**. 2012. Dissertação (Mestrado) - Universidade Estadual do Oeste do Paraná, Foz do Iguaçu, 2012. f. 129.

PACHECO, André. **K vizinhos mais próximos - KNN**. [*S. l.*], 2017. Disponível em: http://computacaointeligente.com.br/algoritmos/k-vizinhos-mais-proximos/. Acesso em: 17 jun. 2022.

FARIA, Maurício Mendes. **Detecção de Intrusões em Redes de Computadores com Base nos Algoritmos KNN, K-Means++ e J48**. 2016. 146 f. Dissertação de Mestrado em Ciência da Computação (Mestrado) - Faculdade Campo Limpo Paulista, [*S. l.*], 2016.

LIMA, RODRIGO LUCIO. **AVALIAÇÃO DO ALGORITMO SVM NA DETECÇÃO DE COMPORTAMENTOS SUSPEITOS EM CENAS DE VÍDEO**. PONTA GROSSA: [*s. n.*], 2014. 60 p.

SOUSA, Alex. **Algoritmo SVM (Máquinade Vetores de Suporte) a partir de exemplos e código (Python e R)**: Data Science, Machine Learning. [*S. l.*], 2019. Disponível em: https://blogdozouza.wordpress.com/2019/04/10/algoritmo-svm-maquina-de-vetores-de-suporte-a-partir-de-exemplos-e-codigo-python-e-r/. Acesso em: 24 jun. 2022.

ARAÚJO, Kevin Martins. UTILIZAÇÃO DO ALGORITMO DE MÁQUINA DE VETORES DE SUPORTE (SVM) PARA PREDIÇÃO DE DADOS CLIMÁTICOS. *In*: ARAÚJO, Kevin Martins. **UTILIZAÇÃO DO ALGORITMO DE MÁQUINA DE VETORES DE SUPORTE (SVM) PARA PREDIÇÃO DE DADOS CLIMÁTICOS**. 2015. Trabalho de Conclusão de Curso (Graduação) - Centro Universitário Luterano de Palmas, Palmas - TO, 2015. f. 96.

SELL, Sair. Utilização da regressão linear como ferramenta de decisão na gestão de custos. **X Congresso Internacional de Custos**, Florianópolis, SC, Brasil, ano 2005, p. 1-14, 2005.

FONSECA, Vania de Matos; SICHIER, Rosely; VEIGA, Glória Valéria. Fatores associados à obesidade em adolescentes: Factors associated with obesity among adolescents. **Revista de Saúde Pública**: Journal of Public Health, Rio de Janeiro - Brasil, ano 1988, ed. 32, p. 541-548, 1988.

SHAW, Laxmi; ROUTRAY, Aurobinda. A Critical Comparison Between SVM and k-SVM in the Classification of Kriya Yoga Meditation State-allied EEG. **IEEE**

International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), [*s. l.*], p. 134-138, 2016.

RODRIGUEZ-AÑEZ, Ciro Romelio. A ANTROPOMETRIA E SUA APLICAÇÃO NA ERGONOMIA: ANTHROPOMETRY AND IT APPLICATION IN ERGONOMICS. **Revista Brasileira de Cineantropometria & Desempenho Humano**, [*s. l.*], v. 3, p. 102-108, 2001.

DALLASTELLA, DJULIANO GUSTAVO. COMPARAÇÃO DO CONSUMO ALIMENTAR E IMC ENTRE ESCOLARES DA REDE PÚBLICA E PARTICULAR DE ENSINO DA CIDADE DE CURITIBA, PR. *In*: DALLASTELLA, DJULIANO GUSTAVO. **COMPARAÇÃO DO CONSUMO ALIMENTAR E IMC ENTRE ESCOLARES DA REDE PÚBLICA E PARTICULAR DE ENSINO DA CIDADE DE CURITIBA, PR**. 2006. Monografia (Bacharelado) - Universidade Federal do Paraná, CURITIBA-PR, 2006. f. 44.

PEREIRA, MANOELA. ASSOCIAÇÃO DE PADRÕES ALIMENTARES EM IDOSOS LONGEVOS E FAIXAS DE IMC. *In*: PEREIRA, MANOELA. **ASSOCIAÇÃO DE PADRÕES ALIMENTARES EM IDOSOS LONGEVOS E FAIXAS DE IMC**. 2022. Trabalho de Conclusão de Curso (Bacharelado) - Universidade Federal do Paraná, Porto Alegre-RS, 2022. f. 30.

VIEIRA, Sarah Aparecida; RIBEIRO, Andréia Queiroz; HERMSDORFF, Helen Hermana Miranda; PEREIRA, Patrícia Feliciano; PRIORE, Silvia Eloiza; FRANCESCHINI, Sylvia do Carmo Castro. ÍNDICE RELAÇÃO CINTURA-ESTATURA PARA PREDIÇÃO DO EXCESSO DE PESO EM CRIANÇAS. **Rev. paul. pediatra.** , [*s. l.*], p. 52-59, 2017.

EICKEMBERG, Michaela; OLIVEIRA, Carolina Cunha; RORIZ, Anna Karla Carneiro; SAMPAIO, Lílian Ramos. Bioimpedância elétrica e sua aplicação em avaliação nutricional. **Revista de Nutrição**, [*s. l.*], p. 883-895, 2011.

CRISPILHO, SHIRLEY FERRAZ. **Impacto do distúrbio mineral e ósseo da doença renal crônica na perda de acurácia da bioimpedância**. 2019. 65 f. Dissertação (Mestrado) - Universidade Nove de Julho, São Paulo, 2019.

CHAPELLE, Olivier; SCHOLKOP, Bernhard. Incorporating Invariances in Nonlinear Support Vector Machines. **NIPS**, Paris - França, ano 2001, 2001.

GHOSH, Sourish; DASGUPTA, Anasuya; SWETAPADMA, Aleena. A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification. **International Conference on Intelligent Sustainable Systems (ICISS 2019)**, [*s. l.*], p. 24-29, 2019.

BAUER, Lidiane. **ESTIMAÇÃO DO COEFICIENTE DE CORRELAÇÃO DE SPEARMAN PONDERADO**. 2007. DISSERTAÇÃO DE MESTRADO (Pós-Graduação) - UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, [*S. l.*], 2007.

CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. *In*: CABRAL, Cleidy Isolete Silva. **Aplicação do Modelo de Regressão Logística num Estudo de Mercado**. Orientador: João José Ferreira Gomes. 2013. Dissertação (Mestrado) - Universidade de Lisboa, [*S. l.*], 2013. f. 59.