# Métodos de Classificação
# Aprendizagem de Máquina
# 2024

**Leonardo B. Balan de Oliveira**[1], **David A. Brocardo**[2]

[1]Centro de Ciências Exatas e Tecnológicas
Campus de Cascavel - UNIOESTE
Caixa Postal 801 – 85.814-110 – Cascavel – PR – Brazil

{leonardo.oliveira23,david.brocardo}@unioeste.br

***Abstract.*** *This paper presents the application of five Machine Learning algorithms (KNN, Decision Tree, Naive Bayes, SVM and MLP) for the classification of obesity levels in a database consisting of 2111 instances, 16 attributes and 7 classes. The data, which is mostly categorical, was analyzed using different parameters between the ML algorithms. Decision Tree obtained the best accuracy (94.70%), standing out for its efficiency in dealing with categorical variables. Multiple classifier systems were also implemented, with the Sum Rule achieving 91.40% accuracy.*

***Resumo.*** *Este trabalho apresenta a aplicação de cinco algoritmos de Aprendizagem de Máquina (KNN, Árvore de Decisão, Naive Bayes, SVM e MLP) para a classificação dos níveis de obesidade em uma base de dados composta por 2111 instâncias, 16 atributos e 7 classes. Os dados, em sua maioria categóricos, foram analisados usando diferentes parâmetros entre os algoritmos de AM. A Árvore de Decisão obteve a melhor acurácia (94,70%), destacando-se pela eficiência em lidar com variáveis categóricas. Sistemas de Múltiplos Classificadores também foram implementados, com a Regra da Soma, que alcançou 91,40% de acurácia.*

## 1. Base de Dados Escolhida

Em primeiro plano, para a realização do trabalho, é necessário uma base de dados. Realizamos uma pesquisa seguindo as especificações estabelecidas, onde a base precisa ter pelo menos 1000 instâncias e permitir a realização de uma tarefa de classificação. Além disso, buscamos em específico por bases que não tem valores ausentes, eliminando assim a necessidade de tratamentos preliminares nos dados. Chegamos a 4 bases interessantes, e, com o auxílio do professor, demos o veredito final em qual utilizaríamos.

Escolhemos a base *Estimation of Obesity Levels Based On Eating Habits and Physical Condition* do [UCI ML Repository 2019]. Ela contém a estimativa dos níveis de obesidade em indivíduos dos países do México, Peru e Colômbia, com base em seus hábitos alimentares e condição física. Mais informações podem ser obtidas pelo artigo publicado sobre a base de dados [Palechor and Manotas 2019]. O Dataset tem por características:

- Valores multivariados (de diversos tipos);
- É da área da Saúde e Medicina;

- Área específica: Obesidade e Risco cardiovascular;
- 2111 instâncias;
- 16 características/atributos;
- 7 Classes;

### 1.1. Atributos

Em relação aos 16 atributos da base de dados, temos a *Tabela 1*.

**Tabela 1. Descrição dos Atributos**

| Atributo | Tipo | Valores Possíveis | Descrição |
|---|---|---|---|
| Gender | Categórico | 0- Female, 1- Male | Sexo da pessoa |
| Age | Float | 14 até 61 anos | Idade da pessoa |
| Height | Float | 1.45m até 1.98m | Altura da pessoa |
| Weight | Float | 39kg até 173kg | Peso da pessoa |
| family_history_with overweight | Binário | 0 - não, 1 - sim | Membro da família sofreu/sufre de excesso de peso? |
| FAVC | Binário | 0 - não, 1 - sim | Come alimentos altamente calóricos com frequência? |
| FCVC | Inteiro | 1 - Nunca, 2 - Às vezes, 3 - Sempre | Costuma comer vegetais nas refeições? |
| NCP | Float | 1, 2, 3 ou 4 - Mais de 3 | Refeições principais diárias |
| CAEC | Categórico | 0 - Não, 1 - Às vezes, 2 - Frequentemente, 3 - Sempre | Come entre as refeições? |
| SMOKE | Binário | 0 - não, 1 - sim | Fuma? |
| CH2O | Float | 1 - Menos de um litro, 2 - Entre 1 e 2 L, 3 - Mais de 2 L | Quantidade de água ingerida diariamente |
| SCC | Binário | 0 - não, 1 - sim | Monitora as calorias ingeridas diariamente? |
| FAF | Float | 0 - Não faço, 1 - 1 ou 2 dias, 2 - 2 ou 4 dias, 3 - 4 ou 5 dias | Frequência de atividade física |
| TUE | Inteiro | 0 - 0–2 horas, 1 - 3–5 horas, 2 - Mais de 5 horas | Tempo de uso de dispositivos tecnológicos |
| CALC | Categórico | 0 - não; 1 - Às vezes; 2 - Frequentemente; 3 - Sempre | Frequência de consumo de álcool |
| MTRANS | Categórico | 0 - Caminhada; 1 - Transporte público; 2 - Automóvel; 3 - Motocicleta; 4 - Bicicleta | Meio de transporte usual |

**Autoria própria**

### 1.2. Atributos Categóricos

Tem-se na base de dados atributos categóricos, sendo preciso o convertimento deles para valores inteiros (melhor de trabalhar e manipular). Os atributos categóricos da base podem ser vistos na própria Tabela 1, onde eles já foram devidamente tratados, adicionando valores para representar cada uma das categorias. Todavia, traduzimos para o português tudo que está na tabela acima para melhor entendimento, mas no banco de dados original os atributos estão em inglês, consideraremos assim para uma exemplificação. Por exemplo o atributo *MTRANS*, onde os valores Walking, Public_Transportation, Automobile, Motorbike e Bike foram convertidos em toda a tabela para: 0, 1, 2, 3, 4, respectivamente.

Utilizamos um algoritmo básico para isso, o mesmo se encontra na pasta entregue na tarefa do teams deste trabalho.

### 1.3. Classes

Para as 7 classes, temos a váriavel de saída *NObeyesdad*, que concluirá o nível de obesidade da pessoa. Os níveis são:

- 0 - Insufficient Weight (Abaixo do peso);
- 1 - Normal Weight (Peso normal);
- 2 - Overweight Level I (Sobrepeso Nível I)
- 3 - Overweight Level II (Sobrepeso Nível II)
- 4 - Obesity Type I (Obesidade Tipo I)
- 5 - Obesity Type II (Obesidade Tipo II)
- 6 - Obesity Type III (Obesidade Tipo III)

### 1.4. Característica dos dados

Interessante resaltar que 77% dos dados que compõem o Dataset foram gerados sinteticamente usando a ferramenta Weka e o filtro SMOTE, e 23% dos dados foram coletados diretamente dos usuários por meio de uma plataforma web [Palechor and Manotas 2019].

## 2. Algoritmos de AM Utilizados

Os algoritmos de Aprendizagem de Máquina para a tarefa de Classificação utilizados foram: o K Vizinhos mais próximos (KNN), Árvore de Decisão (AD), Naive Bayes (NB), Máquina de Vetor de Suporte (SVM) e o Multilayer Perceptron (MLP).

### 2.1. Variação dos Parâmetros

Abaixo segue a faixa de valores que variamos os parâmetros dos algoritmos:

**KNN**

- K: *[1 a 50]*
- distance: *[distance e uniform]*

**Árvore de Decisão**

- criterion: *[entropy e gini]*
- max_depth: *[1 a 11]*
- min_samples_split: *[2 a 16]*

- min_samples_leaf: *[1 a 11]*
- splitter: *[best e random]*

**SVM**

- kernel: *[linear, poly, rbf e sigmoid]*
- custo: *[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]*

**NB**

- *Sem parâmetros*

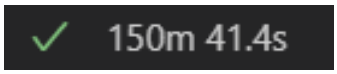
**MLP**

- hidden_layer_sizes: *[5, 6, 10, 12]*
- activation: *[identity, logistic, tanh e relu]*
- max_iter: *[50, 100, 150, 300, 500, 1000]*
- learning_rate: *[constant, invscaling e adaptive]*

## 3. Análise / Resultados

Para a coleta dos dados, foram desenvolvidos algoritmos utilizando a linguagem de programação Python, e executados na ferramenta Visual Studio Code. A escolha dessa ferramenta deu-se pela possibilidade de realizar a execução localmente, minimizando transtornos como a queda de conexão com a internet, por exemplo. Os algoritmos foram executados três vezes, apresentando tempos de execução consistentes entre as rodadas. A Figura 1 exibe o tempo da última execução, sendo também os dados apresentados neste artigo referentes a essa execução.

**Figura 1. Tempo de execução**




Fonte: Autoria própria

Com os resultados de acurácia visualizados abaixo, notamos uma hierarquização em questão de acerto entre os classificadores. Sendo esses os melhores valores:

**Abordagem monolítico**

- 1º **AD**: 0.9470%
- 2º **KNN**: 0.9015%
- 3º **SVM**: 0.8807%
- 4º **MLP**: 0.8693%
- 5º **NB**: 0.6458%

No tópico *4. Análise Estatística*, podemos visualizar todas as acurácias em relação as 20 repetições, bem como a média delas.

Os parâmetros de cada uma dessas melhores execuções foram:

- **AD**
  Criterion: *entropy*; Max_Depth: *8*; Min_Samples_Split: *1*; Min_Samples_Leaf: *5*; Splitter: *best*;
- **KNN**
  Melhor_K: *1*; Melhor_métrica: *distance*;
- **SVM**
  Kernel: *linear*; C: *1.0*;
- **MLP**
  Hidden_Layers_size: *10*; Learning_Rate: *adaptive*; Max_Iterations: *1000*; Activation: *logistic*;
- **NB**
  *Sem parâmetros*.

Acreditamos que a Árvore de Decisão acerta mais / tem maior acurácia que os outros classificadores pois tem uma maior capacidade de lidar de forma eficiente com variáveis categóricas, e a base de dados é predominantemente categórica. As AD são especialmente eficazes em dividir o espaço de atributos em regiões homogêneas em relação à classe alvo, aproveitando a natureza discreta dos dados categóricos para construir critérios de decisão que maximizam a separação entre as classes [Suthaharan 2016]. Isso permite que o modelo capture relações complexas entre os atributos e as classes, resultando em uma acurácia superior.

Já o Naive Bayes, por outro lado, teve uma péssima acurácia nos múltiplos testes feitos. Pensamos que isso ocorre devido ao fato de que, em uma base de dados predominantemente categórica, o desempenho do NB é diretamente afetado pela sua principal suposição de independência condicional entre os atributos [Webb et al. 2010]. Na prática, essa suposição raramente se sustenta, especialmente em conjuntos de dados categóricos onde os atributos podem estar altamente correlacionados, como é o caso desta base.

O K-Nearest Neighbors (KNN), Support Vector Machine (SVM) e Multi-Layer Perceptron (MLP) tendem a se manter equilibrados na acurácia final (melhor precisão obtida entre todas as repetições). Possivelmente, esse desempenho uniforme pode ser atribuído à capacidade desses modelos de lidar de forma eficaz com diversos tipos de variáveis (variaveis categóricas entram aqui) e suas interações complexas.

O KNN, por exemplo, é robusto em identificar padrões locais e classificar instâncias com base em seus vizinhos mais próximos, o que se mostra útil em dados categóricos com relações não lineares. O MLP, com sua estrutura de rede neural, é capaz de capturar complexidades nas relações entre atributos e classes através de suas camadas ocultas, enquanto o SVM, com sua capacidade de encontrar boas margens de separação, lida bem com dados que podem ser não linearmente separáveis [Brun 2024].

**Sistemas de Múltiplos Classificadores**

- 1º **Regra do Soma**: 0.9140%
- 2º **Voto Majoritário**: 0.9035%
- 3º **Borda Count**: 0.9035%

Podemos visualizar acima as médias dos múltiplos classificadores, onde o método da Regra da Soma, que consiste em somar as pontuações atribuídas por cada classificador a cada classe e escolher a classe com a maior soma, demonstrou um desempenho melhor em relação aos outros 2. Os resultados variaram pouco nas 20 repetições, com valores de 0,8845 a 0,9394. Um desempenho consistente e robusto.

O sistema de Voto Majoritário, também mostrou um desempenho robusto, ele se baseia no princípio de que a classe mais frequentemente escolhida pelos classificadores é a mais provável. Suas acurácias variaram de 0,8864 a 0,9394.

O Borda Count não fugiu muito das médias dos 2 acima, no caso obteve a mesma média de acurácia do Voto Majoritário, ele utiliza uma abordagem de pontuação onde cada classificador atribui uma pontuação a cada classe e a classe com a maior pontuação total é escolhida, apresentou uma acurácia média de 0,9035, com uma variação de 0.8864 a 0,9394).

## 4. Análise Estatística

**Tabela 2. Resultado das execs dos Sistemas Monolíticos**

| Repetição | KNN | NB | AD | SVM | MLP |
|---|---|---|---|---|---|
| 1 | 0.8826 | 0.6174 | 0.9470 | 0.8807 | 0.8674 |
| 2 | 0.8769 | 0.5682 | 0.9394 | 0.8561 | 0.8390 |
| 3 | 0.8561 | 0.5928 | 0.9280 | 0.8466 | 0.8598 |
| 4 | 0.8731 | 0.5852 | 0.8977 | 0.8523 | 0.8333 |
| 5 | 0.8655 | 0.6061 | 0.9242 | 0.8258 | 0.8220 |
| 6 | 0.8864 | 0.6061 | 0.9375 | 0.8750 | 0.8523 |
| 7 | 0.8902 | 0.5682 | 0.9470 | 0.8390 | 0.7973 |
| 8 | 0.9015 | 0.6117 | 0.9432 | 0.8523 | 0.8674 |
| 9 | 0.8674 | 0.5890 | 0.9167 | 0.8447 | 0.7666 |
| 10 | 0.8693 | 0.5777 | 0.9451 | 0.8561 | 0.8523 |
| 11 | 0.8769 | 0.6061 | 0.9186 | 0.8390 | 0.8333 |
| 12 | 0.8580 | 0.5530 | 0.9394 | 0.8580 | 0.8485 |
| 13 | 0.8864 | 0.6212 | 0.9432 | 0.8542 | 0.8447 |
| 14 | 0.8750 | 0.5777 | 0.9223 | 0.8655 | 0.8239 |
| 15 | 0.8731 | 0.5795 | 0.9148 | 0.8580 | 0.8693 |
| 16 | 0.8826 | 0.6458 | 0.9318 | 0.8542 | 0.8295 |
| 17 | 0.8902 | 0.5530 | 0.9375 | 0.8750 | 0.8674 |
| 18 | 0.8655 | 0.5833 | 0.9242 | 0.8523 | 0.8277 |
| 19 | 0.8580 | 0.5890 | 0.9432 | 0.8295 | 0.7557 |
| 20 | 0.8258 | 0.6439 | 0.9451 | 0.8674 | 0.8144 |
| **Média** | 0.8730 | 0.5938 | 0.9323 | 0.8541 | 0.7741 |

**Fonte: Autoria própria**

**Tabela 3. Resultado das execs dos Sistemas de Múltiplos Classificadores**

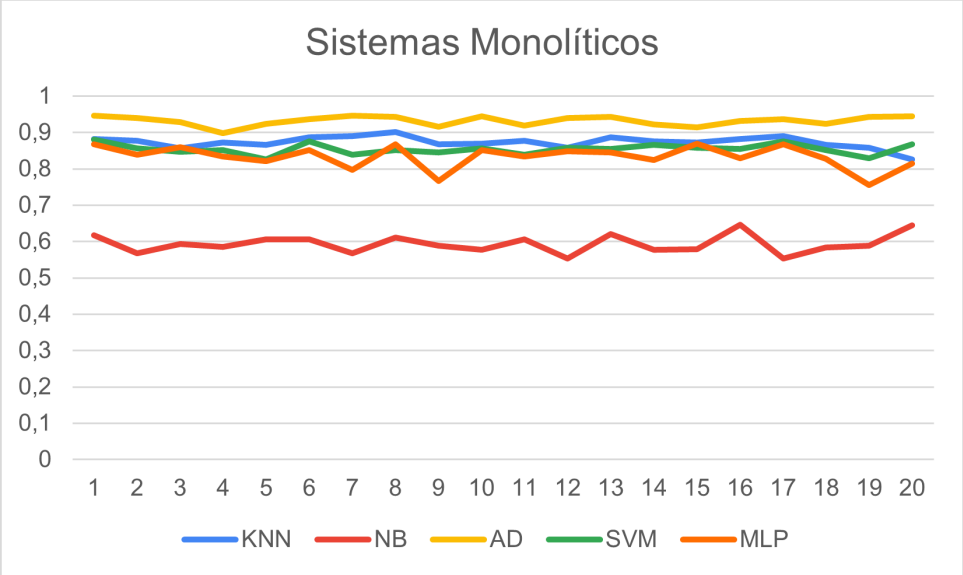
| Repetição | Regra do Soma | Voto Majoritário | Borda Count |
|---|---|---|---|
| 1 | 0.9394 | 0.9394 | 0.9394 |
| 2 | 0.9110 | 0.9091 | 0.9072 |
| 3 | 0.9129 | 0.9072 | 0.9110 |
| 4 | 0.9148 | 0.8996 | 0.8996 |
| 5 | 0.8845 | 0.8864 | 0.8864 |
| 6 | 0.9280 | 0.9223 | 0.9186 |
| 7 | 0.9148 | 0.9015 | 0.9053 |
| 8 | 0.9261 | 0.9129 | 0.9167 |
| 9 | 0.9129 | 0.8996 | 0.8977 |
| 10 | 0.8939 | 0.8864 | 0.8883 |
| 11 | 0.9072 | 0.8958 | 0.8920 |
| 12 | 0.9072 | 0.9015 | 0.9053 |
| 13 | 0.9223 | 0.9110 | 0.9072 |
| 14 | 0.9091 | 0.9015 | 0.8939 |
| 15 | 0.9167 | 0.8939 | 0.8939 |
| 16 | 0.9205 | 0.9053 | 0.9034 |
| 17 | 0.9261 | 0.9053 | 0.9034 |
| 18 | 0.9091 | 0.9034 | 0.9034 |
| 19 | 0.9205 | 0.8958 | 0.8996 |
| 20 | 0.9034 | 0.8920 | 0.8977 |
| **Média** | 0.9140 | 0.9035 | 0.9035 |

**Fonte: Autoria própria**

Realizamos uma análise estatística mais aprofundada dos dados das 20 execuções, como apresentado nas Tabelas 2 e 3 acima, que correspondem, respectivamente, aos resultados dos Sistemas Monolíticos e dos Sistemas de Múltiplos Classificadores. Essa análise pode ser feita de duas maneiras: a primeira é uma análise visual, observando o comportamento ao longo das repetições, ou através da aplicação de testes estatísticos, como Kruskal-Wallis e Mann-Whitney, para fornecer uma interpretação mais rigorosa e quantitativa dos resultados.

Ao visualizarmos as tabelas acima e analisar, é possível destacar algumas observações importantes e devidamente básicas de se fazer. Na Tabela 2, dos Sistemas Monolíticos, vemos que o classificador Naive Bayes apresentou um desempenho constante e significativamente inferior em comparação com os demais classificadores, como foi descrito também ao fim do tópico *3. Análise / Resultados* . Em contraste, como também descrito no tópico 3, a Árvore de Decisão se destacou consistentemente, demonstrando um desempenho superior em relação aos outros modelos monolíticos analisados. Podemos ver isso na *figura 2* abaixo, que destaca visualmente essas diferenças.

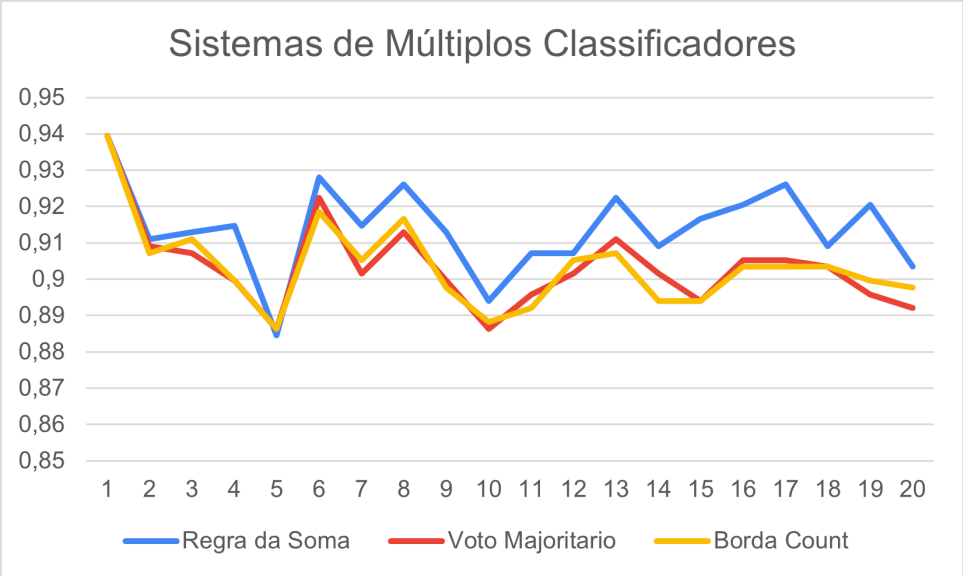
**Figura 2**




Fonte: Autoria própria

Uma análise interessante a ser feita em relação aos Múltiplos Classificadores é o comportamento similar observado ao longo de todo o processo. Por mais que tenham diferenças entre os classificadores, eles apresentaram padrões de desempenho que seguem tendências semelhantes, sugerindo que fatores comuns estão influenciando seus resultados. Como pode ser visto na figura 3.

**Figura 3**

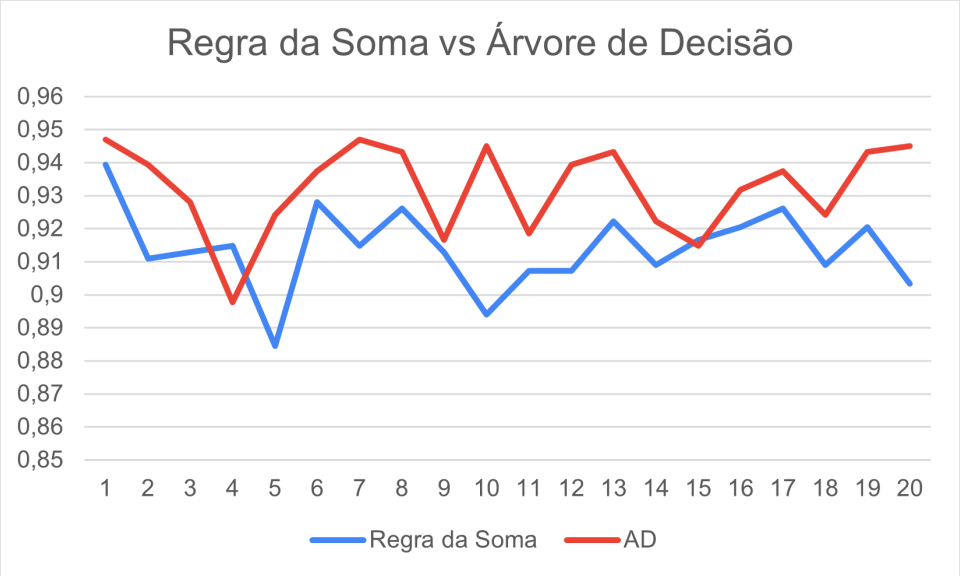



Fonte: Autoria própria

Por fim, ao comparar os melhores classificadores, observamos que a Árvore de Decisão

se manteve à frente durante a maior parte das execuções. Apenas em duas ocasiões a sua acurácia foi inferior à da Regra da Soma. Podemos ver isso na *figura 4*, logo abaixo.

**Figura 4**




Fonte: Autoria própria

Realizando agora a segunda analise, utilizando os testes de Kruskal-Wallis e Mann-Whitney, que são métodos não paramétricos usados para comparar grupos de dados sem a necessidade de assumir uma distribuição normal. Sendo o Kruskal-Wallis uma extensão do teste de Mann-Whitney que serve para poder comparar mais entre dois grupos. Como é possível de analisar nas tabelas 4 a 7 logo abaixo, duas informações são geradas por estes algoritmos: A estatística que irá quantifica a diferença observada entre os grupos, e a significância, ao qual indica a probabilidade de que essa diferença seja significativa. Neste trabalho foi adotado uma significância pré-definida de 5% (0.05), caso o valor seja menor que esse, indica que existe uma diferença significativa entre os métodos, caso contrário não.

Como vemos nas tabelas, praticamente em todos os testes um método se destacava perante ao outro, com exceção do teste Mann-Whitney para os sistema de Múltiplos Classificadores, onde os mesmos não apresentaram uma diferença significativa entre os grupos.

Com estes testes podemos destacar os melhores sistemas para a base de dados trabalhada, sendo o melhor Monolítico a Arvore de Decisão e o melhor Múltiplo Classificador a Regra de Soma. Aplicando o teste de Mann-Whitney, sobre ambos os modelos, a Árvore de decisão se mostrou mais adequada para o problema, conforme a *tabela 8*.

**Tabela 4. Kruskal-Wallis dos Sistemas Monolíticos**

| Estatística | Significância | Resultado |
|---|---|---|
| 82.4217 | $5.3428 \times e^{-17}$ | Rejeitamos a H0. Há uma diferença significativa entre os métodos. |

**Fonte: Autoria própria**

**Tabela 5. Mann-Whitney dos Sistemas Monolíticos**

| Métodos | Estatística | Significância | Resultado |
|---|---|---|---|
| KNN e NB | 400.0 | $6.6531 \times e^{-08}$ | Rejeitamos a H0. O KNN é superior ao NB. |
| KNN e AD | 1.0 | $7.7232 \times e^{-08}$ | Rejeitamos a H0. O AD é superior ao KNN. |
| KNN e SVM | 335.0 | 0.0003 | Rejeitamos a H0.O KNN é superior ao SVM. |
| KNN e MLP | 360.0 | $1.5691 \times e^{-05}$ | Rejeitamos a H0. O KNN é superior ao MLP. |
| NB e AD | 0.0 | $6.6344 \times e^{-08}$ | Rejeitamos a H0. O AD é superior ao NB. |
| NB e SVM | 0.0 | $6.6344 \times e^{-08}$ | Rejeitamos a H0. O SVM é superior ao NB. |
| NB e MLP | 40.0 | $1.5778 \times e^{-05}$ | Rejeitamos a H0. O MLP é superior ao NB. |
| AD e SVM | 400.0 | $6.6250 \times e^{-08}$ | Rejeitamos a H0. O AD é superior ao SVM. |
| AD e MLP | 400.0 | $6.6438 \times e^{-08}$ | Rejeitamos a H0. O AD é superior ao MLP. |
| SVM e MLP | 283.0 | 0.0245 | Rejeitamos a H0. O SVM é superior ao MLP. |

**Fonte: Autoria própria**

**Tabela 6. Kruskal-Wallis dos Sistemas de Múltiplos Classificadores**

| Estatística | Significância | Resultado |
|---|---|---|
| 41.6056 | 0.0021 | Rejeitamos a H0. Há uma diferença significativa entre os métodos. |

**Fonte: Autoria própria**

**Tabela 7. Mann-Whitney dos Sistemas de Múltiplos Classificadores**

| Métodos | Estatística | Significância | Resultado |
|---|---|---|---|
| Regra da Soma e Voto majoritário | 311.5 | 0.0026 | Rejeitamos a H0. Regra da Soma é superior ao Voto Majoritário. |
| Regra da Soma e Borda Count | 312 | 0.0025 | Rejeitamos a H0. Regra da Soma é superior ao Borda Count. |
| Voto majoritário e Borda Count | 200 | 1.0 | Rejeitamos H1. Não há diferença significativa entre os grupos . |

**Fonte: Autoria própria**

**Tabela 8. Mann-Whitney dos Melhores**

| Métodos | Estatística | Significância | Resultado |
|---|---|---|---|
| Regra da Soma e Arvore de Decisão | 59.5 | 0.0001 | Rejeitamos a H0. A Arvore de Decisão é superior ao Regra da Soma |

**Fonte: Autoria própria**

## Referências

Brun, A. L. (2024). Slides de AM. Slides apresentados em aula, [Unioeste - Ciência da Computação], [Cascavel, Brasil].

Palechor, F. M. and Manotas, A. D. L. H. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico.

Suthaharan, S. (2016). Decision tree learning. In *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, pages 237–269. Springer.

UCI ML Repository (2019). Estimation of obesity levels based on eating habits and physical condition. Acessado em: 20/08/2024.

Webb, G. I., Keogh, E., and Miikkulainen, R. (2010). Naïve bayes. In *Encyclopedia of Machine Learning*, volume 15, pages 713–714. Springer.