Universidade do Rio de Janeiro
Tópicos Especiais em Matemática e Computação
Trabalho 3

Gustavo Dias de Oliveira

01/07/2024

1 Descrição dos Dados

1.1 Informações relevantes

O banco de dados selecionado foi ”Estimation of Obesity Levels Based On Eating Habits and Physical Condition”, encontrado no repositório UCI, esta base de dados apresenta dados para a estimativa dos níveis de obesidade em indivíduos dos países México, Peru e Colômbia, com base em seus hábitos alimentares e condição física. As amostras são rotuladas com a variável de classe NObesity (Nível de Obesidade), que permite a classificação dos dados utilizando os valores de Peso Insuficiente, Peso Normal, Sobrepeso Nível I, Sobrepeso Nível II, Obesidade Tipo I , Obesidade Tipo II e Obesidade Tipo III. 77% dos dados foram gerados de forma sintética utilizando a ferramenta Weka e o filtro SMOTE, 23% dos dados foram coletados diretamente dos usuários através de uma plataforma web.

1.2 Número de amostras

Os dados contêm 17 atributos e 2111 amostras.

1.3 Atributos

Descrição geral dos atributos presentes neste conjunto de dados:

1. **Gender:** Gênero do participante.
   Resposta: Masculino(Male), Feminino(Female).
2. **Age:** Idade do participante.
   Resposta: Numérica (anos).
3. **Height:** Altura do participante.
   Resposta: Numérica (metros).
4. **Weight:** Peso do participante.
   Resposta: Numérica (quilogramas).
5. **family_history_with_overweight:** Histórico familiar de sobrepeso.
   Resposta: Sim(yes), Não(no).

6. **FAVC:** Consumo de alimentos com alto teor de calorias.

   Resposta: Sim(yes), Não(no).

7. **FCVC:** Consumo de vegetais.

   Resposta: Numérica (1-Nunca, 2-Às vezes, 3-Sempre).

8. **NCP:** Número de refeições principais por dia.

   Resposta: Numérica (1-Uma, 2-Duas ,3-Três , 4-Mais de três).

9. **CAEC:** Consumo de comida entre as refeições.

   Resposta: Não(No), Às vezes(Sometimes), Frequentemente(Frequently), Sempre(Always).

10. **SMOKE:** Hábito de fumar.

    Resposta: Sim(yes), Não(no).

11. **CH2O:** Consumo de água diário.

    Resposta: Numérica (1-Menos de um litro, 2-Entre 1 e 2 litros, 3-mais de 2 litros).

12. **SCC:** Monitora o consumo de calorias.

    Resposta: Sim(yes), Não(no).

13. **FAF:** Frequência de prática de atividade física.

    Resposta: Numérica (0-não pratica, 1-um ou dois dias, 2-dois ou quatro dias, 3-quatro ou cinco dias).

14. **TUE:** Tempo gasto com dispositivos tecnológicos como celular, videogame, televisão, computador e outros.

    Resposta: Numérica (0-zero a duas horas, 1-três a cinco horas, 2-mais de cinco horas).

15. **CALC:** Consumo de álcool.

    Resposta: Não(No), Às vezes(Sometimes), Frequentemente(Frequently), Sempre(Always).

16. **MTRANS:** Meio de transporte usado pelo participante.

    Resposta: Transporte Publico(Public_Transportation), Automóvel(Automobile), Bicicleta(Bicycle), A pé(Walking), Moto(Motorbike).

17. **NObeyesdad:** Nível de obesidade do participante.

    Resposta: Insuficientemente Pesado(Insufficient_Weight), Normal(Normal_Weight), Excesso de Peso Nivel I(Overweight_Level_I), Excesso de Peso Nivel II(Overweight_Level_II), Obesidade Tipo I(Obesity_Type_I), Obesidade Tipo II(Obesity_Type_II), Obesidade Tipo III(Obesity_Type_III).

1.4 Classes

A classe 'NObeyesdad' é considerada o atributo alvo neste conjunto de dados, sendo fundamental para a análise, pois representa o nível de obesidade dos participantes. É a variável que estamos interessados em prever ou estimar com base nas outras características disponíveis, como gênero, idade, hábitos alimentares e estilo de vida.

1.5 Balanceamento

Para avaliar a balanceamento da classe, comparamos o número de amostras em cada categoria, se o número de amostras em cada categoria for parcialmente igual, então a classe é considerada balanceada.

- Obesity_Type_III: 324 amostras
- Obesity_Type_II: 297 amostras
- Obesity_Type_I: 351 amostras
- Overweight_Level_II: 290 amostras
- Overweight_Level_I: 290 amostras
- Normal_Weight: 287 amostras
- Insufficient_Weight: 272 amostras

Neste caso, não há uma grande disparidade entre o número de amostras em cada categoria. Embora haja algumas diferenças, a variação não é significativa. Portanto, pode-se dizer que a classe "NObeyesdad"está relativamente balanceada.

1.6 Valores Nulos ou Faltantes

Utilizando o comando 'dados.isnull().sum()', que retorna a contagem de valores nulos em cada coluna dos dados, foi observado que os atributos não possuiam nenhum valor nulo ou faltante.

1.7 Valores Duplicados

Usando o comando 'dados[dados.duplicated() == True]', que retorna a tabela com todas as linhas duplicadas encontradas nos dados, pudemos analisar a tabela com todos os valores duplicados, e vimos que eles são existentes.

1.8 Outliers (Boxplot)

Um boxplot é um gráfico que representa a distribuição de um conjunto de dados. Ele é composto por uma caixa que mostra a mediana e os quartis dos dados, com linhas que se estendem a partir da caixa para indicar a variabilidade dos dados. Pontos fora dessa linha são considerados outliers, ou seja, valores incomuns. O boxplot é uma ferramenta eficaz para identificar padrões e discrepâncias em conjuntos de dados.

Na Figura 1 abaixo, conseguimos observar que Age(Idade), Height(Altura), Weight(Peso) e NCP(Numero de refeições principais por dia) possuem outliers.

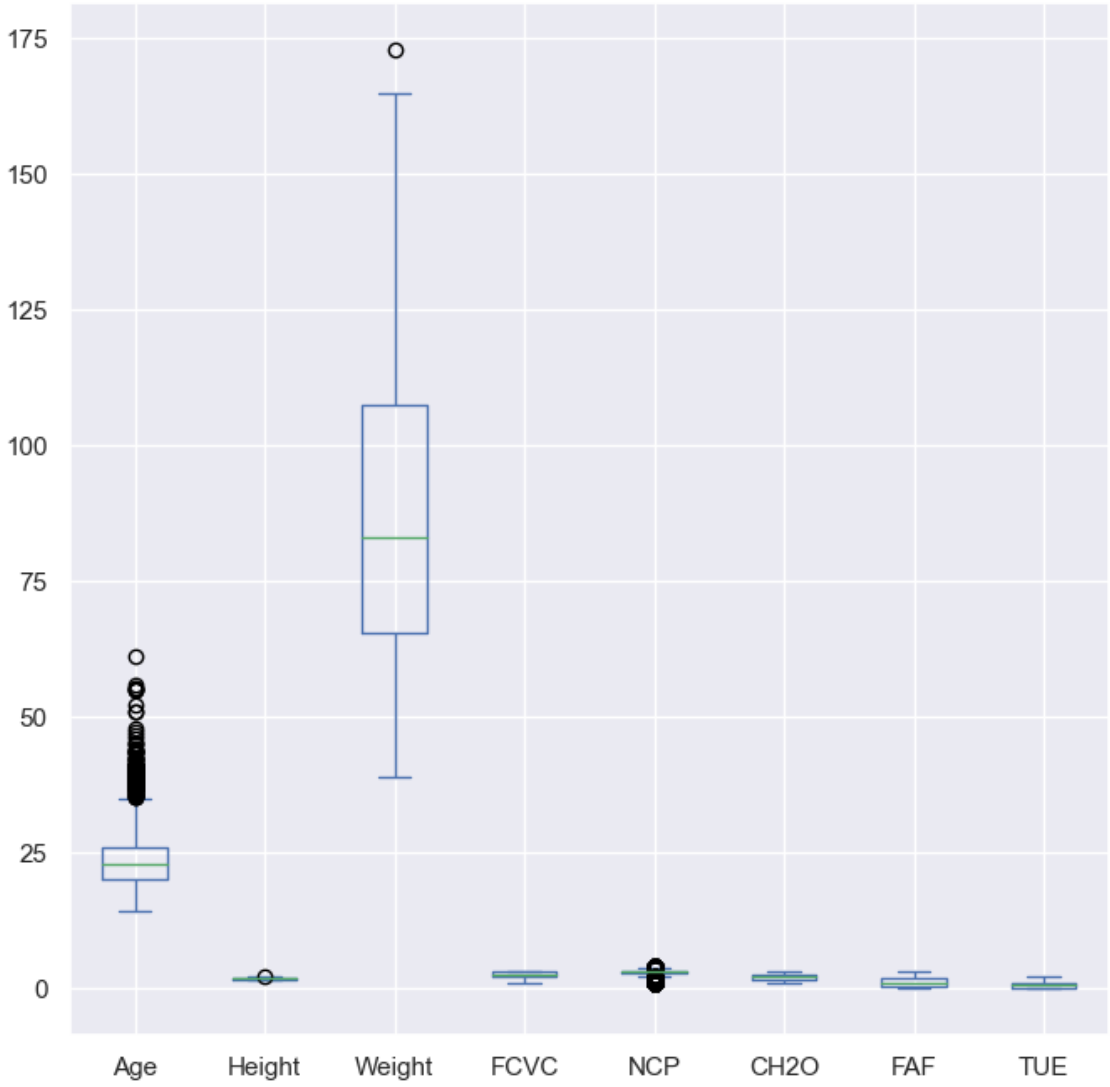

Figura 1: Boxplot dos dados numéricos

2 Análise dos Dados

Será apresentada a média, o desvio padrão, a variância, a matriz de correlação, o valor mínimo e máximo de cada atributo do nosso banco de dados.

2.1 Média

Na tabela 1 abaixo, será apresentados os atributos numéricos e suas determinadas médias.

Tabela 1: Valores Médios dos Atributos

Atributo	**Média**
Idade	24.31
Altura	1.70
Peso	86.59
FCVC	2.42
NCP	2.69
CH2O	2.01
FAF	1.01
TUE	0.66

2.2 Desvio Padrão

O desvio padrão é uma medida estatística que indica o quanto os valores de um conjunto de dados estão dispersos em relação à média.
Na tabela 2 abaixo, será apresentados os atributos numéricos e seus determinados desvios padrão.

Tabela 2: Desvios Padrão dos Atributos

Atributo	**Desvio Padrão**
Idade	6.35
Altura	0.09
Peso	26.19
FCVC	0.53
NCP	0.78
CH2O	0.61
FAF	0.85
TUE	0.61

2.3 Variância

A variância é uma medida estatística que descreve a dispersão dos valores de um conjunto de dados em relação à média
Na tabela 3 abaixo, será apresentados os atributos numéricos e suas determinadas variância.

Tabela 3: Variâncias dos Atributos

Atributo	**Variância**
Idade	40.27
Altura	0.01
Peso	685.98
FCVC	0.29
NCP	0.61
CH2O	0.38
FAF	0.72
TUE	0.37

2.4 Matriz de correlação

A matriz de correlação é uma tabela que mostra como cada variável em um conjunto de dados está relacionada às outras. Ela usa coeficientes de correlação para representar a força e a direção dessas relações. Valores próximos de 1 indicam uma forte correlação positiva, valores próximos de -1 indicam uma forte correlação negativa, e valores próximos de 0 indicam pouca ou nenhuma correlação.

Na Figura 2 abaixo, será apresentado a matriz de correlação dos dados numéricos, nela observamos que as variáveis que possuem maior correlação são Height(Altura) e Weight(Peso) com correlação de 0.46, as outras variáveis possuem correlação de 0.3 para menos, acredito que não tenha grande relevância.

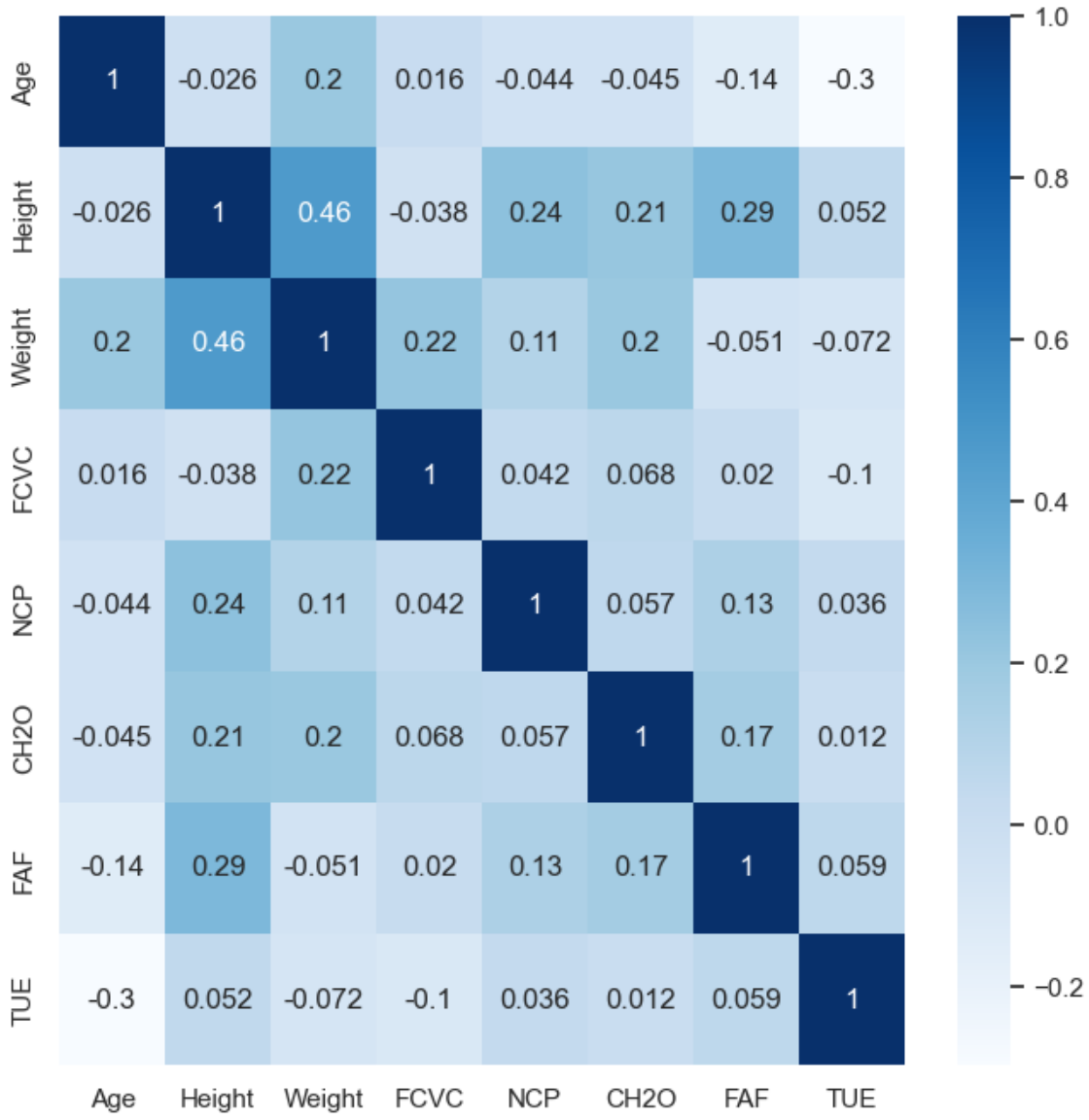

Figura 2: Matriz de correlação

2.5 Valor mínimo

Na tabela 4 abaixo, será apresentados os atributos numéricos e seus determinados valores mínimos.

Tabela 4: Valores Mínimos dos Atributos

Atributo	**Valor Mínimo**
Idade	14
Altura	1.45
Peso	39
FCVC	1
NCP	1
CH2O	1
FAF	0
TUE	0

2.6 Valor Máximo

Na tabela 5 abaixo, será apresentados os atributos numéricos e seus determinados valores maximos.

Tabela 5: Valores Máximos dos Atributos

Atributo	**Valor Máximo**
Idade	61
Altura	1.98
Peso	173
FCVC	3
NCP	4
CH2O	3
FAF	3
TUE	2

3 Metodologia

O pré-processamento dos dados, que consistiu em remover os valores duplicados, remover os outliers dos atibutos [Age(Idade), Height(Altura), Weight(Peso), NCP(Numero de refeições principais por dia)] e transformar as variáveis qualitativas[Gender(Gênero), family_history_with_overweight (Histórico familiar de sobrepeso), FAVC(Consumo de alimentos calóricos com frequência), CAEC (Consumo de alimentos entre refeições), SMOKE(Fumante), SCC(Monitora o consumo de calorias), CALC(Consumo de alcool), MTRANS(Meio de transporte), NObeyesdad(Nível de obesidade)] para quantitativas.

Após o término do pré-processamento foi dado inicio de fato as metodologias.

3.1 Método SelectPercentile

O *SelectPercentile* é uma técnica de seleção de características baseada em percentil. Ele é utilizado para selecionar um percentual das características mais relevantes de um conjunto de dados com base em uma estatística univariada.

3.2 Método K-fold Cross-Validation

O método K-fold cross-validation é uma técnica para avaliar modelos de machine learning dividindo os dados em K subconjuntos (folds), treinando o modelo K vezes usando cada fold como conjunto de validação uma vez, e calculando a média das métricas de desempenho obtidas. Isso ajuda a reduzir a variabilidade na estimativa de desempenho do modelo e utiliza os dados de forma eficiente.

3.3 Método K-Vizinhos Mais Próximos (KNN)

O algoritmo KNN (K-Nearest Neighbors) é um método de aprendizado supervisionado utilizado tanto para classificação quanto para regressão em problemas de machine learning. Sua abordagem é simples e intuitiva: Treinamento, Previsão, Seleção dos Vizinhos, Classificação ou Regressão. O parâmetro ”k”representa o número de vizinhos a serem considerados e é crucial para a eficácia do algoritmo. Um valor baixo de ”k”pode levar a uma variância alta e sensibilidade a ruído, enquanto um valor alto de ”k”pode suavizar fronteiras de decisão e mascarar padrões locais.

3.4 Método de Evolução Diferencial (DE)

O Método de Evolução Diferencial (DE) é uma técnica de otimização global estocástica, utilizada principalmente para resolver problemas de otimização contínua. É uma variante dos algoritmos genéticos e pertence à categoria de algoritmos de otimização evolucionária.

3.5 Método de Enxame de Partículas (PSO)

O Método de Enxame de Partículas (Particle Swarm Optimization - PSO) é uma técnica de otimização computacional baseada em modelos de comportamento social de enxames, como o movimento de bandos de pássaros ou cardumes de peixes.

4 Experimentos Computacionais

O objetivo é encontrar os melhores parâmetros para o classificador KNN utilizando o algoritmo de otimização Differential Evolution, avaliando seu desempenho com o KFold em diferentes conjuntos de dados. O processo é repetido 30 vezes para garantir a robustez dos resultados. Os resultados finais incluem as melhores configurações de parâmetros, bem como as métricas de desempenho correspondentes.

4.1 SelectPercentile

Foi realizado um processo de seleção de características utilizando o SelectPercentile nos dados pré-processados para obter as 10 características mais relevantes de acordo com o método aplicado, visto que os dados já estavam relativamente balanceados.

As 10 características mais relevantes foram:

- **Gender:** Gênero do participante.
- **Age:** Idade do participante.
- **Height:** Altura do participante.
- **Weight:** Peso do participante.
- **family_history_with_overweight:** Histórico familiar de sobrepeso.
- **FAVC:** Consumo de alimentos com alto teor de calorias.
- **FCVC:** Consumo de vegetais.
- **NCP:** Número de refeições principais por dia.
- **CAEC:** Consumo de comida entre as refeições.

4.2 K-fold Cross-Validations com o K-Vizinhos Mais Próximos

Foi aplicado o Método K-fold cross-validation integrado com o método K-Vizinhos Mais Próximos (KNN) para dividir o conjunto de dados, treinar e avaliar o modelo .

Onde os parâmetros utilizados pelo K-Fold seguem na tabela 6 abaixo:

Tabela 6: Parâmetros para o K-fold

Modelo	**Parâmetro**	**Valor**
K-Fold	n_splits	5
	shuffle	True
	random_state	run*10+100

- n_splits: O número de divisões no K-Fold. Foi definido como 5, o que significa que o conjunto de dados será dividido em 5 partes iguais.
- shuffle: Indica se os dados devem ser embaralhados antes de serem divididos. Está definido como True para garantir que as divisões sejam aleatórias.
- random_state: A semente aleatória para reprodutibilidade. Este valor é calculado como run*10+100, variando a cada iteração para garantir diferentes sementes em diferentes execuções.

4.3 Resultados

4.3.1 Evolução Diferencial (DE)

Foi realizado 30 iterações utilizando os parametros da tabela 7 a seguir:

Tabela 7: Limites dos parâmetros do KNN e Parâmetros da DE

Parâmetro KNN	**Limite Inferior**	**Limite Superior**
n_neighbors	1	30
weights	0	1
p	1	3
Parâmetros da DE	**Nome Parâmetro**	**Valor**
Número de iterações	maxiter	10
Tamanho	popsize	10
Imprimir mensagens	disp	True
Otimização local	polish	False
Semente para o gerador de números aleatórios	seed	run*10+100

- **n_neighbors**: O número de vizinhos a serem considerados.
- **weights**: A função de ponderação dos vizinhos.
- **p**: O parâmetro da métrica de distância.

Onde foi obtido os seguintes resultados da tabela 8:

Tabela 8: Resultados Médios e Desvios Padrão para DE

Métrica	**Tipo de Dados**	**Média ± Desvio Padrão**
ACCURACY	Originais	0.9301 ± 0.0026
F1	Originais	0.9274 ± 0.0028
RECALL	Originais	0.9301 ± 0.0026
ACCURACY	Pré-processados	0.9215 ± 0.0031
F1	Pré-processados	0.9197 ± 0.0032
RECALL	Pré-processados	0.9215 ± 0.0031

Os resultados médios e os desvios padrão são fornecidos para dar uma ideia da variabilidade do desempenho do modelo. Observa-se uma consistência nos resultados, com variações muito pequenas, indicando que o modelo é estável.

A seguir, temos os melhores parametros do método conforme a tabela 9 abaixo:

Tabela 9: Melhores Parâmetros e Métricas de Desempenho

Método	**Melhores Parâmetros**	**Acurácia**	**F1**	**Recall**
DE (Originais)	'p': 1, 'n_neighbors': 2, 'weights': 'distance'	93%	93%	93%
DE (Pré-Processados)	'p': 1, 'n_neighbors': 2, 'weights': 'distance'	92%	92%	92%

Os melhores parâmetros para ambas as configurações (dados originais e pré-processados) foram os mesmos, sugerindo que a seleção dos parâmetros KNN é robusta independentemente do pré-processamento dos dados.

Os dados originais tendem a ter um desempenho ligeiramente melhor em termos de acurácia, F1 e recall em comparação com os dados pré-processados, mas em termos de tempo, os dados pré-processados reduziu mais de 50% em relação aos dados originais.

4.3.2 Enxame de Partículas (PSO)

Foi realizado 30 iterações utilizando os parametros da tabela 10 a seguir:

Tabela 10: Limites dos parâmetros do KNN e Parâmetros do PSO

Parâmetro KNN	**Limite Inferior**	**Limite Superior**
n_neighbors	1	30
weights	0	1
p	1	3
Parâmetros do PSO	**Nome Parâmetro**	**Valor**
Número de iterações	maxiter	10
Tamanho do enxame	swarmsize	10
Fator de inércia	omega	0.5
Componente cognitiva	phip	0.5
Componente social	phig	0.5
Informações detalhadas	debug	True

- **n_neighbors**: O número de vizinhos a serem considerados.
- **weights**: A função de ponderação dos vizinhos.
- **p**: O parâmetro da métrica de distância.

Onde foi obtido os seguintes resultados, conforme a tabela 11:

Tabela 11: Resultados Médios e Desvios Padrão para PSO

Métrica	**Tipo de Dados**	**Média ± Desvio Padrão**
ACCURACY	Originais	0.9269 ± 0.0093
F1	Originais	0.9239 ± 0.0099
RECALL	Originais	0.9269 ± 0.0093
ACCURACY	Pré-processados	0.9207 ± 0.0045
F1	Pré-processados	0.9189 ± 0.0046
RECALL	Pré-processados	0.9207 ± 0.0045

Os resultados médios e os desvios padrão são relativamente baixos, especialmente para os dados pré-processados, indicando uma menor variabilidade e maior estabilidade do modelo para esses dados.

Embora a acurácia e o recall dos dados originais sejam ligeiramente superiores, a variação dos resultados para os dados originais é maior (desvio padrão de 0.0093) comparada aos dados pré-processados (desvio padrão de 0.0045).

Segue os melhores parametros do método, conforme a tabela 12 abaixo:

Tabela 12: Melhores Parâmetros e Métricas de Desempenho

Método	**Melhores Parâmetros**	**Acurácia**	**F1**	**Recall**
PSO (Originais)	'p': 1, 'n_neighbors': 1, 'weights': 'uniform'	93%	92%	93%
PSO (Pré-Processados)	'p': 1, 'n_neighbors': 1, 'weights': 'uniform'	92%	92%	92%

Os dados originais apresentam uma leve vantagem em termos de acurácia e recall, enquanto a métrica F1 é igual para ambas as configurações, mas em termos de tempo, os dados pré-processados reduziu mais de 50% em relação aos dados originais.

A consistência dos melhores parâmetros entre os dados originais e pré-processados indica que o PSO é robusto e eficiente na busca de parâmetros ótimos para o KNN.

4.4 Comparação dos Métodos DE e PSO

Comparando os modelos da Evolução Diferencial (DE com dados originais e DE com dados pré-processados) e Enxame de Partículas (PSO com dados originais e PSO com dados pré-processados),

- Em ambas as técnicas (DE e PSO), os dados originais geralmente apresentam métricas ligeiramente melhores em termos de acurácia, F1 e recall em comparação com os dados pré-processados. No entanto, a diferença no desempenho é relativamente pequena, com as métricas dos dados pré-processados sendo muito próximas às dos dados originais.
- Os dados pré-processados tendem a apresentar desvios padrão menores, indicando maior estabilidade e consistência nos resultados, especialmente com PSO. A menor variação nos resultados dos dados pré-processados sugere que o modelo é mais robusto e menos sensível a pequenas mudanças nos dados.
- Os dados pré-processados reduzem o tempo de processamento em mais de 50%, o que é um ganho significativo em eficiência computacional. Considerando o tempo de processamento reduzido e a mínima perda de desempenho, os dados pré-processados representam uma escolha vantajosa em cenários onde a eficiência computacional é crucial.

5 Conclusão

Neste trabalho, foi realizada uma análise comparativa entre os algoritmos de Evolução Diferencial (DE) e Otimização por Enxame de Partículas (PSO), aplicados em dados originais e pré-processados. Nossos resultados mostram que o pré-processamento dos dados tem um impacto significativo no desempenho de ambos os algoritmos, mantendo o desempenho do modelo, e também reduzindo o tempo de execução em mais de 50

Para o algoritmo DE, observamos que a acurácia, F1 e recall para os dados originais foram de 93.01%, 92.74% e 93.01%, respectivamente. Já para os dados pré-processados, os mesmos indicadores foram ligeiramente inferiores, com valores de 92.15%, 91.97% e 92.15%. Apesar da leve queda nos indicadores de desempenho, a redução significativa no tempo de execução torna o pré-processamento uma etapa vantajosa no contexto de aplicações onde o tempo é um fator crítico.

No caso do algoritmo PSO, os resultados seguiram uma tendência semelhante. Para os dados originais, a acurácia, F1 e recall foram de 93%, 92% e 93%, respectivamente. Para os dados pré-processados, os valores foram 92%, 92% e 92%. Novamente, a pequena diminuição no desempenho é compensada pela considerável redução no tempo de processamento.

Portanto, concluímos que, embora o pré-processamento dos dados possa resultar em uma leve diminuição nas métricas de desempenho, a significativa redução no tempo de execução torna esta prática altamente recomendável, especialmente em cenários onde a eficiência temporal é crucial. Caso não seja o caso, pelo ganho de tempo com o pré-processamento dos dados, poderia ser feito um ajuste nos parametros dos métodos para alcançar melhores resultados.

Referências

- https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition
- https://doi.org/10.1016/j.dib.2019.104344
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*.
- Liu, H., & Motoda, H. (Eds.). (2007). *Computational methods of feature selection*. CRC Press.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- Storn, R. and Price, K. (1997). Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. Journal of Global Optimization
- Kennedy, J., & Eberhart, R. C. (1995). Particle Swarm Optimization. Proceedings of IEEE International Conference on Neural Networks