Aplicação de Machine Learning na Medição do Índice de Massa Corporal(BMI)

Danilo Manoel Cáceres Niz RA:2094746

Engenharia de Software– Universidade Tecnológica Federal do Paraná (UTFPR)

Fundamentos de Sistemas Inteligentes

Resumo

Em resumo, este relatório explorou a aplicação de modelos de aprendizado de máquina para prever o Índice de Massa Corporal (BMI) com base em atributos como idade, sexo, altura, peso, atividade física e consumo diário de água. Foram utilizados diversos algoritmos, incluindo KNN, Bernoulli Naive Bayes, Árvore de Decisão e Máquinas de Vetores de Suporte (SVM).

Os resultados indicaram uma variação significativa no desempenho dos modelos. O KNN, sensível ao número de vizinhos, apresentou acurácia entre 86.02% e 94.69%. O modelo Bernoulli Naive Bayes teve uma acurácia de 73.47%, com dificuldades específicas na classe 'Normal'. Em contrapartida, os modelos de Árvore de Decisão (RandomForestClassifier e DecisionTreeClassifier) e SVM atingiram uma acurácia perfeita de 100.00%, sugerindo a possibilidade de overfitting.

A escolha do modelo ideal dependerá do equilíbrio entre acurácia e capacidade de generalização. É crucial validar esses modelos com conjuntos de dados de teste independentes para confirmar sua capacidade de generalização e seu desempenho no mundo real. Além disso, melhorias e ajustes podem ser necessários, especialmente para o modelo Bernoulli Naive Bayes, que teve dificuldades em classificar a classe 'Normal'.

Em conclusão, enquanto todos os modelos foram capazes de prever o BMI com alguma precisão, a pesquisa futura pode se concentrar na otimização dos modelos para melhorar o desempenho em classes específicas e na exploração de técnicas para evitar o overfitting. A validação contínua com conjuntos de dados independentes será essencial para a aplicação prática desses modelos.

Introdução

O aprendizado de máquina é um subcampo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que permitem que os computadores aprendam a partir de dados para fazer previsões ou decisões.

Sendo uma ferramenta poderosa que tem sido aplicada em uma variedade de campos, desde a medicina até o marketing. No contexto da saúde, o aprendizado de máquina pode ser usado para prever resultados de saúde, identificar fatores de risco e ajudar na tomada de decisões clínicas.

Neste relatório, exploramos como o aprendizado de máquina pode ser aplicado para medir o Índice de Massa Corporal (BMI). O BMI é uma medida simples de peso em relação à altura que é comumente usada para avaliar se uma pessoa tem um peso saudável para a sua altura. No entanto, a medição do BMI pode ser influenciada por uma série de fatores, incluindo idade, sexo, altura, peso, nível de atividade física e consumo diário de água.

Ao aplicar técnicas de aprendizado de máquina, podemos criar um modelo que leva em conta todos esses fatores para fornecer uma medição mais precisa do BMI. Por exemplo, podemos usar um algoritmo de classificação ou regressão para prever o BMI com base na idade, sexo, altura, peso, nível de atividade física e consumo diário de água. O modelo pode ser treinado em um conjunto de dados de indivíduos para os quais essas informações estão disponíveis, e então pode ser usado para prever o BMI de novos indivíduos.

Além disso, o aprendizado de máquina também pode ser usado para identificar padrões e tendências nos dados que podem não ser imediatamente aparentes. Por exemplo, pode ser possível identificar grupos de indivíduos com características semelhantes ou identificar fatores que são particularmente importantes na previsão do BMI.

Metodologia

Neste estudo, foi utilizado um conjunto de dados próprio baseado no Diabetes Prediction Dataset. O conjunto de dados contém os seguintes atributos:

- **Idades**: Este atributo representa a idade de uma pessoa em anos. As opções disponíveis são 18, 20, 26, 30, 35, 40, 47, 55, 60, 65.
- **Sexos**: Este atributo representa o sexo de uma pessoa. As opções disponíveis são ‘Feminino’ e ‘Masculino’.
- **Alturas**: Este atributo representa a altura de uma pessoa em metros. As opções disponíveis são 1.50, 1.60, 1.70, 1.80, 1.90, 2.00.
- **Pesos**: Este atributo representa o peso de uma pessoa em quilogramas. As opções disponíveis são 60.5, 70.0, 80.0, 90.0, 100.0, 110.0.
- **Atividades Físicas**: Este atributo representa o nível de atividade física de uma pessoa. As opções disponíveis são ‘Moderada’, ‘Intensa’ e ‘Leve’.
- **Consumos de Água Diários**: Este atributo representa a quantidade de água que uma pessoa consome diariamente. As opções disponíveis são ‘1L’, ‘1.5L’, ‘2L’, ‘2.5L’, ‘3L’.
- **BMI**: O Índice de Massa Corporal (BMI, do inglês Body Mass Index) é uma medida usada para determinar se uma pessoa tem um peso saudável em relação à sua altura. É calculado dividindo o peso de uma pessoa (em quilogramas) pelo quadrado de sua altura (em metros).

Utilizando algoritmos de aprendizado supervisionado para treinar um modelo preditivo capaz de prever o BMI de uma pessoa com base nos valores desses

atributos. O modelo foi treinado em um conjunto de dados de treinamento rotulados, onde cada exemplo de treinamento consiste nos valores desses atributos para uma pessoa específica, juntamente com o valor real do BMI dessa pessoa.

Após o treinamento, o modelo pode ser usado para prever o BMI de novas pessoas com base nos valores de seus atributos. Isso pode ser útil em uma variedade de aplicações, como a avaliação do estado de saúde de uma pessoa ou a identificação de pessoas em risco de desenvolver condições de saúde relacionadas ao peso.

Para a implementação do modelo, utilizamos a biblioteca scikit-learn, uma biblioteca da linguagem Python para machine learning. A biblioteca scikit-learn fornece uma variedade de algoritmos de aprendizado supervisionado que podem ser usados para treinar o modelo, bem como ferramentas para dividir os dados em conjuntos de treinamento e teste, avaliar o desempenho do modelo e ajustar os parâmetros do modelo.

Os algoritmos de aprendizado supervisionado utilizados neste estudo incluem:

- **Aprendizagem Baseada em Instâncias (KNN)**: O algoritmo KNN é um método simples e eficaz que classifica cada objeto com base em seus vizinhos mais próximos no espaço de atributos. Neste estudo, utilizamos o KNeighborsClassifier do scikit-learn com 15 vizinhos.
- **Árvore de Decisão**: As árvores de decisão são modelos preditivos que usam um conjunto de regras de decisão binárias para prever o valor de uma variável alvo. Neste estudo, utilizamos o DecisionTreeClassifier do scikit-learn.
- **SVM - Máquinas de Vetores de Suporte**: O SVM é um algoritmo poderoso e flexível que pode ser usado para tarefas de classificação e regressão. Ele funciona encontrando o hiperplano que melhor divide os dados em duas classes. Neste estudo, utilizamos o SVC do scikit-learn.

Resultados

Após treinar o modelo KNN com n_neighbors=7 e metric='minkowski' com p=2, obtivemos os seguintes resultados:

Matriz de confusão :

	Magresa	**Normal**	**Obesidade**
Magresa	62	14	0
Normal	1	241	18
Obesidade	0	19	625

A matriz de confusão mostra a distribuição das previsões do modelo e dos verdadeiros valores, permitindo ver facilmente quantas previsões foram feitas corretamente e quantas foram feitas incorretamente.
Relatório de classificação:

	Precisão	**Recall**	**F1-Score**	**Suporte**
Magresa	0.9841	0.8158	0.8921	76
Normal	0.8796	0.9269	0.9026	260
Obesidade	0.9720	0.9705	0.9713	644
Média/Total	0.9484	0.9469	0.9469	980

- A acurácia do modelo no conjunto de teste foi de **94.69%**, indicando uma alta taxa de previsões corretas.
- A matriz de confusão revelou que a maioria das previsões correspondia aos valores reais, com algumas exceções.
- O relatório de classificação mostrou que o modelo teve um desempenho excepcional na previsão de todas as três classes: ‘Magresa’, ‘Normal’ e ‘Obesidade’.
- Uma validação cruzada foi realizada para fornecer uma estimativa mais robusta do desempenho do modelo. A acurácia média na validação cruzada foi de **95.96%**.
- Esses resultados indicam que o modelo KNN treinado é capaz de prever com precisão o BMI com base nos atributos fornecidos. No entanto, é sempre importante validar o modelo com um conjunto de dados de teste independente para confirmar sua capacidade de generalização.

Treinando o modelo KNN com n_neighbors=15 e metric='minkowski' com p=2, obtivemos os seguintes resultados:
Matriz de confusão :

	Magresa	**Normal**	**Obesidade**
Magresa	40	36	0
Normal	4	214	42
Obesidade	0	55	589

O relatório de classificação para o modelo no conjunto de teste foi o seguinte:

	Precisão	**Recall**	**F1-Score**	**Suporte**
Magresa	0.9091	0.5263	0.6667	76
Normal	0.7016	0.8231	0.7575	260
Obesidade	0.9334	0.9146	0.9239	644
Média/Total	0.8701	0.8602	0.8598	980

- A acurácia do modelo no conjunto de teste foi de **86.02%**, indicando uma alta taxa de previsões corretas.
- A matriz de confusão revelou que a maioria das previsões correspondia aos valores reais, com algumas exceções.
- O relatório de classificação mostrou que o modelo teve um desempenho excepcional na previsão de todas as três classes: 'Magresa', 'Normal' e 'Obesidade'.
- Uma validação cruzada foi realizada para fornecer uma estimativa mais robusta do desempenho do modelo. A acurácia média na validação cruzada foi de **100.00%**.

Esses resultados indicam que o modelo KNN treinado é capaz de prever com precisão o BMI com base nos atributos fornecidos. No entanto, é sempre importante validar o modelo com um conjunto de dados de teste independente para confirmar sua capacidade de generalização.

Treinando o modelo Bernoulli Naive Bayes, obteve se os seguintes resultados:
Matriz de confusão:

Tabela

	Magresa	**Normal**	**Obesidade**
Magresa	76	0	0

Normal	0	0	260
Obesidade	0	0	644

Relatório de classificação:

Tabela

	Precisão	**Recall**	**F1-Score**	**Suporte**
Magresa	1.0000	1.0000	1.0000	76
Normal	0.0000	0.0000	0.0000	260
Obesidade	0.7124	1.0000	0.8320	644
Média/Total	0.5457	0.7347	0.6243	980

- A acurácia do modelo no conjunto de teste foi de **73.47%**, indicando uma taxa moderada de previsões corretas.
- A matriz de confusão revelou que o modelo teve um desempenho perfeito na previsão da classe 'Magresa', mas não conseguiu prever corretamente nenhuma instância da classe 'Normal'.
- O relatório de classificação mostrou que, embora o modelo tenha tido um desempenho perfeito na previsão da classe 'Magresa' e um bom desempenho na previsão da classe 'Obesidade', ele falhou completamente na previsão da classe 'Normal'.
- Uma validação cruzada foi realizada para fornecer uma estimativa mais robusta do desempenho do modelo. A acurácia média na validação cruzada foi de **72.22%**.

Esses resultados indicam que o modelo Bernoulli Naive Bayes treinado é capaz de prever com precisão o BMI com base nos atributos fornecidos. No entanto, o modelo teve dificuldade em classificar corretamente a classe 'Normal', sugerindo que podem ser necessárias melhorias ou ajustes adicionais no modelo para melhorar seu desempenho para essa classe específica.

Após treinar o modelo de Árvore de Decisão usando o algoritmo RandomForestClassifier com `n_estimators=5`, e obteve se os seguintes resultados:
Matriz de confusão :

Tabela

	Magresa	**Normal**	**Obesidade**

Magresa	76	0	0
Normal	0	260	0
Obesidade	0	0	644

Relatório de classificação :

Tabela

	Precisão	**Recall**	**F1-Score**	**Suporte**
Magresa	1.0000	1.0000	1.0000	76
Normal	1.0000	1.0000	1.0000	260
Obesidade	1.0000	1.0000	1.0000	644
Média/Total	1.0000	1.0000	1.0000	980

- A acurácia no conjunto de teste foi de **100.00%**, indicando uma previsão correta do BMI em todos os casos.
- A matriz de confusão mostrou que todas as previsões correspondiam aos valores reais.
- O relatório de classificação revelou uma precisão, recall e F1-Score perfeitos para todas as classes: 'Magresa', 'Normal' e 'Obesidade'.
- A validação cruzada, que fornece uma estimativa mais robusta do desempenho do modelo, resultou em uma acurácia média de **100.00%**.

Esses resultados indicam que o modelo de Árvore de Decisão treinado é altamente preciso na previsão do BMI com base nos atributos fornecidos. No entanto, uma acurácia de 100% pode sugerir overfitting, ou seja, o modelo pode estar muito bem ajustado aos dados de treinamento e pode não generalizar bem para novos dados. Portanto, é crucial validar o modelo com um conjunto de dados de teste independente para confirmar sua capacidade de generalização.

Treinando o modelo de Árvore de Decisão usando o algoritmo DecisionTreeClassifier com `criterion='entropy'`, `max_depth=10` e `random_state=0` :

Matriz de confusão:

Tabela

	Magresa	Normal	Obesidade
Magresa	76	0	0
Normal	0	260	0
Obesidade	0	0	644

Relatório de classificação :

Tabela

	Precisão	Recall	F1-Score	Suporte
Magresa	1.0000	1.0000	1.0000	76
Normal	1.0000	1.0000	1.0000	260
Obesidade	1.0000	1.0000	1.0000	644
Média/Total	1.0000	1.0000	1.0000	980

- A acurácia no conjunto de teste foi de **100.00%**, indicando uma previsão correta do BMI em todos os casos.
- A matriz de confusão mostrou que todas as previsões correspondiam aos valores reais.
- O relatório de classificação revelou uma precisão, recall e F1-Score perfeitos para todas as classes: 'Magresa', 'Normal' e 'Obesidade'.
- A validação cruzada, que fornece uma estimativa mais robusta do desempenho do modelo, resultou em uma acurácia média de **100.00%**.

Esses resultados indicam que o modelo de Árvore de Decisão treinado é altamente preciso na previsão do BMI com base nos atributos fornecidos. No entanto, uma acurácia de 100% pode sugerir overfitting, ou seja, o modelo pode estar muito bem ajustado aos dados de treinamento e pode não generalizar bem para novos dados. Portanto, é crucial validar o modelo com um conjunto de dados de teste independente para confirmar sua capacidade de generalização.

Treinando o modelo de Máquinas de Vetores de Suporte (SVM) usando o algoritmo SVC com `kernel='rbf'`, `random_state=0` e `C=40`:

Matriz de confusão :

Tabela

	Magresa	Normal	Obesidade
Magresa	76	0	0
Normal	0	260	0
Obesidade	0	0	644

Relatório de classificação :

Tabela

	Precisão	Recall	F1-Score	Suporte
Magresa	1.0000	1.0000	1.0000	76
Normal	1.0000	1.0000	1.0000	260
Obesidade	1.0000	1.0000	1.0000	644
Média/Total	1.0000	1.0000	1.0000	980

- A acurácia no conjunto de teste foi de **100.00%**, indicando uma previsão correta do BMI em todos os casos.
- A matriz de confusão mostrou que todas as previsões correspondiam aos valores reais.
- O relatório de classificação revelou uma precisão, recall e F1-Score perfeitos para todas as classes: ‘Magresa’, ‘Normal’ e ‘Obesidade’.
- A validação cruzada, que fornece uma estimativa mais robusta do desempenho do modelo, resultou em uma acurácia média de **100.00%**.

Esses resultados indicam que o modelo SVM treinado é altamente preciso na previsão do BMI com base nos atributos fornecidos. No entanto, uma acurácia de 100% pode sugerir overfitting, ou seja, o modelo pode estar muito bem ajustado aos dados de treinamento e pode não generalizar bem para novos dados. Portanto, é crucial validar o modelo com um conjunto de dados de teste independente para confirmar sua capacidade de generalização.

Discussão

Os resultados dos modelos de aprendizado de máquina treinados para prever o Índice de Massa Corporal (BMI) com base nos atributos fornecidos são bastante variados.

O modelo KNN com `n_neighbors=7` e `metric='minkowski'` com `p=2` teve uma acurácia de 94.69% no conjunto de teste e 95.96% na validação cruzada. Quando o número de vizinhos foi aumentado para 15, a acurácia no conjunto de teste caiu para 86.02%, mas a acurácia na validação cruzada permaneceu em 100.00%. Isso sugere que o modelo KNN pode ser sensível ao número de vizinhos escolhidos.

O modelo Bernoulli Naive Bayes teve uma acurácia de 73.47% no conjunto de teste e 72.22% na validação cruzada. Este modelo teve dificuldade em classificar corretamente a classe 'Normal', o que sugere que podem ser necessárias melhorias ou ajustes adicionais no modelo para melhorar seu desempenho para essa classe específica.

Os modelos de Árvore de Decisão treinados com os algoritmos RandomForestClassifier e DecisionTreeClassifier tiveram uma acurácia de 100.00% tanto no conjunto de teste quanto na validação cruzada. No entanto, uma acurácia de 100% pode indicar que o modelo pode estar sofrendo de overfitting, ou seja, ele pode estar muito bem ajustado aos dados de treinamento e pode não generalizar bem para novos dados.

O modelo de Máquinas de Vetores de Suporte (SVM), treinado com o algoritmo SVC e parâmetros `kernel='rbf'`, `random_state=0` e `C=40`, também teve uma acurácia de 100.00% tanto no conjunto de teste quanto na validação cruzada.

Em resumo, todos os modelos treinados foram capazes de prever com precisão o BMI com base nos atributos fornecidos, embora o modelo Bernoulli Naive Bayes tenha tido um desempenho inferior em comparação com os outros modelos. A acurácia de 100% alcançada por alguns dos modelos sugere que eles podem estar sofrendo de overfitting, e é crucial validar esses modelos com um conjunto de dados de teste independente para confirmar sua capacidade de generalização.

Conclusão

Os modelos de aprendizado de máquina treinados para prever o Índice de Massa Corporal (BMI) apresentaram resultados variados. O modelo KNN, dependendo do número de vizinhos escolhidos, teve uma acurácia entre 86.02% e 94.69%. O modelo Bernoulli Naive Bayes teve uma acurácia de 73.47%, com dificuldades específicas na classificação da classe 'Normal'.

Por outro lado, os modelos de Árvore de Decisão (tanto RandomForestClassifier quanto DecisionTreeClassifier) e o modelo de Máquinas de Vetores de Suporte (SVM) alcançaram uma acurácia perfeita de 100.00%. No entanto, essa perfeição

pode ser um indicativo de overfitting, sugerindo que esses modelos podem estar muito bem ajustados aos dados de treinamento e podem não generalizar bem para novos dados.

Em resumo, enquanto todos os modelos foram capazes de prever o BMI com alguma precisão, a escolha do modelo ideal dependerá do equilíbrio entre acurácia e a capacidade de generalização para novos dados. Além disso, para alguns modelos, podem ser necessárias melhorias ou ajustes adicionais para melhorar o desempenho em classes específicas. A validação contínua desses modelos com conjuntos de dados de teste independentes será crucial para confirmar sua capacidade de generalização e desempenho no mundo real.

A conclusão deve resumir os principais pontos do relatório, reiterar os resultados e sugerir direções para pesquisas futuras.

Referências

SciELO - Brasil - Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências

Como funciona o KNN (K-nearest neighbors) (didatica.tech)

Bernoulli Naive Bayes - GeeksforGeeks

Como funciona o algoritmo de Árvore de Decisão (Decision Tree) (didatica.tech)

Máquina de vetores de suporte (SVM) - Aprender Ciência de Dados (aprenderdatascience.com)

ML | Underfitting e Overfitting – Acervo Lima

Validação Cruzada de Modelos de Machine Learning | by Kizzy Terra | Programação Dinâmica | Medium

A influência de dados correlacionados em modelos de Aprendizado de Máquina - Um estudo... (usp.br)

Tipos de Aprendizado de Máquina #3 - DEV Community

LINK PRO COLAB:

https://colab.research.google.com/drive/1umE6bpBMfOeSJ_vRYXCLIvKMVn-hw9X-?usp=sharing