ATIVIDADE PRÁTICA – K-MEANS

Jadson Goulart de Matos (21103270)[1] and Régis Nyland Bloemer (20102404)[2]

[1]DEC0014-06655 (20231) - Inteligência Artificial e Computacional, UFSC

27 de maio de 2023

Resumo

O k-means é um tipo de algoritmo de aprendizado de máquina não supervisionado que agrupa os dados com base em suas características. Neste caso, é usado para criar dois grupos (ou 'clusters') de pacientes, na tentativa de distinguir entre aqueles que têm diabetes e aqueles que não têm. No entanto, como os resultados mostram, a árvore de decisão, que é um algoritmo de aprendizado de máquina supervisionado, foi capaz de alcançar um desempenho significativamente melhor nesta tarefa.

Lista de Figuras

Lista de Tabelas

1 Introdução

Os dados foram divididos em conjuntos de treinamento e teste, e dois modelos de aprendizado de máquina foram criados e treinados: um modelo de clustering usando k-means e um modelo de árvore de decisão.

K-Means é um tipo de algoritmo de aprendizado de máquina não supervisionado. Nessa atividade foi usado a biblioteca [2] scikit-learn para implementar o modelo cluster em Python. Os dados foram obtidos do Kaggle: [1].

2 Dados

Os dados contêm várias variáveis de entrada, incluindo gênero, idade, hipertensão, doença cardíaca, histórico de tabagismo, Índice de Massa Corporal (BMI), nível de Hemoglobina Glicada (HbA1c), e nível de glicose no sangue. A variável de saída, ou alvo, é 'diabetes', indicando se o paciente tem ou não diabetes.

Algumas dessas variáveis são categóricas (como gênero e histórico de tabagismo), e foram codificadas numericamente para serem utilizadas nos algoritmos de aprendizado de máquina.

1. **gender**: Este é o gênero do paciente, que pode ser masculino ou feminino.
2. **age**: Esta é a idade do paciente, em anos.
3. **hypertension**: Esta é uma indicação de se o paciente tem hipertensão ou não. "1"indica que o paciente tem hipertensão e "0"indica que o paciente não tem hipertensão.
4. **heart_disease**: Esta é uma indicação de se o paciente tem doença cardíaca ou não. "1"indica que o paciente tem doença cardíaca e "0"indica que o paciente não tem doença cardíaca.
5. **smoking_history**: Esta coluna detalha o histórico de tabagismo do paciente. As opções incluem "never"(nunca), "current"(atualmente é fumante) e "No Info"(sem informação).
6. **bmi**: Esta é a medida do Índice de Massa Corporal do paciente.
7. **HbA1c_level**: Esta é a medida do nível de Hemoglobina Glicada (HbA1c) do paciente, um indicador importante do controle de glicose a longo prazo.
8. **blood_glucose_level**: Este é o nível de glicose no sangue do paciente.
9. **diabetes**: Esta é a variável de destino que estamos tentando prever. "1"indica que o paciente tem diabetes e "0"indica que o paciente não tem diabetes.

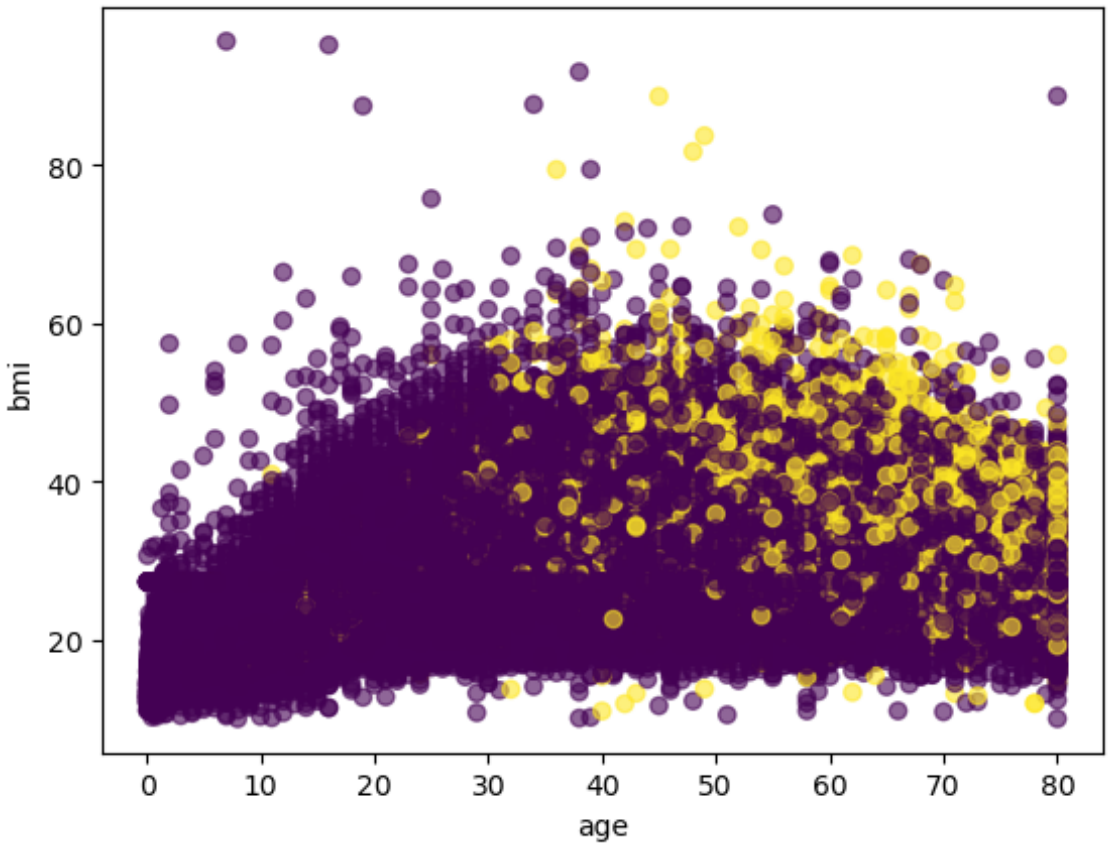
Podemos ver que há algumas colunas que são do tipo object, que significa que são variáveis categóricas, como 'gender', 'smoking_history'. Uma árvore de decisão pode lidar com variáveis categóricas diretamente, mas para facilitar a implementação em Python, vamos usar o LabelEncoder para codificar essas variáveis para valores numéricos.

2.1 Visualização dos dados

Antes de prosseguir para a criação e treinamento do modelo de árvore de decisão, é útil visualizar os dados para entender melhor as relações entre as variáveis.

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Female	80.0	0	1	never	25.19	6.6	140	0
Female	54.0	0	0	No Info	27.32	6.6	80	0
Male	28.0	0	0	never	27.32	5.7	158	0
Female	36.0	0	0	current	23.45	5.0	155	0
Male	76.0	1	1	current	20.14	4.8	155	0
Female	20.0	0	0	never	27.32	6.6	85	0
Female	44.0	0	0	never	19.31	6.5	200	1
Female	79.0	0	0	No Info	23.86	5.7	85	0
Male	42.0	0	0	never	33.64	4.8	145	0

Tabela 1: Amostra de todos os dados do CSV



Examinando o gráfico de dispersão, podemos analisar a relação entre essas variáveis. Parece que não há uma relação linear clara entre 'age' e 'bmi', uma vez que os pontos de dados estão espalhados por todo o gráfico.

Figura 1: Gráfico de dispersão

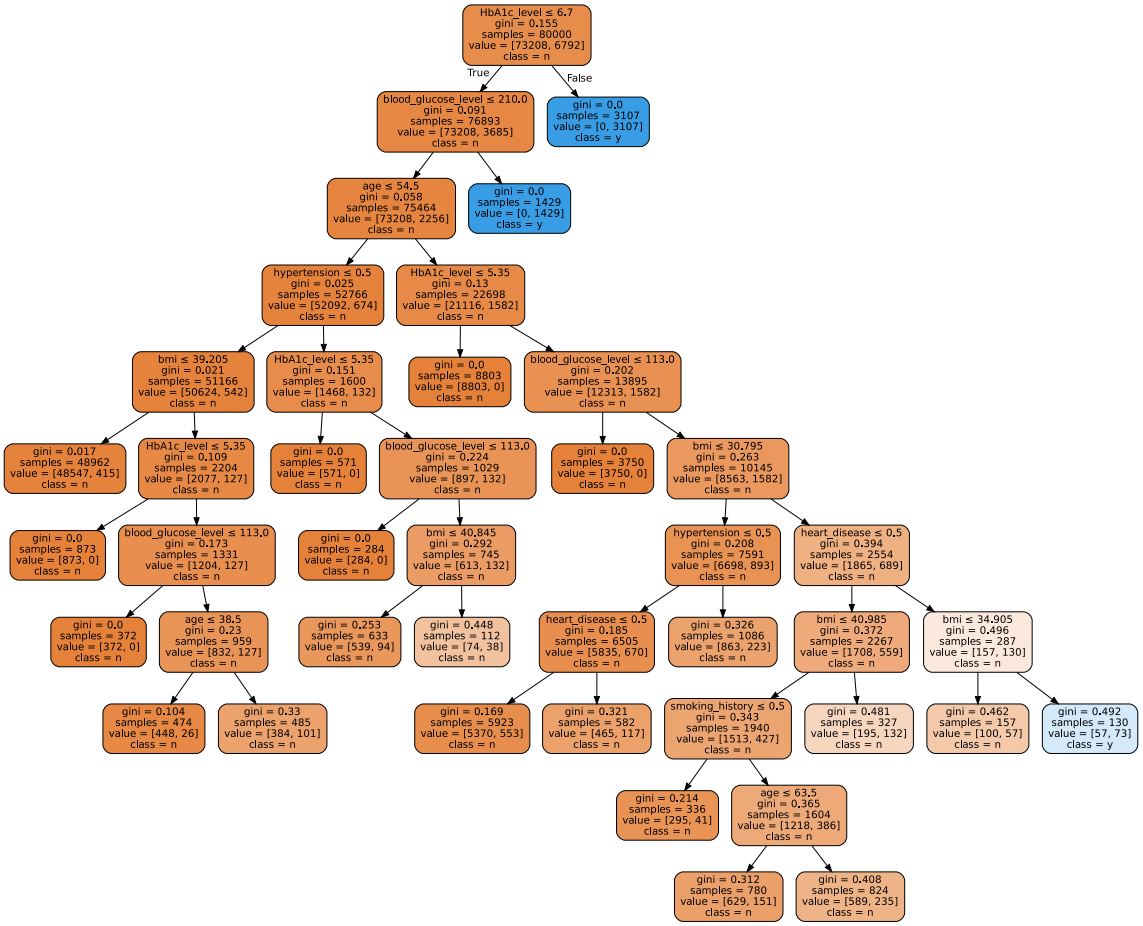


Figura 2: Gráfico da árvore de decisão

3 Metodologia

Os dados foram divididos em conjuntos de treinamento e teste, e dois modelos de aprendizado de máquina foram criados e treinados: um modelo de clustering usando k-means e um modelo de árvore de decisão.

4 Resultados

Os modelos foram avaliados em termos de sua precisão nos dados de treinamento. O modelo da árvore de decisão alcançou uma alta precisão de 97%, enquanto o modelo k-means teve uma precisão de 58%.

Os resultados desta atividade prática mostram que tanto o k-means quanto a árvore de decisão podem ser utilizados para prever diabetes com base em várias características de saúde do paciente. No entanto, neste caso, a árvore de decisão apresentou um desempenho significativamente melhor do que o k-means, indicando que pode ser uma escolha mais apropriada para esta tarefa específica de previsão.

Referências

[1] MUSTAFÁ, M. Diabetes prediction dataset - kaggle.com. [Accessed 27-May-2023].

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.