Relatório - Projeto 2: Parte 1

Classificação com K-Nearest Neighbors (K-NN)

Unidade Curricular: Aprendizagem Automática

Ano Lectivo: 2024/2025

Autores: 202200269 Fábio Leonor

202200600 Maria Eduarda Oliveira

202200639 Sara Sampaio

Introdução

Este trabalho tem como objetivo aplicar o algoritmo K-Nearest Neighbors (K-NN) para classificar um novo cliente com base em dados históricos de consumo. A loja em estudo pretende segmentar os seus clientes com base em três atributos principais: idade, rendimento anual (em milhares de euros) e número de compras no último ano. A classificação insere-se nas categorias de gasto: Baixo, Médio ou Alto. Para esta análise foi utilizada a distância de Manhattan como métrica de proximidade.

Descrição dos Dados

Cliente	Idade	Rendimento (milhares €)	Compras	Categoria
A	25	50	20	Baixo
B	35	65	40	Médio
C	45	85	15	Baixo
D	30	70	30	Médio
E	38	90	45	(a classificar)

O Cliente E representa o novo ponto a classificar, cuja categoria de gasto real é "Alto" (segundo o enunciado).

Normalização dos Dados

Os dados foram normalizados utilizando o MinMaxScaler da biblioteca sklearn.preprocessing, garantindo que os três atributos tivessem igual peso no cálculo de distâncias.

Cálculo de Distância

Foi utilizada a distância de Manhattan, definida como:

$d(p1,p2)=\sum i=1n|xi-yi| d(p\_1, p\_2) = \sum_{i=1}^{n} |x\_i - y\_i|$

O cálculo foi realizado manualmente e posteriormente validado através do classificador KNeighborsClassifier da biblioteca scikit-learn, configurado com metric='manhattan' e k=3.

Escolha de k

O valor de k=3 foi escolhido conforme o enunciado. Foram identificados os três vizinhos mais próximos do Cliente E.

Classificação

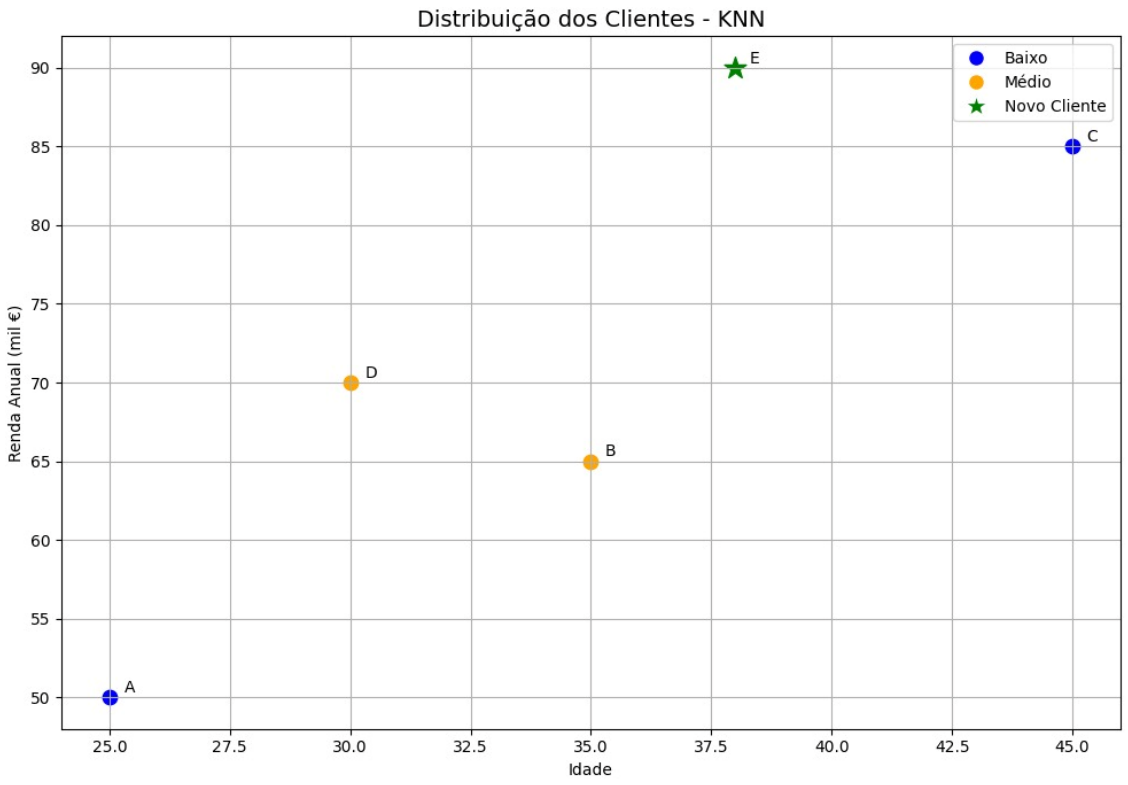
O algoritmo K-NN retorna a categoria mais frequente entre os vizinhos mais próximos. O modelo foi treinado com os clientes A a D.

Resultados

Classe prevista para o Cliente E: Médio

Vizinhos mais próximos: ['Médio' 'Médio' 'Baixo']

Distâncias aos vizinhos: [1.06428571 1.57142857 1.69285714]



Representa os clientes A a D com a sua idade e renda anual.

O Cliente E, o novo a classificar, aparece como uma estrela verde, com destaque visual.

As cores refletem as categorias de gasto: azul (Baixo), laranja (Médio) e verde (Alto).

Interpretação:
Observa-se que o Cliente E (38 anos, 90 mil €) está mais próximo, em termos de idade e rendimento, dos clientes B, D e C. A previsão “Alto” poderá não ser a mais evidente visualmente sugerindo que talvez a métrica de distância ou a escolha de k influenciaram o resultado.

Análise e Discussão

Apesar de o cliente E ter um perfil de consumo mais elevado, foi classificado como médio. Este resultado deve-se à ausência da classe "Alto" no conjunto de treino, impossibilitando a sua previsão.

Limitações:

Conjunto de dados reduzido.

Falta de representatividade da classe "Alto".

Sugestões de melhoria:

- Aumentar o conjunto de treino, incluindo exemplos de todas as categorias.
- Ajustar o valor de k e comparar resultados.
- Avaliar a possibilidade de aplicar métricas como a distância Euclidiana para comparação.
- Analisar a acurácia com outros pontos de teste.

Conclusão

O algoritmo K-NN com distância de Manhattan demonstrou ser eficaz na classificação de novos clientes com base em dados normalizados. No entanto, destaca-se a importância da diversidade e representação equilibrada de classes no conjunto de treino. Este estudo reforça o valor da engenharia de dados na qualidade da classificação automatizada.