Tarefa 1 - Clustering Application

Alexandre Esposte Santos

1 de setembro de 2023

1 Apresentação da base da dados

Para essa tarefa foi escolhido um conjunto de dados da plataforma Kaggle, onde a base escolhida é denominada por "Star Type Classification / NASA". Cada instância desses dados corresponde a uma estrela, em que para cada uma dessas há uma série de variáveis, como exemplo a luminosidade da estrela, a temperatura, o raio , a cor, a classe espectral e etc... Além dessas variáveis há também o tipo da estrela, tais como anã branca, anã vermelha, gigante e assim por diante. Ao todo essa base contém sete variáveis, sendo quatro delas variáveis numéricas, duas categóricas e uma correspondendo ao rótulo.

2 Objetivo

Originalmente esse conjunto de dados foi elaborado para o desenvolvimento de modelos de classificação. Entretanto, estaremos utilizando essa base para o agrupamento dos dados. Neste contexto, utilizaremos os rótulos das estrelas para comparar os agrupamentos alcançados pelo algoritmo K-means.

3 Desenvolvimento

Na lista a seguir apresentamos os passos efetuados durante a tarefa.

1. Verificação da integridade da base, isto é, verificação de valores ausentes/nulos e instâncias duplicadas.
2. Elaborou-se melhor o nome das variáveis na tabela para que dessa forma as informações fossem mais claras.
3. Aplicou-se um encoder nas variáveis categóricas. Dessa maneira, conseguimos codificar/transformar as variáveis categóricas em variáveis numéricas.
4. Padronizou-se toda a base através do z-score.
5. Com a base tratada aplicamos o método PCA.
6. Analisamos as componentes obtidas. Dessa forma, observamos o quanto cada componente principal explica os nossos dados.
7. Selecionou-se as três componentes mais explicativas conforme solicitado no escopo da tarefa.
8. Aplicou-se o agrupamento K-means para vários clusters e determinamos qual deveria ser a quantidade mais adequada.
9. Analisou-se os agrupamentos obtidos, além disso comparamos os agrupamentos com os rótulos originais.

Os passos listados acima são apenas um resumo do que foi feito. No notebook enviado há mais detalhes acerca de cada passo executado além das conclusões ao final da tarefa.