Mini Project: Unsupervised Machine Learning

the DEVELOPER Academy

Brief

Select a dataset of your choice and create an unsupervised machine learning algorithm to provide further insight into the data.

Requirements

Description in general with detailed list of requirements.

1. Start with some exploration of the data – what are the fields, how are they distributed, will they require any transformations?
2. Prepare the dataset for machine learning
3. Create an unsupervised machine learning model
4. Measure the accuracy of your model

Deliverables

You should submit your code along with some analysis. Your analysis can be presented as presentation slides, or using something like Jupyter Notebooks.

Data

There are three datasets for you to choose from. Pick the one you find most interesting or have an idea for how machine learning could be applied. I have given you some examples of the kind of analysis you could do – but feel free to come up with your own!

Dataset 1 – Customer Data

This dataset contains information about customers of a car company.

Example insight: Can you find groups of customers who are similar? Can you come up with any recommendations that would be useful to a marketing team?

File name: customers.csv
Find out more about this dataset here: https://www.kaggle.com/vetrirah/customer

Dataset 2 – Wine Quality

This dataset contains information about various types of wine.

Example insight: Find similarities between different wines. How many groups are there? Are there any wines which stand out as unique to the rest?

File name: winequality-red.csv and winequality-white.csv (you can use both datasets or just the one)
Find out more about this dataset here: https://www.kaggle.com/maitree/wine-quality-selection

Dataset 3 – Country Data for Charity

This dataset contains socio-economic and health factors that determine the overall development of countries. The purpose of the dataset is to help a charity fighting poverty to identify groups of countries where it is needed most.

Example insight: Which countries can be grouped together and have the highest need of help from the charity?

File name: country-data.csv
Find out more about this dataset here:
https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data

Extension

To extend the project, try using two different unsupervised learning algorithms and use methods to measure their accuracy. Which model would you choose? Why?