Machine Learning Final Project

Name	ID
Malak Mahmoud Aref	20221445867
Bassant Mohamed	20221376715
Nureen Ehab	20221465124
Zainab Mohamed Abdallah	20221310251

Intelligent Systems Department

Problem 1: Sales prediction

In this problem we will build a machine learning model for sales prediction by simple linear regression.
We used the dummy Shampoo dataset but we faced a problem with the month column because it's not numeric so will add extra column called (convert) to replace each month to numbers beginning from 0. Now, we could use convert instead of month to do operations.
Model Building: split the dataset into training and testing data it's usually a good practice to keep 70% of the data in your train dataset and the rest 30% in your test dataset we will fit the regression line on the train dataset.
Model Evaluation: we will make some predictions on the test data and we calculate the r-squared.

Problem 2: Patient's Sickness Prediction System

In this problem we will use the heart dataset and will apply clustering by k-means, PCA and t-SNE.

PCA vs. t-SNE

Although both PCA and t-SNE have their own advantages and disadvantages, some key differences between PCA and t-SNE can be noted as follows:

- t-SNE is computationally expensive and can take several hours on million-sample datasets where PCA will finish in seconds or minutes.
- PCA it is a mathematical technique, but t-SNE is a probabilistic one.
- Linear dimensionality reduction algorithms, like PCA, concentrate on placing dissimilar data points far apart in a lower dimension representation. But in order to represent high dimension data on low dimension, non-linear manifold, it is essential that similar data points must be represented close together, which is something t-SNE does not PCA.
- Sometimes in t-SNE different runs with the same hyperparameters may produce different results hence multiple plots must be observed before making any assessment with t-SNE, while this is not the case with PCA.
- Since PCA is a linear algorithm, it will not be able to interpret the complex polynomial relationship between features while t-SNE is made to capture exactly that.

Clustering by K-Means

K-means clustering measures similarity using ordinary straight-line distance (Euclidean distance). It creates clusters by placing a number of points (centroids), inside the feature-space. Each point in the dataset is assigned to the cluster of whichever centroid it's closest to.

Problem 3: Student University Recommendation (Bonus)

We used the Movielens dataset because it's easy to handle as it contains only two files which is the movies and the ratings. We **cleaned the title** on the data to remove the extra characters like dashes and parentheses to make the search easier
That algorithm manily consists of:

1. Built a **search engine** to search movie titles.
2. Created a **recommendation engine** to recommend us movies based on that we had watched before.

Reference of helper resources

https://www.kaggle.com/code/ashydv/sales-prediction-simple-linear-regression/notebook
https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html
https://www.youtube.com/watch?v=eyEabQRBMQA