Report: Machine learning(ex: chatgpt, translation alorithm, ai chat bots)

Machine learning is a subfield of artificial intelligence that focuses on studying data and algorithms to mimic human learning to predict outcomes and patterns and optimize the fitting of these outcomes by utilizing a dataset. The results are gradually improved as the program learns to imitate and put to use learnt information. It is merely a way for computers to learn without the explicit need to be programmed. Instead, computers learn to program themselves by repetitively processing information.

Supervised learning: The most common type, including constant information input to help the "learning" process. It has an input and output dataset labelled (learn from previous usage). This permits the models to become more prone to accuracy over time.

Unsupervised learning: The algorithm is not provided with labelled data. It consists of a way to discover patterns and optimize solutions. There are no labelled target outputs; these are only created by the algorithm. It is mainly used to analyze and cluster unlabeled datasets. (K means clustering is an example). The machine analyzes different patterns and trends found in a cluster of data to optimize usage/consumption (any prompt used) — for example, in the case of Matlab, kcluster=2 after further analysis (therefore a graph of 2 curves is necessary enough to provide information about electricity consumption over the year.

Semi-supervised learning: medium between supervised and unsupervised learning. Train the initial model on labelled samples to apply it to another number of unlabelled data.

Reinforcement machine learning: This type of machine learning uses validation of outcomes to learn. For example, training a model by telling it when it made the right decision, which over time the machine will learn about the actions it should take.

CLUSTER FUNCTIONS: CALINSKI-HARABASZ METHOD
KEY TERMS:

Calinski-Harabasz Index(CH): variance ratio criterion. It measure the cluster differences from each other and how compact data points are with each clusters. Higher values of CH indicate better clustering more similarity between data and groups.

CH index:
$CH=Sb\ /(k-1)\ /\ Sw/(n-k),$ where

Sb is the dispersion of cluster centroids from the global centroid (within cluster dispersion)

Sw is the dispersion of points from their respective cluster centroids. (Between cluster dispersion)

N is the total number of data points

K is the k-cluster number

Cluster dispersion is the spread of data points in a given space. The higher the CH index, the more accurately the data points are represented in their respective clusters.

Within cluster dispersion: measures the dispersion of data points within each cluster — the lower this index is, the more similar the datas are in the specific cluster.

Between-cluster dispersion is the distance between cluster centroids, which demonstrates how different each cluster is from the others. The higher this value, the greater the difference and dispersion between clusters.

The output plots different Ch index values for different numbers of clusters. Optimal value of clusters (for k) if when the CH index reaches its max, indicating better clustering.