# EPOCH TASK SUBMISSION

Roshan Y Singh

## Task 1: Classification Model

**Methods Used:**

**Data Loading and Cleaning:**

- Loaded the data from a CSV file using pandas.
- Converted the 'Latitude' and 'Longitude' columns to numeric, coercing errors.
- Filtered data to include only records from the state of Karnataka.
- Dropped rows with NaN values in 'Latitude' and 'Longitude'.
- Filtered latitude and longitude to be within valid ranges for Karnataka.

**K-Means Clustering:**

- **Initialization of Centroids:** Randomly selected initial centroids from the data.
- **Cluster Assignment:** Assigned each data point to the nearest centroid using Euclidean distance.
- **Centroid Update:** Recomputed centroids as the mean of data points in each cluster.
- **Convergence Check:** Repeated assignment and update steps until centroids no longer changed or max iterations reached.
- **Within-Cluster Sum of Squares (WCSS):** Calculated to evaluate cluster compactness.

- **Silhouette Score:** Measured how similar a data point is to its own cluster compared to other clusters.
- **Davies-Bouldin Index:** Evaluated cluster separation, with lower values indicating better clustering.

**Evaluation Metrics:**

- **Elbow Method:** Plotted WCSS against the number of clusters to identify the optimal k.
- **Silhouette Score:** Plotted to determine the quality of clustering.
- **Davies-Bouldin Index:** Plotted to evaluate cluster compactness and separation.

**Visualization:**

- Plotted the Elbow method, Silhouette score, and Davies-Bouldin Index to determine the optimal number of clusters.
- Visualized the final clusters and centroids on a scatter plot.

**Inferences:**

**Optimal Number of Clusters:**

- The Elbow method, Silhouette score, and Davies-Bouldin Index are crucial for determining the appropriate number of clusters.
- Chose an optimal number of clusters based on the evaluation metrics.
- Also let the number of clusters reflect the number of districts in the state.

### Density of Pin Codes:

- These pin codes seemed to well-outline my home state of Karnataka.
- There were a greater number of regions included around the cities, such as Bengaluru and Mysuru, whereas those near the borders with other states and the regions around the Western Ghats were sparse.
- Letting the number of clusters as 4, also lets us in on the four divisions of Karnataka- Bengaluru, Mysuru, Belagavi and Kalaburgi.

### Use of Additional Methods:

- Used scales apart from the Elbow plot to choose the optimum number of clusters.
- Also, the centroids seem to fall towards the district centers of particular districts

### Reference Links:

- Statquest Video Resources given in the document.
- Scikit-learn Clustering Documentation
- Silhouette Score - Wikipedia
- Davies-Bouldin Index - Wikipedia
- Use of AI tool - ChatGPT

## Task 2: Classification Model

### Code Overview

This project involves two main tasks: Optical Character Recognition (OCR) using a Convolutional Neural Network (CNN) and sentiment analysis using a Logistic Regression model with TF-IDF vectorization. The objective is to recognize text from images, and then analyze the sentiment of the recognized text.

### 1. OCR using CNN

### Steps Involved:

1. **Data Cleaning and Preparation**:
    - Load the alphabet dataset (`alphabets_28x28.csv`) which contains 28x28 pixel images of letters.
    - Clean the dataset by removing rows with corrupted data.
    - Convert the data into a format suitable for training a CNN.
2. **Building the CNN Model**:
    - Define a CNN with two convolutional layers followed by max-pooling layers, a flattening layer, and two dense layers.
    - Compile the model using the Adam optimizer and categorical cross-entropy loss.
3. **Training the CNN Model**:
    - Split the data into training and testing sets.
    - Train the model and evaluate its accuracy on the test set.
4. **Segmenting and Predicting Letters**:

    - Segment input images into smaller images representing individual letters.
    - Use the trained CNN to predict each letter.
    - Reconstruct the recognized text from individual letter predictions.

**2. Sentiment Analysis**

**Steps Involved:**

1. **Data Preparation**:
    - Load the sentiment analysis dataset which contains text lines and their associated sentiment labels.
    - Preprocess the text by converting it to lowercase and removing punctuation.
2. **Training the Sentiment Analysis Model**:
    - Use TF-IDF vectorization to convert text data into numerical features.
    - Train a Logistic Regression model using these features.
3. **Evaluating the Sentiment Analysis Model**:
    - Split the data into training and testing sets.
    - Evaluate the model's accuracy on the test set.
4. **Predicting Sentiment on OCR Text**:
    - Use the trained sentiment analysis model to predict the sentiment of the text recognized by the OCR system.

**Inferences:**

- Observed good accuracies with the CNN model(Around 99%)
- Relatively good accuracies from the Sentiment Analysis Model (83.33%)

- The Epochs took some time to work and process, and so did the images, however, eventually, the sentences observed were clean and correct
- The use on TensorFlow module was a keynote
- The integrated system's overall sentiment analysis accuracy on OCR predictions was approximately 75%, indicating that the pipeline from text recognition to sentiment classification performed reasonably well but had some room for improvement.

**Future Improvements:**

1. **Improved Segmentation**: Enhance image segmentation techniques to better handle varied character sizes and fonts.
2. **Advanced Preprocessing**: Implement more sophisticated preprocessing steps for both image and text data.
3. **Model Optimization**: Experiment with different CNN architectures and hyperparameters to further improve accuracy.
4. **Additional Metrics**: Evaluate models using additional metrics such as precision, recall, and F1-score to gain more insights into performance.

**References:**

1. TensorFlow Documentation: TensorFlow
2. OpenCV Documentation: OpenCV
3. Scikit-Learn Documentation: Scikit-Learn
4. Video Resources given in the document, such as StatQuest and NeuralNine videos were very informative.
5. Use of AI Tool - ChatGPT