Plant Disease Classification using K-Nearest Neighbors (KNN)

This project implements a K-Nearest Neighbors (KNN) classifier for identifying plant diseases based on leaf images from the PlantVillage dataset. The pipeline covers the complete process from data preprocessing to model evaluation and prediction.

---

1. Objective

The goal is to accurately classify plant diseases using image data through the KNN algorithm. We aim to:

- Load and preprocess the PlantVillage dataset
- Flatten and normalize image data
- Train a KNN classifier
- Evaluate its performance
- Tune hyperparameters (k-value)
- Predict diseases from new images

---

2. Dataset Description

- Source: Kaggle dataset "emmarex/plantdisease"
- Structure: Directory-based where each sub-folder represents a plant disease class
- Input: Colored leaf images
- Output: Disease label (e.g., "Tomato___Early_blight")

---

3. Preprocessing Steps

- Resize each image to 64x64 pixels using OpenCV
- Flatten images into 1D arrays

- Encode labels using `LabelEncoder`
- Normalize the data using `StandardScaler`
- Reduce dimensionality using `PCA` (50 components)

---

4. Model Training

- Algorithm: K-Nearest Neighbors (KNN)
- Parameters:
    - Number of neighbors: 12
    - Distance metric: Cosine
- Dataset split: 80% training, 20% testing
- Trained on the PCA-transformed data

---

5. Evaluation Results

- Accuracy: ~68%
- Classification Report: Precision, recall, and F1-score for each disease class
- Confusion Matrix: Visual representation using Seaborn heatmap
- Labels restored to original disease names for interpretability

---

6. Hyperparameter Tuning

- Method: 5-fold cross-validation on a subset of 1000 samples
- Explored k-values from 1 to 19
- Best performing K: Displayed along with its cross-validated accuracy

- Visualization: Line plot of accuracy versus K

---

7. Image Prediction Utility

A custom function `predict_image(image_path)` was implemented to classify any new image using the trained pipeline. It applies the same preprocessing steps (resize, flatten, scale, PCA) before predicting the disease label.

---

8. Conclusion

This KNN-based approach provides a decent baseline for plant disease classification with moderate accuracy. Despite its simplicity, the model performs well on a diverse set of plant disease images. Future improvements could include using more complex classifiers (e.g., CNNs), data augmentation, and deeper feature extraction.

Libraries Used:

- `opencv-python`
- `scikit-learn`
- `matplotlib`, `seaborn`
- `kagglehub` for dataset download

Note: The entire pipeline is reproducible and can be extended further for advanced classification tasks using deep learning.