# Data Preparation/Feature Engineering

## Overview

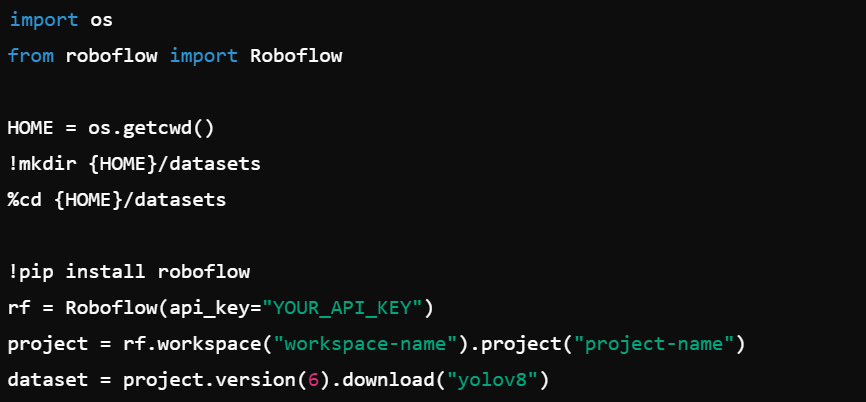
Data preparation and feature engineering are vital steps in any machine learning project. They involve collecting, cleaning, transforming, and organizing raw data into a suitable format for model training. These steps ensure the data's quality and relevance, improving the machine learning model's accuracy and performance.

## Data Collection

Source: The dataset was collected from Roboflow, specifically from the project "Tomato Leaf Disease" version 6.

**Preprocessing Steps During Collection:**

* Auto-Orient: Ensured all images were correctly oriented.
* Resize: All images were resized to 512x512 pixels for uniformity.



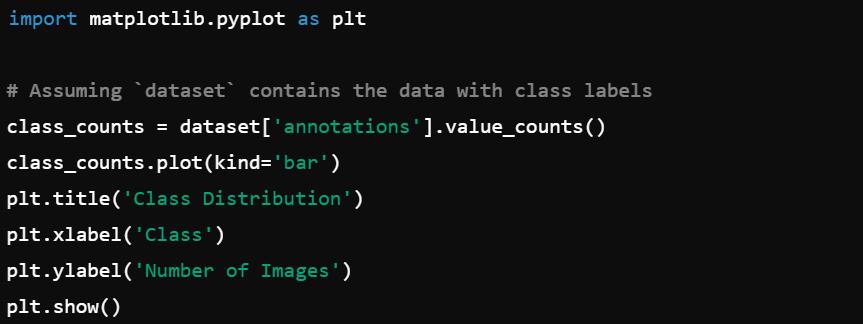
## Data Cleaning

* Handling Missing Values: Ensured all images in the dataset were present and correctly labeled.
* Outliers: Visual inspection of images was performed to remove any that were not relevant (e.g., mislabeled or corrupted images).
* Other Data Quality Issues: Ensured consistency in image dimensions and formats.

## Exploratory Data Analysis (EDA)

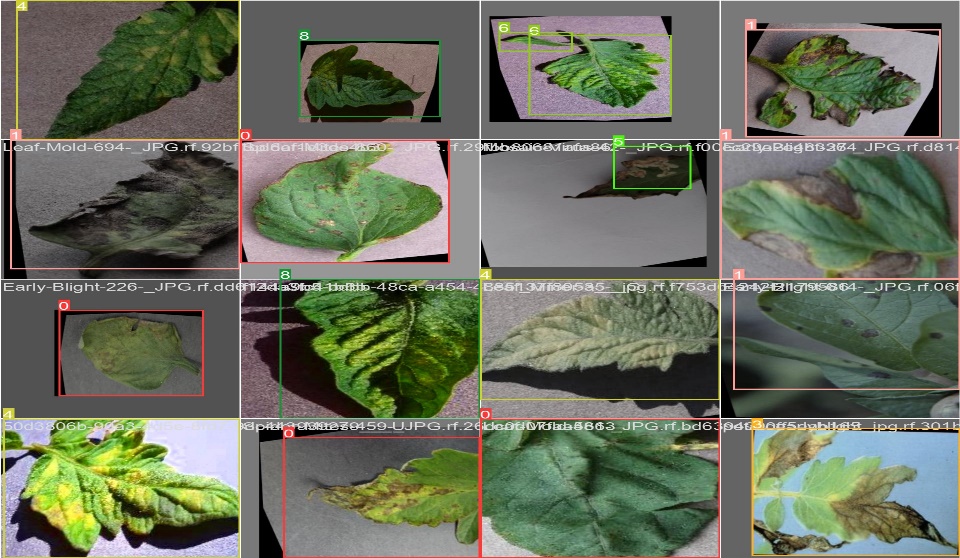
Exploratory Data Analysis involves visualizing and understanding the data to extract key insights and identify patterns. Below are some key insights and visualizations:

**Class Distribution:** Understanding the number of images per class.



**Image Samples:** Visualizing samples from different classes.

**Augmentation Impact**: Visualizing the effects of augmentations such as flipping, rotation, and exposure adjustments.

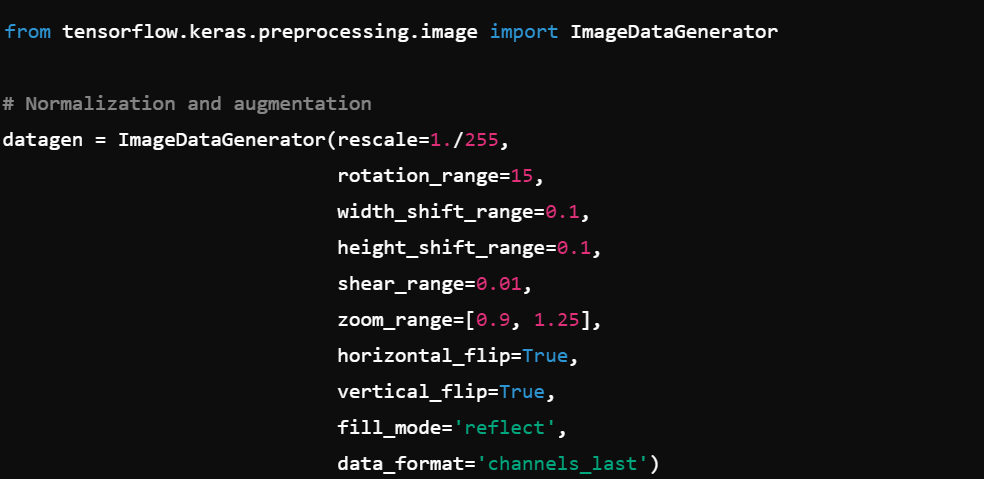


## Feature Engineering

Feature engineering involves creating new features or transforming existing ones to improve model performance. In this project, the main features are the pixel values of the images. Augmentation techniques such as flipping, rotating, and exposure adjustments were applied to create more training examples and improve the model's robustness.

### Augmentations:

* Flip: Horizontal and vertical flips.
* Rotate: Random rotations between -15° and +15°.
* Shear: Horizontal and vertical shearing.
* Exposure: Adjustments to simulate varying lighting conditions.



## Data Transformation

* Scaling and Normalization: Image pixel values were normalized to the range [0, 1].
* Encoding: Labels were encoded into numerical format for model training.

A computer screen with white text and colorful text

Description automatically generated

## Model Exploration

### Model Selection

#### Chosen Model: YOLO v8 (You Only Look Once version 8)

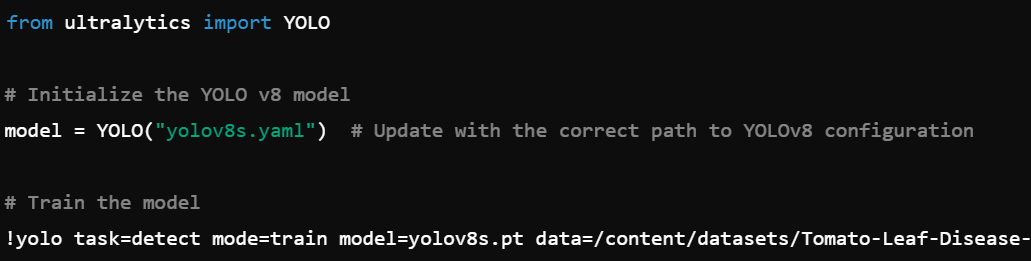
**Rationale:**

* Strengths: High accuracy, real-time processing capabilities, efficient computation, robust performance on large datasets.
* Weaknesses: Requires significant computational resources for training.

### Model Training

#### Training Details:

* Hyperparameters: Initial learning rate, batch size, number of epochs.
* Cross-Validation: K-fold cross-validation was used to validate model performance across different subsets of the data.



## Model Evaluation

### Evaluation Metrics:

Precision, Recall, and F1-Score: Used to evaluate the classification performance.

A graph of a graph

Description automatically generated with medium confidence

A chart with blue squares

Description automatically generatedConfusion Matrix: To visualize true positives, false positives, true negatives, and false negatives.

ROC Curve: To evaluate the trade-off between true positive rate and false positive rate at various threshold settings.

A graph showing different colored lines

Description automatically generated

## Code Implementation: A screen shot of a computer program Description automatically generated

