Fruits 360 Dataset Analysis

By

Group 7

Chiwen Shi

Shuhao Xia

Wensong Liu

Ying Lu

Overview

In most cases, when we try to log in or register a new account in a system, we have to enter verification code or select some specific kind of pictures to certify that we are not robots. Trying to train the model on how to distinguish a certain kind of pictures can help it verify whether the user's verification information is correct in a more accurate way.

Objectives

We have a dataset of images containing fruits, there are 81 kinds of fruits in total. Our goals for the analysis of the fruit dataset is as follows:

- Choose 2 kinds fruits to fit our model, during model building, we will try some different models, including convolutional neural network(CNN), then pick the best model for the further work.
- Given a specific picture, the model can identify which kind of fruits is and give the name of the fruits.

Dataset

- Data sources: https://www.kaggle.com/moltean/fruits

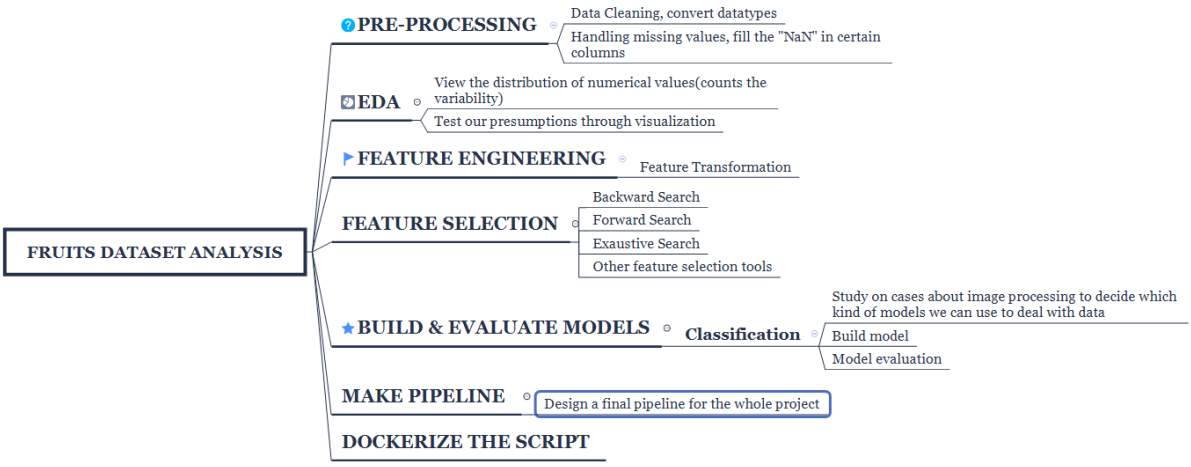
The fruits dataset contains 55244 images in total, and 41332 images(one fruit per image) for training set, 13877 images(one fruit per image) for test set. There are about 45 images that contain more than one fruit per image.

Use Cases

Web systems and mobile applications: all the web systems and mobile applications that requires log in verification can use our model. It can help them programmatically verify whether the user is a robot or not.

Workflow

Basically we have split our data analysis process into four parts, the first part is to preprocessing the dataset, and then do the exploratory data analysis part. After that, we will finish model building and evaluating part, and then make a pipeline as well as dockerize the whole project. Our workflow is as follows:



- Data Wrangling: data cleaning, handle missing values, convert datatypes
- Exploratory Data Analysis: view the correlation between different features and test our presumptions through visualization
- Feature Engineering & Feature Selection: To decide whether we need to do feature transformation using different methods and feature engineering tools, such as backward, forward search. Filter out important features through feature selection.
- Model Building: build models and evaluate the models.
- Final Pipeline: design a final pipeline for the whole project and dockerize the script.

Timeline

Timeframe	Work
2018.12.01-2018.12.03	Data Preprocessing, Exploratory Data Analysis
2018.12.04	Feature Engineering & Feature Selection
2018.12.05 -2018.12.07	Model Building, Evaluating and Selection
2018.12.08 -2018.12.09	Final Pipeline and Dockerize the whole project