# Homework 1 Decision Tree 460 G

**Sairam Sri Vatsavai**

## Synthetic dataset and Pokémon dataset

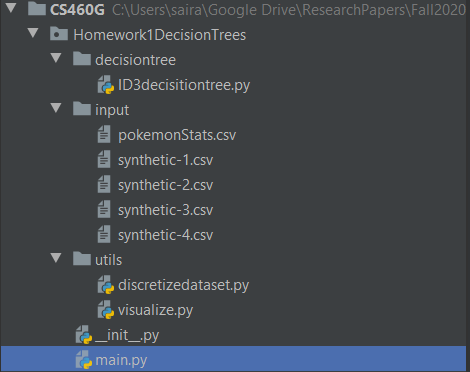
**Dataset Modifications** : I have added column names to each synthetic dataset. The features are named as ‘x1’ , ‘x2’ and target column as ‘class’. I combined the legendary column to pokemonStats.csv for Pokémon classifier.

**Discretization** : I discretized all the dataset features with equidistant binning and named the bins bin0,bin1 and so on. I did try different number of bins, but finally fixed the number of bins to 4 bins for all the datasets.

**Decision Tree**: I used information gain and entropy to find the best feature to split the dataset. I used ID3 algorithm to construct the tree. I used a python dictionary to create the tree and my implementation is inspired from <https://www.python-course.eu/Decision_Trees.php>.

**Implementation Details:** I created a python project shown in Figure 1.

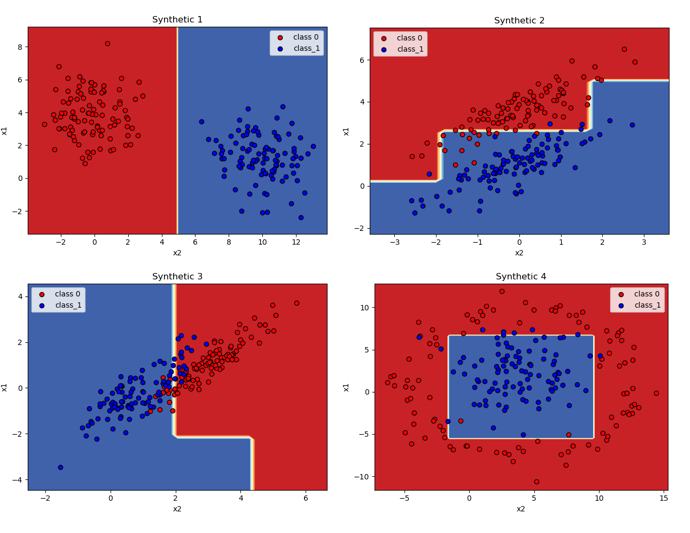
* It contains ID3 class for decision tree implementation including Entropy, Information Gain calculations function and ID3 implementation.
* Input folder contains the modified datasets
* Utils folder contains functions for discretization and visualization of decision boundary
* main.py contains calls for training all the datasets to create decision trees and testing the accuracy or error for each of them. It also creates plots of decision tree boundary and data points.



**Figure 1. Project Structure**

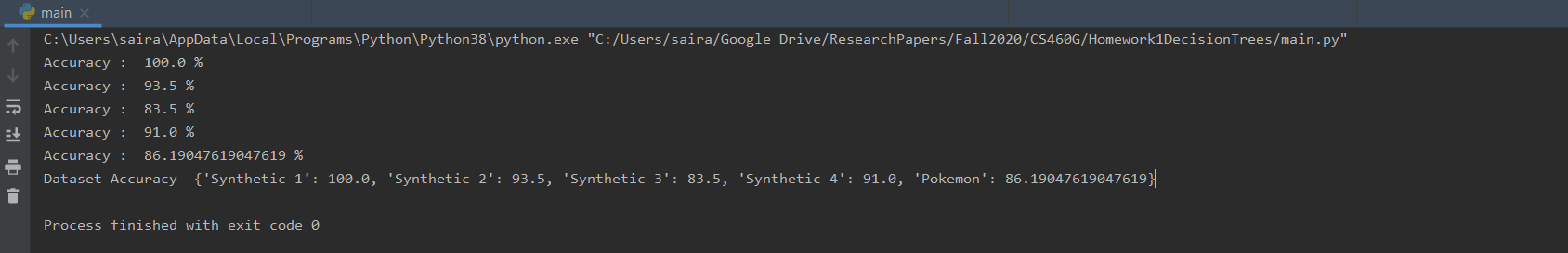
**Visualization :**  Figure 2. contains the plots of trained decision trees for each synthetic datasets. The visualization implementation is adapted from :

<https://scikit-learn.org/stable/auto_examples/tree/plot_iris_dtc.html>.



**Figure 2. Decision Tree boundary plots for each synthetic dataset**

**Accuracy Details:** I am reporting the accuracy details of each dataset in Figure 3.



**Figure 3. Accuracy details for synthetic datasets and Pokémon dataset**

**Libraries Utilized:**

* **matplotlib**
* **pandas**
* **numpy**