Machine Learning - Assignment 1

Files present

`src` directory:

- `diabetes.csv` - Contains the data for training the decision tree.
- `requirements.txt` - Contains all the necessary dependencies with their versions.
- `decision_tree.py` - Contains the decision tree model.
- `utils.py` - Contains all helper functions required.
- `main.py` - Main file for completing all tasks required.

`plots` directory:

- `before_pruning.gv.png` - Image of the decision tree obtained after Q2, i.e., the best tree after 10 random splits.
- `best_depth.gv.png` - Image of the decision tree obtained after Q3, i.e., the tree with best depth limit.
- `after_pruning.gv.png` - Image of the decision tree obtained after Q4, i.e., after pruning.
- `depth_accuracy.png` - Plot of test accuracy v/s maximum depth.
- `nodes_accuracy.png` - Plot of test accuracy v/s number of nodes in the tree.

Instructions to run the code

- Navigate to the `src` directory.
- Ensure you are using a latest version of Python3, and install all dependencies.
  `pip install -r requirements.txt`
- Execute the file `main.py`
  `python main.py`
- Images of the three decision trees and the two plots will be created in the same directory.

Giving specific values to parameters

- To run using all default values for parameters
  `python main.py`
- To pass a specific maximum depth for Q1 and Q2 (say 6)
  `python main.py --depth 6`
- To pass a specific impurity measure for Q2 and Q3 ('ig' = information gain, 'gini' = gini index), say 'ig'
  `python main.py --measure ig`
- To pass a path for the data file `python main.py --file <path_to_file>`
- For more help regarding any of these
  `python main.py --help`