Group Member(s):

- Irsyad Nabil – 1406546134
- Meta Andrini Utari – 1406546153

Programming Assignment

ID3 Decision Tree

Algorithm

The code that generates the decision tree is comprised of 2 main classes: Dataset and Node. The Dataset class largely concerns itself with handling and presenting input data, whereas the Node class is used to generate the representation of the decision tree.

Dataset (Decision Tree)

The methods contained in this part of the Dataset class are related to the algorithm that generates an ID3 Decision Tree given a training set.

```
populate_data()
```

Populates the dataset with the given pre-processed list of lines read from the .txt file. Each tuple is stored as a namedtuple.

```
entropy()
```

Calculates the entropy of an attribute r, or attributes r and s, with r acting as the target attribute and s as the splitting attribute. Entropy is required to measure the homogeneity of the training set, as well as its information gain. A training set where the values are similar or identical (i.e. homogeneous) has an entropy of 0. On the other hand, a training set where the values are different and equally distributed (i.e. non-homogeneous) has an entropy of 1.

The formulae to calculate entropy, both with or without a splitting attribute, are derived from online sources.

```
information_gain()
```

Calculates the information gain (decrease in entropy) after splitting target attribute r with splitting attribute s. The information gain can be calculated by subtracting `entropy(r, s)` from `entropy(r)`.

```
_id3()
```

Recursively constructs a decision tree given a target attribute, a list of predicting attributes, and an example dataset. Returns a tree of Node objects. The root node is determined by the attribute with the largest information gain. Once the root node is initialized, it would spawn child nodes for each unique value associated with the root node attribute. The code runs recursively that way until the leaf nodes are reached, after which it returns a single node in which the label for the value is either a

‘yes’ or a ‘no’.

```
build_tree()
```

Returns an ID3 Decision Tree based on a given data according to the attributes and target attributes.

Dataset (Training Set Accuracy)

The methods contained in this part of the Dataset class are related to the algorithm that determines how much accuracy can be expected from the generated decision tree given a training set.

```
derive_ruleset()
```

Derives a rule set from the generated decision tree, where each rule is defined by the path it takes from the root node to a leaf node.

```
eval_condition()
```

Evaluates a condition based on a list of tuples of attribute-value pairs that define a condition derived from a ruleset.

```
print_ruleset()
```

Prints every rule in the rule set in a more human-readable format.

```
accuracy()
```

Calculates the training set accuracy by calculating the ratio of all the correct predictions made from each rule versus the total of all predictions made on the training set.

```
paths()
```

Computes every possible unique paths from the root node to a leaf node. Used to derive a ruleset from a decision tree.

Node

Provides the representation of a decision tree node. Basic functionalities such as node type, arbitrary value, and number of children are included.

User Manual

The program can be run on the `bash` terminal by simply typing in `python id3.py [.txt file of input data]`. For example, `python id3.py data_titanic.txt`. On Mac's Terminal, `python` should be substituted for `python3`.