1. Q1. Refer to function generate_data() in generate_data.py file
2. Q2. Refer to decision_tree() in decision_trees.py for the function that fits the data to a decision tree

   Refer to error() in decision_trees.py for the function that returns the error of a decision tree on a data set.
3. One tree generated for k=4, m=30 is shown in Fig. 1. No, the generated tree does not make any sense. There does not seem to be any relation between the ordering of the variables of the tree and the dependency of the variables to each other or the class label Y. Moreover, on multiple trials with the same parameters k=4, m=30, the trees generated were widely different. No common trend was apparent across these trees.

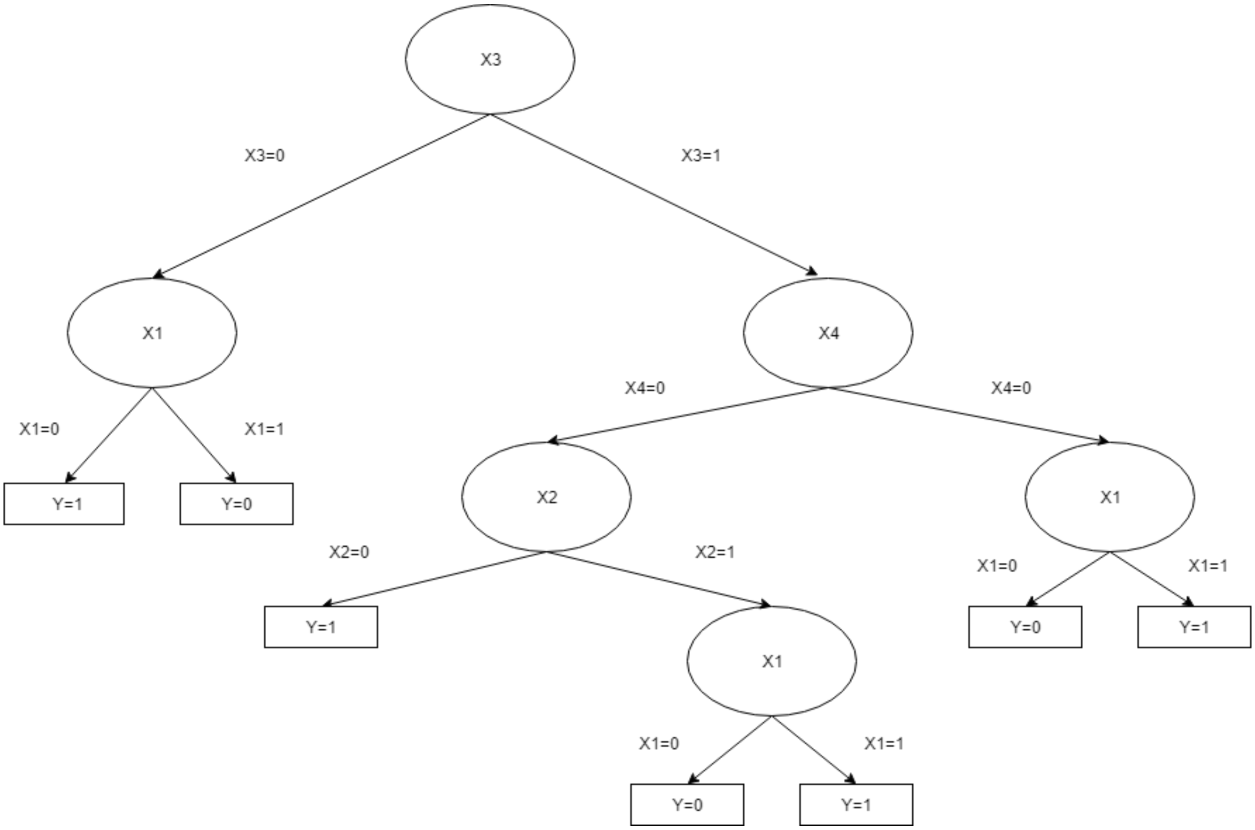

Fig 1. Tree generated for k=4, m=30

4. The tree generated above using k=4, m=30 was tested against a set of 1000 data points. The average classification error on this data was equal to 3.4%.
5. To solve this problem, m was varied between 50 and 500 with a step of 10 and the number of test samples was kept constant at 1000. The observed values of $|err_{train}(f) - err(f)|$ vs m is shown in Fig 2. As can be observed from the graph, the marginal value of addition data w.r.t $|err_{train}(f) - err(f)|$ which can be given by dy/dx decreases as the value of m increases.

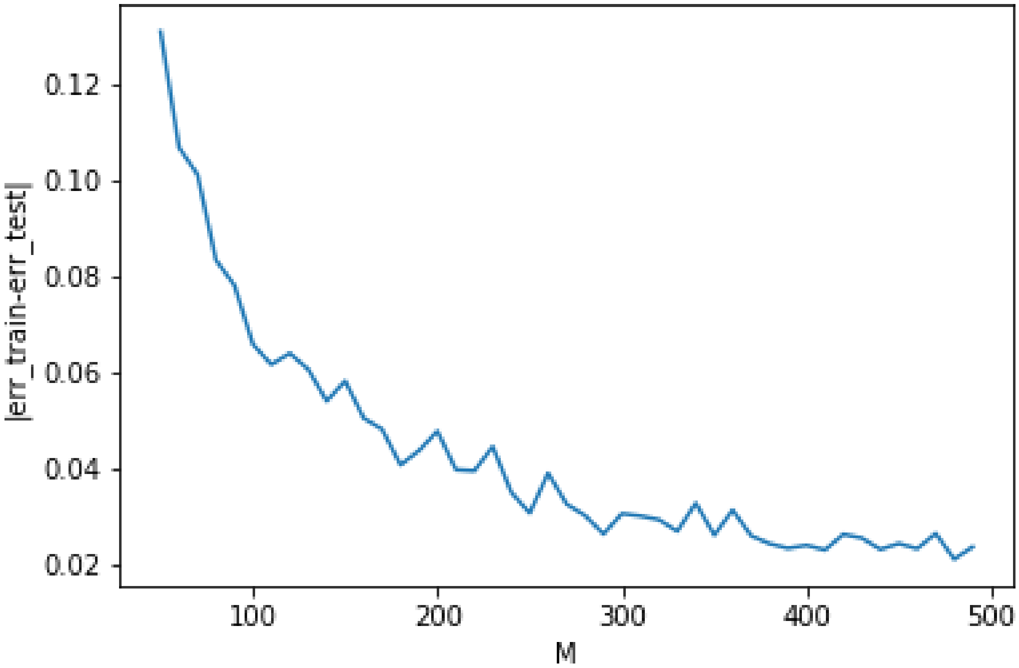

Fig 2. $|err_{train}(f) - err(f)|$ vs m

6. An alternative metric I(X) that can be used instead of information gain IG(X) is described value. As with information gain, the variable X used to split the data is the one with maximum I(X).

$$I(X) = \sum_{x} f(X = x) * |\, f(Y = 1|X = x) - f(Y = 0|X = x)\,|$$

where f(X=x) is the frequency of data points with value of the random variable X equal to x. f(Y=y|X=x) is the frequency of data points where Y=y given that X=x.

Here, the absolute difference term is an indicator of the distinguishing power of variable X when it is equal to x. and the other multiplication term f(X=x) performs the function of the magnitude by which the difference term is multiplied.

The comparison of the performance of the 2 metrics, information gain and the alternative, with respect to the marginal value of the variable is shown in Fig 3. As can be seen, the alternative metric actually performs better than IG. With a lower marginal value with the alternative, the performance of the decision tree trained on smaller datasets can be thought of as more accurate.

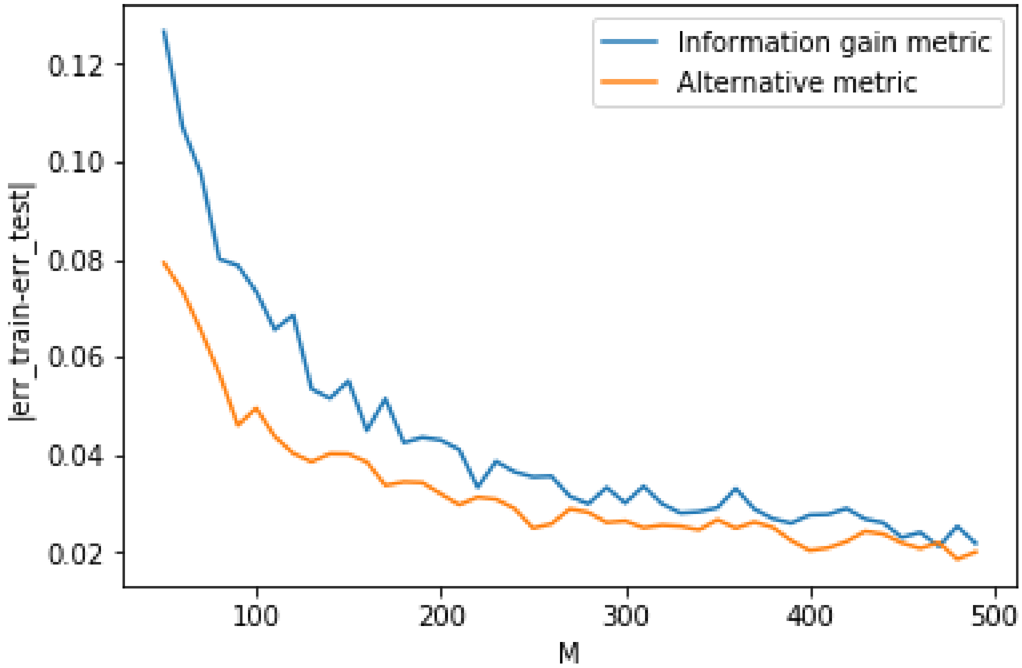

Fig 3. Information gain vs alternative metric