1.

a. For generation of data, pls refer to function generate_data() in generate_data.py file.
b. For fitting a decision tree on the data, pls refer to function decision_tree() in decision_trees.py file.
c. For calculating error, pls refer to function error() in decision_trees.py file.

The graph for error rate v/s m is given in Fig 1. The testing was done on M values between 0,15000 with 20 repetitions and the test data size was 1000. Yes, the observed data agrees with the intuition that as the training data size increases, the tree is able to perform better on the test data.

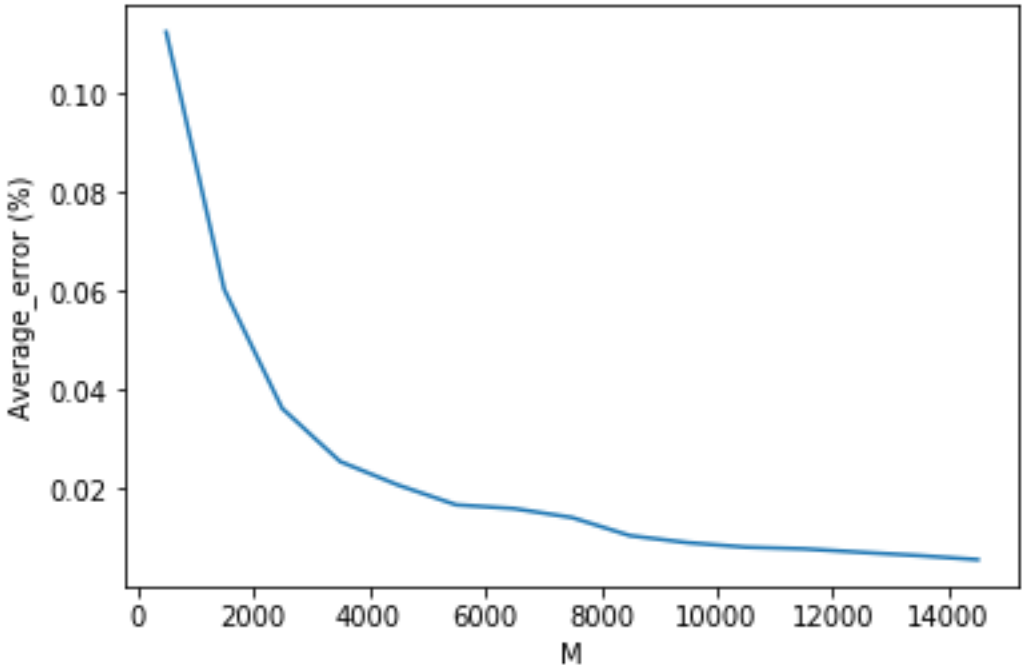



Fig 1. Average error v/s sample size M

2. Values for M were varied between 10,000 and 150,000 and the average number of irrelevant variables in the tree was observed over 20 trials. The data observed is shown in Fig 2. As can be expected, the number of irrelevant variables included in the tree increases with an increase in m. This trend is, however, not consistent and is mostly likely due to limited number of trials. The final value of the y-variable is expected to eventually reach 0 once all possible data points have been used for fitting.

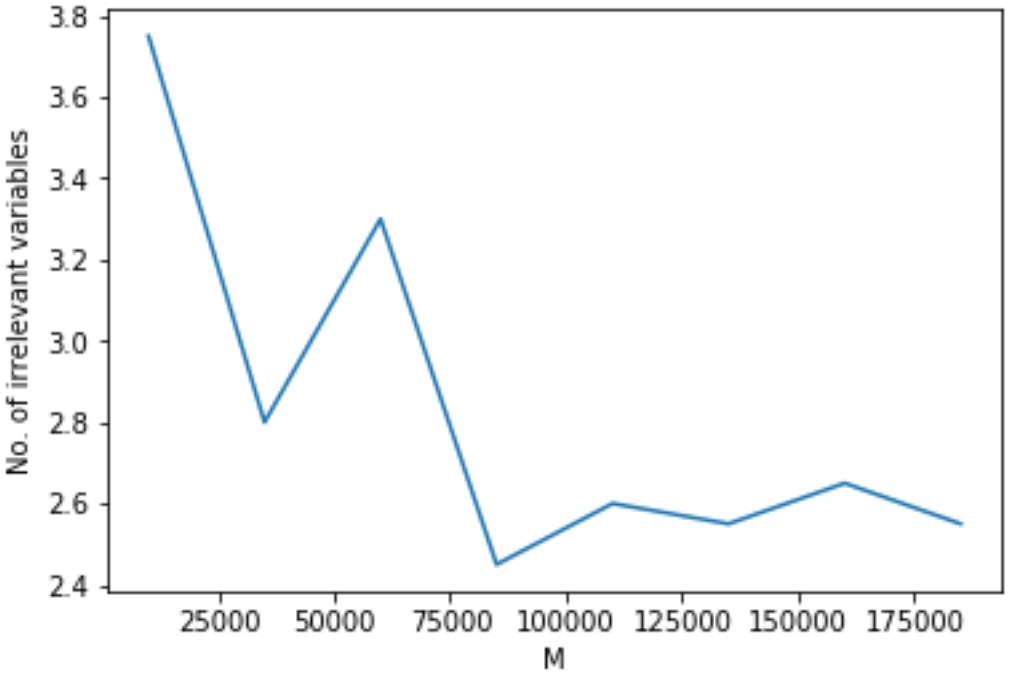



Fig 2. No. of irrelevant variables v/s m

3. a) The value of depth d was varied over the whole range 0 and 20 with testing performed on 20 repetitions. The error rate as a function of depth d is depicted in Fig 3. The data suggests value of d as 9, after which point the error rate becomes more or less constant.

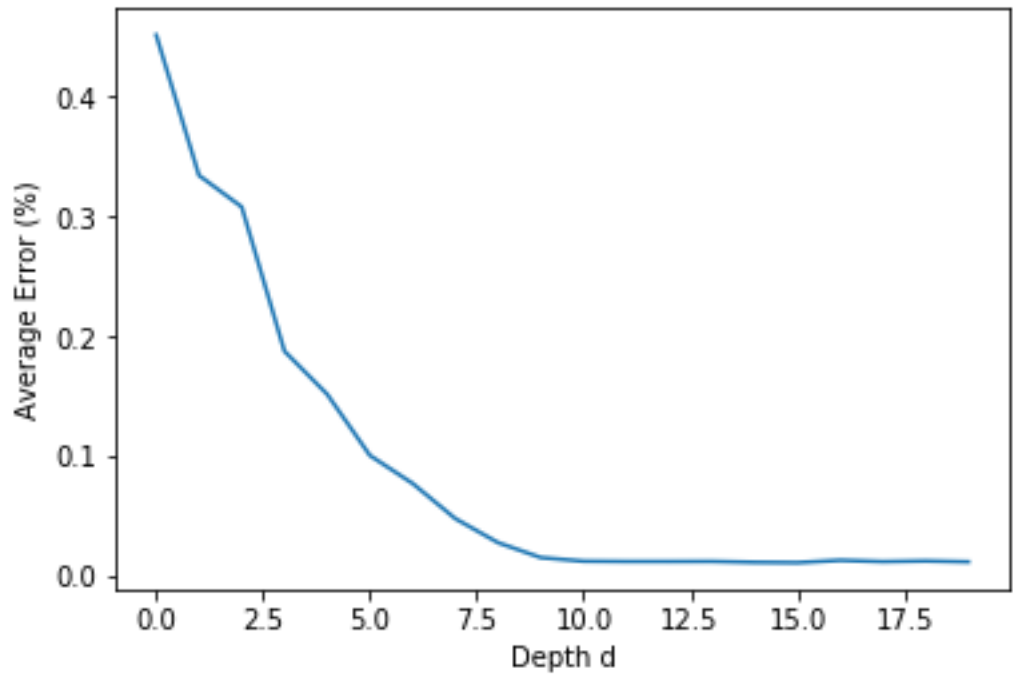
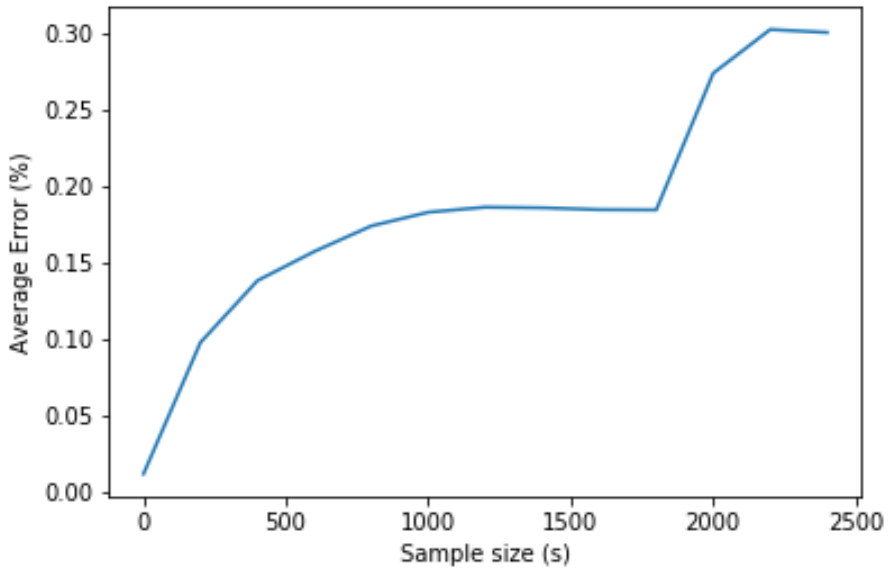
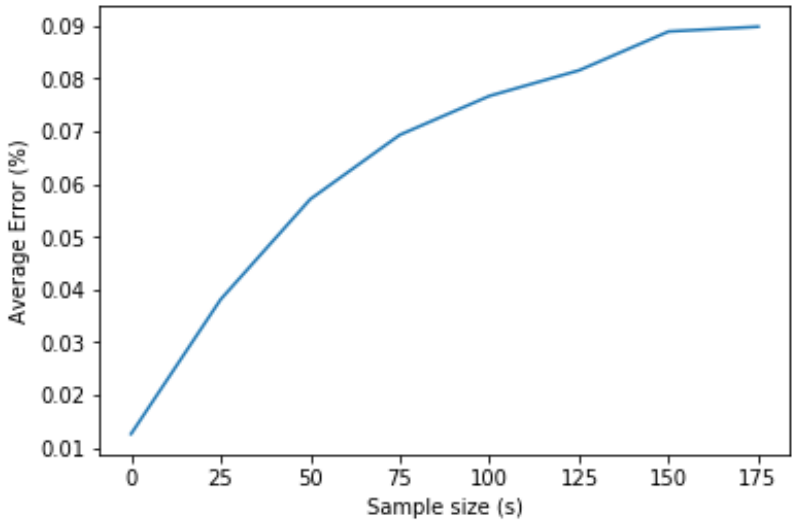


Fig 3. Average error vs depth of tree

b) The sample size was varied between 0 and 5000. The results of average error rate over 20 repetitions is shown in Fig 4a. As expected, smaller sample size gives lower error rates. However, the observed results show a good point for s as 1500, for a good part before which the error rate stays constant. Using such a value for s would give better total training time without compromising too much on performance. (The error rate v/s sample size for a different range of s is also shown in Fig 4b.)




(a)



(b)

Fig 4. Average error rates v/s minimum samples size s

c. The value of T was varied between 1 and 10. The avg. error over 20 trials is shown in Fig 5. The graph suggests a value of T0 as 1 at which the error rate appears to have a global minima. This property could, however, change with an increase in no. of trials. This testing could not be performed due to limited time.

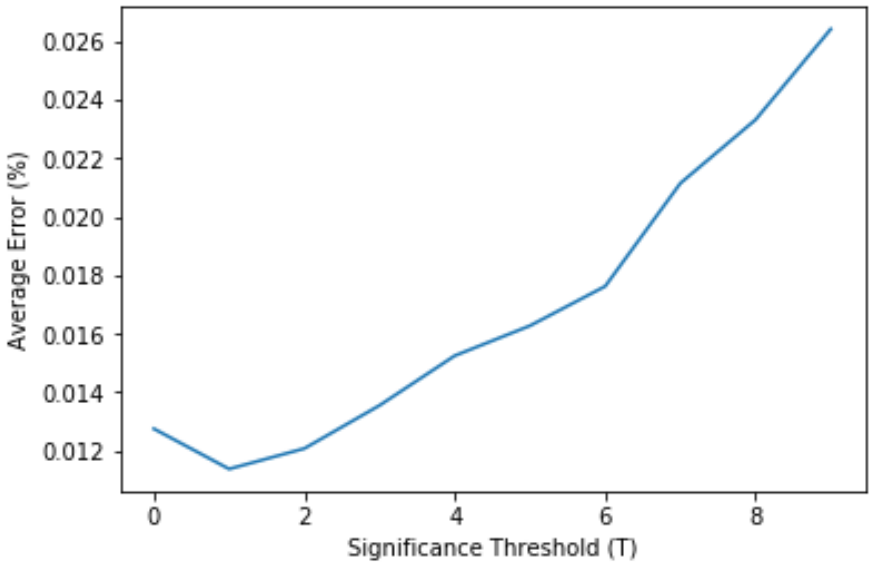



Fig 5. Variable Independence test. Error rate v/s Significance threshold T.

5. All parameters and conditions were retained from Q2, and the maximum allowed depth of the tree d was fixed at d=9 as was observed in Q3a. The average number of irrelevant variables included in the tree vary w.r.t. m in a manner like what was observed in Q2. The absolute number of irrelevant variables, however, was significantly lower.

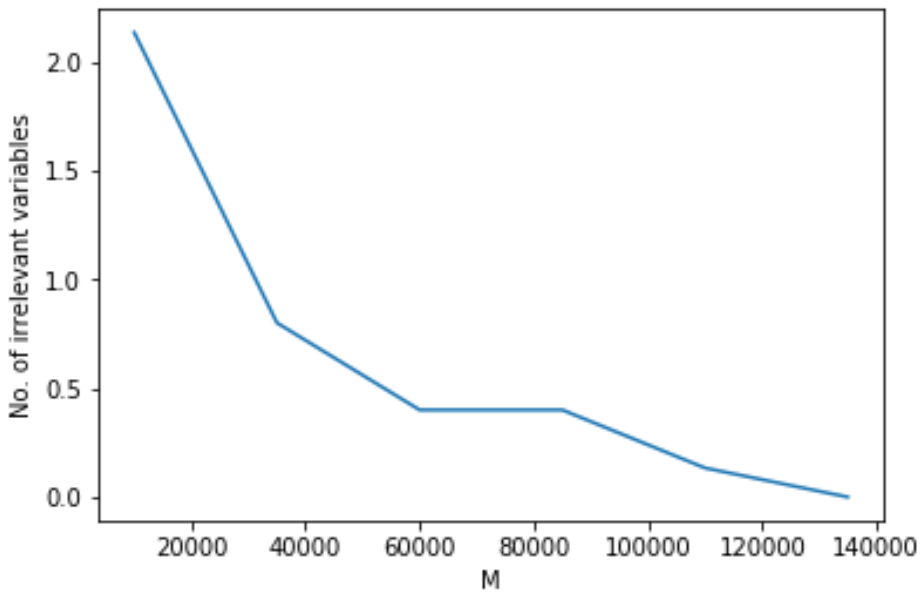



Fig 6. No. of irrelevant variables in tree with max. depth d v/s m

6. Once again, keeping the parameters similar to Q2, the value of minimum sample size s allowed to split the tree was kept fixed at 1500. Using this s resulted in having 0 irrelevant variables being included in the tree independent of the value of m, which is mildly interesting. This observation is shown in Fig 7.

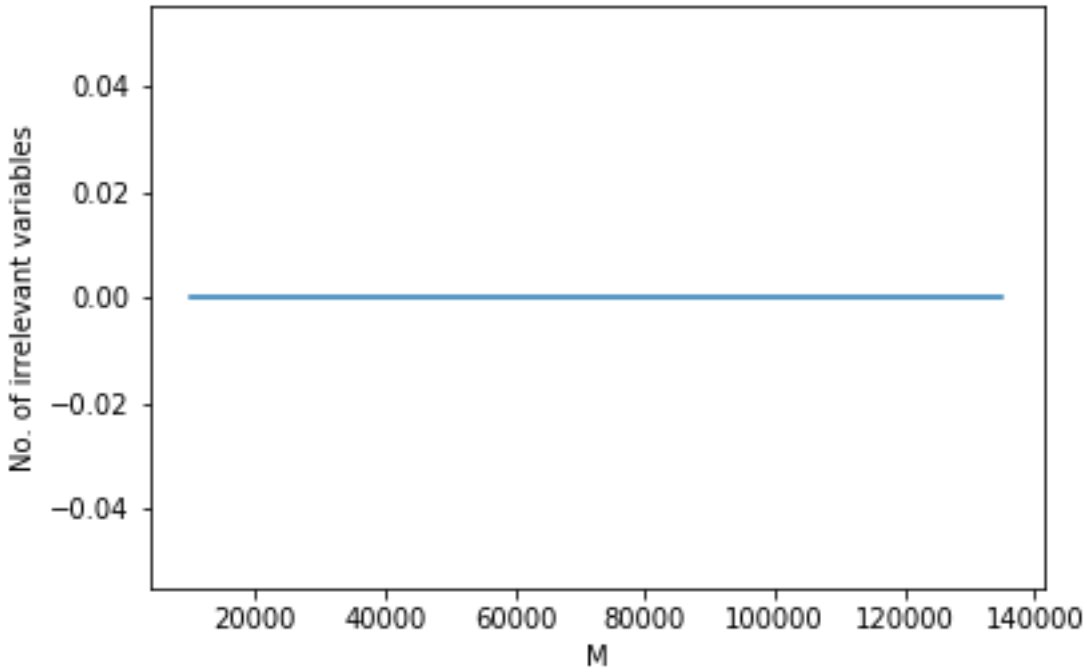



Fig 7. No. of irrelevant variables in tree vs m with minimum allowed samples size s=1500

7. With the same testing conditions, a significance threshold of T0=1 was used as indicated by the experiment in Q3c. The results observed are depicted in Fig 8. The trend observed, while inconsistent, is mostly a downward slope, which is what is expected by intuition and previous experiments. And similar to what was seen in Q2a, the absolute number of variables itself is lower than what was observed when the Chi-square test was not used.

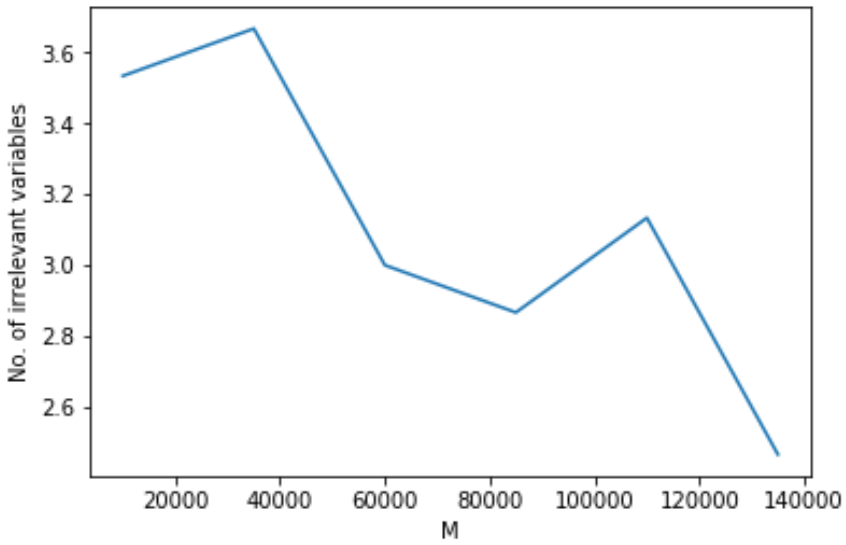



Fig 8. Number of irrelevant variables v/s m with significance threshold T0=1