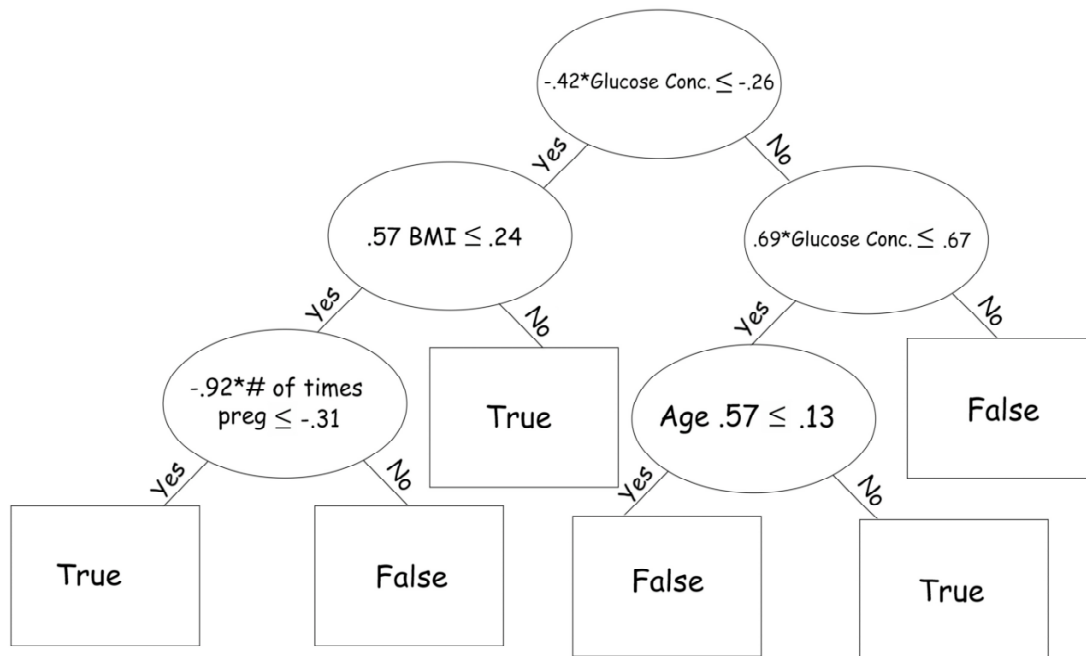


A sample tree produced by running the program is shown below



This decision tree has an accuracy of 76% on the PIMA Indians diabetes database, which is comparable with other data mining technique's accuracies of 65 – 83% on the same database. In comparison, using Weka's[46] J48 classification tree inducer (C4.5) without pruning and 5-fold cross validation (as the genetic program used here had) on the same dataset, gave a decision tree with 71.3542 % accuracy and 43 nodes.

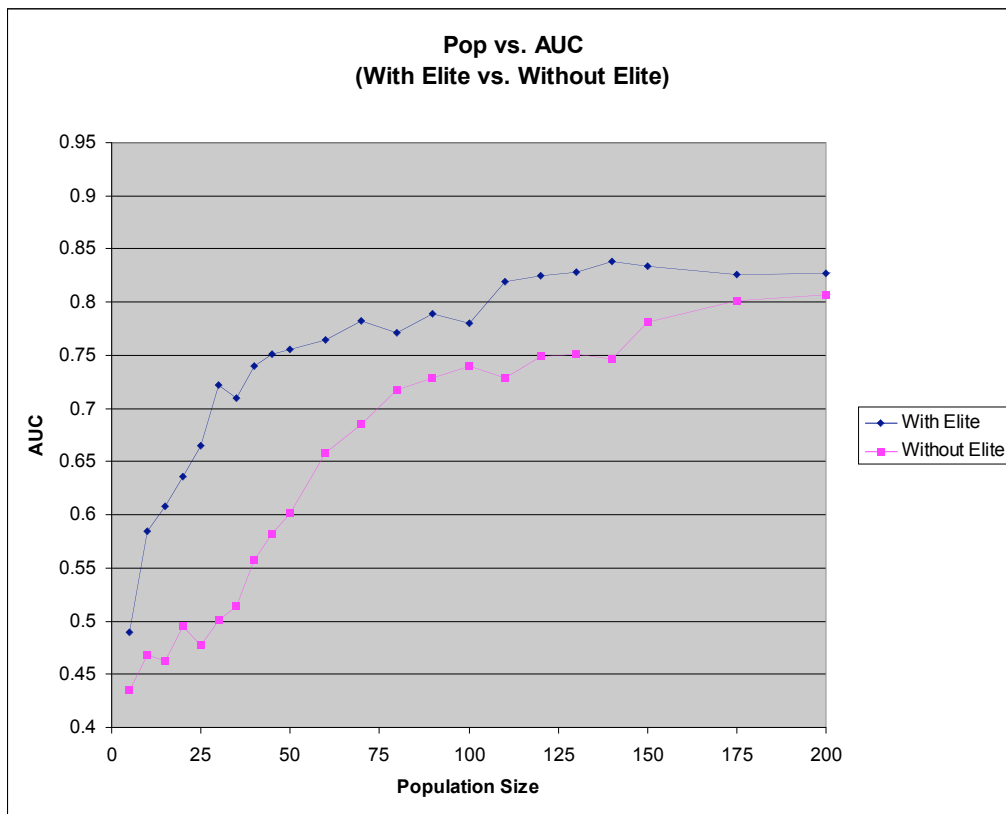
In this tree, the top node splits on the concentration of glucose. Fasting blood glucose level is known for being used as the standard for diagnosing diabetes [32, 36], so using it as the first split of the tree makes sense. The equation is the equivalent of: if Glucose concentration $\leq .565$. Since the attributes were linearly normalized between 0 and 1, the threshold must also be converted back to the original range for the Glucose attribute. Glucose values ranged from a low of 0 to a high of 199. $199 * .565$ gives 112.5. So, the first test is asking if the glucose concentration was ≤ 112.5 . Blood glucose levels of up to 100 are considered normal. Levels between 100 and 126 mg/dl are considered to be risk factors for type 2 diabetes and its complications. [34]. So the split threshold makes sense, considering current medical knowledge.

Evaluating the left child, the equation is the equivalent of asking if the Body Mass Index (BMI) is $\leq .42$. Converting this back to the original range (BMI values in the dataset ranged from 0 to 67.1) gives us: if BMI ≤ 28 . A BMI of 25 to 30 is considered overweight by the American Heart Organization [33], and weight is considered one of the largest risk factors to developing diabetes [37]. So the split and threshold in this node are understandable as well, and confirms that being overweight is a risk factor for diabetes.

Chapter 4.

4. Results

4.1 With and without Elitism



A plot of the GP run both with and without the elite sets, for a variety of population sizes, plotting the area under the ROC curve (AUC) for the pareto set at the end of 1,000 generations.

On the next few pages, the plot shows how the AUC changes as the generations progress during a single run, with several different population sizes, with and without the Elite sets.

Following that are plots of the three different types of mutation rates,