Data Mining - Assignment 1 - Classification Trees, Bagging and Random Forests

## Data Breakdown

We examined Eclipse's bug reports dataset comprising of 1,472 samples in this analysis. As shown in [Fig. 1](#Fig1), the reported bugs are from three different releases of Eclipse: 2.0, 2.1, and 3.0, and the majority of reported bugs are from the later release, with 661 bugs from release 3.0, while the earliest version in the records has the fewest bugs, with only 377.

|  |
| --- |
|  |
| **Fig. 1.** **Bug Report Distribution by Release in Eclipse Dataset.** This bar plot shows bug report counts for different Eclipse version releases in the dataset. The number of defects in each group is written in white on the right side of each bar. |

## Results

|  |
| --- |
| A diagram of a tree  Description automatically generated |
| **Fig. 2. Illustration of the first three splits in the single decision tree.** Each node displays the condition for the split at the top or indicates that it is a leaf node. The numbers on the right side of each node (with a blue background) represent the samples in that node belonging to class 0. The numbers on the left side (with a red background) represent the samples belonging to class 1. |

|  |
| --- |
|  |
| **Fig. 3. Confusion matrices of the three different trees constructed from the training dataset** |

## Discussion