Question 6: Performance

Dummy Dataset 1

The tree size of the dummy dataset 1 is 3, and the classification rate is 1.0. The tree size is small because we only have 1 attribute and 20 examples in the dataset. In addition, this attribute alone (5) classifies the dataset perfectly from the decision tree. This is the reason that results in the perfect classification rate.

Dummy Dataset 2

The tree size of the dummy dataset 2 is 11, and the classification rate is 0.65. From the decision tree, we can see that it splits on 5 attributes. However, we only have 20 examples in the dataset, which is too small to capture the usefulness of classification in decision tree. Hence, it can lead to the situation that there is not a deciding attribute that splits the dataset nicely. Therefore, we have a low classification rate.

Car Dataset

The tree size of the dataset is 408, and the average classification rate is 0.94625. The dataset is relatively small, leading to a smaller size of the decision tree. In this dataset, there are 6 attributes. 3 of these attributes have 4 distinct values, and the other have 3 distinct values. In total, there are 1728 potential combinations based on these attributes, which is the same as the number of examples within the dataset. It means that even one attribute has a great influence classifying the dataset into different pieces. After all, the dataset has a relatively high classification rate maybe due to the fact that the dataset covers most of the combinations of attributes, resulting in better classification.

Connect4 Dataset

The tree size of the dataset is 41521, and the average classification rate is 0.7675. The dataset has 43 attributes, which is really big. In addition, we have 67557 examples within the dataset. These are the factors that make the tree size so huge. The decision tree does not do a good job classifying the data, which somewhat makes sense since there are so many possible permutations based on 43 attributes. Furthermore, it is not appropriate to use positions of the board as attributes because they are not independent. A player’s move might influence another player’s move, leading to potential interdependence between board positions. Thus, this factor adds some noise during the classification process, leading to a low classification rate.

Question 7: Applications

Car Dataset

In the Car Dataset, we use several attributes, such as safety, price, and etc. to determine the value of a car. Similarly, in the medical field, we can use this method in the same manner. We use a couple of symptoms to determine the disease. For a specific disease, the potential symptoms might include cough, flu, and so on. Decision Trees can be used on this type of dataset to decide the illness of a patient in a logical way.

Connect4 Dataset

One of the algorithms that we can utilize to improve the Connect4 playing bot is the minimax algorithm. The condition of this game satisfies the requirement of the minimax algorithm (adversarial search): there are two players, all the players know the current state of the board, and there is a winner and a loser at the end. However, as we use this algorithm, we should limit the number of steps that the max player can look ahead because there are so many attributes in the dataset, leading to so much time spent searching for the result. The number of attributes is still the same, and we can use the minmax to determine the move to take in the decision tree. In some sense, it feels like “pruning” the decision tree since using minimax allows us to eliminate a lot of unnecessary instances.

Question 8: Novel Dataset

Breast Cancer Dataset

Performance

The tree size of the breast cancer dataset is 485, and the average classification rate is 0.64775. The tree size is not really huge because we have only 9 attributes to split on. However, the classification rate is pretty low due to the fact that there are only 277 examples in the dataset. Since we have 505440 combinations out of these 9 attributes, more data are needed in order to make the decision tree effective.

Applications

Although more data are needed for this dataset to become effective, this type of dataset can have a great influence in the medical field. Doctors can utilize the decision tree generated by the dataset to determine if there is a recurrent event in a patient in a logical manner. In addition, the decision tree can also be used to inform the general public. People can use this information to know the general trend of this illness.