**Decision Tree and Ensemble Learning**

**Task 1**

For the Titanic challenge (https://www.kaggle.com/c/titanic), we need to guess whether the

Individuals from the test dataset had survived or not. Please:

**1) Preprocess your Titanic training data**

I have done some preprocessing to the Titanic data to get it ready for modeling.

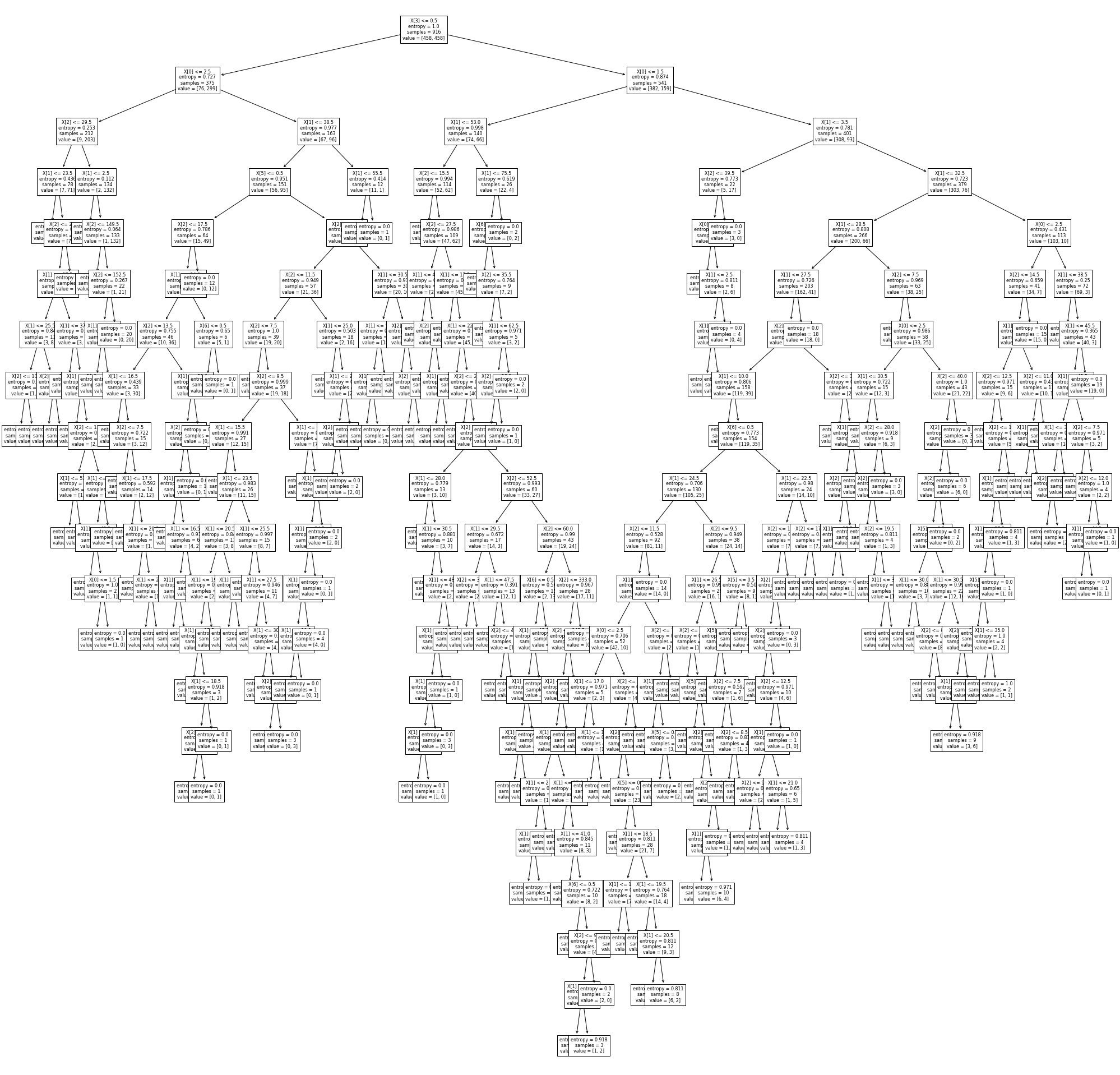
1. First I removed the columns PassengerID, Name, Ticket for having too many categories which distort the model. I also removed the Cabin for having too many missing values
2. The categorical data was transformed to have a numeric tag
3. The missing values in the Embarked and the Age were being imputed by the KNN method
4. For the categorical variables of Sex and Embarked, dummy (0,1) variables created
5. Variables are selected using the Pearson correlation with target and kept the variables that had more than 7% correlation with target
6. Imbalanced data is being treated by synthesis oversampling of the categories with lower representation
7. Once the dataset became tidy and ready to be modeled I applied the decision tree and random forest bagging method and compared them with each other

**2) Select a set of important features. Please show your selected features and explain how you perform feature selection.**

1. Features are selected after I created the dummy variables of the categorical data of Sex and Embarked.
2. 'pclass', 'age', 'fare', 'sex\_0.0', 'sex\_1.0', 'embarked\_0.0', 'embarked\_1.0' variables were being selected (sex\_1 🡪 Female , sex\_0 🡪 Male, embarked\_0 🡪 S , embarked\_1 🡪C )
3. I used the Pearson correlation of the variable with the target and selected the columns that had a correlation more than 7 %

**3) Learn and fine-tune a decision tree model with the Titanic training data, plot your Decision tree**

1. I ran the Grid Search Cross Validation over the parameters of decision tree and the Grid search found the best accuracy on using the criterion of Entropy and max-\_depth of 20
2. The plot for the decision tree is as below:



**4) Apply the five-fold cross validation of your fine-tuned decision tree learning model to the Titanic training data to extract average classification accuracy**

1. The fivefold cross validation was applied to the dataset and the average prediction accuracy of the decision tree model is **0.8068**

**5) Apply the five-fold cross validation of your fine-tuned random forest learning model to the Titanic training data to extract average classification accuracy**

1. I used the grid search cross validation to find the best set of attributes for the random forest model and the mixture bootstrap=False, criterion='entropy', max\_depth=20, min\_samples\_leaf=2, min\_samples\_split=5, n\_estimators=400 was the best tune for the model.
2. Fivefold cross validation was applied and the average classification accuracy improved by 2 point to **0.8264**

**6) Which algorithm is better, Decision Tree or Random Forest?**

1. The random forest improved the classification accuracy and since the algorithm is using a randomized feature selection bagging method I am choosing this model for this prediction

**7) What are your observations and conclusions from the algorithm comparison and analysis?**

1. This is the comparison between how decision tree and random forest are able to model a classification problem in 7 dimensions space. The observation is that although the Random Forest seems to be a more robust bagging method, but still the decision tree performs very well in a classification problem like what we have. The other observation is that with the same dataset some more advanced model may only be able to improve the classification accuracy by a few points and the higher the accuracy, the harder to improve the prediction accuracy

**The link to the code is as below:**

[**https://github.com/nimaarvin83/CAP5610/blob/main/HW2\_Titanic/titanic.ipynb**](https://github.com/nimaarvin83/CAP5610/blob/main/HW2_Titanic/titanic.ipynb)

**Task 2**

**Consider the decision tree shown in the diagram below. The counts shown in the leaf nodes correspond to the number of training records associated with the nodes.**

1. **What is the training error rate for the tree? Explain how you get the answer?**

Number of leaf nodes (K) = 6

The optimistic training error

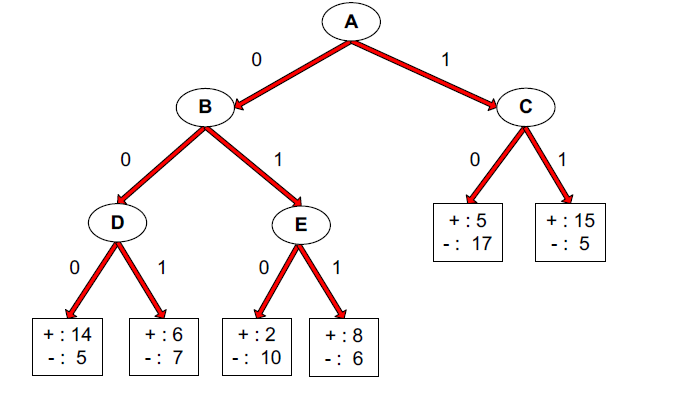
The pessimistic training error

For example in the node ABD0 which we have +14 & -5 , means that the -5 is incorrectly classified because the data points were mostly classified with plus but 5 of them were being classified incorrectly as negative

1. **Given a test instance T={A=0, B=1, C=1, D=1, E=0}, what class would the decision tree above assign to T? Explain how you get the answer?**

For the data point T we just follow the path that the decision tree leads us. Since A=0 then the tree leads us from A to B. Since, B=1 the tree leads us to E. and Since E=0 then the tree leads us to the leaf node of +2 & -10.

Since in that node the data points are mostly classified as negatives, the tree will assign the data point T to class Negative (-).



**Task 3**

Consider the following data set for a binary class problem.

The formula for entropy is:

**Q1: What is the overall entropy before splitting?**

Since we have 10 data points, 4 from class of + and 6 from class of - , then we calculate before splitting Entropy as below:

**Q2: What is the gain in entropy after splitting on A?**

Entropy after splitting on A :

Gain after splitting on A = 0.97095-0.68964 = **0.28131**

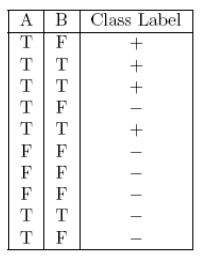
**Q3: What is the gain in entropy after splitting on B:**

Entropy after splitting on B :

Gain after splitting on B = -0.7144= **0.2565**

**Q4: Which attribute would the decision tree choose?** The decision tree will choose the attribute A because it can have more information gain

**Q5: Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations. (We want to split first on the variable which maximizes the information gain until there are no nodes with two class labels. )**



**Task 4:** Please answer and explain.

Q1: Are decision trees a linear classifier?

Q2: What are the weaknesses of decision trees?

Q3: Is Misclassification errors better than Gini index as the splitting criteria for decision trees?