# Problem 1 – Classification tree (35%)

The dataset ‘titanic.csv’ included in the assignment contains information about passengers onboard RMS Titanic on its maiden voyage. Your goal is to build a classification tree to predict which passengers survived based on the provided information.

## 15% credit. Grow a classification tree on the entire training set.

Visualize the tree (e.g., as in L11, slide 24), discuss which variables are selected at each split and explain the meaning of those decisions, analyze node purity, etc. This part of the exercise is open-ended. Explore the model and see what you can find out.

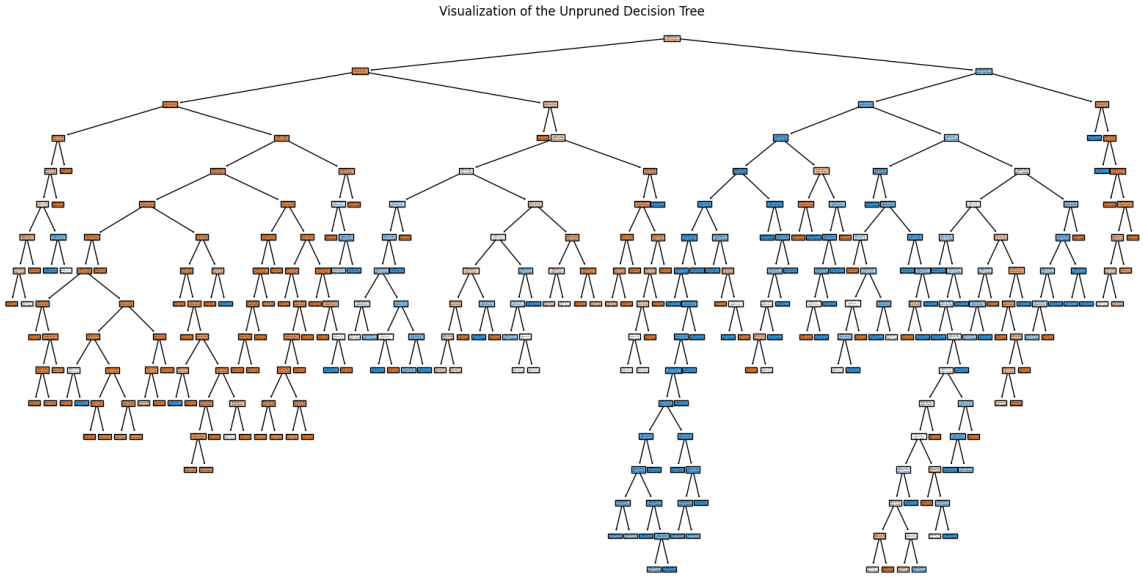


*Figure 1. Unpruned Decision Tree Grown on Entire Training Set*

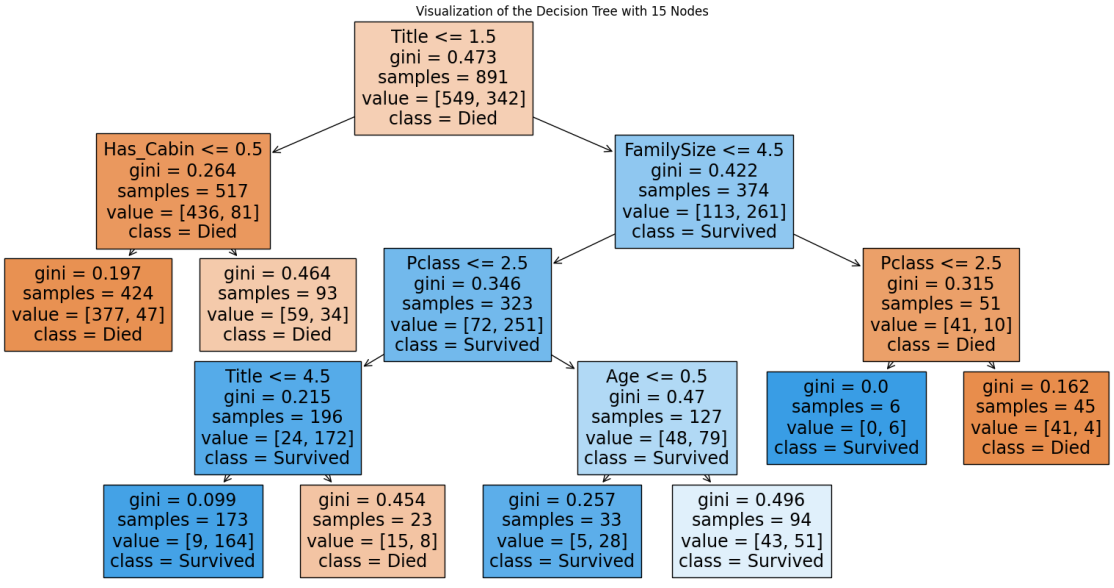


*Figure 2. Decision Tree with Pruning Parameter 0.005.*

Discussion of findings:

We grew two classification trees on the entire training set. One, which was unpruned, and another in which we limit the number of splits by applying an alpha value. The reason why is because the unpruned tree contains so many splits that it is difficult to interpret and discuss the variables selected at each split and the reasons behind decisions.

***Discussion of Features Picked in Figure 2:***

First, let's discuss some of the features that were picked in the above tree displayed in Figure 1.

*Features:*

- Title is a feature that retains information about the passenger's name. One reason why we do this is because there could be relevant information drawn from a name that can indicate class, wealth, or privilege. In this case, 1 refers to those with the Mr. title, 2 is the Master title, 3 is the Mrs title, 4 is the Miss title, and 5 refers to those with rare titles of privilege, such as Dr, Lady, Countess, and etc.
    - Perhaps more privileged passengers had a greater likelihood of escape, due to the perception of increased wealth or status. This assumption is based on literature review that women were more likely to survive than males, and that those with higher status had a higher chance of escaping (based on this Research Article).
- Has_Cabin is a feature that states whether or not a passenger had a cabin aboard the Titanic. In the dataset, some passengers did not have any cabin listed, and I interpreted this as individuals who may not have had an individual cabin to themselves (Hypothesis referenced from this document that makes the same assumption).
    - Perhaps those with cabins potentially had a higher chance of escape, as having a personal cabin is indicative of increased escape. Assuming that having a cabin implies that you are not a stowaway, or perhaps that a passenger was a crew member, perhaps not having a cabin made you more likely to die, as evacuation teams would place less importance on such passengers. This is further backed up by this Wikipedia article, which breaks down the casualties for the Titanic and takes note of the high casualty rate of staff members aboard the Titanic. Perhaps this is because crew members were assisting other passengers in escape, and had to remain on the Titanic longer.
- FamilySize is a feature that combines that of SibSp and Parch. This new feature counts the total number of relations, including the passengers themselves, aboard the Titanic. One thing of note is that there were some large families aboard the Titanic, as referenced by this Wikipedia article.
    - Perhaps larger families evacuated together aboard the Titanic, to avoid the splitting of families. Another reason why having a large family could potentially indicate successful evacuation is that large families often include many children. A common phrase often quoted when speaking on the topic of the Titanic is "Women and Children first". It would not be a stretch to believe that having a large family could be indicative of higher survival rates simply because it could

imply the existence of children, who did have higher survival rates based on the above articles.

- Pclass is a feature in the original titanic dataset that described the class of a passenger's ticket (i.e. first, second, or third).
    - Not only can Pclass potentially determine the privilege and status of a passenger (as richer passengers could afford the first class ticket) but in addition the the prior, Pclass also determined the location of passengers on the Titanic.
    - Based on this article, Third class passenger casualties were high, which may have been in part due to the location of the cabins. According to the article, "Ship's regulations were designed to keep third-class passengers confined to their area of the ship [...] leaving many of the confused third-class passengers stuck below decks" (Wikipedia). Thus, it is plausible to think that Pclass is a good indicator of survival.

***Discussion of Splits and Meaning of Decisions:***

For the sake of succinctness, we only discuss the Gini Indexes of the leaves of the decision tree, or notable splits that either reduced or increased the Gini Indexes of the resulting nodes. All of the following interpretations are based on Figure 2.

**Leaves of Has_Cabin <=0.5:**

For the first leaf with Gini Index value 0.197, the model classifies the 377 of the 424 nodes as class "Died", and is the result of splitting on Title <= 1.5 and splitting on Has_Cabin <=0.5. Thus, the first node predicted that those who did not have a cabin, and also had the title "Mr." Died. The Gini-index value is somewhat low, and on further inspection of the values within the node, we see that 377 samples were classified as dead, but 47 survived, which tells us that the node achieves better purity then that of its parent node.

The second node with a gini-index value of 0.464 is the result of splitting on Title <= 1.5 and splitting on Has_Cabin <=0.5 (i.e. those with the title Mr. and had a cabin). This node doesn't have a high purity, and on closer inspection of the values within the node, we see that 59 samples were classified as dead, but 34 survived. These values are almost equal to each other, hence the high Gini-Index value, as the split actually introduces more impurity from that of the parent node.

What is notable is that for both this leaf and the second leaf with Gini-Index 0.464, though both split on Has_Cabin <=0.5, both outcomes of the leaves predict Death.

Overall, this classification aligns with our understanding of what occurred during the evacuation of the Titanic, as the title Mr. not only denotes gender as male, but also denotes a lower class (as men could have titles that depicted more privilege, such as Dr. Col, Capt, or etc.) . From what we know, men and those with lower status were more likely to die.

**Leaves of Title <= 4.5:**

The first leaf has a gini index 0.099, which is a good score considering that the node contains 9 samples classified as Dead, and 164 samples who survived. The individuals in this node include individuals who do not have the “Mr.” Title or a “Rare” Title, have a family size that is less than 4, either a first or second class passenger. This tells us that those with the titles “Master”, “Mrs” and “Miss” with the above criteria were more likely to survive. This aligns with our understanding of the Titanic Disaster, as those with higher status, those who classified as Women (as indicated by Title) and were of first and second class passenger status were more likely to survive.

An interpretation that makes a little less sense is that of the sibling node of the above. The sibling node includes individuals who have the “Rare” Title, were of first or second class passenger status, and had a family size of less than 4. Based on the Titanic disaster background, one would think that those with “Rare” titles were more likely to survive as “Rare” titles indicate more status and privilege. However, the Gini-Index value is observed to be 0.454; as the Gini-Index approaches 0.5, the node gets closer to almost equal distribution of both classes. Upon further analysis, we see that in this node, 15 of the 23 samples were classified as Died, but 8 samples survived. This node doesn’t have as high purity as desired, so perhaps further splits at this node can increase purity of the children nodes.

**Leaves of Age<=0.5:**

The children of this split includes individuals who don’t have the “Mr” title, have a family size of less than 4, and are considered Third class passengers. Both of the children nodes are classified as “Survived”; however, both contain differing levels of purity.

The first child, with a gini-Index value of 0.257, includes individuals who had the above characteristics, and were either 16 or below in age. This observation makes sense, as children were more likely to survive the Titanic disaster, but third class passengers had a less likely rate of survival due to location. The node does not achieve perfect purity, as 5 individuals out of the 33 were classified as Died in the original dataset, but has a higher node then that of its sibling node.

Its sibling node has a gini-index value of 0.496, which indicates a high level of impurity. Samples within this node have the above characteristics as stated in the first paragraph, but were older than 16. Upon observation of the samples, though testing samples with the described traits would be classified as Survived, we see that 43 samples had Died in the Titanic, and 51 had survived. These two numbers are close to each other, so perhaps another split is necessary to increase separation of the two classes. Thus, this impurity aligns with our understanding of the Titanic disaster as we know that those who were not classified as “children” were less likely to survive.

**Leaves of Pclass<=0.5:**

The children of this split includes individuals who don't have the "Mr" title and had a family size greater than 4. The children nodes differ in the following ways.

The first node has a gini-index value of 0 and includes the individuals who were either first or second class passengers.This tells us that this node is pure, and all samples used to train this model that have the above traits all survived. Upon further inspection though, we see that though this node is pure, there are only 6 samples that fell into this category, which raises the question: does this split truly achieve purity and the above selected features are good indicators of guaranteed survival or will the model perform differently and misclassify unseen data and this node is pure just by luck? Overall, this interpretation aligns with understanding of the Titanic disaster, as those of higher class were more likely to survive. Perhaps those with larger families were more likely to survive as they were more likely to stick together and evacuate together, especially with the added privilege of being a higher class passenger.

The second node also has a relatively low gini-index value, and includes individuals who were third class passengers with the above features in the first paragraph. Samples within this node are classified as Died, and the node achieves relatively high purity, as there are 41 samples within this node from the training set who were correctly identified as Died, and 4 of the samples survived. This aligns with understanding of the Titanic Disaster, as there were many large third class families who perished in the Titanic, such as the Sage Family, and the Andersson Family. Whether this implies status inequality on the ship due to refusal to allow these families to evacuate, or because of location on the ship, interpreting this information is somewhat difficult without further understanding of procedures on the Titanic.

***Summary (If you feel reading the above is too lengthy):***

- In short, based on the above decision tree, if the observation includes a "Mr" Title, then it is classified as "Died".
- If the sample does not have a "Mr" Title, has a family size greater than 5, and is a second or first class passenger, then they are classified as "Survived" with a Gini-Index Purity of 0.0, which implies that this classification is pure and most likely doesn't have exceptions (at least based on the tree. This doesn't say anything about testing and validation sets!).
- Else, if the sample has a family size greater than 5, but is a third class passenger, they are classified as "Died".
- If a passenger has a family size less than 4, and is a third class citizen, they are classified as "Survived".
- If a passenger has a family size less than 4, and does not have a "Rare" title (which indicates privilege and status) they are classified as "Survived".
- If a passenger has a family size less than 4, and has a "Rare" title (which indicates privilege and status) they are classified as "Died".

## 10% credit. Repeat the previous process, but this time using different values for the pruning hyperparameter $\alpha$ (see L11, slide 20).

Analyze and discuss the results.

For the sake of this problem, we tested with alphas `[0, 0.005, 0.01 , 0.015, 0.02 , 0.025, 0.03 , 0.035]`.

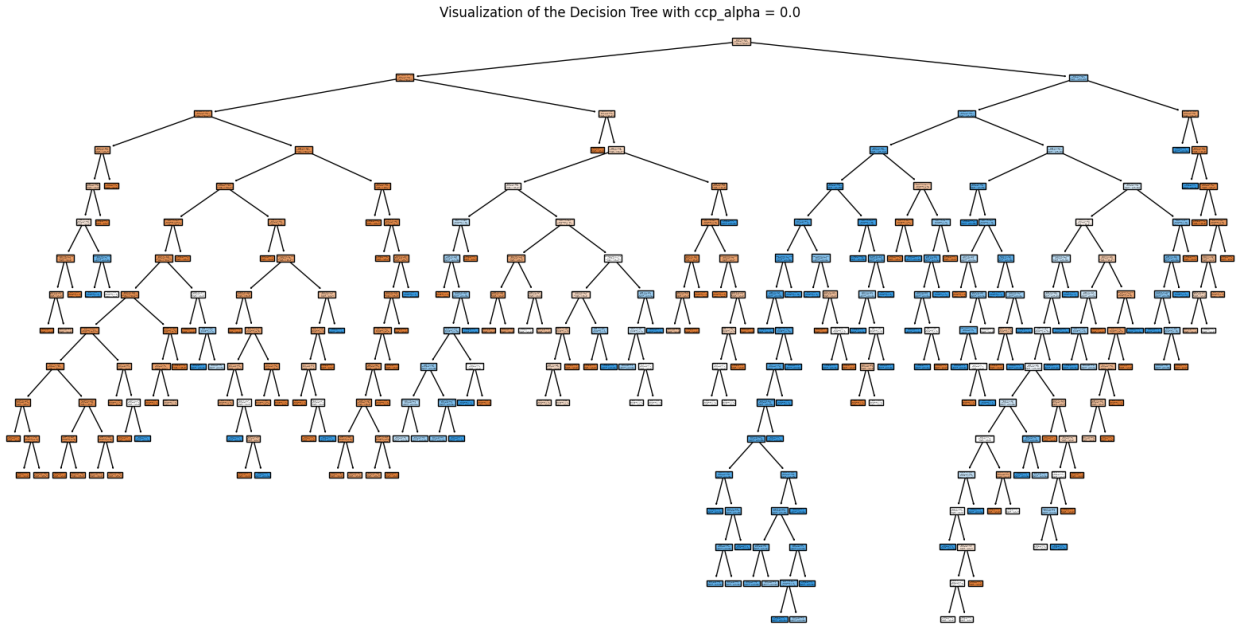


*Figure 3. Decision Tree with Pruning Parameter 0*

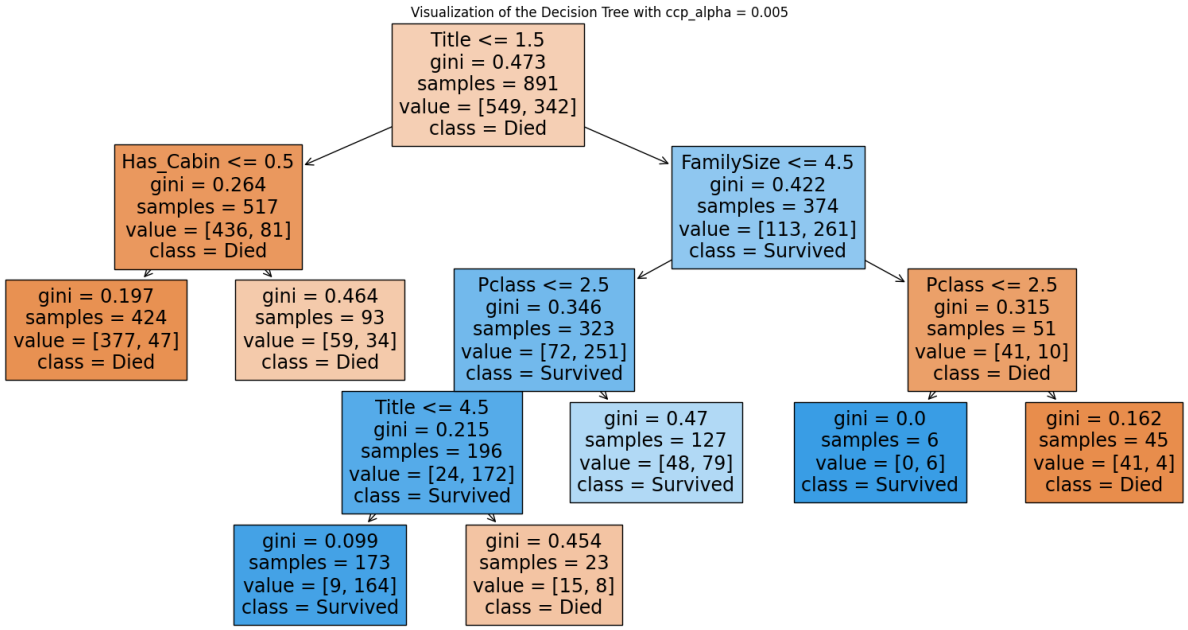


*Figure 4. Decision Tree with Pruning Parameter 0.005*

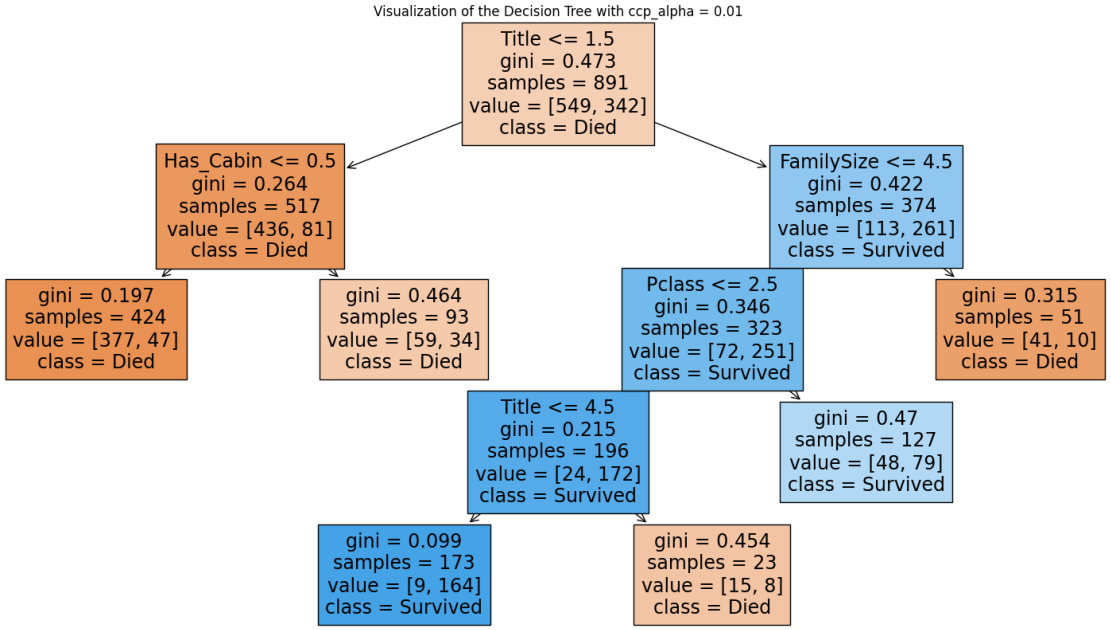


*Figure 4. Decision Tree with Pruning Parameter 0.01*

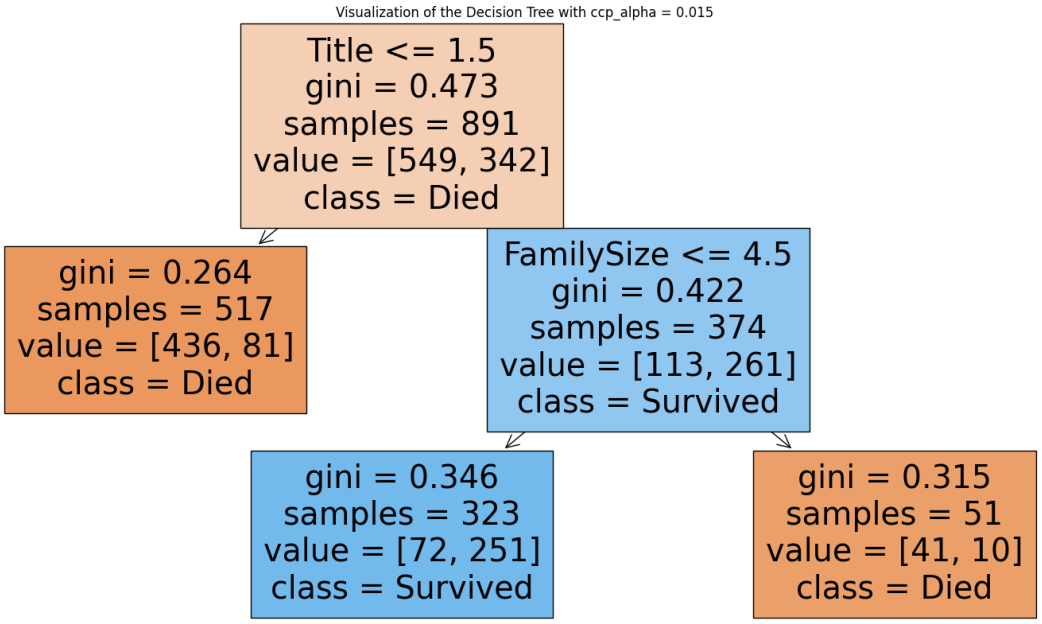


*Figure 5. Decision Tree with Pruning Parameter 0.015*

*Note: To save space, we removed the visualizations for trees with Pruning Parameter 0.02, 0025, and 0.03 as choosing these pruning parameters results in a tree that looks like the above. No further nodes were pruned following selection of these alphas.*

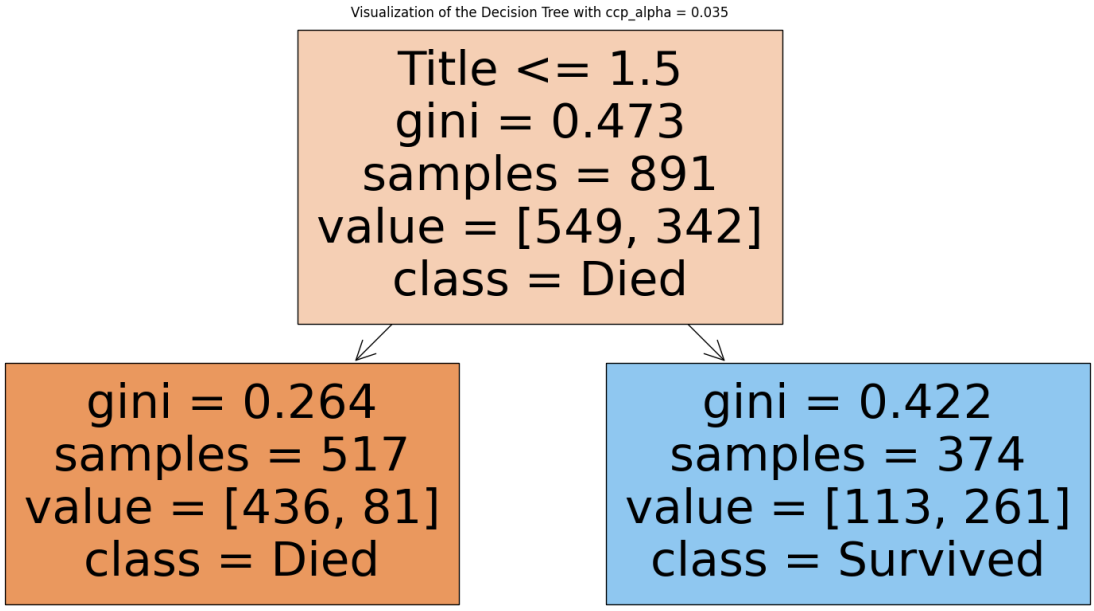


*Figure 6. Decision Tree with Pruning Parameter 0.035*

Discussion of findings:

Alpha allows for us to control the tradeoff between the model complexity and fit to the data by preventing overfitting. Thus, based on the above findings and visualizations of Decision Trees grown with different pruning value parameters, we see that the Figures above align with our understanding of Tree pruning.

For example, in Figure 3, the decision tree with pruning parameter 0 creates a tree that is analogous with the unpruned decision tree. However, upon setting the pruning hyperparameter $\alpha$ to 0.005, we get Figure 4, which results in a significantly pruned down subtree. This is interesting, as the change from Figure 3 to Figure 4 is very drastic. This is most likely because the Tree in Figure 3 includes very many splits, which means that the tree will most likely produce good predictions on the training set; however, it overfits and provides poor prediction on test data.

In Figure 4, which uses a pruning parameter value of 0.01, reduces the number of nodes again in the tree by getting rid of the splits Age<=0.5 and Pclass <=2.5. This results in a subtree that is 4 nodes smaller than that of Figure 3.

In Figure 5, we get a tree model that uses pruning parameter 0.015, which reduces the number of nodes from Figure 4 by 7 nodes. This is a much more drastic pruning change then that of the prior model, as it removes the splits Has_Cabin <= 0.05, Pclass <= 2.5, and Title<=4.5. This tells us that this pruning parameter creates more dramatic change then that of 0.01 when we compare the model using pruning parameter 0.01 to 0.005.

Finally in Figure 6, we get a tree model that contains only one split, which uses the pruning parameter value of 0.035. What is notable about this model is that increasing alpha values from 0.015 by increments of 0.005 to 0.03 results in a tree model that looks identical to that of Figure 5. This makes sense, as in order to penalize more nodes in a tree, a larger alpha value is needed to prune.

## 10% credit. Estimate classification performance using 10-fold cross validation.

Discuss which pruning criteria you used and how it impacted classification performance on unseen data.

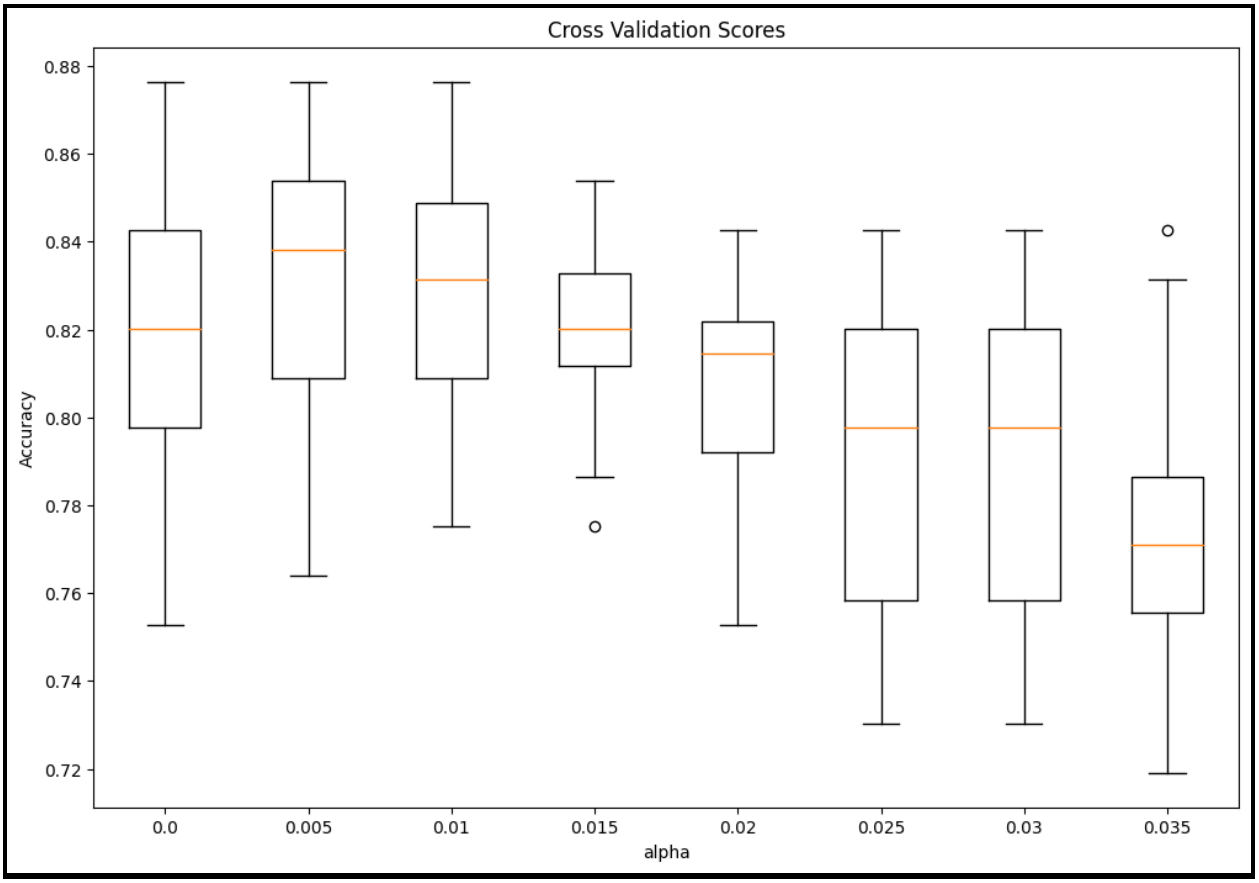


*Figure 7. Cross Validation Classification Performance Against Pruning Parameters*

Discussion of findings:

After estimating classification performance for the decision tree model using 10 cross fold validation for pruning criteria alphas 0.0, 0.005, 0.01, 0.015, 0.02, 0.025, 0.03, and 0.035, we perform ANOVA to see if the classification performance is statistically significant.

Our null hypothesis is that there is no significant difference between the different model classification performances, and that all group means are equal. The alternative hypothesis is that there is a significant difference between at least one pair of the models in the set of all models.

We choose a significance level of 0.05 and get a resulting p-value of 0.0361. This tells us there is a significant difference between the classification performances of the models given that the alphas, or pruning criteria, change.

In terms of the “best performing alpha” on average, the ideal pruning parameter was found to be 0.005, or the tree shown in . This is notable as larger alphas act as a penalty term for more nodes, which tells us that for an ideal decision tree using our pre-processed data, while an unpruned decision tree is not the ideal model, in turn, a heavily pruned decision tree is also unideal. In fact, a model with a selected alpha value of 0.035, which results in Figure 6, which only contains a single split is notably the worst performing on average compared to the other pruning parameters. This aligns with the model displayed in Figure 6, as the resulting leaves of the decision tree have very large sample sizes, and relatively high Gini Index values (especially that of the second leaf.

In general, introducing pruning criteria helped to increase classification performance on unseen data, as alpha values  0.005, 0.01, and 0.015 had either improved, or the same classification performance on average as the unpruned decision tree.

## References:


- *Introduction to Decision Trees (Titanic dataset) | Kaggle*
    - Data pre-processing information was taken from this notebook,  which helped a lot in retaining vital information and converting some columns into usable information:
        - Changing Sex to binary value.
        - Mapping different values of Fare to values between 0-3 based on the fare price. (higher fares mapped to larger values)
        - Mapping age to values between 0-3.
        - Determining family size of people.
        - Determining whether or not a passenger is alone on the trip
        - Mapping the titles so we can retain some information regarding names.
            - (i.e. those with titles such as `'Lady', 'Countess','Capt', 'Col','Don', 'Dr', 'Major', 'Rev', 'Sir', 'Jonkheer', 'Dona' are mapped to 5 as those titles indicate more privilege.`
    - I chose this notebook because I found it interesting how they were able to retain name information. It really helped me to get a better understanding of how people pre process string data, as if I were to do this without any reference, I would have dropped the name column entirely.

- *Hyper-parameter Tuning using GridSearchCV | Decision Trees Part 8*
- *Hyperparameter Tuning of Decision Tree Classifier Using GridSearchCV*
- *Decision Trees Explained — Entropy, Information Gain, Gini Index, CCP Pruning | by Shailey Dash | Towards Data Science*
- Decision Trees Explained — Entropy, Information Gain, Gini Index, CCP Pruning | by Shailey Dash | Towards Data Science.