**Question 1.**

|  |  |
| --- | --- |
| **TV Shows** | **Accuracy on 49 States** |
| GrunkleStan | 81.63% |
| StatisticianGrandmas | 81.63% |
| KeepingUpWithGranpas | 79.59% |
| DataMinerDynasty | 77.55% |
| Gonzo | 59.18% |
| JustAteAFrog | 57.14% |
| SmartCayote | 57.14% |
| KermitTheFrog | 55.10% |
| MahnaMahna | 53.06% |
| TheRealGrandmas | 53.06% |

**Question 2.**

Diagram

Description automatically generated

*For Statistician Grandmas- the above is the resulting decision tree that has used the MPV metric to make a split (came through Gini Index). The cutoff Minute Per Voter was 11.131 minutes.*

*For the left leaf, If the Minutes per Voter of Statistician Grandmas in a state is greater than 11.131 minutes, then we can be 82.14% (23/28) confident that the state will vote for a Democrat.*

*For the Right leaf, if the Minutes Per Voter of Statistician Grandmas in a state is less than 11.131 minutes then we can be 85.71% confident that a state will vote for Republican.*

Diagram

Description automatically generated

*For Keeping up with Grandpas- - the above is the resulting decision tree that has used the POF metric to make a split. The cutoff Percentage of fans was 7.160%.*

*For the left leaf, if the Percentage of fans of Keeping up with Grandpas in a state is greater than 7.160% then we can be 94% (17/18) confident that the state will vote for Republican.*

*For the right leaf, if the Percentage of fans of Keeping up with Grandpas in a state is less than 7.160% then, we can be 81% (25/31) confident that the state will vote for Democrat.*

**Question 3.**

|  |  |
| --- | --- |
| **TV Shows** | **Accuracy on 39 Safe States** |
| StatisticianGrandmas | 82.05% |
| KeepingUpWithGranpas | 82.05% |
| GrunkleStan | 76.92% |
| Gonzo | 71.79% |
| JustAteAFrog | 71.79% |
| DataMinerDynasty | 66.67% |
| SmartCayote | 66.67% |
| TheRealGrandmas | 58.97% |
| KermitTheFrog | 56.41% |
| MahnaMahna | 56.41% |

**Question 4.**

|  |  |  |
| --- | --- | --- |
| **TV Shows** | **Accuracy on 49 States** | **Accuracy on 39 Safe States** |
| GrunkleStan | 81.63% | 76.92% |
| StatisticianGrandmas | 81.63% | 82.05% |
| KeepingUpWithGranpas | 79.59% | 82.05% |
| DataMinerDynasty | 77.55% | 66.67% |
| Gonzo | 59.18% | 71.79% |
| JustAteFrog | 57.14% | 71.79% |
| SmartCayote | 57.14% | 66.67% |
| KermitTheFrog | 55.10% | 56.41% |
| MahnaMahna | 53.06% | 56.41% |
| TheRealGrandmas | 53.06% | 58.97% |

**A.** From the table in Question 4 you should see that in most cases (i.e., for most shows), the accuracy is higher when you limit your analysis to the safe states. **Explain why?**

*We have access to the database of what TV shows have been watched by which household in which state for how long. Based on this database we have an accuracy rate about which state will vote for whom. But other data are coming from the fact that the presidential election outcome in the USA is determined by how people in the Swing States (Colorado, Florida, Iowa, Nevada, New Hampshire, North Carolina, Ohio, Virginia, Wisconsin, and Pennsylvania) vote. The outcome of the Swing State determines the outcome of the election. Their voting patterns cannot be predicted with certainty as with Safe States.*

*The accuracy of TV Shows based on MPV and POF is higher when the data is limited to Safe States due to its nature of being safe. These 39 states are called safe because the outcome of their presidential voting is the predictable way in advance before the actual election or the voting results are out. As we know that California will vote for Democrats and states like Wyoming or Alaska will vote for Republicans. These states are safe. So, when we run the analysis on 39 safe states only, the accuracy tends to go higher compared to the states with swing states as we have the fact that these states have voting patterns that can be easily predicted with certainty. With this data tested on known safe states, it can be assumed that such TV shows’ viewership showcases the panel’s interest. (i.e., using a specific show on safe states does not require the decision tree to be trained ex-post-election).*

**B.** Explain **what could cause those few (counter-intuitive) cases where the accuracy is higher in 49 states than in 39 states?**

*Out of 10 shows, 2 shows have decreased in accuracy rate in 39 safe states- Grunkle Stan and Data Miner Dynasty. One of the possible reasons could be the false discovery that is making it possible to find models that have accuracies greater than the baselines by pure chance. The overall accuracy increased in the 39-state model in all but two cases. In those two cases, there are many reasons the accuracy could have been higher in a 49-state model, but it is most likely due to chance or luck. Given that only two shows decreased within the 10 swing states in accuracy, it is easier and more likely to be that those two shows viewership just happened to be in line with the viewership habits of the voting base in the safe states, thereby increasing the overall accuracy in the 49-state model. This also supports the overall accuracy of this model and its miscalling of two of ten swing states, providing an overall accuracy of 80%.*

**Question 5.**

**Diagram

Description automatically generated**

*For Statistician Grandmas- the above is the resulting decision tree that has used the MPV metric to make a split. The cutoff Minute Per Voter was 11.131 minutes. The cut-off didn’t change if we remove Swing States from our Observations but,*

*For Keeping up with Grandpas-*

**Diagram

Description automatically generated**

*The above is the decision tree that has used the MPV metric to make a split. The cutoff Minute Per Voter was 14.643 minutes.*

*Of the two TV shows observed, the only change in cutoff occurred with* ***Keeping Up with Grandpas****. When we build a tree using this show based on 49 states, its cutoff was* ***7.16 POF****, but when removed the swing states from our observation and build a tree on only safe states the cutoff changed to* ***14.643 MPV****. This change could be due to the removal of data points from the original tree. When those data points were removed and the model was run again, a change occurred at the first split of the tree. That initial change resulted in the total change of cutoff criteria in the tree, thus changing the entire appearance of the tree and the cut-off metric changing from POF to MPV.*

**Question 6.**

**Table

Description automatically generated**

***Nevada and Wisconsin*** *were two Swing States that have been miscalled by the predictions. The MPV cut of Statistician Grandmas was* ***11.131*** *and for State Nevada, MPV is* ***7.537*** *which is less than 11.131, we predicted Nevada to be voted for Republican and the actual winner is a Democrat.*

*Similarly, the MPV for Wisconsin the MPV is* ***10.829*** *which is less than 11.131, we predicted Wisconsin will vote for Republicans, and they voted for Democrats. They were both predicted to vote Republican, but both voted Democrat. This results in an 80% accuracy of the prediction model. Out of 10, 2 of the predictions were wrong.*