Decision Tree

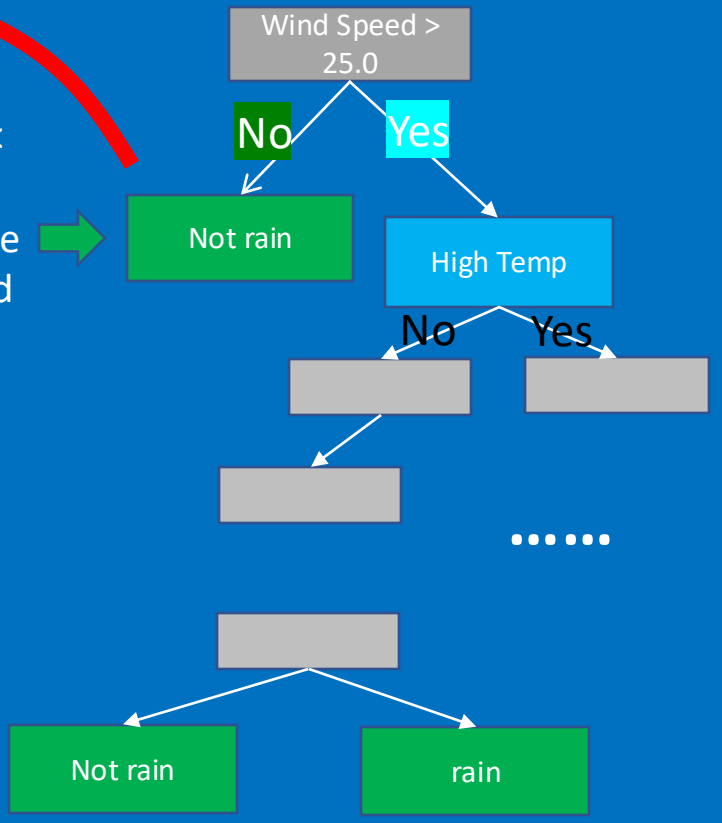
How to avoid overfitting

Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	No	No	10.0	No
Yes	Yes	Yes	30.0	Yes
Yes	Yes	No	20.0	No
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Yes

No

Yes

In this example, if we start from root with “wind speed < 25.0”, there is only one “no-rain” sample for us to use. The model can be easily overfitted

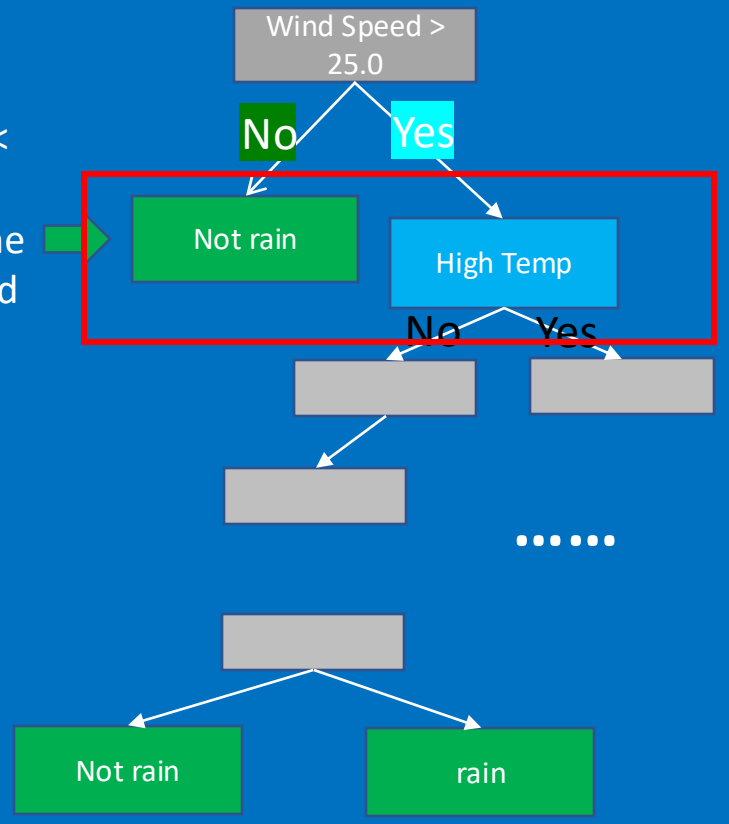


Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	No	No	10.0	No
Yes	Yes	Yes	30.0	Yes
Yes	Yes	No	20.0	No
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Yes

No

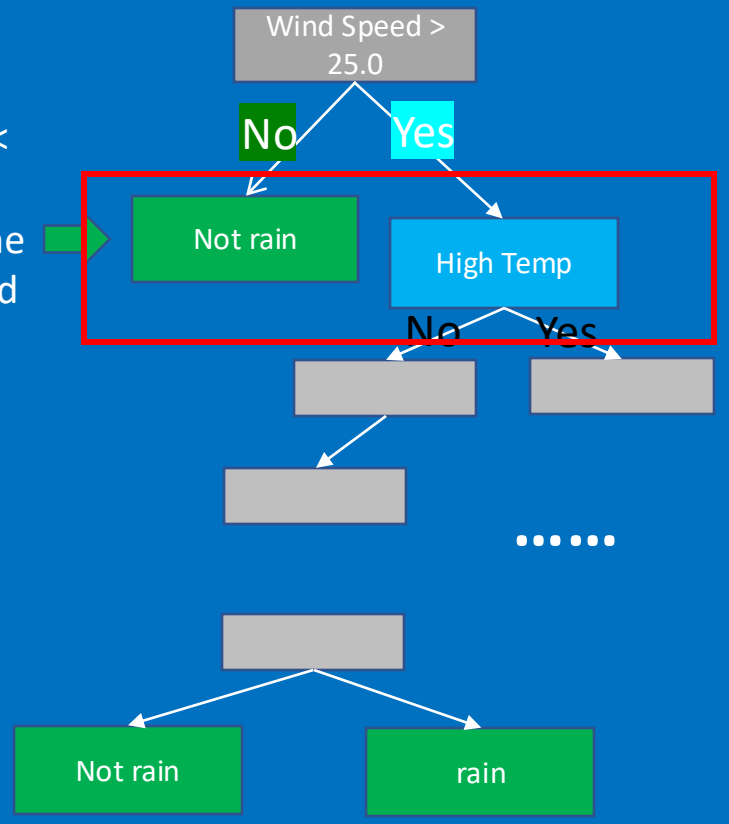
Yes

In this example, if we start from root with “wind speed < 25.0”, there is only one “no-rain” sample for us to use. The model can be easily overfitted (since very few case for this leaf, it is difficult to have confidence for this level of split for future data …)

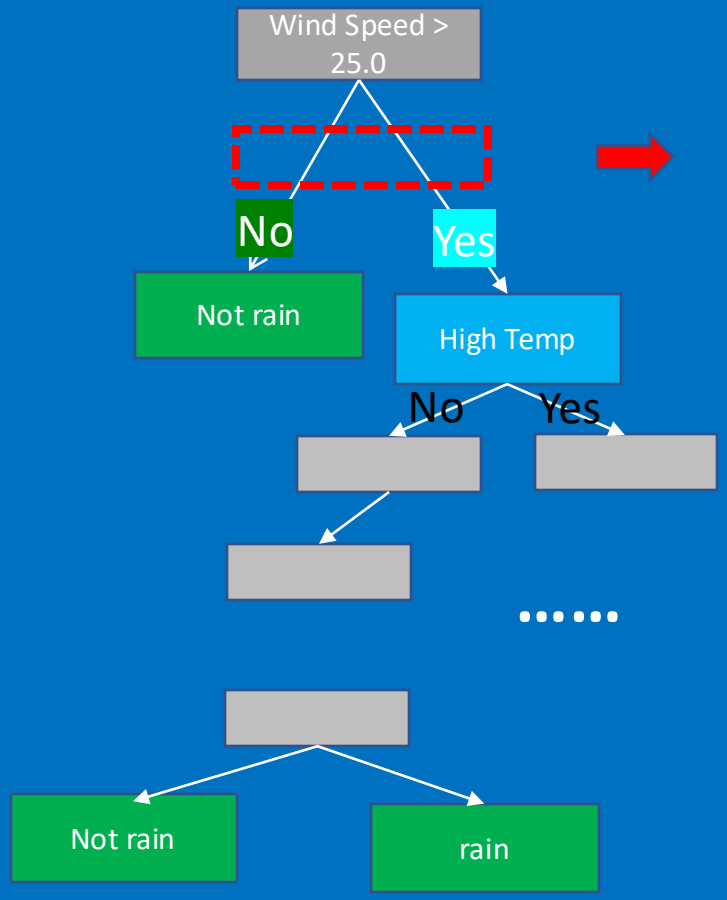


Low pressure	High Temperature	High humidity	Wind Speed	Rain	
No	No	No	10.0	No	No
Yes	Yes	Yes	30.0	Yes	
Yes	Yes	No	20.0	No	
Yes	No	Yes	50.0	No	Yes
No	No	Yes	70.0	Yes	

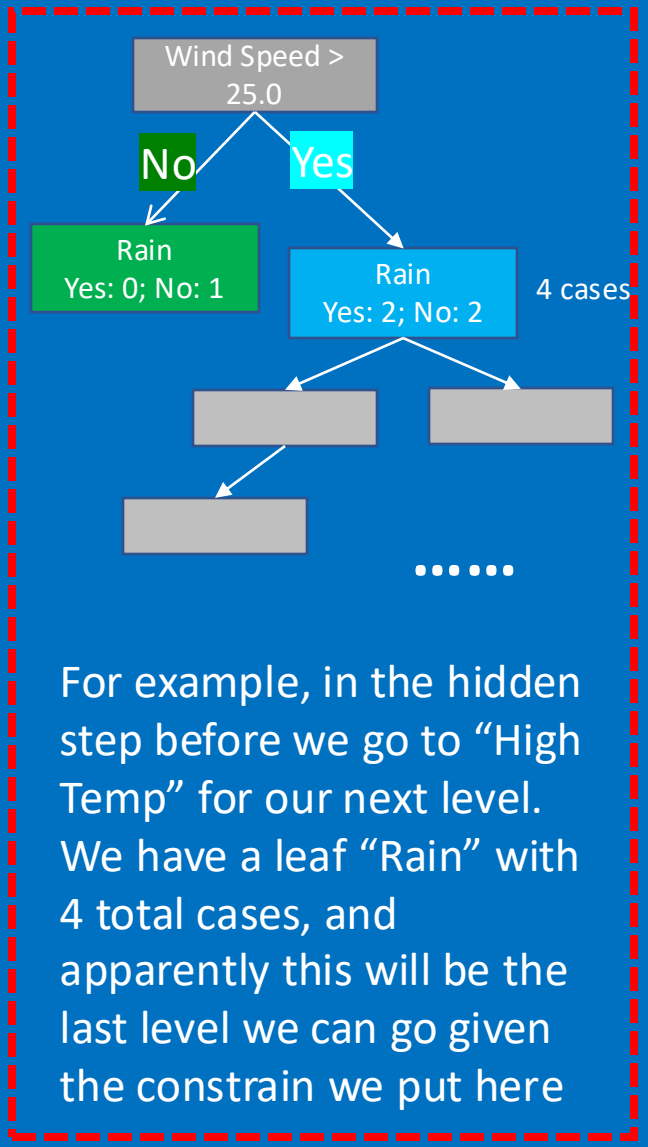
In this example, if we start from root with “wind speed < 25.0”, there is only one “no-rain” sample for us to use. The model can be easily overfitted (since very few case for this leaf, it is difficult to have confidence for this level of split for future data …)



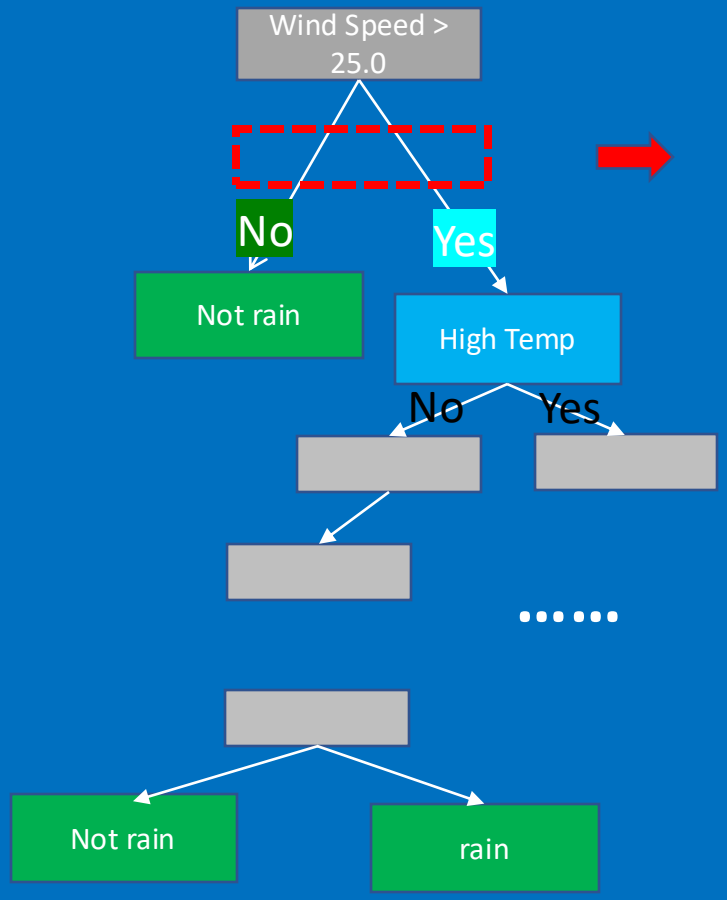
In order to address this, we can limit how deep (how many levels) the decision tree want to go. For example, if we can require it must have at least 4 samples in the level for the tree to grow



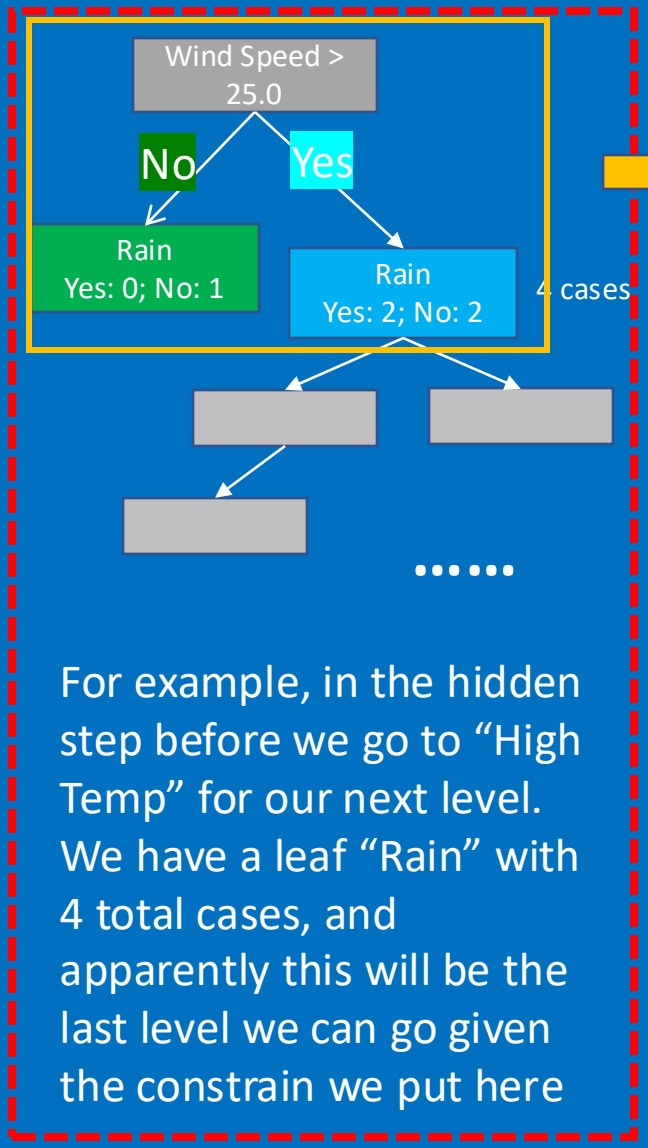
In order to address this, we can limit how deep (how many levels) the decision tree want to go. For example, if we can require it must have at least 4 cases in the level for the tree to grow



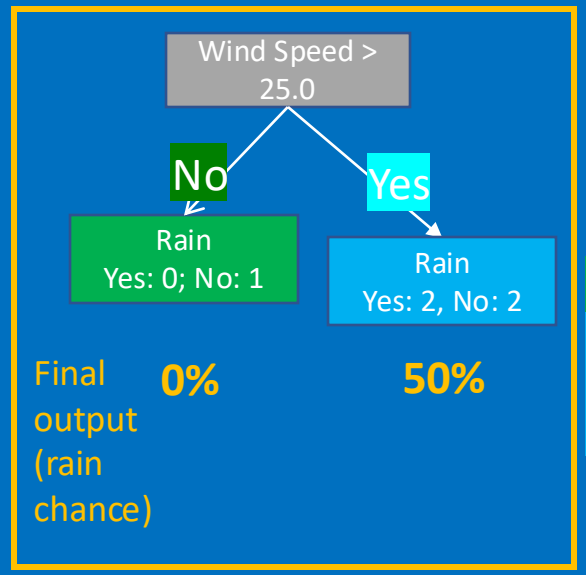
For example, in the hidden step before we go to “High Temp” for our next level. We have a leaf “Rain” with 4 total cases, and apparently this will be the last level we can go given the constrain we put here



In order to address this, we can limit how deep (how many levels) the decision tree want to go. For example, if we can require it must have at least 4 cases in the level for the tree to grow

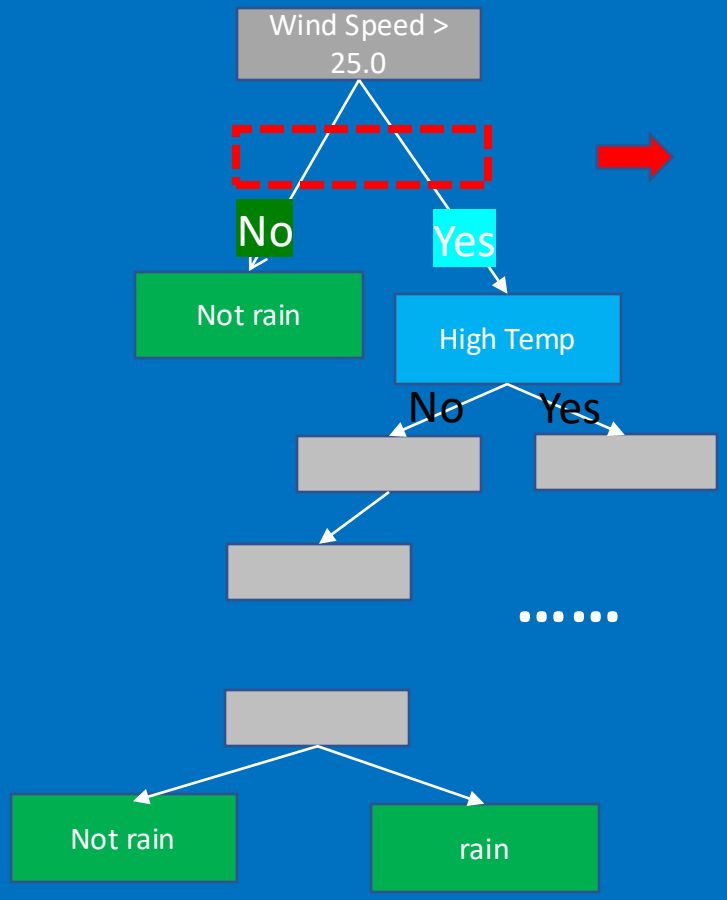


For example, in the hidden step before we go to "High Temp" for our next level. We have a leaf "Rain" with 4 total cases, and apparently this will be the last level we can go given the constrain we put here

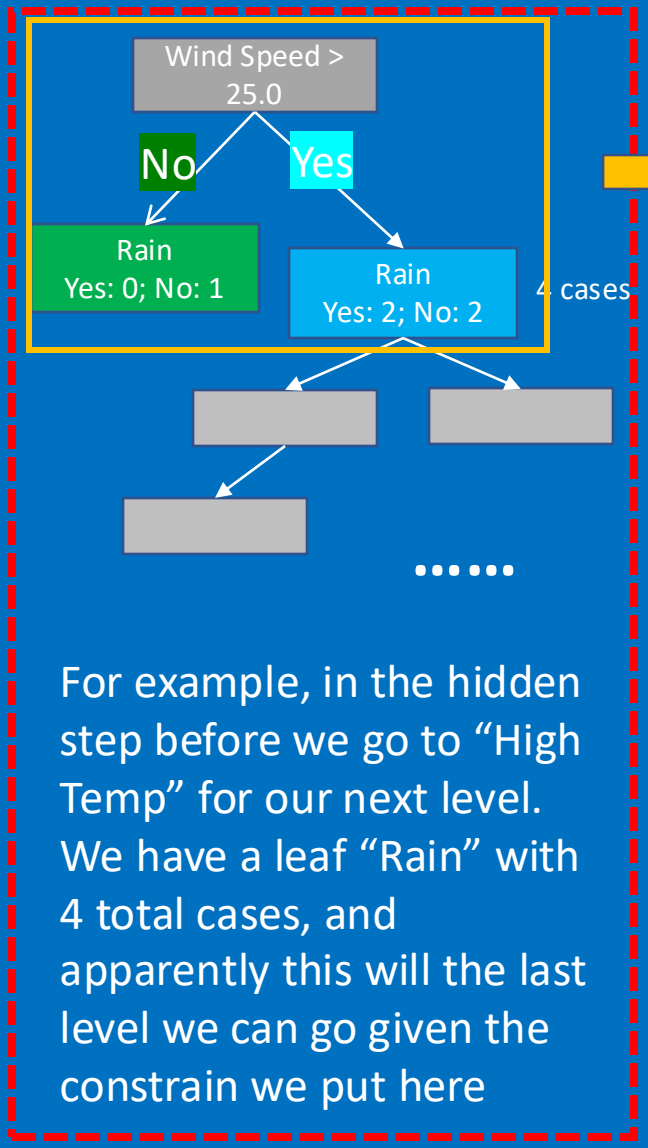


So here we see that by limiting the number of levels (only one level) the tree can grow, we get less determinant output

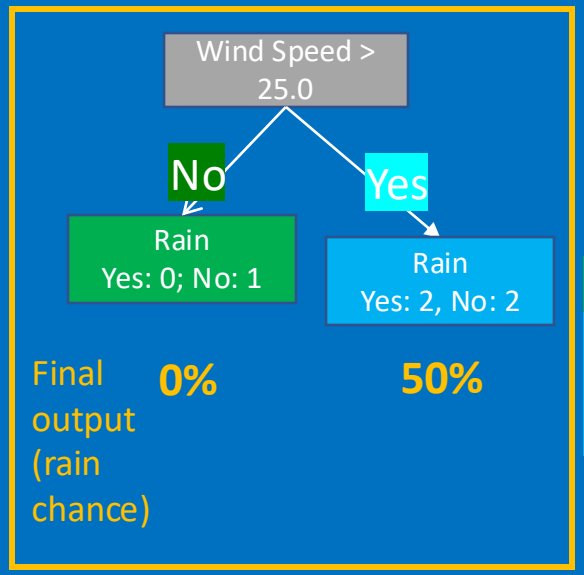
Low pressu	High Tempe	High humidi	Wind Speed	Rain
No	No	No	10.0	No
Yes	Yes	Yes	30.0	Yes
Yes	Yes	No	20.0	No
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Yes



In order to address this, we can limit how deep (how many levels) the decision tree want to go. For example, if we can require it must have at least 4 cases in the level for the tree to grow



For example, in the hidden step before we go to "High Temp" for our next level. We have a leaf "Rain" with 4 total cases, and apparently this will the last level we can go given the constrain we put here



So here we see that by limiting the number of levels (only one level) the tree can grow, we get less determinant output

In reality, we can split training/testing dataset and do a bunch of cross-validations to determine how deep we want our tree to grow

Low pressu	High Tempe	High humidi	Wind Speed	Rain
No	No	No	10.0	No
Yes	Yes	Yes	30.0	Yes
Yes	Yes	No	20.0	No
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Yes