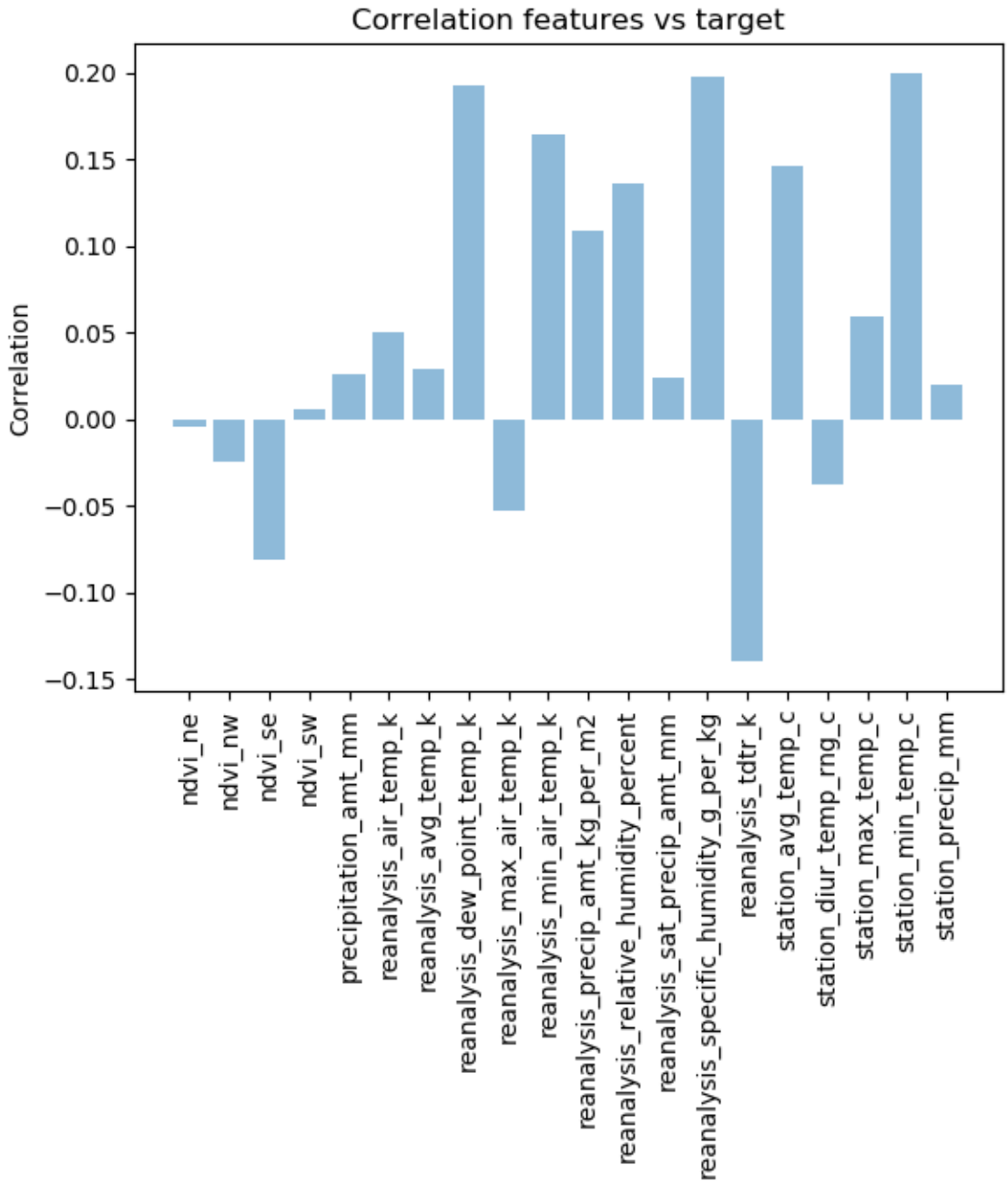
Our data is from Iquitos 2004 – 2007
Only week deleted was the 2005-01-01, because it didn't have any data.
We did some changes to the dataset regarding empty spaces and strange explained in task 1 and 2. In retrospective we should have left the database as it was and used pandas internally to modify it while executing the script, that way we could modify our decisions on the go without having to redo the whole database cleaning. We will keep it in mind for future scripts.

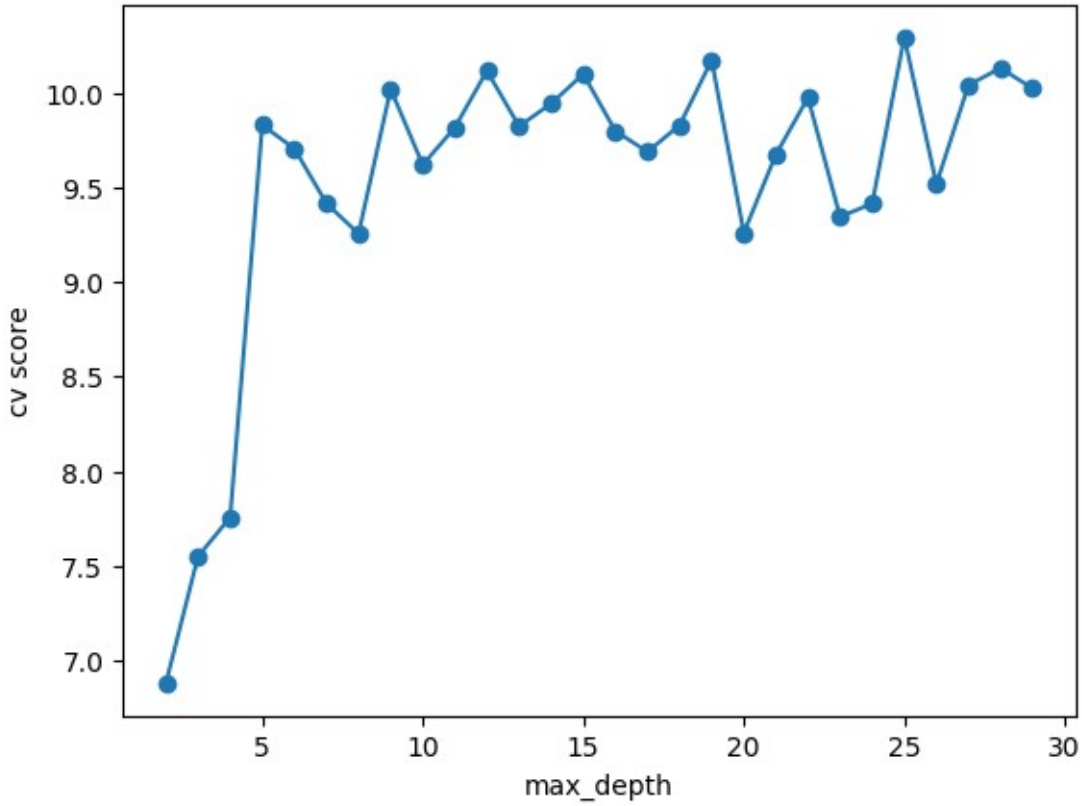
In the study of the training data in the previous task we established that certain features are useless for the this problem. For this task, we are going to return to the original database and get what are the useless features by the decision trees. Then we are going to check with our previous studies and see how correct we were or weren't.

First of all, we did the correlation between the features and our target, which is the total number of cases detected:



Our features don’t have a very high correlation with the target, as we can see with a mere 0.2 correlation as the maximum. We also have some that don’t even get to 0.05, we studied the possibility of deleting those but we decided against it in the end since the algorithm that we are going to use for this task (decision tree) chooses by itself the relevant features. We are going to get back to this point later on this document.

After all we did a cross-validation test to check what depth should out decision tree have.



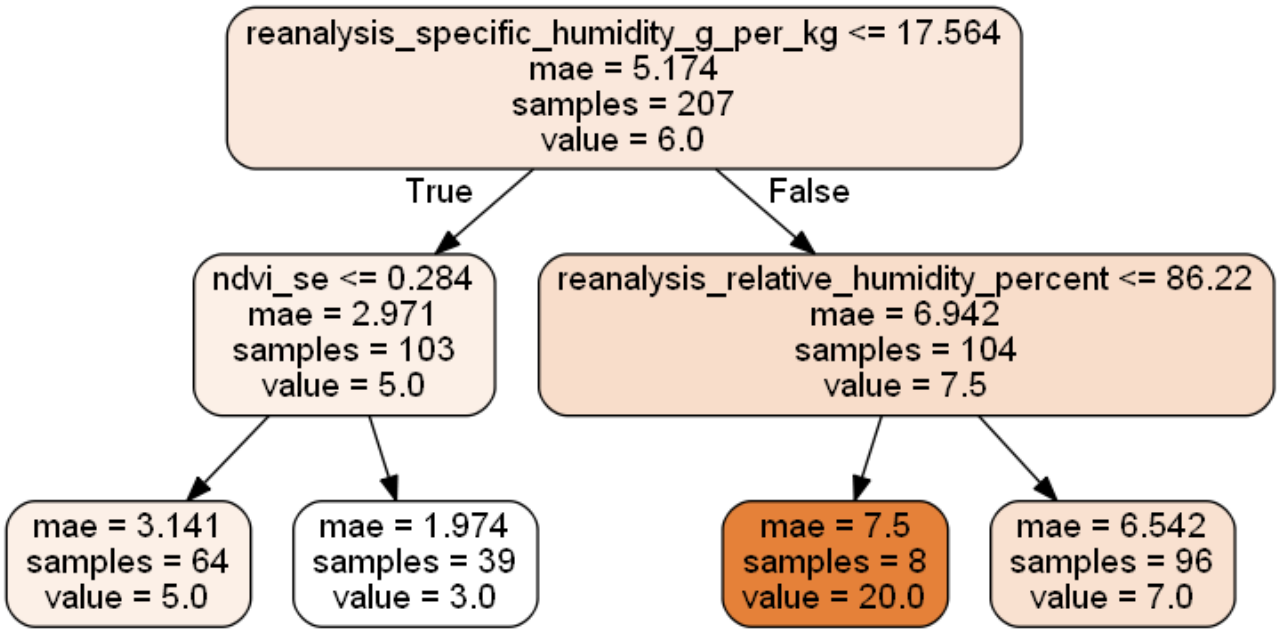
As we can see, we get the best results with depth = 2.

After that we compute the features relevancy with this depth.

ndvi_ne	0
ndvi_nw	0
ndvi_se	0.266667
ndvi_sw	0
precipitation_amt_mm	0
reanalysis_air_temp_k	0
reanalysis_avg_temp_k	0
reanalysis_dew_point_temp_k	0
reanalysis_max_air_temp_k	0
reanalysis_min_air_temp_k	0
reanalysis_precip_amt_kg_per_m2	0
reanalysis_relative_humidity_percent	0.32381
reanalysis_sat_precip_amt_mm	0
reanalysis_specific_humidity_g_per_kg	0.409524
reanalysis_tdtr_k	0
station_avg_temp_c	0
station_diur_temp_rng_c	0
station_max_temp_c	0
station_min_temp_c	0
station_precip_mm	0

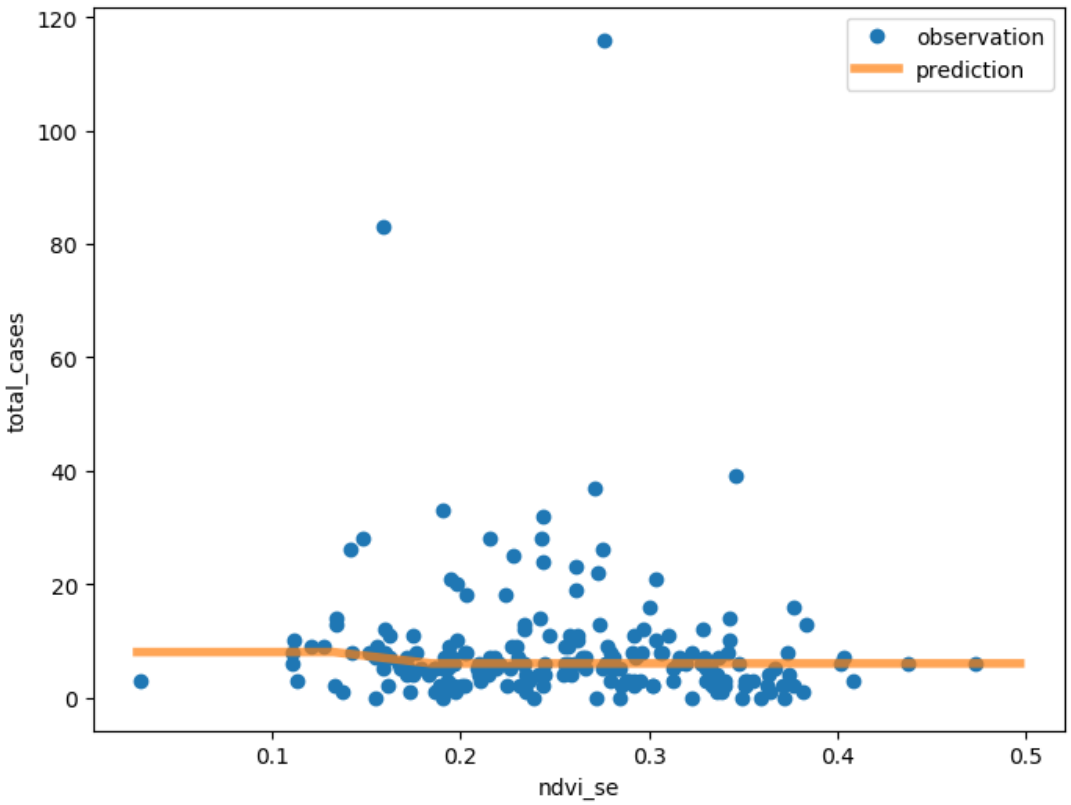
As a note, we can see that the correlation is not very related with the relevancies, since two variables that are used for our decision tree (reanalysis_relative_humidity_percent and ndvi_se) are not the one that are the most correlated, sure they are one of the most correlated ones still, but it's important to note this for future tasks were will we have to manually delete some useless features.

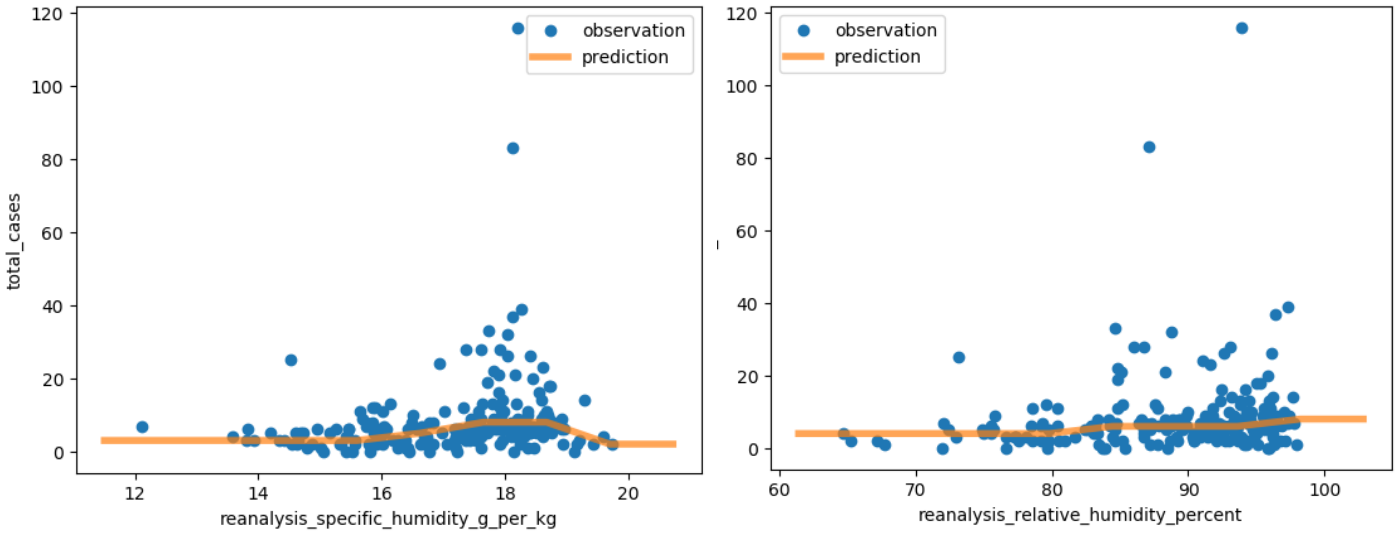
This is the resulting tree.



We tried with other depths to study them, but the overfitting got much much worse so we left it at 2.

Here's the cross-validation test to check visually if the 3 features used in our tree.





Here we can see they fit the model nicely.