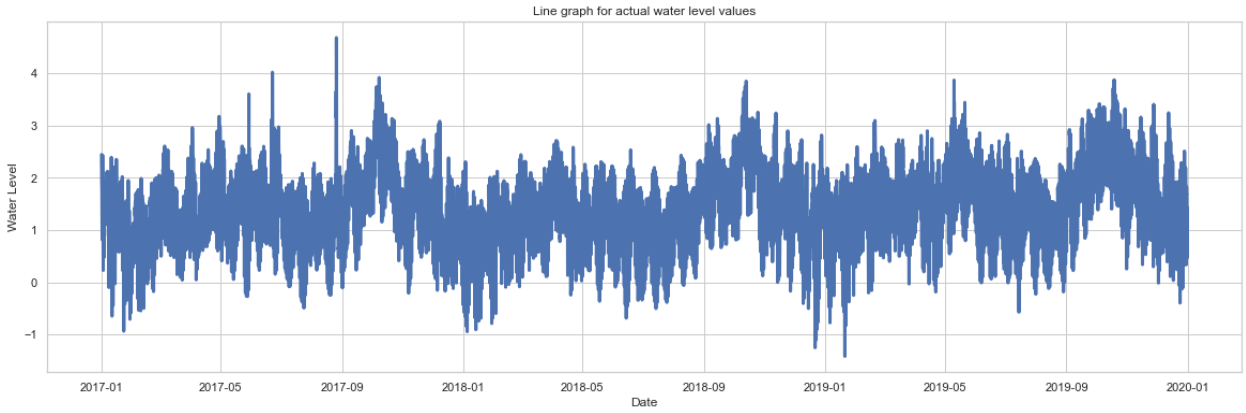
Assignment Report

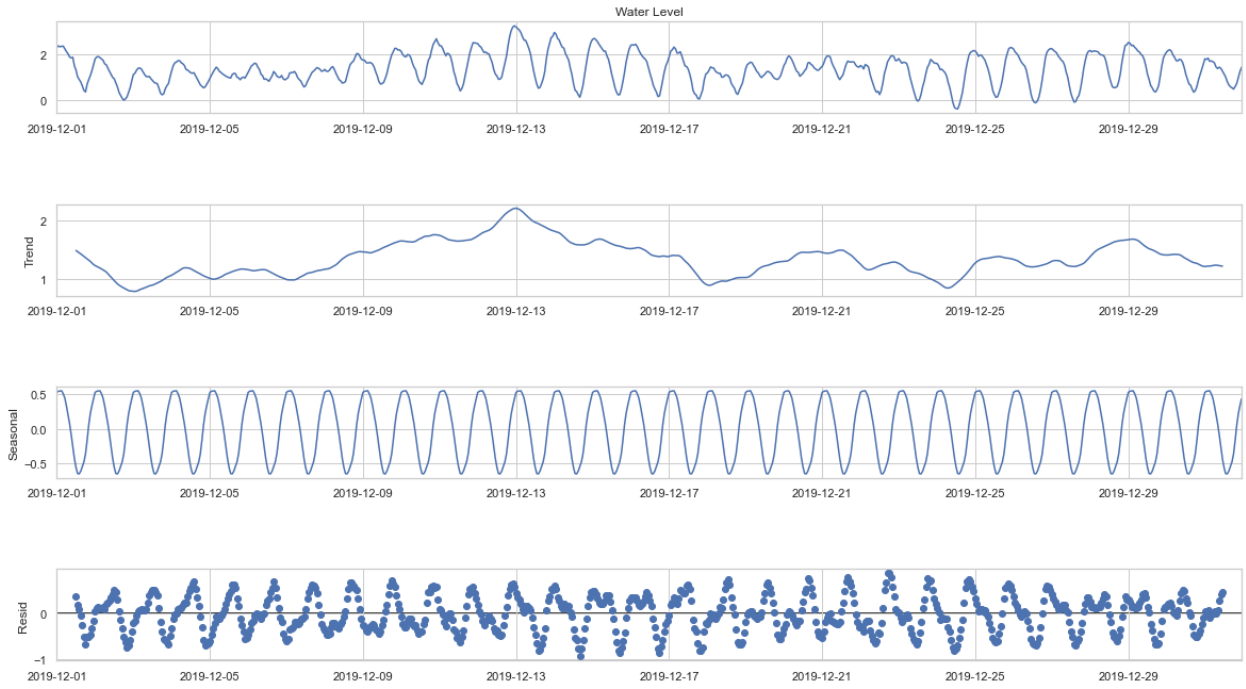
Name: Apurva Anand **A#:** 04285700

I took 3 approaches to predicting water levels.

- Decision Tree (5 features)
- Decision Tree (3 features)
- Random Forest

Training data:
Here I plot some basic graphs to understand training data. We get a clue of highs and lows as well as a clue of what kind of trends are present.





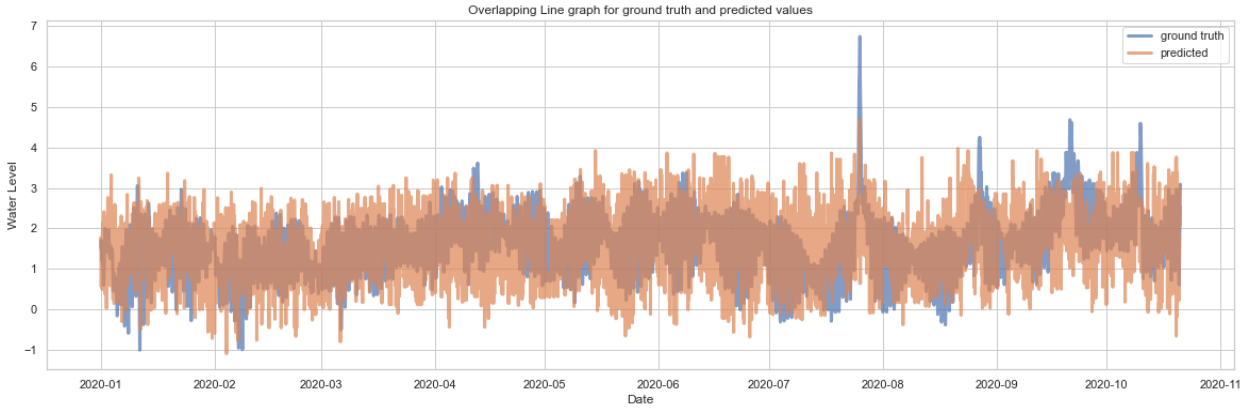
Decision Tree:

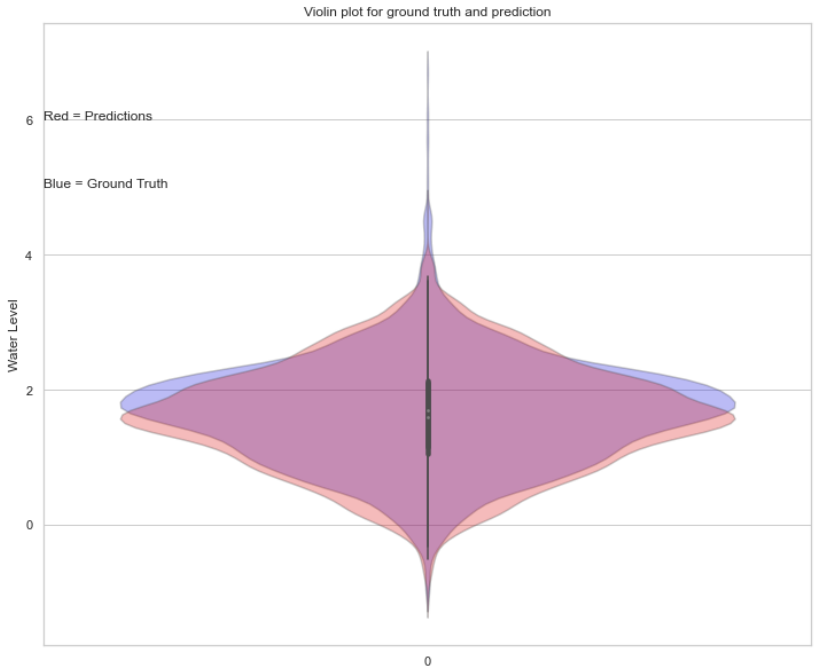
Decision trees learn how to best split the dataset into separate branches, allowing it to learn non-linear relationships.It builds many individual trees, pooling their predictions. As they use a collection of results to make a final decision, they are referred to as “Ensemble techniques”.

Decision tree with 5 features

('Air Temperature', 'Water Level Sigma', 'Wind Speed', 'Wind Direction', 'Wind Gust')

MSE	RMSE	R2	MAE	MedAE
1.040	1.019	-0.700	0.803	0.666

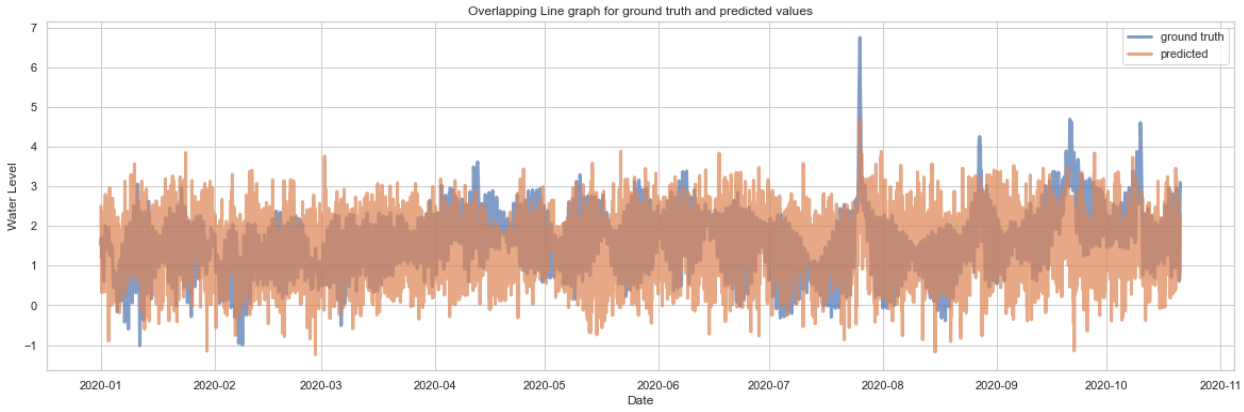


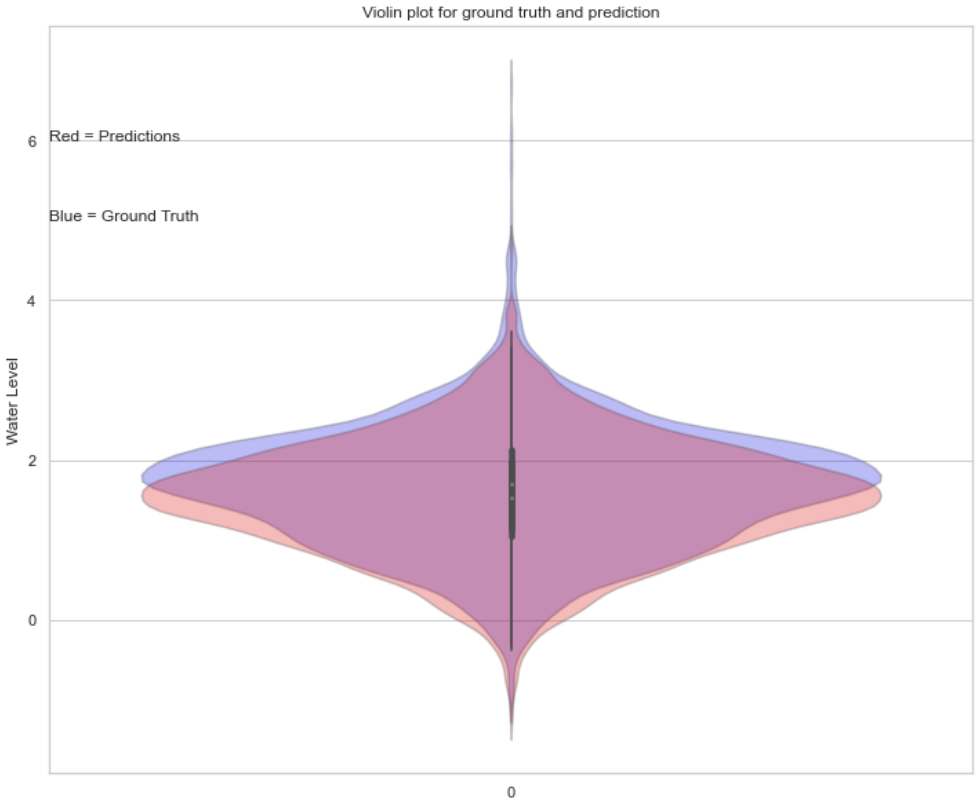


Decision tree with 3 features

('Water Level Sigma', 'Wind Speed', 'Wind Gust')

MSE	RMSE	R2	MAE	MedAE
1.054	1.027	-0.724	0.814	0.678

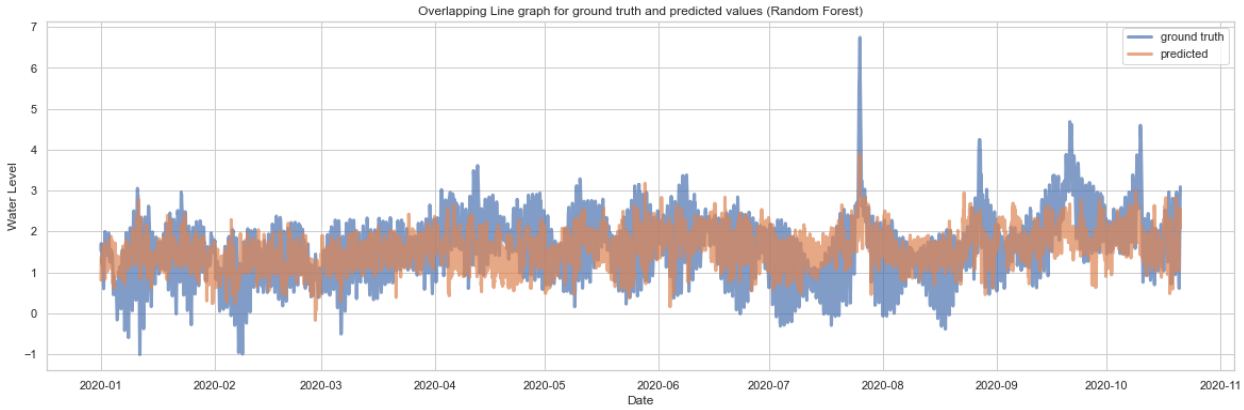


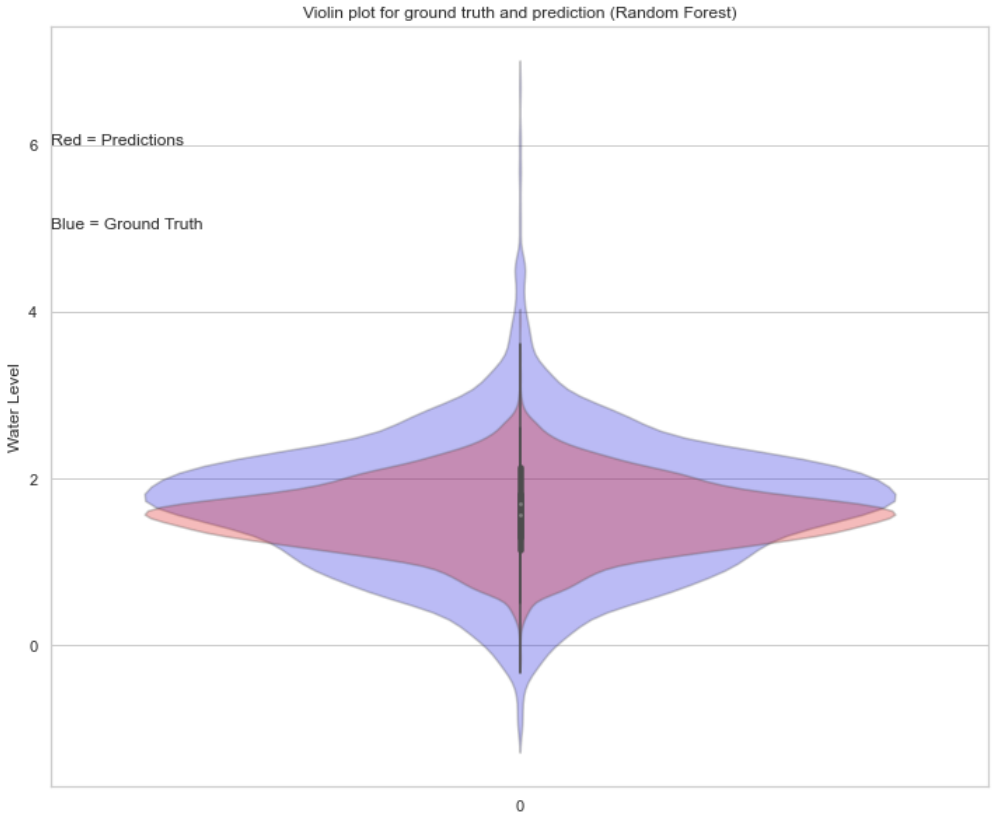


Random Forest 5 features

('Air Temperature', 'Water Level Sigma', 'Wind Speed', 'Wind Direction', 'Wind Gust')

MSE	RMSE	R2	MAE	MedAE
0.578	0.760	0.054	0.592	0.486





We see that Random Forest performs slightly better than Decision Tree.

Team Comparison

	MSE	RMSE	R2	MAE	MedAE
Decision Tree (Apurva)	1.054	1.027	-0.724	0.814	0.678
ARIMA (Viren)	0.055	0.235	0.366	0.189	0.167
Random forest (David)	0.456	-1.220	0.502	0.584	0.526
LSTM (Manidhar)	0.694	0.720	-0.190	0.680	0.615

Insights:

The biggest issue I faced was prepping the data to be properly used by a model. It included dealing with missing values, making sure the shape of the data frame is equal, splicing values correctly and dealing with DateTime.

I made overlapping graphs (line and violin) to get a visual depiction of how my predictions were doing. Those graphs narrate an interesting story of how well we performed.