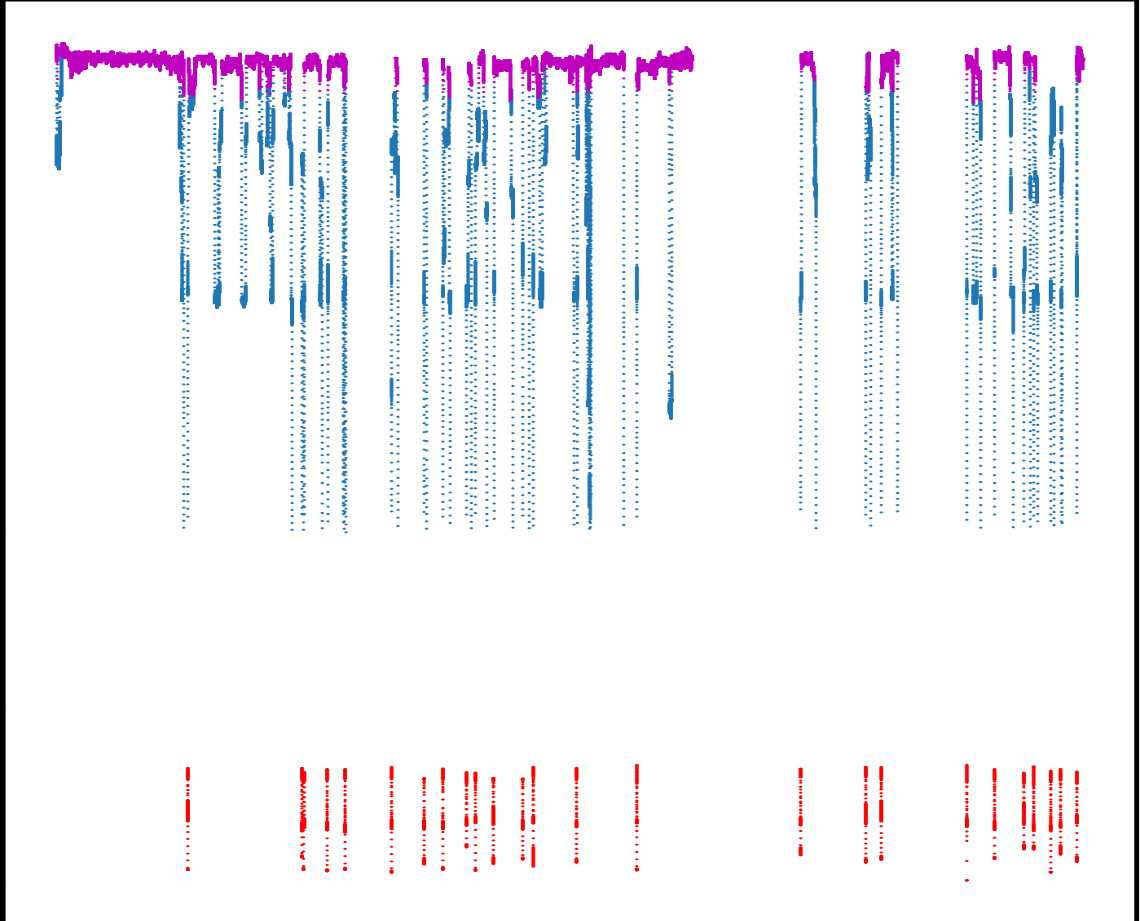
# Data Exploration

Interesting findings:

* Modes : start and operation
* Temperature changes due to seasons
* Regulations of Power affect the data, decided to make a new mode called regulation
* Some data is not affecting much based on Expert opinion – Reactive Power
* Some data is nearly constant and thus not needed – Turbine Rotational Speed and vibrations
  + However, it can be a good opportunity to investigate in the future, especially vibration

We augmented the dataset by splitting in up in three categories, by mode, in addition to ranges where power was changed during operation mode. Below an example of splitting into three modes, a representation of Guide Vane Opening. Red = start, Blue = Regulation, Pink = Operation



# Model Consideration

We decided to use three models with random forest regressor. A random forest is an ensamble estimator that fits a number of regression decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.  
Following hyperparameter can be adjusted

* n\_estimators = number of trees in the foreset
* max\_features = max number of features considered for splitting a node
* max\_depth = max number of levels in each decision tree
* min\_samples\_split = min number of data points placed in a node before the node is split
* min\_samples\_leaf = min number of data points allowed in a leaf node

An ensemble model was used since it’s a shotgun approach where weak single regressors can be powerful together.

In training the model we used a 0.33/0.67 test/train split that gave a good result without cross-validation.

Skcitlearn library was used for Random Forest Algorithm. Below are prediction for each mode plotted against the test-valus of Bolt1 tension.

Operation Mode

Chart

Description automatically generated

Regulation Mode

Chart, line chart

Description automatically generated

Start Mode

Chart

Description automatically generated