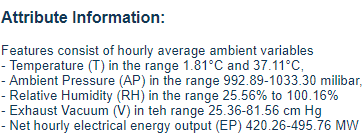
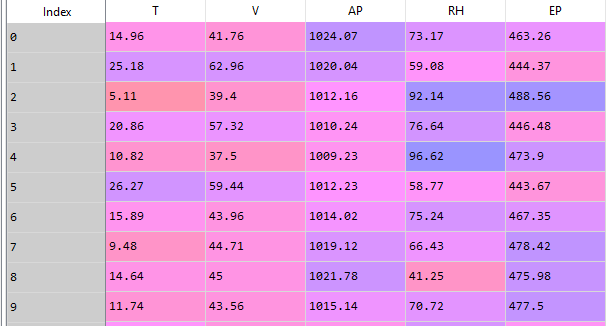
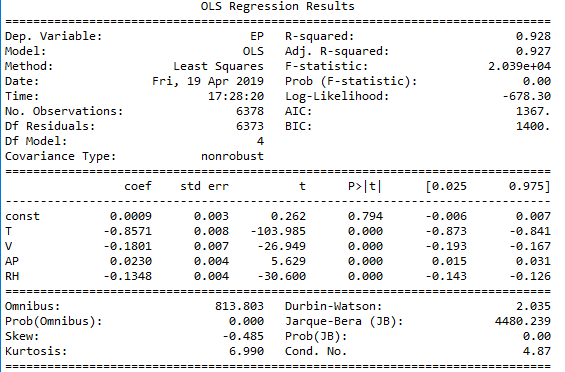
**Regression**

The dataset contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to **predict the net hourly electrical energy output (EP) of the plant.**



**(EP is Y-variable)**

All features are all significant according to the OLS Regression report below (P values close to 0) or back elimination method. Then I did a standard scaling for the whole dataset for better modeling. The full report for the 4 algorithms (linear regression, KNN,SVM, ***random forest***) I used for this regression model are the followings: (clearly, we can see that random forest is the best algorithm for the prediction of this model. It gives the highest adjusted-R2 score of 0.96, and the lowest RMSE value of 0.2 (low prediction error). according to the overall R2 score and RMSE value, we can conclude that these X variables are good predictors to predict the net hourly electrical energy output (EP) of the plant.)



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear regression | KNN | SVM | ***random forest*** |
| R2 | 0.9310 | 0.9563 | 0.9475 | ***0.9604*** |
| CV (full df) | 0.9285 | 0.9556 | 0.9444 | ***0.9605*** |
| Adjusted R2 | 0.9309 | 0.9562 | 0.9475 | ***0.9604*** |
| RMSE | 0.2629 | 0.2092 | 0.2292 | ***0.1991*** |
| CV (full df) | 0.2672 | 0.2104 | 0.2355 | ***0.1983*** |