Regresi Dua Objektif menggunakan Stacked Gradient Boosting dengan Ensemble Univariate Cubic Spline Smoothing

Abyoso Hapsoro Nurhadi, Edbert Theda, Valentino Prasetya

1. Latar Belakang

Hazim sedang mengerjakan project kuliah yaitu merakit dua buah quadcopter. Quadcopter adalah helikopter multirotor yang diangkat dan didorong oleh empat rotor (baling-baling). Quadcopter yang dirakit Hazim diatur untuk terbang dengan dua tipe rute lintasan secara otomatis (autopilot) serta dapat dipantau melalui remote control ataupun komputer. Pada quadcopter tersebut terdapat beberapa sensor, diantaranya adalah global positioning system (GPS), global position accuracy (GPA), inertial measurement unit (IMU), dan magnetometer (Kompas). Setelah tahap perakitan selesai, Hazim akan melakukan uji coba penerbangan pada quadcopter rakitan nya untuk diambil data nya. Melalui kedua quadcopter tersebut, dilakukan uji coba terbang dengan mengikuti koordinat longitude dan latitude pada GPS quadcopter sesuai dengan rute yang telah diatur sebelumnya. Pada akhirnya, Hazim ingin mengetahui bagaimana pergerakan dari Quadcopter.

2. Tujuan dan Manfaat

Tujuan penulisan makalah ini adalah:

i) Memprediksi trajektori Quadcopter dari posisi ke-X hingga posisi terakhir untuk setiap percobaan

ii) Memberikan analisa data mengenai trajektori Quadcopter

Manfaat penulisan makalah ini adalah:

i) Hasil penelitian dapat dijadikan sebagai sarana untuk menyusun strategi pengembangan trajektori Quadcopter.

ii) Menambah wawasan dan kemampuan berpikir mengenai Analisa data pada pergerakan Quadcopter

3. Batasan

Tidak digunakan pendekatan Runtun Waktu.

4. Metode Analisa Data

- Tahap Pra-pengolahan Data

Pada Dataset diberikan 93 variabel, dimana dua diantaranya merupakan variabel dependen yaitu koordinat longitude dan latitude dari Quadcopter. Dataset tersebut juga beris 8791 data dimana 6440 data merupakan sampel untuk *train data* dan 2351 sisanya merupakan *test data*. Dari 91 fitur terdapat 11 fitur yang memiliki nilai yang konstan atau bernilai sama persis dengan kolom lain, yaitu:

DSAlt	MOFsX	MOfsX_2	SAlt
GWk	MOfsY	MOfsY_2	SMS
MOfsZ	MOfsZ_2	VV	

Kesebelas fitur ini akan kita hapus dari 93 fitur pada awal dataset, dikarenakan tidak dapat memberikan informasi pada model. Kemudian kita akan mengecek korelasi antar fitur dari fitur-fitur yang tersisa. Didapatkan 35 fitur yang mempunyai korelasi tinggi dengan fitur lainnya, yaitu

AccX_2	GyrZ_2	TimeUS_IMU2	OfsY_2	LineNo_IMU2
AccY_2	HDop	TimeUS_MAG	OfsZ_2	LineNo_MAG
Alt	TimeUS_ATT	TimeUS_MAG2	ThO	LineNo_MAG2

BAlt	TimeUS_CTUN	TimeUS_RCOU	LineNo_CTUN	LineNo_RCOU
C4	TimeUS_GPA	MagX_2	LineNo_GPA	OfsX
GyrX_2	TimeUS_GPS	MagY_2	LineNo_GPS	OfsY
GyrY_2	TimeUS_IMU	OfsX_2	LineNo_IMU	OfsZ

35 fitur ini akan kita hapus dari fitur-fitur yang tersisa sehingga hanya tersisa 45 fitur. Dri 45 fitur ini, variabel id tidak akan dijadikan variabel independen sehingga terdapat 44 fitur yang akan dijadikan variabel independen.

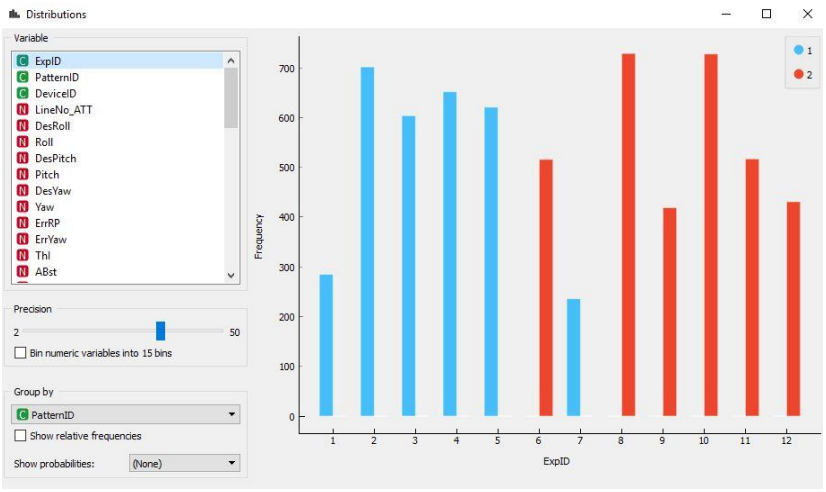
Untuk outlier, dari analisa yang dilakukan kita dapatkan beberapa kondisi jika data tersebut outlier:

- Jika ExpID = 3 dan LineNo_ATT <= 9158 or >= 14968
- Jika ExpID = 4 dan LineNo_ATT <= 2076
- Jika ExpID = 10 dan LineNo_ATT <= 8857

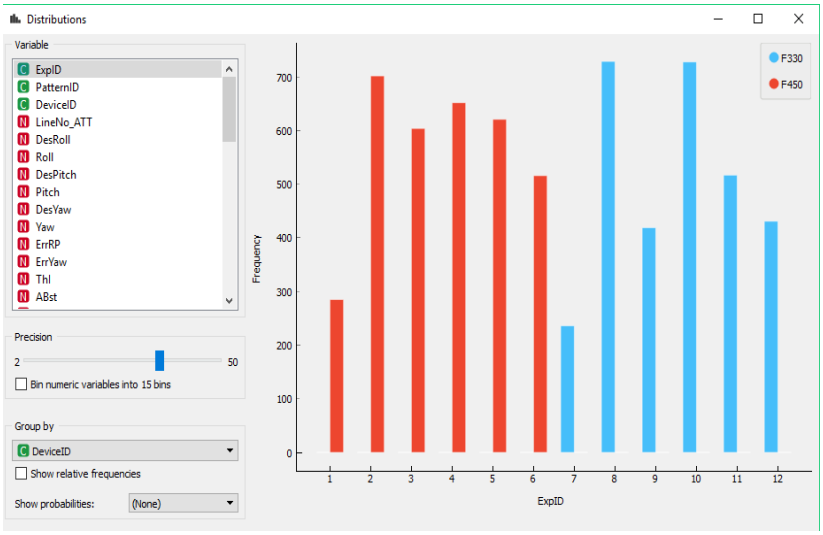
Sehingga data yang akan kita gunakan untuk model berukuran (6317,47)

- Visualisasi Data

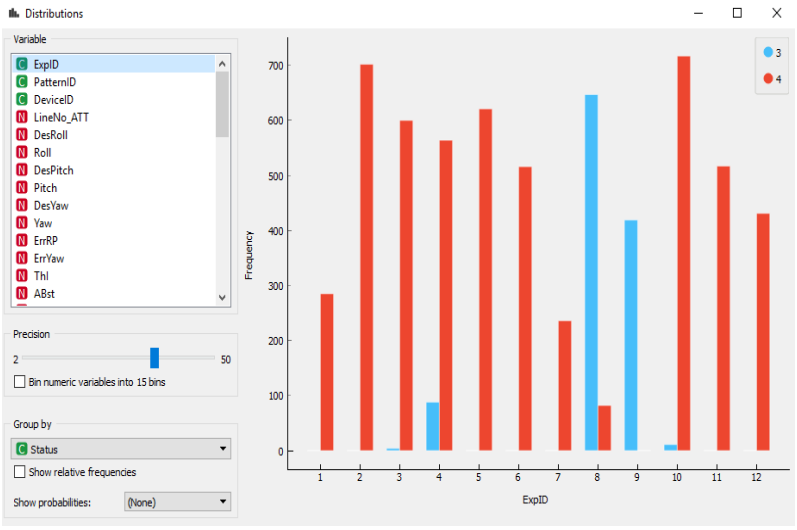
  Visualisasi Data dilakukan dengan software orange, Plot dan Distribusi Data :



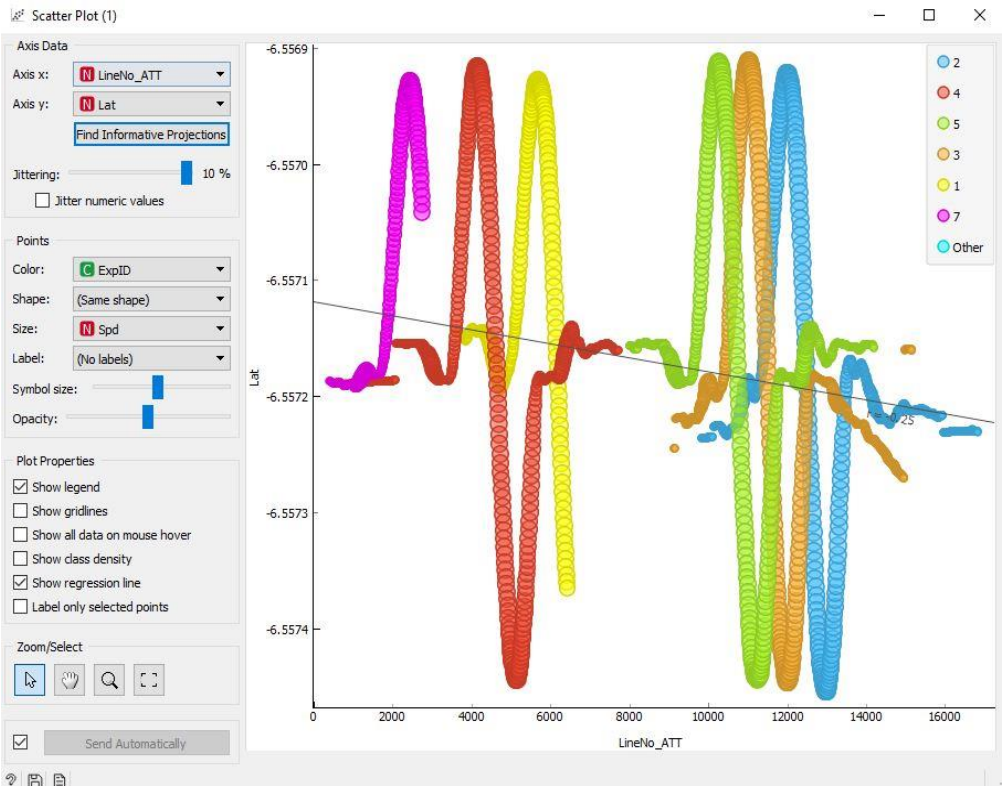
Up : Tabel PatternID-ExpID



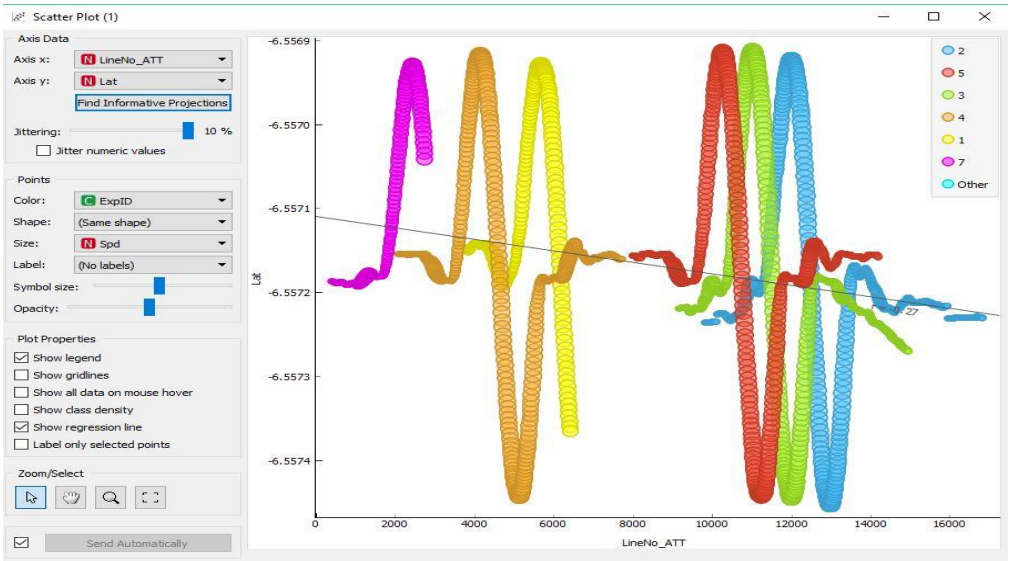
Up : Tabel DeviceID - ExpID



Up: Tabel Status - ExpID



Grafik Waktu terhadap Latitude tiap ExpID untuk PatternID = 1 pada data Train



Grafik Waktu terhadap Latitude tiap ExpID untuk PatternID=2 pada DataTrain

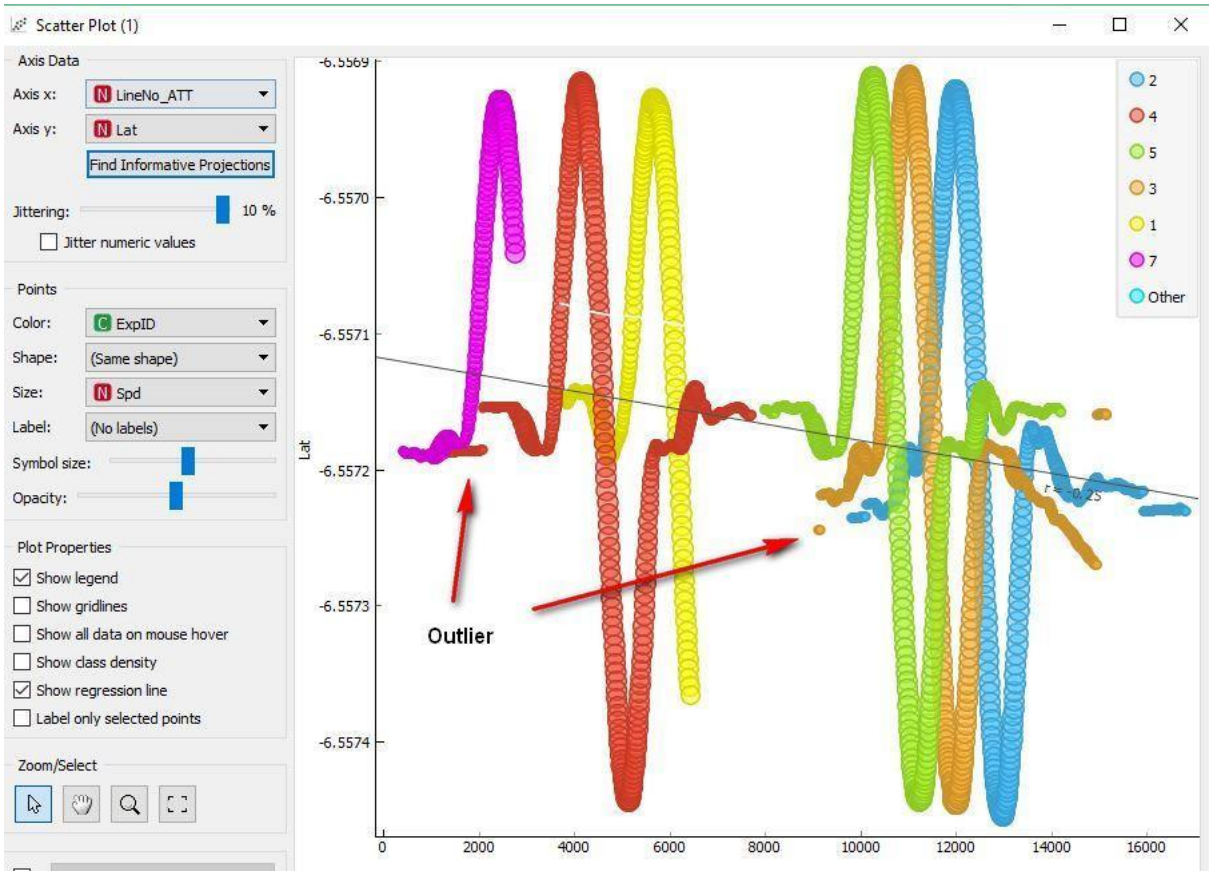
Dari grafik di atas, dapat dilihat bahwa ExpID dengan pattern yang sama memiliki pola pergerakan Latitude dan Langitude yang hampir sama.

Ekstraksi Data

Dari analisa diatas, dapat dilihat bahwa ExpID yang memiliki nilai Latitude dan Langitude yang hampir sama. Untuk itu kita buat variabel baru TimeLag dari ExpID yang dihitung dari waktu dimana nilai latitude drone mulai bergerak naik. Kemudian dibentuk variabel baru Time yang diperoleh dari selisih antara TimeLag dengan LineNo_ATT yang menggambarkan lama waktu setelah drone mulai bergerak. Variabel Time ini kemudian digunakan sebagai salah satu variabel dalam pencarian model menggantikan ExpID dan LineNo ATT

Deteksi Outlier

Deteksi Outlier dilakukan dengan melihat titik-titik pada grafik waktu-target



Outlier yang terdeteksi kemudian tidak digunakan dalam proses mencari model

5. Desain dan Implementasi

Tahap Pengembangan Model:

Karena metrik yang diinginkan adalah metrik yang spesifik, maka pertama didefinisikan fungsi skoring untuk *Mean Haversine Distance*. Kemudian dilakukan 7-fold *cross validation* dengan fungsi skoring ini. Sebagai *baseline*, dilakukan percobaan dengan beberapa model, seperti:

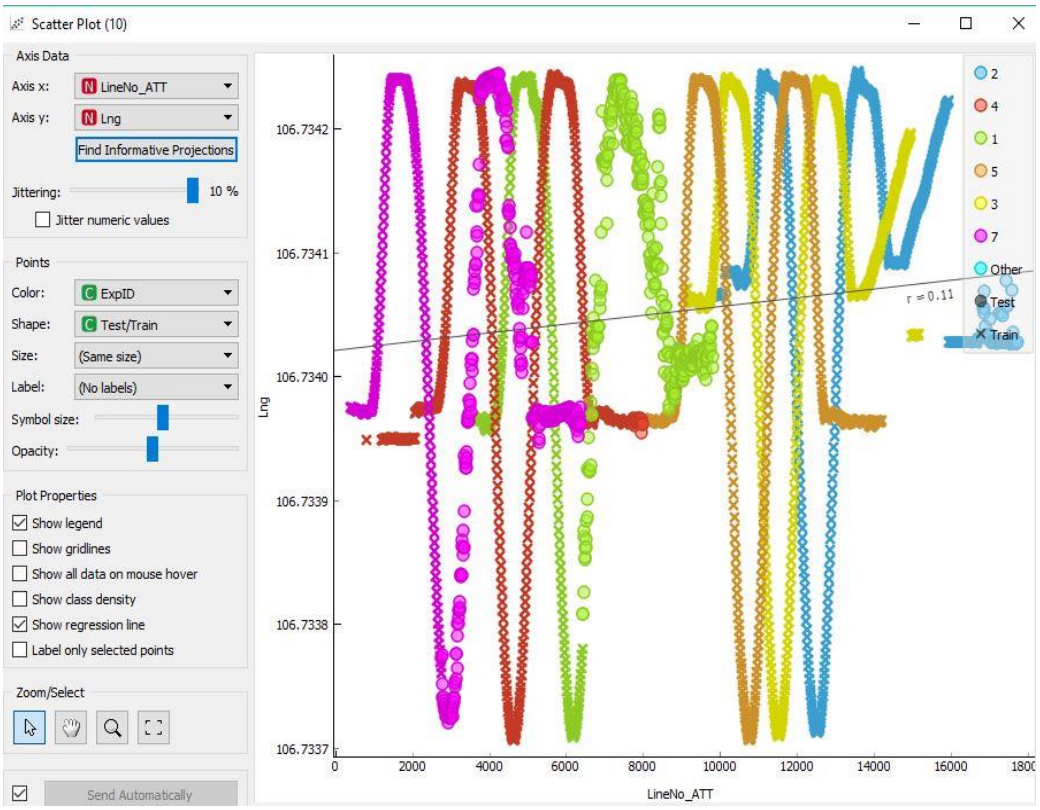
- Linear Regression
- Ridge Regression
- Orthogonal Matching Pursuit
- K-Nearest Neighbors Regressor
- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- Histogram-based Gradient Boosting Regressor
- Adaboost Regressor
- Bagging Regressor
- XGB Regressor

Data diolah dengan beberapa transformasi seperti MinMax Scaling, Maximum Absolute Scaling, dan Normalisasi. Ditemukan metode transformasi yang paling cocok untuk data yang diolah adalah MinMax Scaling karena performa beberapa model naik cukup tinggi. Setelah dikaji lebih lanjut, diperoleh bahwa K-Nearest Neighbors (KNN) dan Histogram-based Gradient Boosting Regressor (Histogram GBR) berperforma dengan paling baik.

Model 1: Stacking Regressor menggunakan Ridge Regressor dan Histogram GBR dengan meta regressor KNN pada data yang masih terdapat outlier namun sudah dilakukan MinMax scaling.

Model 2: Histogram GBR yang dioptimisasi menggunakan GridSearch pada data yang sudah tidak ada outlier serta terdapat lag data dan sudah dilakukan MinMax scaling.

Pengembangan Model dengan Smoothing:



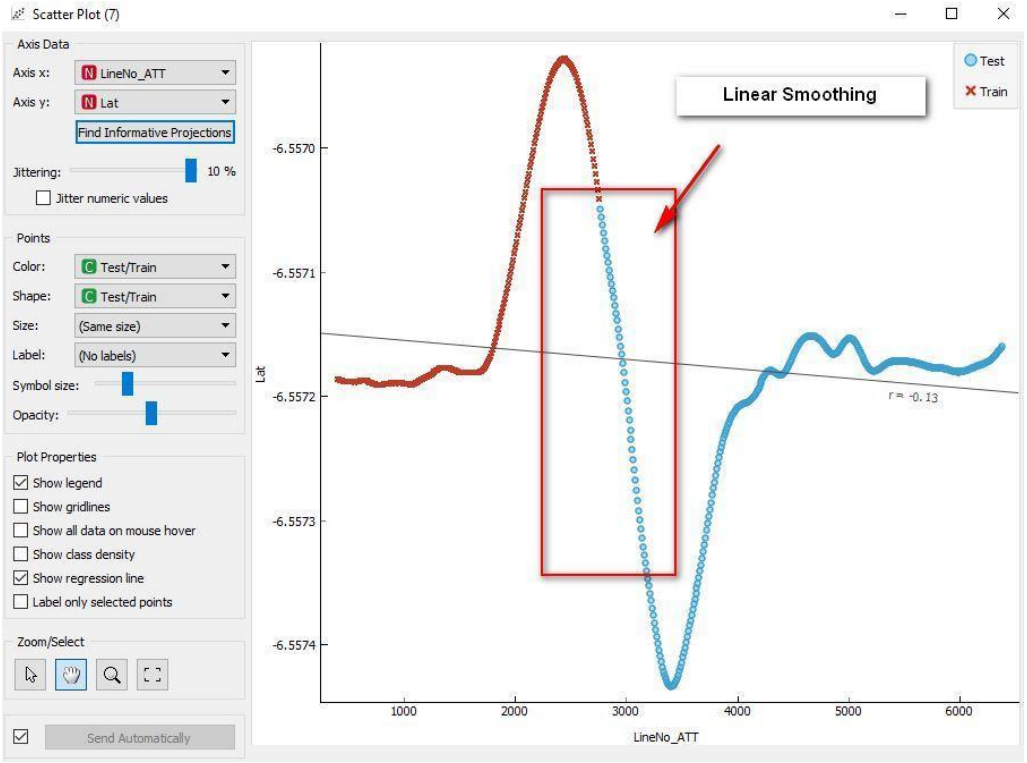
Up : Hasil prediksi Langitude menurut model HistGradientBoosting

Bentuk (x/o) menunjukan data train / prediksi pada data test

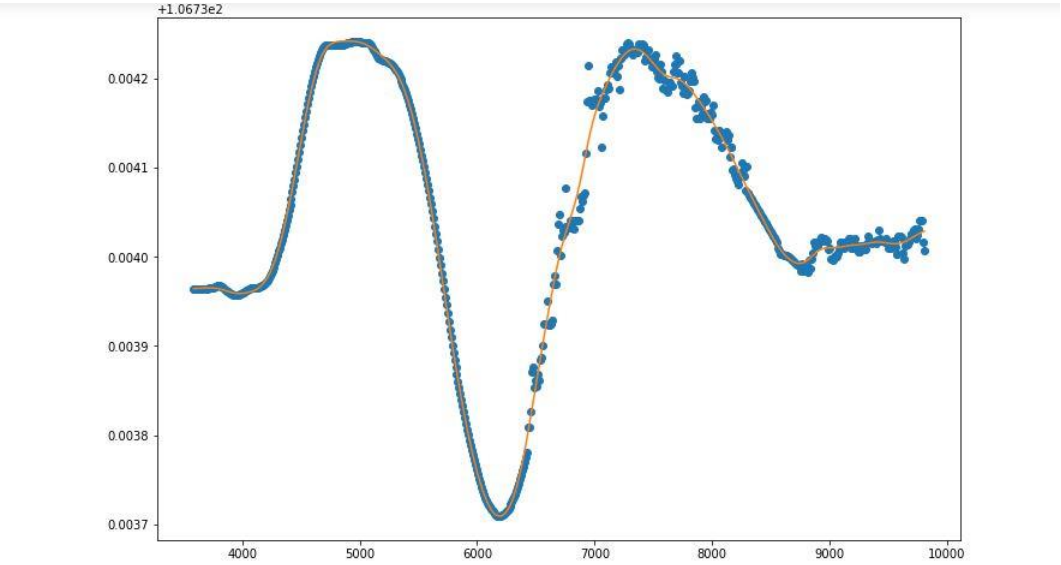
Dari model yang diperoleh kemudian dibuat grafiknya pada tabel waktu-target. Hasil yang didapatkan dari model HistGradientBoosting ternyata memberikan trayektori Quadcopter yang tidak terlalu *smooth* . Padahal gerakan drone merupakan suatu gerakan yang seharusnya kontinu dan smooth. Untuk itu, kami melakukan proses smoothing pada hasil prediksi yang diperoleh

Metode Smoothing yang dipilih yaitu linear smoothing dan Univariate Cubic Smoothing Spline. Linear Smoothing adalah proses Smoothing Data yang dilakukan dengan mendekatkan data pada suatu model linear. Hal ini sangat berguna pada model ini karena pada beberapa part pergerakan latitude dan langitude, terdapat gerakan-gerakan yang mendekati linear (misalnya pada saat kenaikan dan penurunan tajam). Sebelum melakukan linear smoothing, data dipotong terlebih dahulu menjadi bagian2 berdasarkan waktu. Kemudian pada

bagian yang seharusnya linear (pada pola kenaikan dan penurunan), dilakukan smoothing secara linear.



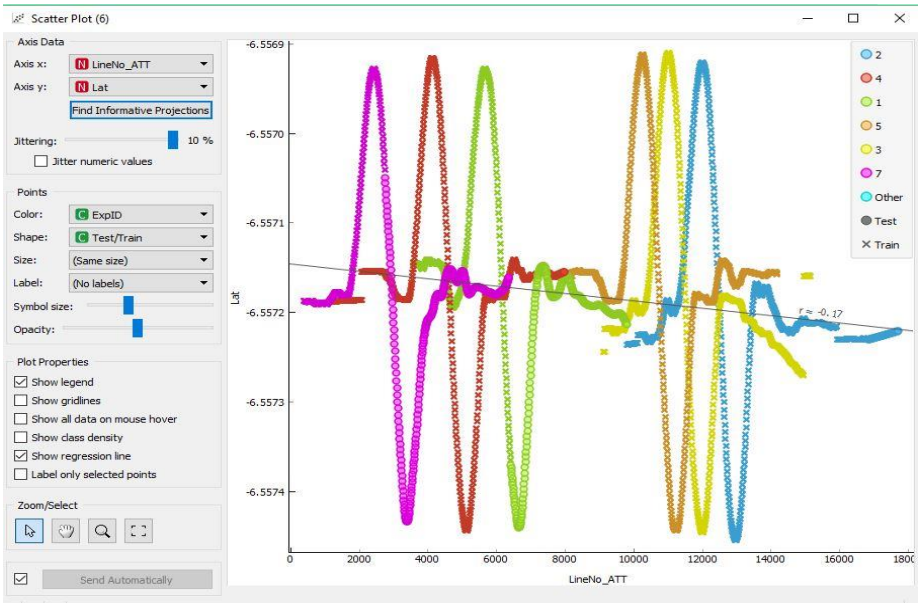
Selain smoothing secara linear, kami juga menggunakan Univariate Cubic Smoothing Spline. Cara kerja metode ini yaitu dengan mendekatkan data terhadap suatu fungsi yang terturunkan secara kontinu. Metode ini membantu mengubah pola data yang sebelumnya masih banyak patahan (?) menjadi lebih halus. Hal ini penting karena pergerakan drone secara teoritis berbentuk mengikuti fungsi yang halus. Implementasi smoothing pada data diatas dibantu dengan fungsi UnivariateCubicSmoothingSpline yang terdapat pada package csaps pada python



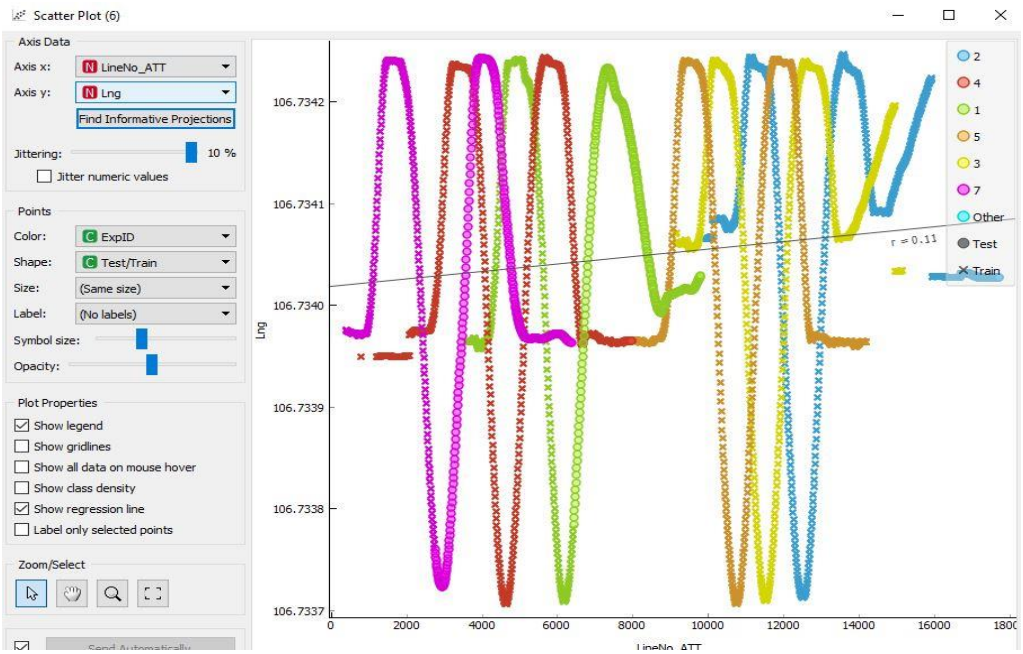
Hasil Smoothing pada grafik Langitude untuk ExpID = 1

Maka dari itu penulis memutuskan untuk mengaplikasikan *Cubic Smoothing Spline* pada hasil yang telah didapatkan. *Cubic Smoothing Spline* adalah teknik *smoothing* menggunakan pendekatan polinomial pangkat 3 (kyknya ini salah deh). Maka hasil prediksi trayektori Quadcopter setelah di-*smoothing* sebagai berikut

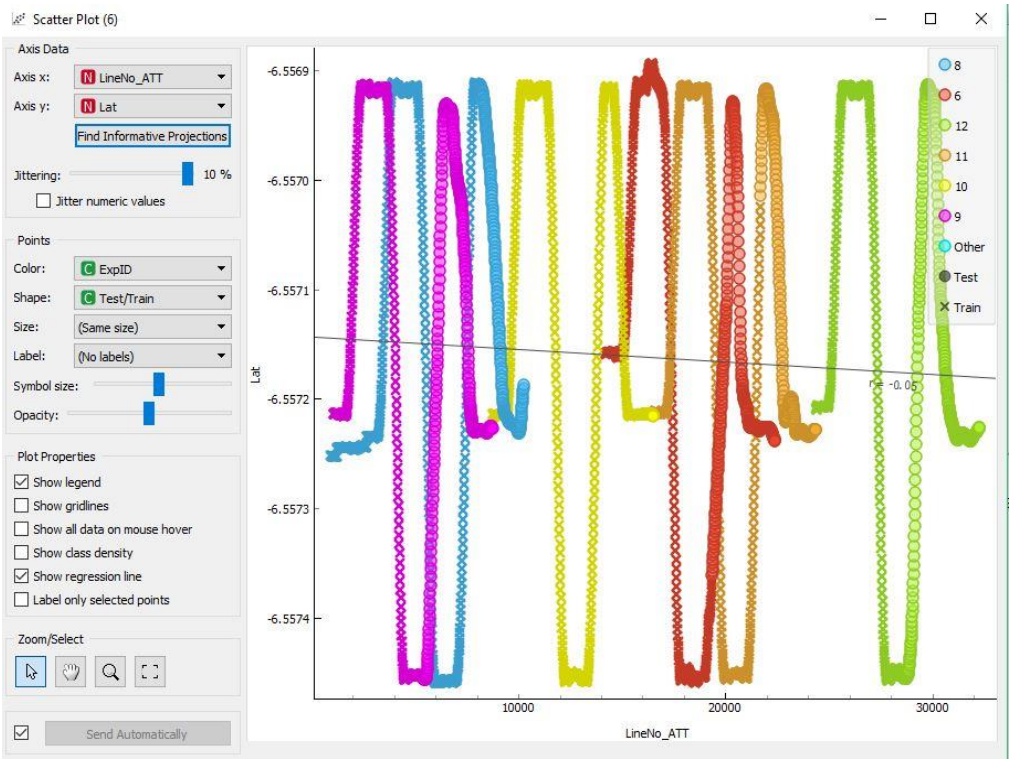
6. Hasil dan Analisa

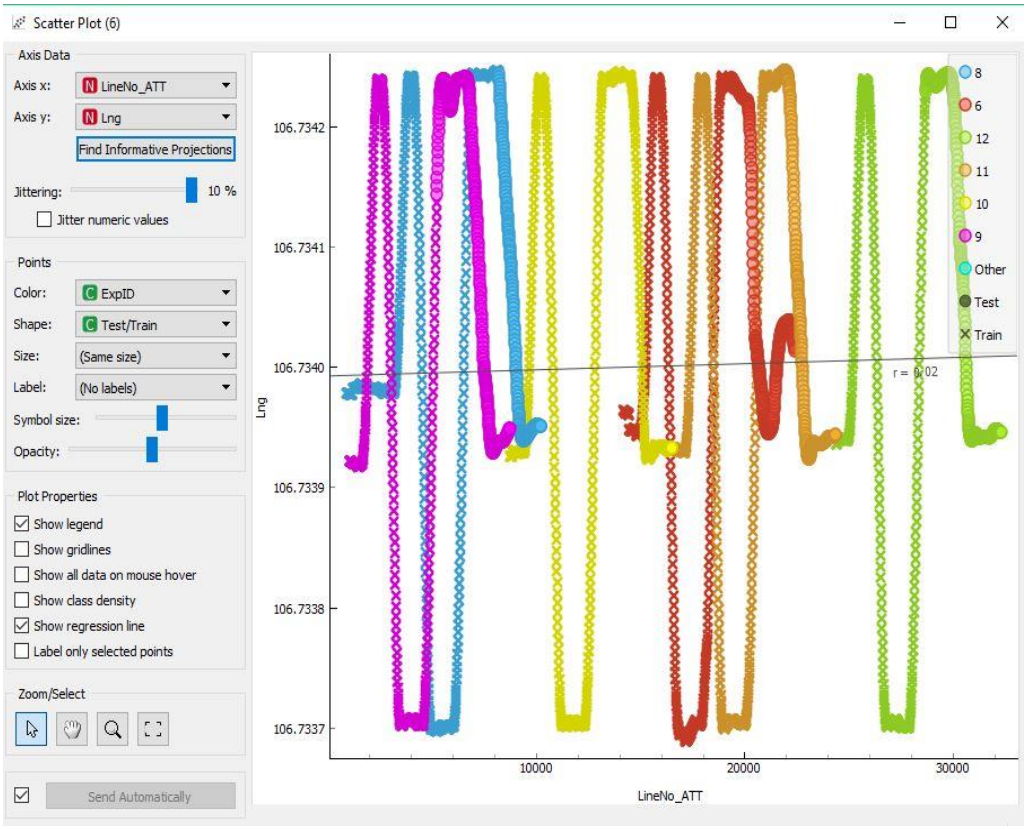


Hasil untuk Pattern ID 1



Hasil untuk Pattern ID 2





Keterangan : Warna menunjukan jenis PatternID, bentuk (x / o) menunjukan data train (nilai yang dari awal diketahui) atau data test (hasil prediksi)

Dapat dilihat bahwa hasil prediksi dari data di atas juga menunjukan pola yang serupa menunjukan bahwa hasil di atas masuk akal

7. Kesimpulan

Setelah membandingkan berbagai macam model, akhirnya kita memilih model HistGradientBoosting dengan Linear dan Univariate Cubic Spline Smoothing sebagai model akhir kita dengan error mean haversine distance dari submission pada Public Leaderboard yaitu sebesar 0.00304.