Faktor Exacta

Vol. 15, No. 3, Sept 2022, pp. 139~150
eISSN: 2502-339X, pISSN: 1979-276X, DOI: https://doi.org/10.30998/faktorexacta.v15i2.13254 ❒ 139

Prediksi Daya Output Sistem Pembangkit Listrik Tenaga Surya (PLTS) Menggunakan Regresi Linear Berganda

Suryo Bramasto[1], Dian Khairiani[2]
[1,2]Informatics Engineering, Institut Teknologi Indonesia

Article Info

Article history:

Received June 12, 2022
Revised Jul 17, 2022
Accepted Jul 26, 2022

Keywords:

Data analytics
Linear regression
Measurement data
Prediction
Solar power

ABSTRACT

The power generated by Pembangkit Listrik Tenaga Surya (PLTS) from time to time is fluctuating due to the influence of weather and other external conditions. This study predicts the output power of PLTS Sumalata in North Gorontalo Regency with data analytics on datasets obtained from measurements at two plants in order to maximize the potential generated solar energy. Multiple linear regression is used with Cross-industry standard for data mining (CRISP-DM) implementation. The equation formed from the prediction of the output power in Plant 1 is Y = -22216632810.1123-771640073.1888X_1+2349039057.8254X_2-25796134709.3552X_3 and for Plant 2 is Y = -2784.107+300.0146X_1–173.7016X_2+21773.3845X_3. The correlation coefficient on the Plant 1 dataset is 0.52 and the Plant 2 dataset is 0.92, so the weather and other external conditions have an effect of 52% on the output power generated at Plant 1 and 92% at Plant 2. It was also found that the prediction for Plant 1 is more accurate than Plant 2. In order to improve the accuracy of the prediction for evaluating the performance of the PLTS system, measurement data with a minimum measurement duration of one year is needed to represent seasonal conditions throughout the year.

Corresponding Author:

Suryo Bramasto,
Informatics Engineering,
Institut Teknologi Indonesia,
Jl. Raya Puspiptek Serpong, Tangerang Selatan.
Email: suryo.bramasto@iti.ac.id

1. PENDAHULUAN

Teknologi *machine learning* mengadopsi proses belajar yang dilakukan oleh manusia, untuk diterapkan pada mesin dengan membuat model matematis yang dapat merefleksikan pola-pola data yang ada. Teknik ini berusaha menciptakan *intelligent agent* untuk mengumpulkan data dari lingkungan kemudian diolah menjadi informasi dan atau pengetahuan baru [1]. Ketersediaan data menjadi salah satu faktor utama yang dibutuhkan pada implementasi teknologi ini. Dengan data, mesin dilatih untuk membaca dan menerjemahkan pola yang ada menggunakan algoritma yang dapat diterapkan [2].

Machine learning dapat digunakan untuk memprediksi masa depan (*unobserved event*) secara kualitatif maupun kuantitatif, baik pada data diskrit maupun kontinyu. Salah satu penerapannya adalah untuk memprediksi jumlah daya listrik yang dihasilkan oleh sistem Pembangkit Listrik Tenaga Surya (PLTS) [3]. Daya output pada sistem PLTS yang bersifat fluktuatif dapat diprediksi menggunakan variabel terkait untuk mengetahui produksi energi pada periode berikutnya.

Beberapa penelitian terkait prediksi daya output pada sistem PLTS telah dilakukan, antara lain yang dilakukan oleh K. Anuradha, et al. pada tahun 2021, dimana penelitian dilakukan dengan pendekatan *machine learning* untuk memprediksi daya yang dihasilkan di seluruh negara bagian India berdasarkan data lingkungan [4]. Metode yang digunakan adalah *Linear Regression* (LR), *Support Vector Machine Regression* (SVMR) dan *Random Forest* (RF). Hasil penelitian menunjukkan bahwa model regresi RF mempunyai performa lebih baik dari model lainnya, dengan akurasi sebesar 94,01%. Hasil prediksi yang diperoleh, tentunya sangat bergantung terhadap keterkaitan antar variabel pada data set yang tersedia dan metode yang digunakan. Oleh karena itu,

dipandang perlu untuk melakukan penelitian prediksi daya output pada sistem PLTS menggunakan metode lain dengan variabel yang berbeda untuk menguji seberapa kuat hubungan antara variabel yang digunakan terhadap hasil prediksi yang didapatkan.

Tujuan dari penelitian ini adalah mengetahui tingkat akurasi dari metode regresi linier menggunakan alat bantu Weka 3.8 untuk melakukan prediksi terhadap produksi energi dari sistem PLTS berdasarkan korelasinya dengan data iradiasi, suhu modul, dan suhu lingkungan. Sehingga dapat diperoleh hasil prediksi jumlah daya yang dihasilkan oleh sistem PLTS pada periode berikutnya, salah satunya untuk dapat digunakan sebagai bahan evaluasi performa sistem. Sedangkan sistem PLTS yang digunakan sebagai studi kasus adalah PLTS Sumalata di Gorontalo Utara yang memiliki kapasitas 2-megawatt peak (MWp).

1.1 Machine Learning

Machine Learning merupakan teknik untuk melakukan inferensi terhadap data dengan pendekatan matematis. Inferensi yang dilakukan menitikberatkan ranah hubungan antar variabel. Inti dari *Machine Learning* adalah untuk membuat model matematis yang merefleksikan pola data [5]. *Machine Learning* memiliki dua tujuan yakni untuk memprediksi masa depan (*unobserved event*) dan atau memperoleh ilmu pengetahuan (*knowledge discovery*) [1].

Guna mencapai tujuan tersebut, digunakan data sampel, kemudian dibuat model untuk generalisasi aturan atau pola data, sehingga dapat digunakan untuk mendapatkan informasi dan atau membuat keputusan. Dari dua sampel yang berbeda, dapat diambil kesimpulan yang berbeda, sehingga teknik pemilihan sampel merupakan hal yang penting [6]. Persyaratannya menggunakan dua kelompok data sampel yang akan digunakan, yakni data latih (*training data*) yang digunakan untuk membangun model dan data uji (*testing data*) yang digunakan untuk menguji kinerja dari model pembelajaran. Dari sisi metode pembelajaran, *machine learning* dapat dikategorikan menjadi *supervised learning, semi-supervised learning, unsupervised learning,* dan *reinforcement learning.*

Supervised learning berarti pembelajaran yang terarah/terawasi. Tujuan pembelajaran ini secara umum adalah untuk memperkirakan fungsi pemetaannya, sehingga ketika terdapat input baru, sistem dapat memprediksi output untuk input tersebut. Pendekatan *supervised learning* mempunyai input dan *desired* output yang dapat dibuat menjadi suatu model hubungan matematis sehingga mampu melakukan prediksi dan klasifikasi berdasarkan data yang telah ada sebelumnya. *Supervised Learning* menggunakan satu set pelatihan untuk mengajarkan model untuk menghasilkan output yang diinginkan. Dataset pelatihan ini mencakup input dan output yang benar, yang memungkinkan model untuk belajar dari waktu ke waktu. Algoritma mengukur akurasi melalui fungsi kerugian (*Loss Function)*, lalu menyesuaikan hingga kesalahan minimal. Beberapa algoritma yang termasuk dalam *supervised learning* antara lain:

a. Regresi Linier;
b. Analisis Deret Waktu;
c. *Decision Tree* dan *Random Forest*;
d. *Naive Bayes Classifier*;
e. *Nearest Neighbour Classifier*;
f. *Artificial Neural Network;* dan
g. *Support Vector Machine.*

Permasalahan pada s*upervised learning* dapat dikelompokkan menjadi masalah klasifikasi (*classification problems*) dan masalah regresi (*regression problem*).

Semi-Supervised Learning pada dasarnya mirip dengan *Supervised Learning*, perbedaan terdapat pada proses pelabelan data. Jika pada *Supervised Learning* terdapat ‘guru’ yang berfungsi memberikan ‘kunci jawaban’ pada proses input-output, maka pada metode *Semi-Supervised Learning* tidak terdapat ‘kunci jawaban’ secara eksplisit yang dibuat oleh ‘guru’. ‘Kunci jawaban’ ini dibuat secara otomatis. Pada metode ini, umumnya data masukan (*input data*) tersedia dalam jumlah besar dan hanya beberapa dari data tersebut yang dilabeli, kemudian diciptakan data tambahan baik menggunakan *supervised* maupun *unsupervised learning*, kemudian membuat model belajar dari data tambahan tersebut.

Pada metode *unsupervised learning*, tidak terdapat ‘guru’ yang mengajar. Pendekatan *unsupervised learning* tidak menggunakan data latih atau data training untuk melakukan pengelompokan berdasarkan *cluster*. Beberapa algoritma pada *unsupervised learning* antara lain:

a. *K-Means*;
b. *Hierarchical Clustering*;
c. DBSCAN;
d. *Fuzzy C-Means*; dan
e. *Self-Organizing Map*.

Berdasarkan model matematisnya, algoritma-algoritma tersebut tidak memiliki target variabel (*desired output*). Salah satu tujuan dari algoritma ini adalah mengelompokkan objek yang hampir sama dalam suatu area

tertentu. Permasalahan pada *unsupervised learning* dapat dikelompokkan menjadi *clustering problems* dan *association problems*.

Reinforcement Learning (RL) adalah pembelajaran terhadap apa yang akan dilakukan dan bagaimana memetakan situasi kedalam aksi untuk mendapatkan *reward* yang maksimal. Pembelajar (*learner*) tidak diberitahu aksi mana yang akan diambil, tetapi lebih pada menemukan aksi mana yang dapat memberikan *reward* yang maksimal dengan mencoba melakukannya. *Learner* dan pengambil keputusan (*decision maker*) disebut agen (*agent*). Segala sesuatu yang berinteraksi dengan agen disebut lingkungan (*environment*). Agen memilih aksi dan lingkungan merespon kepada aksi tersebut dan memberikan keadaan (*state*) baru kepada agen. Lingkungan juga menghasilkan *reward*, nilai numerik tertentu yang coba dimaksimalkan agen tiap waktu.

1.2 Regresi

Regresi merupakan salah satu teknik untuk meramalkan data di masa yang akan datang menggunakan variabel *independent* dan variabel penjelas/bebas). Berbeda dengan klasifikasi yang memprediksi nilai variabel yang bersifat diskret, regresi melakukan fungsi pembelajaran yang memetakan sebuah unsur data ke sebuah variabel prediksi bernilai nyata, yang digunakan untuk memprediksi nilai variabel yang bersifat kontinyu. Terdapat beberapa jenis regresi antara lain:

a. Regresi Linier [7], merupakan sebuah pendekatan untuk memodelkan hubungan antara variabel terikat Y dan satu atau lebih variabel X yang merupakan variabel bebas, dimana seluruh variabelnya adalah data kuantitatif. Disebut linier karena setiap estimasi atas nilai diharapkan memgalami peningkatan atau penurunan mengikuti garis lurus. Metode ini digunakan untuk mengetahui bagaimana variabel dependent dapat diprediksikan melalui variabel independent atau variabel prediktor. Dampak dari penggunaan regresi dapat digunakan untuk memutuskan apakah naik dan menurunnya variabel dependent dapat dilakukan melalui menaikan dan menurunkan keadaan variabel independent, atau meningkatkan keadaan variabel *dependent* dapat dilakukan dengan meningkatkan variabel *independent* dan atau sebaliknya. Berdasarkan jumlah variabel bebas X, regresi linier dibagi menjadi dua jenis yakni regresi linier sederhana dan regresi linier berganda (*multiple linear regression).*
b. Regresi Non-Linier, merupakan hubungan antara variabel Y dan X yang tidak linier. Tidak linier artinya, laju perubahan Y akibat laju perubahan X tidak konstan untuk nilai-nilai X tertentu. Seperti regresi kuadratik, kubik. Misal: produksi padi akan meningkat saat diberi pupuk taraf rendah ke sedang. Namun apabila diberi pupuk dengan taraf tinggi, maka tingkat produksinya malah semakin menurun.
c. Regresi *Dummy*, merupakan hubungan antara variabel y (data kuantitatif) dan variabel x (data kualitatif). Misal: melihat pengaruh kemasan terhadap harga jual makanan. Dengan kode '1' mewakili kemasan menarik dan '0' jika kemasan tidak menarik. Kode '1' dan '0' adalah variabel *dummy*.
d. Regresi Logistik, merupakan hubungan antara variabel y (data kualitatif) dan variabel x (data kuantitatif). Misal: Apabila ingin diketahui apakah konsumen akan membeli makanan di rumah makan berdasarkan penilaian konsumen terhadap lokasi, pelayanan, pendapatan. Dalam kasus ini hanya ada 2 kemungkinan respon konsumen, yaitu konsumen membeli (1) dan tidak membeli (0).

Beberapa algoritma yang dapat digunakan pada metode regresi antara lain:

a. *Simple linear regression*
b. *Multiple linear regression*
c. *Polynomial regression*
d. *Support vector regression*
e. *Decision tree regression*
f. *Random forest regression*

1.3 Regresi Linear Berganda

Regresi linier berganda merupakan model persamaan yang menjelaskan hubungan satu variabel tak bebas/response (Y) dengan dua atau lebih variabel bebas/predictor (X_1, X_2,…, X_n) [8]. Regresi linier diformulasikan menggunakan persamaan:

$$\mathbf{Y} = \boldsymbol{a} + \boldsymbol{b}\mathbf{X} \quad (1)$$

dimana:

Y : Variabel tak bebas (nilai variabel yang akan diprediksi)
X : Variabel bebas
A : Konstanta
B : Nilai koefisien regresi

Maka pada regresi linier berganda, secara matematis diekspresikan dengan:

$$\mathbf{Y = a + b_1X_1 + b_2X_2 + \ldots\ldots + b_nX_n} \quad (\textbf{Error! No text of specified style in document.})$$

Keadaan-keadaan bila koefisien-koefisien regresi yaitu b_1, b_2 dst mempunyai nilai:

- Nilai=0, maka dalam hal ini variabel Y tidak dipengaruh oleh X_1 dan X_2
- Nilai negatif, maka terjadi hubungan dengan arah terbalik antara variabel tak bebas Y dengan variabel-variabel X_1 dan X_2
- Nilai positif, maka terjadi hubungan yang searah antara variabel tak bebas Y dengan variabel-variabel X_1 dan X_2

Menentukan a, b_1, b_2, …, b_n dapat menggunakan metode kuadrat terkecil melalui persamaan normal sebagai berikut:

$$\begin{bmatrix} \boldsymbol{n} & \Sigma\boldsymbol{X_1} & \Sigma\boldsymbol{X_2} & \Sigma\boldsymbol{X_3} & \ldots & \Sigma\boldsymbol{X_n} \\ \Sigma\boldsymbol{X_1} & \Sigma\boldsymbol{X_1}^2 & \Sigma\boldsymbol{X_1X_2} & \Sigma\boldsymbol{X_1X_3} & \ldots & \Sigma\boldsymbol{X_1X_n} \\ \Sigma\boldsymbol{X_2} & \Sigma\boldsymbol{X_1X_2} & \Sigma\boldsymbol{X_2}^2 & \Sigma\boldsymbol{X_2X_3} & \ldots & \Sigma\boldsymbol{X_2X_n} \\ \Sigma\boldsymbol{X_3} & \Sigma\boldsymbol{X_1X_3} & \Sigma\boldsymbol{X_2X_3} & \Sigma\boldsymbol{X_3}^2 & \ldots & \Sigma\boldsymbol{X_3X_n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \Sigma\boldsymbol{X_n} & \Sigma\boldsymbol{X_1X_n} & \Sigma\boldsymbol{X_2X_n} & \Sigma\boldsymbol{X_3X_n} & \ldots & \Sigma\boldsymbol{X_n}^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b_1} \\ \boldsymbol{b_2} \\ \boldsymbol{b_3} \\ \ldots \\ \boldsymbol{b_n} \end{bmatrix} = \begin{bmatrix} \Sigma\boldsymbol{Y} \\ \Sigma\boldsymbol{X_1Y} \\ \Sigma\boldsymbol{X_2Y} \\ \Sigma\boldsymbol{X_3Y} \\ \ldots \\ \Sigma\boldsymbol{X_nY} \end{bmatrix} \quad (2)$$

Bentuk persamaan matriks di atas termasuk ke dalam suatu sistem persamaan linier. Mencari atau menentukan a, b_1, b_2, b_3, …, b_n berarti mencari atau menentukan solusi dari sistem persamaan linier (SPL). Mencari solusi SPL ada berbagai macam cara, diantaranya ialah Metode Eliminasi Gauss, Metode Invers (Metode Matriks yang diperbesar dan Metode Matriks Adjoin), dan Metode Cramer [9].

Metode Cramer merupakan metode yang paling populer dalam menentukan suatu solusi SPL karena sifatnya yang mudah dipelajari dan sederhana. Menurut Cramer jika terdapat SPL sebagai berikut:

$$\underbrace{\begin{bmatrix} \boldsymbol{a_{11}} & \boldsymbol{a_{12}} & \boldsymbol{a_{13}} & \ldots & \boldsymbol{a_{1n}} \\ \boldsymbol{a_{21}} & \boldsymbol{a_{22}} & \boldsymbol{a_{23}} & \ldots & \boldsymbol{a_{2n}} \\ \boldsymbol{a_{31}} & \boldsymbol{a_{32}} & \boldsymbol{a_{33}} & \ldots & \boldsymbol{a_{3n}} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{a_{n1}} & \boldsymbol{a_{n2}} & \boldsymbol{a_{n3}} & \ldots & \boldsymbol{a_{nn}} \end{bmatrix}}_{{}_nA_n} \underbrace{\begin{bmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \\ \boldsymbol{x_3} \\ \ldots \\ \boldsymbol{x_n} \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \boldsymbol{y_3} \\ \ldots \\ \boldsymbol{y_n} \end{bmatrix}}_{y} \quad (3)$$

Maka x_1, x_2, x_3, …, x_n dapat langsung dicari dengan membagi determinan matriks A_j dengan determinan matriks koefisien A. Dimana A_j = matriks A yang kolom ke- j nya diganti dengan matriks Y. Contoh:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \ldots & a_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn} \end{bmatrix}$$

$$A_1 = \begin{bmatrix} Y_1 & a_{12} & a_{13} & \ldots & a_{1n} \\ Y_2 & a_{22} & a_{23} & \ldots & a_{2n} \\ Y_3 & a_{32} & a_{33} & \ldots & a_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ Y_n & a_{n2} & a_{n3} & \ldots & a_{nn} \end{bmatrix} \quad A_2 = \begin{bmatrix} a_{11} & Y_1 & a_{13} & \ldots & a_{1n} \\ a_{21} & Y_2 & a_{23} & \ldots & a_{2n} \\ a_{31} & Y_3 & a_{33} & \ldots & a_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & Y_n & a_{n3} & \ldots & a_{nn} \end{bmatrix} \text{ dst}$$

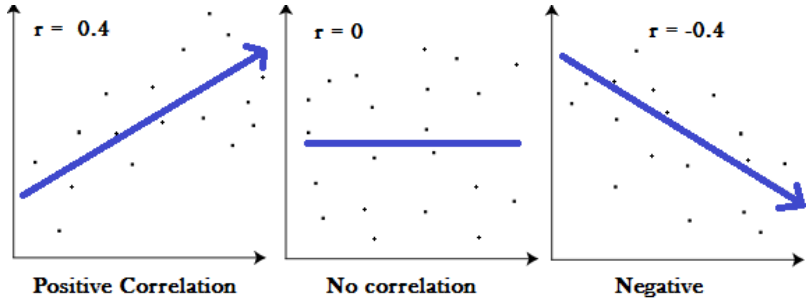
$$x_j = \frac{|A_j|}{|A|} \; dengan\; j = 1,2,3, \ldots dst,$$

sehingga $x_1 = \frac{|A_1|}{|A|}$ $x_2 = \frac{|A_2|}{|A|}$ $x_3 = \frac{|A_3|}{|A|}$, dst (4)

1.4 Evaluasi Tingkat Prediksi

Setelah mendapatkan hasil prediksi dari data uji, kita perlu melakukan evaluasi lebih lanjut terhadap hasil tersebut. Metode evaluasi yang akan diterapkan untuk menghitung tingkat akurasi hasil prediksi daya output PLTS menggunakan enam cara pengukuran, yakni Correlation Coefficient (CC), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), dan Root Relative Squared Error (RRSE) [5].

a. *Correlation Coefficient* (CC), digunakan untuk menemukan seberapa kuat hubungan antara data. Nilai CC berkisar antara -1 dan 1, dengan ketentuan
   i. CC bernilai positif (+) menunjukkan hubungan positif yang kuat;
   ii. CC bernilai negatif (-) menunjukkan hubungan negatif yang kuat;
   iii. Nilai CC 0 menunjukkan tidak ada hubungan sama sekali.



Gambar 1. Ilustrasi nilai *correlation coefficient*

CC diperoleh menggunakan formula:

$$CC = \frac{n(\sum XY)-(\sum X)(\sum Y)}{\sqrt{[n(\sum X^2)-(\sum X)^2][n(\sum Y^2)-(\sum Y)^2]}} \quad (5)$$

b. *Mean Absolute Error* (MAE) adalah rata-rata nilai *Absolute Error* dari kesalahan prediksi, dengan mengabaikan tanda positif ataupun negatifnya. MAE dihitung menggunakan formula:

$$MAE = \frac{\sum_{i=1}^{n} |Y'_i - Y_i|}{n} \quad (6)$$

c. *Root Mean Squared Error* (RMSE), merupakan nilai rata-rata dari jumlah kuadrat kesalahan, juga dapat menyatakan ukuran besarnya kesalahan yang dihasilkan oleh suatu model prediksi. RMSE di-peroleh dengan perhitungan berikut:

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n} (Y'_i - Y_i)^2}}{n} \quad (7)$$

d. *Relative Absolute Error* (RAE), merupakan nilai mutlak dari kesalahan absolut total terhadap nilai mutlak kesalahan absolut total prediktor sederhana. RAE diperoleh dengan perhitungan berikut

$$RAE = \frac{\sum_{i=1}^{n} |Y'_i - Y_i|}{\sum_{i=1}^{n} |Y_i - \frac{\sum_{i=1}^{n} Y_i}{n}|} \quad (8)$$

e. *Root Relative Squared Error* (RRSE), merupakan kesalahan kuadrat total dibagi dengan kesalahan kuadrat total dari prediktor sederhana. Dengan menghitung RRSE, dapat mengurangi kesalahan menuju dimensi yang sama dengan nilai yang diprediksi. RRSE dihitung dengan formula berikut

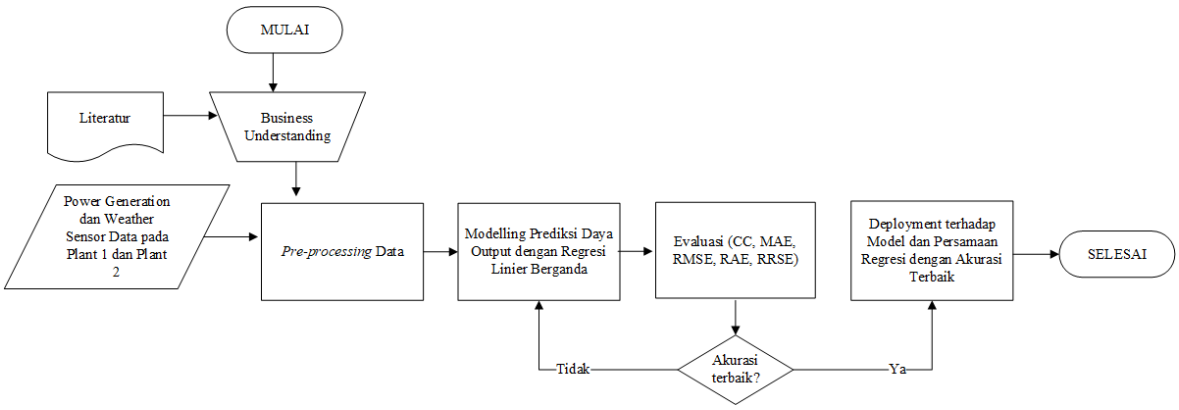
$$\mathbf{RRSE} = \sqrt{\frac{\sum_{i=1}^{n}(Y'_i - Y_i)^2}{\sum_{i=1}^{n}[Y'_i - (\frac{\sum_{i=1}^{n} Y_i}{n})]^2}} \quad (9)$$

2. METODE

Metode yang digunakan dalam penelitian ini menggunakan metode standar CRISP-DM yang disesuaikan dengan kondisi penelitian. Cross-Industry Standard Process for Data Mining (CRISP-DM) dikembangkan pada tahun 1996 oleh analis yang mewakili Daimler Chrysler, SPSS, dan NCR. CRISP menyediakan proses standar yang non-proprietary dan tersedia secara bebas untuk menyesuaikan data mining ke dalam strategi pemecahan masalah umum dari bisnis atau unit penelitian [10]. Menurut CRISP-DM, siklus proyek data mining memiliki enam fase [10], yakni:

a. Business Understanding
b. Data Understanding
c. Data Preparation
d. Modelling
e. Evaluation
f. Deployment

Penerapan CRISP-DM pada prosedur kerja untuk penilitian ini diilustrasikan pada Gambar 2.



Gambar 2. Metode penelitian

2.1 Business Understanding

Tahapan pekerjaan dalam pemahaman penelitian untuk mengembangkan model prediksi daya output PLTS dengan perumusan identifikasi masalah, batasan penelitian dan tujuan penelitian.

2.2 Data Understanding

Data dikumpulkan dari dua *plant* pada PLTS Sumalata selama periode 34 hari berturut-turut dengan interval 15 menit. Data terdiri atas dua pasang *file*, yang setiap pasangnya memiliki satu set data pembangkit listrik dan satu set data pembacaan sensor. Kumpulan data pembangkit listrik diambil dari inverter, di mana setiap inverter memiliki beberapa baris panel surya yang terpasang. Sementara data sensor dikumpulkan dari pembangkit, dengan rangkaian sensor tunggal yang dipasang secara optimal [9]. Rekaman data set pada *plant* 1 dan *plant* 2 dimulai pada tanggal 15 Mei 2021 pukul 00:00:00 sampai dengan 17 Juni 2021 pukul 23:45:00. Data pembangkitan terdiri atas 7 kolom, dengan 68.778 baris pada *plant* 1 dan 67.698 baris pada *plant* 2. Sementara data sensor terdiri dari 6 kolom, dengan 3.182 baris pada *plant* 1 dan 3.259 pada *plant* 2. Selanjutnya identifikasi dilakukan terhadap masing-masing atribut/kolom pada setiap data set menggunakan aplikasi Weka 3.8, dengan *resume* ditunjukkan pada tabel 1.

Tabel 1. Status Atribut pada Dataset Plant_1_Generation_Data.csv

NO	ATRIBUT	TIPE	MISSING	DISTINCT	UNIQUE
1.	DATE_TIME	Nominal	0 (0%)	3158	0 (0%)
2.	PLANT_ID	Numeric	0 (0%)	1	0 (0%)
3.	SOURCE_KEY	Nominal	0 (0%)	22	0 (0%)
4.	AC_POWER	Numeric	0 (0%)	32909	29390 (43%)
5.	DC_POWER	Numeric	0 (0%)	32686	28973 (42%)
6.	DAILY_YIELD	Numeric	0 (0%)	29900	24344 (35%)
7.	TOTAL_YIELD	Numeric	0 (0%)	37267	36422 (53%)

Berdasar tabel 1, diperoleh beberapa informasi sebagai berikut:

a. Tidak ditemui *field* yang kosong (*missing value*) maupun nilai negatif pada seluruh atribut dalam data set Plant_1_Generation_Data.csv

b. Atribut DATE_TIME terdapat 22 duplikasi data pada tiap *instance.*
c. Atribut PLANT_ID hanya berisi satu buah data untuk seluruh *instance*, menunjukkan ID dari pembangkit tersebut.
d. Atribut SOURCE_KEY terdiri dari 22 label dengan perulangan pada kisaran 3.100 kali untuk masing-masing label.

Sedangkan *resume* status atribut data set Plant_1_Weather_Sensor_Data.csv ditunjukkan pada tabel 2.

Tabel 2. Status Atribut pada Dataset Plant_1_Weather_Sensor_Data.csv

NO	ATRIBUT	TIPE	MISSING	DISTINCT	UNIQUE
1.	DATE_TIME	Nominal	0 (0%)	3182	3182 (100%)
2.	PLANT_ID	Numeric	0 (0%)	1	0 (0%)
3.	SOURCE_KEY	Nominal	0 (0%)	1	0 (0%)
4.	AMBIENT_TEMPERATURE	Numeric	0 (0%)	3182	3182 (100%)
5.	MODULE_TEMPERATURE	Numeric	0 (0%)	3182	3182 (100%)
6.	IRRADIATION	Numeric	0 (0%)	1758	1757 (55%)

Berdasar tabel 1, diperoleh beberapa informasi sebagai berikut:

a. Tidak ditemui *field* yang kosong (*missing value*) maupun maupun nilai negatif pada seluruh atribut dalam data set Plant_1_Weather_Sensor_Data.csv
b. Atribut SOURCE_KEY hanya berisi satu buah data untuk seluruh *instance*, menunjukkan ID dari data logger yang digunakan.

Kemudian resume atribut data set Plant_2_Generation_Data.csv ditunjukkan pada tabel 3.

Tabel 3. Status Atribut pada Dataset Plant_2_Generation_Data.csv

NO	ATRIBUT	TIPE	MISSING	DISTINCT	UNIQUE
1.	DATE_TIME	Nominal	0 (0%)	3259	0 (0%)
2.	PLANT_ID	Numeric	0 (0%)	1	0 (0%)
3.	SOURCE_KEY	Nominal	0 (0%)	22	0 (0%)
4.	AC_POWER	Numeric	0 (0%)	31067	30164 (45%)
5.	DC_POWER	Numeric	0 (0%)	31012	30063 (44%)
6.	DAILY_YIELD	Numeric	0 (0%)	30918	28187 (42%)
7.	TOTAL_YIELD	Numeric	0 (0%)	33118	32043 (47%)

Berdasar tabel 3, diperoleh beberapa informasi yang serupa dengan yang ditemui pada data set Plant_1_Generation_Data.csv sebagai berikut:

a. Tidak ditemui *field* yang kosong (*missing value*) maupun nilai negatif pada seluruh atribut dalam data set Plant_2_Generation_Data.csv
b. Atribut DATE_TIME terdapat 22 duplikasi data pada tiap *instance.*
c. Atribut PLANT_ID hanya berisi satu buah data untuk seluruh *instance*, menunjukkan ID dari pembangkit tersebut.
d. Atribut SOURCE_KEY terdiri dari 22 label dengan perulangan 3.259 kali untuk masing-masing label.

Resume atribut data set Plant_2_Weather_Sensor_Data.csv ditunjukkan pada tabel 4.

Tabel 4. Status Atribut pada Dataset Plant_2_Weather_Sensor_Data.csv

NO	ATRIBUT	TIPE	MISSING	DISTINCT	UNIQUE
1.	DATE_TIME	Nominal	0 (0%)	3259	3259 (100%)
2.	PLANT_ID	Numeric	0 (0%)	1	0 (0%)
3.	SOURCE_KEY	Nominal	0 (0%)	1	0 (0%)
4.	AMBIENT_TEMPERATURE	Numeric	0 (0%)	3259	3259 (100%)
5.	MODULE_TEMPERATURE	Numeric	0 (0%)	3259	3259 (100%)
6.	IRRADIATION	Numeric	0 (0%)	1863	1862 (57%)

Berdasar tabel 4, diperoleh informasi yakni:

a. Tidak ditemui *field* yang kosong (*missing value*) maupun maupun nilai negatif pada seluruh atribut dalam data set Plant_2_Weather_Sensor_Data.csv
b. Atribut SOURCE_KEY hanya berisi satu buah data untuk seluruh *instance*, menunjukkan ID dari data logger yang digunakan.

2.3 Data pre-processing

Pada Plant_1_Generation_Data ditemukan duplikasi atribut DATE_TIME sejumlah 22 duplikasi pada setiap *instance.* Kemudian atribut SOURCE_KEY terdiri atas 22 label dengan perulangan pada kisaran 3.100

kali untuk masing-masing label. Hal ini dikarenakan, pencatatan data pembangkitan dilakukan melalui 22 inverter yang terpasang. Untuk itu, perlu dilakukan *data cleaning* yang bertujuan untuk menghilangkan duplikasi data. Proses *data cleaning* dilakukan menggunakan bahasa Python pada Jupyter Notebook, dimana melalui proses tersebut data set Plant_1_Generation_Data yang semula terdiri dari 68.778 *instance* dan 7 *attribute* menjadi 3.158 *instance* dan 5 *attributes. Attribute* yang tidak diikutsertakan lagi adalah PLANT_ID dan SOURCE_KEY yang berisi data tentang identifikasi pembangkit dan inverter. Hasil preprocessing pada tahap ini disimpan dalam file Plant_1_GenDat_Grouped.csv. Begitu pula dengan data set Plant_2_Generation_Data, *data cleaning* dan *data reduction* dilakukan melalui proses yang sama. Dataset yang semula terdiri dari 67.698 *instance* dan 7 *attribute* menjadi 3.259 *instance* dan 5 *attributes.* Hasil *preprocessing* pada tahap ini kemudian disimpan dalam file Plant_2_GenDat_Grouped.csv.

Field yang terdapat pada empat file dataset yang akan digunakan telah memiliki tipe data yang sesuai untuk proses regresi. Namun karena akan dilakukan integrasi dataset (antara Plant_1_GenDat_Grouped dengan Plant_1_Weather_Sensor_Data dan Plant_2_GenDat_Grouped dengan Plant_2_Weather_Sensor_Data), maka perlu dilakukan *data transformation* dimana tipe field DATE_TIME perlu diubah dari 'nominal' menjadi 'date'. Format 'date' yang akan digunakan adalah yyyy-MM-dd HH:mm:ss. Transformasi data dilakukan dengan memilih data set yang akan diubah, kemudian mengaktifkan opsi '*Invoke options dialog*' pada Weka 3.0 dan menentukan format *date* yang akan digunakan. Selanjutnya, dataset disimpan dalam format berekstensi .arff untuk kemudian dapat diproses lebih lanjut.

Tahap *preprocessing* selanjutnya adalah *data integration.* Tahapan ini akan menggabungkan antara dataset Plant_1_GenDat_Grouped.arff dengan Plant_1_Weather_Sensor_Data.arff dan Plant_2_GenDat_Grouped.arff dengan Plant_2_Weather_Sensor_Data.arff. Sebelum dilakukan *data integration*, terlebih dahulu dilakukan identifikasi terhadap data set yang akan digabungkan. Keterangan dari dataset yang hendak digabungkan ditunjukkan pada tabel 5.

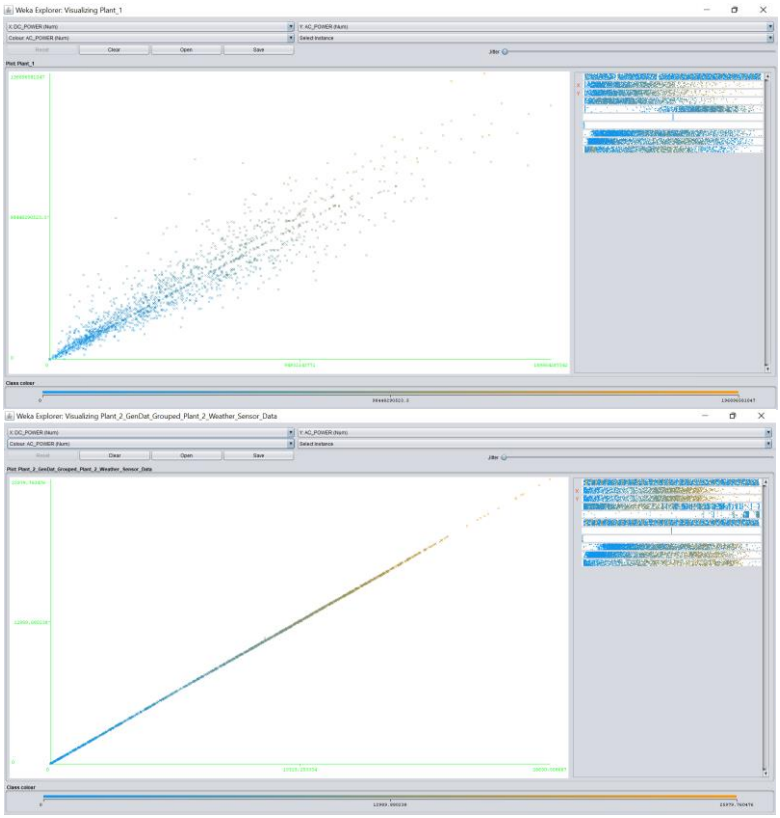
Tabel 5. Dataset yang Akan Digabungkan

NO	DATA SET AWAL	JUMLAH INSTANCES	DATA SET GABUNGAN
1.	Plant_1_GenDat_Grouped.arff	**3.158**	Plant_1.arff
2.	Plant_1_Weather_Sensor_Data.arff	**3.182**	
3.	Plant_2_GenDat_Grouped.arff	3.259	Plant_2.arff
4.	Plant_2_Weather_Sensor_Data.arff	3.259	

Pada tabel 5, terlihat masih terdapat jumlah *instance* yang berbeda antara dataset 1 dan 2. Untuk itu, evaluasi dilakukan menggunakan menggunakan bahasa Python pada Jupyter Notebook. Dataset Plant_1_GenDat_Grouped.arff yang semula memiliki 3.158 *instances* dan Plant_1_Weather_Sensor_Data.arff dengan 3182 *instances* telah digabungkan menjadi Plant_1.csv yang terdiri dari 3157 *instances.* Selanjutnya file dikonversi menjadi Plant_1.arff. Sementara untuk dataset 3 dan 4, oleh karena jumlah *instance* sudah sama, maka proses *data integration* dapat dilakukan. Terdapat *attribute* yang sama pada kedua data set, yakni DATE_TIME, sehingga *attribute* DATE_TIME pada Plant_2_Weather_Sensor_Data.arff perlu diubah terlebih dahulu. Kemudian langkah terkahir pada tahap *data integration* adalah menggunakan perintah *merge* di menu SimpleCLI pada Weka 3.8. Sebelum dilakukan *modelling* guna prediksi daya output, dilakukan *data reduction* untuk mengeliminir beberapa atribut pada data set Plant_1.arff dan Plant_2.arff. Atribut DC_POWER ditentukan sebagai *class* yang akan diprediksi.

Berdasarkan evaluasi terhadap sebaran data tiap atribut, diperoleh informasi sebagai berikut:

i. Pada atribut PLANT_ID dan SOURCE_KEY hanya berisi satu data yang sama untuk seluruh *instances*, sementara pada atribut DATE_TIME sebaran data bersifat konstan.
ii. Atribut DC_POWER memiliki pengaruh yang kuat terhadap DC_POWER sebagai *class* yang akan diprediksi, ditinjau dari visualisasi pada Gambar . Apabila atribut ini dipertahankan, maka akan menimbulkan potensi *overfitting.*
iii. Terdapat atribut DAILY_YIELD dan TOTAL_YIELD yang merupakan rekaman data jumlah energi yang dihasilkan secara periodik (harian dan total), yang dalam kasus ini tidak diperlukan untuk memprediksi daya output (DC_POWER).



Gambar 3. Visualisasi korelasi atribut DC_POWER terhadap DC_POWER

Dengan demikian, tersisa atribut AMBIENT_TEMPERATURE, MODULE_TEMPERATURE dan IRRADIATION serta DC_POWER sebagai *class* yang akan diprediksi.

2.4. Modelling Prediksi Daya Output PLTS dengan Regresi Linear Berganda

Pada tahap pemodelan ini pemilihan algoritma *machine learning* yang digunakan adalah Regresi Linear Berganda [9]. Data hasil preprocessing menjadi input model dari algoritma tersebut.

2.5 Evaluasi

Tahap evaluasi terhadap model dalam melakukan prediksi menggunakan metrik pengujian CC, MAE, RMSE, RAE, dan RRSE [5].

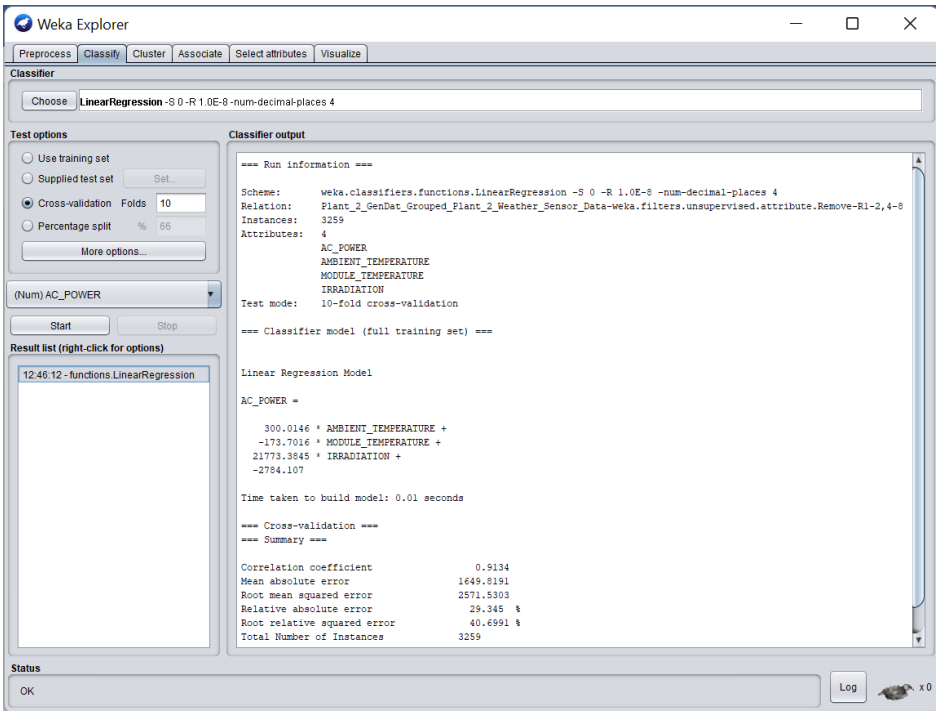
2.6 Deployment

Tahap pemanfaatan model yang telah dibuat sehingga model tersebut selanjutnya dapat dikembangkan dalam program aplikasi sebagai sebuah *function* atau *method* untuk memprediksi daya output PLTS kedepannya untuk kasus di PLTS Sumalata.

3. HASIL DAN PEMBAHASAN

3.1 Proses Prediksi

Model regresi linear berganda digunakan pada Weka 3.8 dengan menggunakan fungsi Classify dan *classifier* LinearRegression. Dataset yang pertama kali diprediksi adalah Plant2.arff, dimana dengan *cross validation* 10 dan atribut target DC_POWER, maka dihasilkan seperti yang ditunjukkan pada gambar 4.



Gambar 4. Hasil prediksi dengan regresi linear berganda

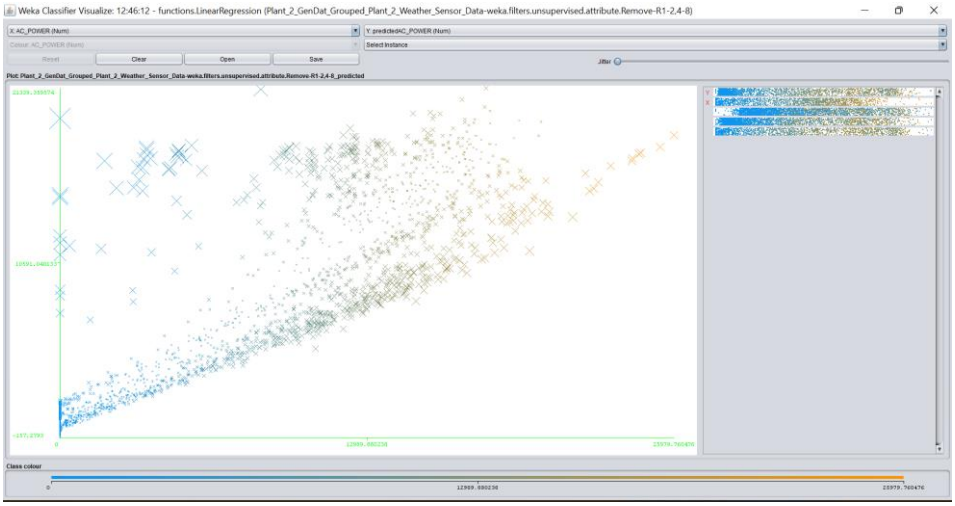
Penjelasan dari gambar 6 tersebut adalah yakni persamaan regresi linear yang diperoleh adalah:

$$Y = -2784.107 + 300.0146\ X_1 - 173.7016\ X_2 + 21773.3845\ X_3 \quad (11)$$

dengan:
Y : DC_POWER sebagai variabel tak bebas
X_1 : AMBIENT_TEMPERATURE
X_2 : MODULE_TEMPERATURE
X_3 : IRRADIATION
dengan evaluasi akurasi prediksi:

a. *Correlation Coefficient* (CC) = 0.9134
b. *Mean Absolute Error* (MAE) = 1649.8191
c. *Root Mean Squared Error* (RMSE) = 2571.5303
d. *Relative Absolute Error* (RAE) = 29.345 %
e. *Root Relative Squared Error* (RRSE) = 40.6991%

Sedangkan visualisasi dari hasil prediksi yang ditunjukkan pada gambar 6, ditampilkan pada gambar 5.



Gambar 5. Visualisasi hasil prediksi

3.2 Hasil Prediksi dengan Regresi Linear Berganda

Proses prediksi untuk masing-masing data set dilakukan dengan menggunakan beberapa variasi komposisi jumlah data latih dan data uji yang digunakan. Resume dari hasil prediksi ditunjukkan pada tabel 6 dan tabel 7.

Tabel 6. Resume Hasil Prediksi terhadap Dataset Plant1.arff

KOMPOSISI / HASIL	*CROSS VALIDATION*		*PERCENTAGE SPLIT* (Data Latih : Data Uji)		
	5	**10**	**50:50**	**66:34**	**90:10**
Persamaan Regresi Linier	$Y = -22216632810.1123 - 771640073.1888\ X_1 + 2349039057.8254\ X_2 - 25796134709.3552\ X_3$				
Correlation Coefficient	0.5579	0.5567	0.5402	0.5182	0.5343
Mean Absolute Error	19965537073.1416	19982539908.4889	20336996186.6248	20946713857.9716	19399074902.1373
Root Mean Squared Error	28517768065.9222	28544620449.6325	29539629877.2906	30589634888.194	28386360309.5179
Relative Absolute Error	71.9108 %	71.9846 %	71.9265 %	74.0124 %	70.5097 %
Root Relative Squared Error	82.9587 %	83.0498 %	84.1714 %	85.5712 %	84.4418 %

Tabel 7. Resume Hasil Prediksi terhadap Dataset Plant2.arff

KOMPOSISI / HASIL	*CROSS VALIDATION*		*PERCENTAGE SPLIT* (Data Latih : Data Uji)		
	5	**10**	**50:50**	**66:34**	**90:10**
Persamaan Regresi Linier	$Y = -2784.107 + 300.0146\ X_1 - 173.7016\ X_2 + 21773.3845\ X_3$				
Correlation Coefficient	0.9134	0.9134	0.9159	0.9084	0.9229
Mean Absolute Error	1649.4425	1649.8191	1621.9268	1645.4102	1620.5938
Root Mean Squared Error	2570.8122	2571.5303	2514.6011	2587.5543	2435.1995
Relative Absolute Error	29.3315 %	29.345 %	28.8342 %	29.6109 %	28.6972 %
Root Relative Squared Error	40.6788%	40.6991%	40.1137 %	41.8407%	38.9546 %

3.3 Pembahasan Hasil Prediksi

Berdasarkan proses prediksi dengan Weka 3.0 terhadap dataset Plant 1 dan Plant 2 berikut hasil yang dirangkum pada tabel 6 dan 7, maka dapat dinyatakan beberapa hal sebagai berikut:

1. Berdasarkan evaluasi hasil prediksi dengan CC, MAE, RMSE, dan RRSE maka diketahui prediksi terhadap Plant 2 lebih akurat dibandingkan dengan Plant 1. Lebih akurat dalam hali ini adalah semakin mendekati garis linear pada visualisasi yang dapat diartikan semakin mendekati hasil pengukuran aktual pada perangkat.
2. Berdasarkan *correlation coefficient* yang diperoleh maka diketahui bahwa cuaca dan kondisi eksternal lainnya lebih berpengaruh terhadap daya yang dihasilkan dari waktu ke waktu pada Plant 2 dibandingkan dengan pada Plant 1.
3. Terjadi perbedaan angka yang signifikan pada variabel-variabel antara persamaan regresi linear yang diperoleh pada Plant 1 dan Plant 2. Plant 1 dan Plant 2 di PLTS Sumalata menggunakan perangkat mesin yang sama jenis, usia dan *merk*nya; baik guna pembangkitan, distribusi, pengkuran, dan monitoring. Lokasi Plant 1 dan Plant 2 PLTS Sumalata adalah saling bersebelahan dengan terpisah jarak kurang lebih 20 meter. Posisi panel surya dalam hal ini kemiringan dan tata letak relatif terhadap matahari antara Plant 1 dan Plant 2 juga sama. Perbedaan antara Plant 1 dan Plant 2 adalah pihak atau orang yang melakukan *setup* dan kalibrasi perangkat, serta operator *plant* setiap harinya. Kemudian sampai saat ini juga belum pernah dilakukan pengukuran/pemantauan secara terus menerus pada DC_POWER *actual* yakni menggunakan perangkat pengukuran yang dipasang pada *input* baterai penyimpan daya. Perangkat pengukuran sampai saat ini hanya ada pada *solar panel*.

4. PENUTUP

Persamaan garis dari hasil prediksi terhadap daya output pada Plant 1 adalah Y = -22216632810.1123 - 771640073.1888X_1 + 2349039057.8254X_2 - 25796134709.3552X_3. Sedangkan persamaan garis dari hasil prediksi pada Plant 2 adalah Y = -2784.107 + 300.0146X_1 – 173.7016X_2 + 21773.3845X_3. Berdasarkan pengujian diperoleh *correlation coefficient* pada dataset Plant 1 sebesar 0,52 dan dataset Plant 2 sebesar 0,92, sehingga dapat disimpulkan bahwa iradiasi, suhu modul, dan suhu lingkungan mempunyai pengaruh yang lebih signifikan terhadap daya output yang dihasilkan Plant 2 dibadingkan dengan pada Plant 1. Kemudian nilai MAE, RMSE, RAE dan RRSE pada dataset Plant 1 lebih tinggi dibandingkan Plant 2, sedangkan keeratan hubungan antara variabel bebas terhadap variabel terikat pada dataset Plant 2 lebih kuat dibandingkan pada dataset Plant 1; dengan demikian dapat disimpulkan juga bahwa prediksi terhadap daya output dari waktu ke waktu pada Plant 2 lebih akurat dibandingkan dengan Plant 1. Akurasi prediksi baik pada Plant 1 maupun Plant 2 dapat ditingkatkan sehingga dapat dipergunakan sebagai bahan evaluasi performa sistem PLTS. Untuk itu dibutuhkan data pengukuran dengan durasi paling tidak satu tahun untuk dapat mewakili kondisi musim sepanjang tahun, seperti musim kemarau, hujan, dan kondisi cuaca ekstrim. Kemudian berdasarkan hasil prediksi yang dilakukan dengan regresi linear berganda, diketahui terdapat perbedaan angka-angka yang signifikan antara Plant 1 dan Plant 2, baik pada atribut maupun variable. Dengan demikian diperlukan penelitian-penelitian lanjutan dengan kondisi dimana semua perangkat pengukuran yang digunakan pada Plant 1 dan Plant 2 terkalibrasi dalam kondisi yang sama. Kemudian juga diperlukan penambahan atribut baru yakni DC_POWER_actual, dimana merupakan data yang dihasilkan dari perangkat pengukuran yang dipasang pada input baterai penyimpan daya, sekaligus juga pendekatan baru dalam proses prediksi daya output PLTS Sumalata tersebut dengan pendekatan yang lebih mutakhir yakni *deep learning*.

DAFTAR PUSTAKA

[1] Jan Wira Gotama Putra, *Pengenalan Konsep Pembelajaran Mesin dan Deep Learning*, 1.4. 2020.

[2] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2014. doi: 10.1002/9781118874059.

[3] C. Wang, D. Yang, S. Gao, S. Zhao, L. Li, and X. Wang, "Research on Electricity Forecasting Method Based on Big Data," in *2022 5th International Conference on Energy, Electrical and Power Engineering (CEEPE)*, 2022, pp. 304–308. doi: 10.1109/CEEPE55110.2022.9783384.

[4] K. Anuradha, D. Erlapally, G. Karuna, V. Srilakshmi, and K. Adilakshmi, "Analysis Of Solar Power Generation Forecasting Using Machine Learning Techniques," *E3S Web of Conferences*, vol. 309, p. 01163, Oct. 2021, doi: 10.1051/e3sconf/202130901163.

[5] F. Aslay and N. S. Ting, "Machine Learning-Based Estimation of Output Current Ripple in PFC-IBC Used in Battery Charger of Electrical Vehicles: A Comparison of LR, RF and ANN Techniques," *IEEE Access*, vol. 10, pp. 50078–50086, 2022, doi: 10.1109/ACCESS.2022.3174100.

[6] Ani Kannal, *Solar Power Generation Data*. 2020.

[7] O. L. R. Albrecht and C. J. Taylor, "A linear regression variable time delay estimation algorithm for the analysis of hydraulic manipulators," in *2022 UKACC 13th International Conference on Control (CONTROL)*, 2022, pp. 148–153. doi: 10.1109/Control55989.2022.9781372.

[8] Frank E. Harrel, Jr., *Regression Modelling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd ed. Springer, 2015.

[9] Z. Zhang, F. Bai, H.-B. Quan, R.-J. Yin, and W.-Q. Tao, "PEMFC Output Voltage Prediction Based on Different Machine Learning Regression Models," in *2022 5th International Conference on Energy, Electrical and Power Engineering (CEEPE)*, 2022, pp. 401–406. doi: 10.1109/CEEPE55110.2022.9783124.

[10] J. S. Saltz, "CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 2337–2344. doi: 10.1109/BigData52589.2021.9671634.