Algoritma *Data Mining* untuk Optimasi Suhu dan Waktu Roasting Nibs Biji Kakao di *Cocoa Teaching Industry* (CTI) UGM

Rizky Alif Ramadhan
Departemen Teknologi Informasi dan Teknik Elektro
Universitas Gadjah Mada
Yogyakarta, Indonesia
rizky.alif.r@ugm.ac.id

Enas Dhuhri Kusuma
Departemen Teknologi Informasi dan Teknik Elektro
Universitas Gadjah Mada
Yogyakarta, Indonesia
enas@ugm.ac.id

Noor Akhmad Setiawan
Departemen Teknologi Informasi dan Teknik Elektro
Universitas Gadjah Mada
Yogyakarta, Indonesia
noorwewe@ugm.ac.id

***Intisari*— Algoritma *data mining* mulai digunakan pada berbagai bidang termasuk industri makanan. Dengan demikian, algoritma *data mining* ini tentunya juga dapat diterapkan pada industri kakao, terutama dalam proses pengolahan biji kakao. Salah satu proses dalam pengolahan biji kakao yang paling menentukan kualitas dari pengolahan biji kakao tersebut adalah proses *roasting.* Proses *roasting* adalah proses mengeluarkan kandungan air pada biji kakao, mengeringkannya, serta mengembangkan biji tersebut agar mendapatkan aroma dan warna yang khas dan sesuai dengan standar. Agar hasil *roasting* bagus, maka suhu dan durasi pemanasan harus optimal. Suhu dan durasi merupakan variabel yang memiliki peran penting dalam proses *roasting*. Tujuan dari penelitian ini membuat rekomendasi suhu pada durasi *roasting* tertentu berdasarkan data-data pada proses *roasting* seperti kapasitas roasting, kadar air, pH, jenis biji dan lain sebagainya. Suhu akan dijadikan target variabel pada penelitian ini. Penulis membandingkan tiga algoritma *data mining*, yaitu *Support Vector Regression (SVR), Multiple Linear Regression, Extreme Learning Machine (ELM),* dan *Particle Swarm Optimization-Extreme Learning Machine (PSO-ELM)* pada penelitian ini. Algoritma *data mining* tersebut dievaluasi menggunakan ukuran standar regresi seperti MAPE dan RMSE untuk nantinya dibandingkan performanya. Setelah melakukan validasi silang, didapatkan algoritma data *mining* terbaik untuk memodelkan suhu yaitu PSO-ELM dengan MAPE dan RMSE masing-masing 4.13% dan 7.36, kemudian SVR dengan MAPE 4.76% dan RMSE 9.17, selanjutnya ada ELM denngan MAPE 4.79% dan RMSE 8.62, dan yang terakhir ada *MLR* dengan MAPE 4.80% dan RMSE 8.63. Diharapkan dengan penelitian ini, operator dapat menentukan suhu dan durasi *roasting* yang optimal agar hasil *roasting* yang dihasilkan baik, sehingga kualitas produksi dari pengolahan biji kakao meningkat.**

Kata kunci— *Data Mining*, Biji Kakao, *Roasting*, Suhu, Durasi, SVR, Regresi Linear, ELM, PSO

I. PENDAHULUAN

Indonesia merupakan salah satu negara penghasil kakao terbesar di dunia. Tercatat pada tahun 2020, Indonesia mampu menghasilkan 659,7 ribu ton kakao [1] dengan nilai ekspor mencapai US$1,21 miliar [2]. Hal ini tentunya menjadikan kakao sebagai salah satu potensi komoditas perkebunan unggulan di Indonesia. Menurut data yang diambil dari situs web Kementerian Pertanian Republik Indonesia, pada tahun 2021, luas perkebunan kakao mencapai 1.497.467 Ha [3]. Selain menjadi komoditas perkebunan unggulan di Indonesia, berdasarkan hal tersebut kakao juga dapat menjadi sumber lapangan kerja bagi masyarakat sekitar perkebunan kakao.

Universitas Gadjah Mada melalui UGM *Cocoa Teaching and Learning Industry* ikut andil dalam menghasilkan produk kakao di Indonesia. UGM *Cocoa Teaching and Learning Industry* bertujuan untuk mendorong dan mempercepat program hilirisasi industri pengolahan kakao dan sebagai wahana produktif berbasis riset dan inovasi untuk mendukung proses pembelajaran yang bersinergi dengan industri. Industri coklat yang terletak di Kabupaten Batang tersebut merupakan hasil kerjasama dari berbagai pihak yaitu dari Universitas Gadjah Mada, Kementerian Perindustrian, Dikti, dan Pemerintah Kabupaten Batang. Industri ini sangat unik karena berlokasi di tengah-tengah perkebunan kakao warga [4].

Menurut wawancara dan kunjungan pabrik yang dilakukan oleh penulis, terdapat dua produk akhir yang dihasilkan pada UGM *Cocoa Teaching and Learning Industry* yaitu *butter* dan coklat bubuk. Sebelum menjadi *butter* dan coklat bubuk, tentunya biji kakao akan melalui beberapa proses produksi. Salah satunya adalah proses *roasting* atau penyangraian. Proses roasting adalah proses mengeluarkan kandungan air pada biji kakao, mengeringkannya, serta mengembangkan biji tersebut agar mendapatkan aroma dan warna yang khas. Variabel yang berpengaruh dalam proses *roasting* adalah waktu dan suhu yang diatur dalam proses r*oasting* [5]. Agar hasil *roasting* bagus, maka suhu dan waktu roasting harus optimal. Pada penelitian ini, dengan menggunakan algoritma data mining, penulis akan membuat model yang dapat digunakan untuk merekomendasikan suhu dan durasi *roasting* berdasarkan data-data pada proses r*oasting* seperti kapasitas r*oasting*, kadar air, pH, jenis biji dan variabel lain. Penulis membandingkan tiga algoritma *data mining*, yaitu *Support Vector Regression* (SVR), *Multiple Linear Regression (MLR), Extreme Learning Machine* (ELM), dan *Particle Swarm Optimization-Extreme Learning Machine* (PSO-ELM) pada penelitian ini. Algoritma *data mining* tersebut dievaluasi menggunakan ukuran standar regresi seperti MAPE dan RMSE untuk nantinya akan dibandingkan performanya. Dengan demikian, diharapkan operator dapat menentukan suhu dan durasi roasting yang optimal agar hasil roasting yang dihasilkan baik, sehingga kualitas produksi dari pengolahan biji kakao memiliki kualitas yang tinggi.

II. METODE

A. Dataset

Dataset yang digunakan adalah dataset pemantauan produksi dari tim laboratorium UGM *Cocoa Teaching and Learning Industry* serta data suhu *roaster* dari MindSphere mulai bulan April 2021 sampai bulan Juli 2022. Terdapat 43 baris data pada dataset tersebut. Data yang digunakan mempunyai beberapa variabel sebagai berikut:

TABLE I. KAMUS DATA PENELITIAN

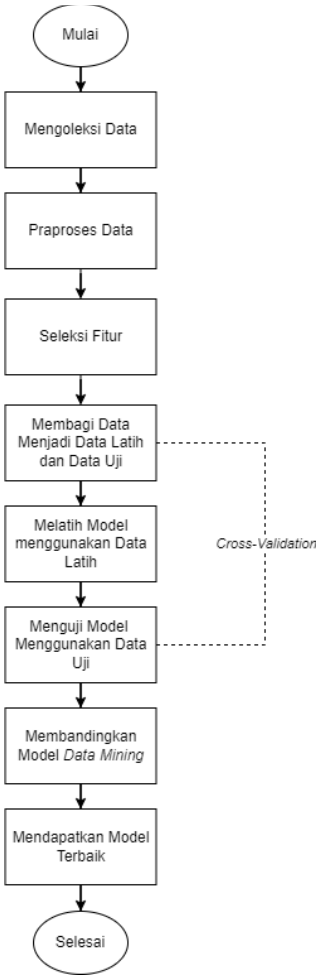
Kolom	Keterangan
nibs_capacity	kapasitas nibs (kg)
solution_load	jumlah air yang dibutuhkan (L)
beans_source	wilayah kebun dari biji tersebut
is_alkalized	alkilasi atau tidak
product_type	tipe produk
durasi_roasting	durasi *roasting*
suhu	suhu pengaturan
pH_0	pH awal biji kakao
pH_N	pH akhir biji kakao
moist_0	kadar air (%) awal biji kakao
moist_N	kadar air (%) akhir biji kakao

Pada penelitian ini, model yang dikembangkan memiliki variabel dependen yaitu suhu dan variabel independent yaitu variabel lainnya selain variabel suhu.

B. Alur Metodologis

Terdapat beberapa tahapan yang ditempuh pada penelitian ini, yaitu:

- Mengoleksi Data
- Pra Proses Data
- Seleksi Fitur
- Membagi Data menjadi Data Latih dan Data Uji
- Melatih Model menggunakan Data Latih
- Menguji Model dengan Data Uji
- Membandingkan Model Data Mining
- Mendapatkan Model yang Terbaik



Gambar 1. Alur Metodologis

C. Pra Proses Data

Pra proses data dilakukan untuk membersihkan data, serta untuk memastikan data siap atau layak di-inputkan ke dalam model. Beberapa tahap yang dilakukan pada pra proses data pada penelitian ini adalah sebagai berikut:

1) One-Hot Encoding

One-Hot Encoding digunakan untuk kolom kategorik, yaitu beans_source dan product_type. *One-Hot Encoding* adalah skema pengkodean yang paling banyak digunakan. *One-Hot Encoding* membandingkan setiap tingkat variabel kategori dengan tingkat referensi tetap. Satu hot encoding mengubah satu variabel dengan n observasi dan d nilai berbeda, menjadi d variabel biner dengan masing-masing n observasi. Setiap pengamatan menunjukkan ada (1) atau tidak adanya (0) dari variabel biner dikotomis [6].

2) MinMax Normalization

MinMax Normalization digunakan untuk melakukan normalisasi pada data sehingga memiliki skala yang sama (0 sampai 1), hal ini diharapkan mengurangi bias pada pemodelan. *MinMax Normalization* adalah metode normalisasi menggunakan transformasi linear pada data asli sehingga menghasilkan data baru yang variasinya seimbang antara satu kolom dengan kolom lainnya [7]. *MinMax Normalization* dapat ditulis dengan perumusan sebagaimana berikut :

$$X_{new} = \frac{X_{old} - Min(X_{old})}{Max(X_{old}) - Min(X_{old})} \quad (1)$$

D. Seleksi Fitur

Seleksi fitur adalah salah satu tahapan dalam proses d*ata mining* untuk memilih variabel yang relevan dengan variabel target dan membuang variabel yang tidak diperlukan atau redundan. Tujuan dari seleksi fitur adalah untuk meningkatkan performa, mempercepat waktu, dan mengurangi beban pada proses prediksi. Terdapat 3 metode populer yang digunakan dalam seleksi fitur yaitu *embedded, filter, dan wrapper* [8]. Seleksi fitur yang digunakan pada penelitian ini adalah seleksi fitur berbasis filter. Teknik seleksi fitur berbasis filter menggunakan metode statistik seperti kemiripan, ketergantungan, informasi, jarak untuk menunjukkan ketergantungan atau korelasi penting antara variabel input dan target [9]. Apabila variabel target numerik dan variabel input numerik, korelasi pearson dapat digunakan. Jika variabel target numerik dan variabel input kategorik maka dapat digunakan ANOVA F-test [10].

E. Support Vector Regression (SVR)

Model pertama yang digunakan pada penelitian ini adalah SVR. SVR adalah penerapan algoritma *Support Vector Machine* untuk kasus regresi. Dalam kasus regresi, output adalah bilangan real atau kontinu. SVR adalah metode yang dapat mengatasi *over-fitting*. Sehingga akan menghasilkan kinerja yang baik. Misalnya N adalah data latih (X, y) dengan menggunakan SVR, pengguna dapat menentukan fungsi $f(X)$. Fungsi tersebut memiliki deviasi terbesar E dari target sebenarnya untuk semua data pelatihan. Jika nilai E sama dengan 0 maka regresi dianggap sempurna. SVR disini bertujuan untuk menemukan fungsi regresi $f(X)$ yang dapat mendekati output ke target aktual, dengan toleransi kesalahan E dan kompleksitas minimal. Fungsi regresi $f(X)$dapat dituliskan dengan :

$$f(X) = w^T \varphi(X) + B \tag{2}$$

$\varphi(X)$ menunjukkan suatu titik dalam ruang fitur berdimensi lebih tinggi dan hasil pemetaan input vektor X dalam ruang fitur berdimensi lebih rendah. Koefisien w dan b akan diestimasi menggunakan persamaan [11] :

$$min \frac{1}{2}\|w\|^2 + C\frac{1}{N}\sum_{i=1}^{N} L_E(y_i, f(X_i)) \tag{3}$$

$$y_i - w\varphi(X_i) - b \leq E \tag{4}$$

$$w\varphi(X_i) - y_i + b \leq E, \qquad i = 1,2,3, \dots, N$$

Dimana,

$$L_E(y_i, f(X_i)) = \begin{cases} |y_i - f(X_i)| - E \\ |y_i - f(X_i)| \end{cases} \tag{5}$$

Pada umumnya, terdapat 3 kernel yang digunakan pada SVR atau SVM. Berikut adalah 3 kernel yang biasa digunakan pada SVR [11] :

- Linear Kernel

$$k(x, y) = x^T y + C \tag{6}$$

- Polynomial Kernel

$$k(x, y) = (ax^T y + C)^d \tag{7}$$

- Radial Basis Function (RBF) Kernel

$$k(x, y) = \exp(-\gamma\|x - y\|^2)^d \tag{8}$$

Langkah pertama yang dilakukan pada pemodelan menggunakan model SVR adalah melakukan *hyperparameter tuning* menggunakan metode *grid search* untuk mencari parameter SVM yang optimal. *Grid search* menguji semua kombinasi dari *hyperparameter* yang diberikan pada konfigurasi model *machine learning* [12]. *Grid search* membagi rentang parameter yang akan dioptimalkan ke dalam grid dan melintasi semua titik untuk mendapatkan parameter yang optimal. *Grid search* mengoptimalkan parameter SVM (dalam kasus ini SVR) menggunakan teknik validasi silang sebagai metrik kinerja [13]. *Grid search* dapat diimplementasikan menggunakan *library* sklearn dengan menggunakan perintah GridSearchCV. Beberapa keuntungan menggunakan grid *search* sebagai metode optimasi adalah penerapannya mudah, dapat menemukan λ yang jauh lebih baik daripada pengoptimalan sekuensial manual, dan keandalan dan dimensinya yang rendah [13].

F. Multiple Linear Regression (MLR)

Analisis regresi digunakan untuk menentukan korelasi antara dua atau lebih variabel yang mempunyai hubungan sebab-akibat. Selain itu analisis regresi juga dapat digunakan sebagai formula untuk memprediksi pada topik terkait menggunakan hubungan antar variabel. Regresi yang menggunakan satu variabel independen disebut regresi univariat, sementara regresi yang menggunakan lebih dari satu variabel independen dusebut regresi multivariat atau *multiple*. Dalam analisis regresi multivariat, dilakukan upaya untuk memperhitungkan variasi variabel independen dalam variabel dependen secara sinkron. Formula dari analisis regresi multivariat dapat dituliskan sebagai :

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{8}$$

Dengan y adalah variabel dependen, X adalah variabel independen, β adalah parameter, dan ε adalah error [14].

G. Extreme Learning Machine (ELM)

ELM adalah *Artificial Neural Network* berbasis *least-square* dan termasuk kedalam *single-layer feed-forward* yang dapat digunakan pada kasus klasifikasi maupun kasus regresi. Jumlah *neuron* atau *nod*e pada *hidden layer* berjumlah besar. ELM dapat diekspresikan sebagai :

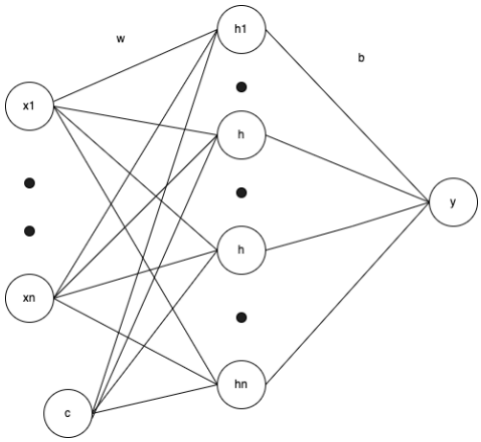
$$e_j = \sum_{i=1}^{H} b_i f(w_i, c_i, x_j), \quad j = 1,2, \dots, N \tag{9}$$

H adalah jumlah node pada hidden layer, b adalah bobot pada output layer, w adalah bobot pada input layer, dan c adalah bias, serta x adalah data input, kemudian N adalah banyaknya data input. $f(w_i, c_i, x_j)$ merupakan fungsi aktivasi. Pada algoritma ELM biasa, nilai w dibuat random berdasarkan distribusi probabilitas. Nilai b yang merupakan bobot pada output layer dapat dicari menggunakan persamaan :

$$b = A^+ Y \tag{10}$$

Dengan Y adalah vektor nilai aktual dan A^+ adalah *pseudo-invers* dari matriks A. Dimana matriks A adalah matriks hasil dari output layer yang komponennya berisi nilai $f(w_i, c_i, x_j)$

[15]. Setelah kita mendapatkan w dan b maka w dan b tersebut bisa diujikan pada data uji. Arsitektur ELM dari penelitian ini adalah sebagai berikut :



Gambar 2. Arsitektur ELM

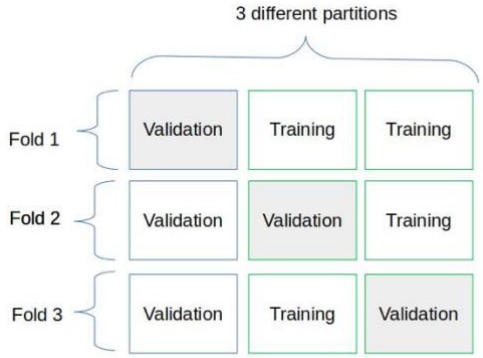
x adalah *input layer*, w adalah bobot input, c adalah bias, h adalah *hidden layer* yang komponennya adalah hasil perhitungan dari fungsi aktivasi ReLU $f(w_i, c_i, x_j)$ dan y adalah *output layer*.

Pada penelitian ini, ELM akan dioptimasi performanya menggunakan *Particle Swarm Optimization* (PSO). PSO termasuk di antara optimasi skohastik. Algoritma PSO mempekerjakan segerombolan partikel yang melintasi ruang pencarian multidimensi untuk mencari nilai optimal. Setiap partikel adalah solusi potensial dan dipengaruhi oleh pengalaman tetangganya serta dirinya sendiri [16].

PSO digunakan untuk mencari nilai bobot input dan bias pada ELM. Fungsi objektif yang digunakan adalah RMSE. PSO-ELM akan terus beriterasi agar menemukan nilai RMSE yang minimal. Pada kondisi tersebutlah ELM berhasil dioptimasi bobot dan biasnya sehingga diharapkan akan memiliki performa yang lebih baik dibanding dengan menginisiasi bobot dan bias dengan nilai random.

H. Ukuran Evaluasi

Model atau algoritma data *mining* akan dilihat performanya menggunakan skema 3-*fold cross-validation*. *Cross-Validation* atau validasi silang adalah metode statistik untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua segmen: satu digunakan untuk mempelajari atau melatih model dan yang lainnya digunakan untuk memvalidasi model. Dalam validasi silang, set pelatihan dan validasi harus saling silang dalam putaran berturut-turut sehingga setiap titik data memiliki peluang untuk divalidasi. Bentuk dasar dari cross-validation adalah *k-fold cross-validation* [17]. Skema dari validasi silang ini dapat dilihat pada gambar berikut ini:



Gambar 3. *Skema 3-Fold Cross-Validation*

Metrics atau ukuran evaluasi yang digunakan adalah MAPE dan RMSE. MAPE secara matematiks dapat ditulis sebagai:

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{A_t - F_t}{A_t}\right| \quad (11)$$

N adalah jumlah data yang diprediksi, A_t adalah nilai aktual dan F_t adalah nilai prediksi. Persamaan (11) harus dikalikan 100 untuk menjadi persentase. MAPE memiliki skala independen dan mudah ditafsirkan, yang membuatnya populer di kalangan praktisi industri [18]. Sementara RMSE secara matematis dapat ditulis sebagai:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - F_i)^2} \quad (12)$$

Dengan N adalah banyaknya data, A_i adalah nilai aktual dan F_i adalah nilai prediksi. RMSE adalah ukuran akurasi yang baik, tetapi hanya untuk membandingkan kesalahan peramalan dari model yang berbeda atau konfigurasi model untuk variabel tertentu [19].

III. HASIL DAN PEMBAHASAN

A. Seleksi Fitur

Untuk variabel numerik, seleksi fitur dilakukan menggunakan metode statistik pearson sementara untuk variabel kategorik, seleksi fitur dilakukan menggunakan metode statistika ANOVA F-score, pada kedua metode tersebut dilakukan uji hipotesis dengan melihat p-value. Penelitian ini menggunakan tingkat kepercayaan 0.01 sehingga apabila p-value < 0.01 maka hipotesis 0 ditolak dan hipotesis 1 diterima. Berikut adalah pernyataan dari hipotesis 0 dan hipotesis 1 :

- H0 : variabel bebas x tidak memiliki hubungan dengan variabel target y (suhu)
- H1 : variabel bebas x memiliki hubungan dengan variabel target y (suhu)

TABLE II. HASIL UJI STATISTIK VARIABEL NUMERIK

Variabel	r_suhu	pvalue_suhu
nibs_capacity	-0.5089	0.0009
solution_load	-0.2962	0.0671
pH_0	0.0392	0.8125
pH_N	0.1251	0.4480
delta_pH	-0.1771	0.2807

moist_0	-0.0248	0.8809
moist_N	-0.5860	0.0001
delta_moist	0.2155	0.1876

Berdasarkan uji statistik variabel numerik tersebut, untuk variabel dependen y yaitu suhu, terdapat 2 variabel independen yang secara signifikan berhubungan dengan variabel dependen suhu yaitu nibs_capacity dan moist_N. Dengan demikian, **variabel numerik yang akan dipilih** pada pemodelan menggunakan algoritma *data mining* adalah **nibs_capacity dan moist_N serta durasi_roasting**, karena penelitian ini juga digunakan untuk mengoptimalisasi suhu dan durasi_roasting**.** untuk mengoptimalisasi durasi *roasting* dan suhu *roasting*, maka durasi dan suhu dihubungkan pada model

TABLE III. HASIL UJI STATISTIK VARIABEL KATEGORIK

Variabel	fscore_suhu	pvalue_suhu
beans_source	29.522289	0.000004
product_type	0.512261	0.478655
is_alkalized	6.490727	0.015131

Berdasarkan uji statistik pada variabel kategorik, didapati bahwa terdapat 1 variabel independen yang secara signifikan mempunyai hubungan dengan variabel dependen suhu yaitu beans_source. Dengan demikian, **variabel kategorik yang akan digunakan** pada tahap pemodelan adalah variabel **beans_source** saja**.** Dari sekian banyak variabel yang ada pada data yang tersedia, variabel independen yang mempunyai tingkat signifikansi yang tinggi dengan variabel suhu adalah sebagai berikut:

- nibs_capacity
- moist_N
- durasi_roasting (untuk mengoptimalisasi durasi *roasting* dan suhu *roasting*, maka durasi dan suhu dihubungkan pada model)
- beans_source

Variabel tersebut akan dijadikan variabel indpenden pada model dan yang akan bertindak sebagai variabel dependen atau variable target adalah suhu.

B. Pemodelan menggunakan SVR

Tahap pertama yang dilakukan adalah mencari *hyperparameter* optimal menggunakan *grid search*. *Hyperparameter* yang dicari adalah C, gamma, epsilon, dan kernel. Berikut adalah *range* atau kumpulan nilai *hyperparameter* yang akan dicari nilai optimalnya :

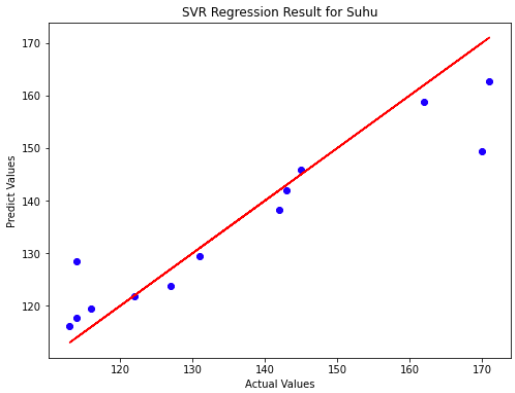
parameter = {'C': [0.1, 0.001, 1, 10, 12, 14, 16, 18, 20, 22],
'gamma': [0.001, 0.01, 0.1, 1, 2, 5],
'epsilon': [0.001, 0.01, 0.1, 1, 2, 4],
'kernel': ("rbf", "poly", "linear")}

Setelah dilakukan pencarian *hyperparameter* yang optimal menggunakan *grid search* dengan 3-fold validasi silang, didapat *hyperparameter* yang optimal sebagai berikut:

{'C': 22, 'epsilon': 1, 'gamma': 1, 'kernel': 'rbf'}

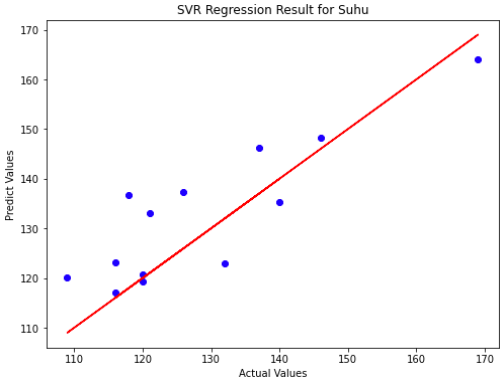
Skor rata-rata RMSE terbaik yang didapat dengan parameter tersebut sebesar 9.17 dan untuk skor rata-rata MAPE sebesar 4.76%. Berikut adalah perbandingan nilai prediksi dan aktual dari model tersebut untuk setiap fold-nya :

- Fold 1



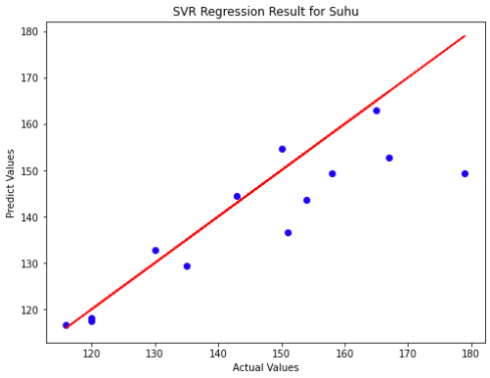
Gambar 4. Hasil Pengujian SVR pada Fold 1

- Fold 2



Gambar 5. Hasil Pengujian SVR pada Fold 2

- Fold 3



Gambar 6. Hasil Pengujian SVR pada Fold 3

Gambar di atas menunjukan perbandingan nilai prediksi dan aktual pada data tes untuk setiap fold pada pemodelan suhu menggunakan SVR. Untuk fold pertama, didapat nilai MAPE sebesar 3.71% dan nilai RMSE sebesar 7.72, pada fold kedua didapat nilai MAPE sebesar 5.73% dan nilai RMSE sebesar 8.85, dan pada fold ketiga didapat nilai MAPE sebesar 4.84% dan nilai RMSE sebesar 10.94. Apabila dirata-ratakan, pada pemodelan menggunakan SVR di dapat rata-rata MAPE sebesar 4.76 (1.01) dan RMSE

sebesar 9.17 (1.63). Berikut adalah tabel lengkap dari hasil pengujian model tersebut:

TABLE IV. HASIL PENGUJIAN MODEL SVR

SVR		
Ukuran	**MAPE**	**RMSE**
Fold 1	3.71	7.72
Fold 2	5.73	8.85
Fold 3	4.84	10.94
Rata-rata	**4.76**	**9.17**
Std Dev	**1.01**	**1.63**

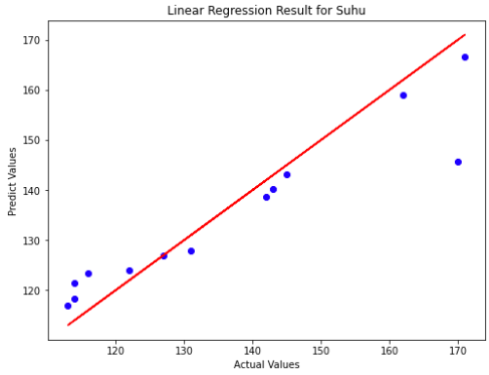
C. Pemodelan menggunakan MLR

Pada penelitian ini, MLR diimplementasikan menggunakan metode *least-square* pada pustaka scikit-learn menggunakan fungsi LinearRegression(). Fungsi yang dibentuk oleh model tersebut adalah sebagai berikut:

$$suhu = 153.84 - 6.36 \cdot nibs\_capacity - 15.99 \cdot moist\_N - 16.97 \cdot durasi\_roasting + categoric(source) \quad (13)$$

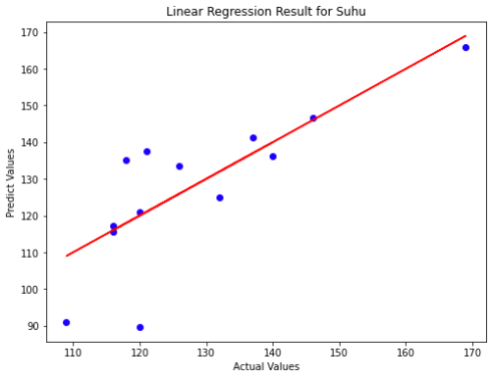
Dengan *categoric* bernilai -18.26 untuk beans *source* ADACHI, 0.10 untuk bean source BONDO, -19.46 untuk beans source GOBEL, -17.18 untuk beans source HMGN, -13.08 untuk beans source papua, 2.34 untuk beans source PGL, 21.89 untuk beans source SGU, 36.42 untuk beans source siklon, dan 7.23 untuk beans source UNID KOREA. Berikut adalah hasil pengujian dari model tersebut untuk setiap fold-nya :

- Fold 1



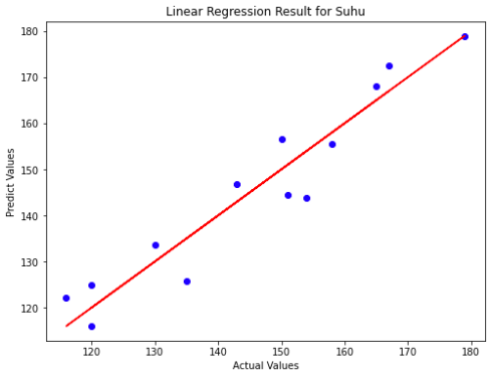
Gambar 7. Hasil Pengujian MLR pada Fold 1

- Fold 2



Gambar 8. Hasil Pengujian MLR pada Fold 2

- Fold 3



Gambar 9. Hasil Pengujian MLR pada Fold 3

Gambar 7 sampai 9 memperlihatkan perbandingan antara nilai aktual dan nilai prediksi dari pemodelan suhu menggunakan MLR. Untuk fold pertama, didapat nilai MAPE sebesar 3.72% dan nilai RMSE sebesar 7.83, pada fold kedua didapat nilai MAPE sebesar 7.06% dan nilai RMSE sebesar 12.32, dan pada fold ketiga didapat nilai MAPE sebesar 3.62% dan nilai RMSE sebesar 5.73. Apabila dirata-ratakan, pada pemodelan menggunakan SVR di dapat rata-rata MAPE sebesar 4.80% (1.96) dan RMSE sebesar 8.63 (3.37). Berikut adalah tabel lengkap dari hasil pengujiannya :

TABLE V. HASIL PENGUJIAN MODEL MLR

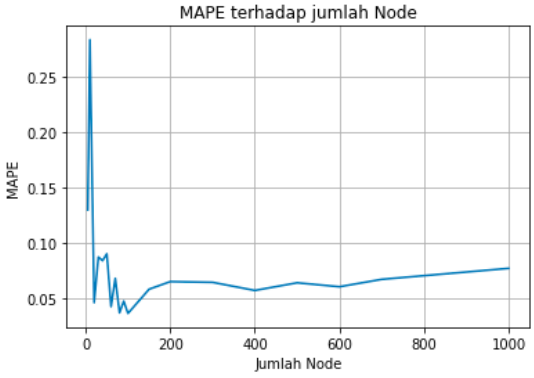
MLR		
Ukuran	**MAPE**	**RMSE**
Fold 1	3.72	7.83
Fold 2	7.06	12.32
Fold 3	3.62	5.73
Rata-rata	**4.8**	**8.63**
Std Dev	**1.96**	**3.37**

Apabila dibandingkan dengan model SVR nilai rata-rata MAPE pada pemodelan suhu menggunakan MLR lebih besar, namun tidak terlalu signifikan perbedaannya. Kemudian untuk nilai rata-rata RMSE, model MLR memiliki nilai yang lebih kecil dibandingkan dengan model SVR. Walaupun model MLR memiliki nilai error rata-rata yang

lebih kecil dibandingkan denghan model SVR, namun model MLR memiliki standar deviasi yang lebih besar. Hal ini menunjukan bahwa model SVR lebih stabil untuk memodelkan suhu pada setiap foldnya dibandingkan dengan model MLR. Terlihat pada fold 2, model MLR kesulitan untuk memodelkan suhu karena memiliki error yang jauh lebih besar dibanding model SVR.

D. Pemodelan menggunakan ELM

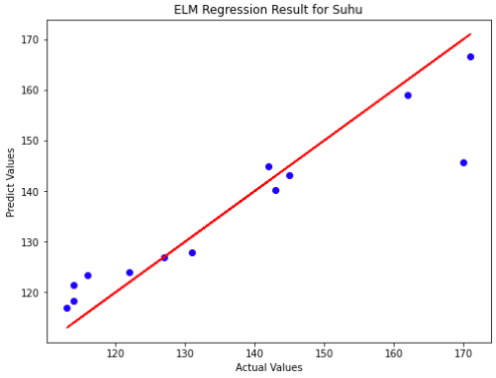
Algoritma ketiga yang diuji adalah ELM. Sama seperti algoritma sebelumnya, ELM diuji menggunakan skema *3-fold*. Pada penelitian ini ELM yang digunakan memiliki 1 *hidden layer* dengan 100 *nodes*. 100 nodes dipilih karena berdasarkan hasil percobaan pada berbagai jumlah *node*s, jumlah *nodes* 100 memiliki MAPE terkecil dengan 3.69%.



Gambar 10. Nilai MAPE pada Percobaan Jumlah Node

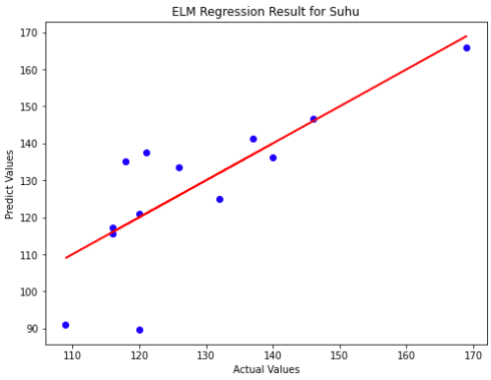
Setelah mendapatkan jumlah *node* dengan MAPE terkecil, selanjutnya akan dilakukan pengujian pada data tes menggunakan skema 3-*fold cross-validation.* Berikut adalah hasil dari pengujian algoritma ELM dengan skema 3-fold untuk setiap foldnya :

- Fold 1



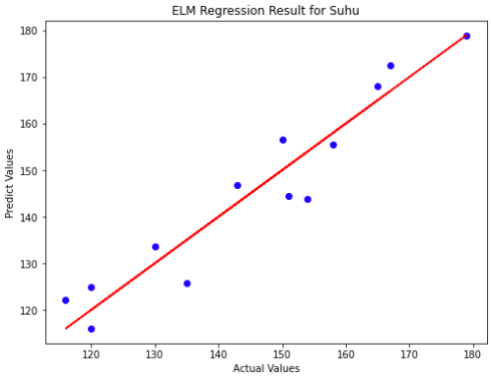
Gambar 11. Hasil Pengujian ELM pada Fold 1

- Fold 2



Gambar 12. Hasil Pengujian ELM pada Fold 2

- Fold 3



Gambar 13. Hasil Pengujian ELM pada Fold 3

Pada gambar 11 sampai 13 dapat dilihat perbandingan antara nilai aktual dan nilai prediksi dari pemodelan suhu menggunakan ELM. Untuk fold pertama, didapat nilai MAPE sebesar 3.69% dan nilai RMSE sebesar 7.81, pada fold kedua didapat nilai MAPE sebesar 7.06% dan nilai RMSE sebesar 12.32, dan pada fold ketiga didapat nilai MAPE sebesar 3.62% dan nilai RMSE sebesar 5.73. Apabila dirata-ratakan, pada pemodelan menggunakan SVR di dapat rata-rata MAPE sebesar 4.79% (1.97) dan RMSE sebesar 8.62 (3.37). Berikut adalah tabel lengkap dari hasil pengujiannya :

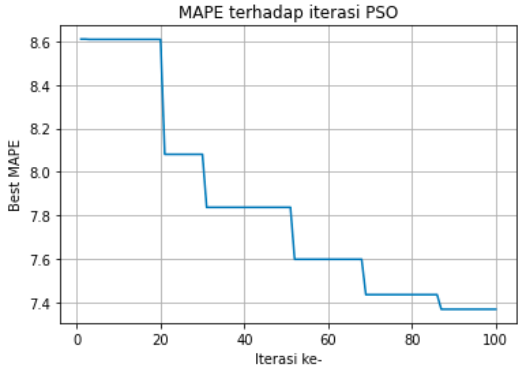
TABLE VI. HASIL PENGUJIAN MODEL ELM

ELM		
Ukuran	**MAPE**	**RMSE**
Fold 1	3.69	7.81
Fold 2	7.06	12.32
Fold 3	3.62	5.73
Rata-rata	**4.79**	**8.62**
Std Dev	**1.97**	**3.37**

Hasil pemodelan menggunakan ELM memiliki performa yang mirip dengan pemodelan menggunakan MLR. Model ELM sedikit lebih baik memprediksi suhu pada fold 1 dibanding dengan model MLR. Sama seperti model MLR, model ELM kesulitan untuk memodelkan suhu pada fold 2.

E. Pemodelan menggunakan PSO-ELM

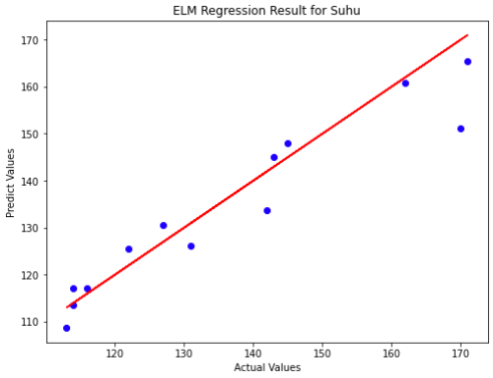
Algoritma terakhir yang diujikan pada penelitian ini adalah PSO-ELM. PSO bertindak sebagai algortima optimasi dari bobot awal dan bias yang nantinya akan diikutkan pada perhitungan dalam model ELM. Fungsi objektif yang digunakan pada PSO adalah nilai RMSE, sehingga PSO dikatakan berhasil atau optimal apabila memiliki nilai MAPE yang paling kecil. PSO akan melakukan iterasi sebanyak 100 kali dengan 10 kandidat bobot awal dan bias. Berikut adalah nilai MAPE untuk setiap iterasi yang dilakukan oleh algoritma PSO :



Gambar 14. Nilai RMSE terhadap iterasi PSO

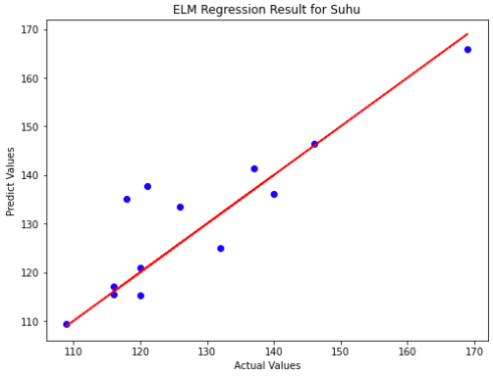
Didapat bahwa sampai iterasi ke-100 terdapat improvisasi nilai RMSE rata-rata menjadi 7.37. Setelah mendapat bobot awal dan bias yang optimal, algoritma tersebut diujikan menggunakan skema *3-fold*. Berikut adalah hasil pengujian dari algoritma PSO-ELM:

- Fold 1



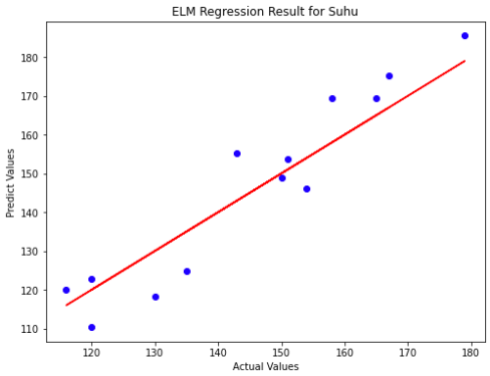
Gambar 15. Hasil Pengujian PSO-ELM pada Fold 1

- Fold 2



Gambar 16. Hasil Pengujian PSO-ELM pada Fold 2

- Fold 3



Gambar 17. Hasil Pengujian PSO-ELM pada Fold 3

Gambar 15 sampai 17 memperlihatkan perbandingan nilai aktual dan prediksi pada pemodelan suhu menggunakan PSO-ELM dengan skema 3-fold. Untuk fold pertama, didapat nilai MAPE sebesar 3.21% dan nilai RMSE sebesar 6.48, pada fold kedua didapat nilai MAPE sebesar 4.91% dan nilai RMSE sebesar 7.59, dan pada fold ketiga didapat nilai MAPE sebesar 5.01% dan nilai RMSE sebesar 8.03. Apabila dirata-ratakan, pada pemodelan menggunakan SVR di dapat rata-rata MAPE sebesar 4.14% (0.90) dan RMSE sebesar 7.37 (0.80). Berikut adalah tabel lengkap dari hasil pengujiannya:

TABLE VII. HASIL PENGUJIAN MODEL PSO-ELM

PSO-ELM		
Ukuran	**MAPE**	**RMSE**
Fold 1	3.21	6.48
Fold 2	4.19	7.59
Fold 3	5.01	8.03
Rata-rata	**4.14**	**7.37**
Std Dev	**0.90**	**0.80**

Hasil pemodelan menggunakan PSO-ELM lebih baik apabila dibandingkan dengan model sebelumnya (SVR, MLR, dan ELM). Untuk setiap foldnya, nilai error dari model PSO-ELM lebih rendah dari model sebelumnya kecuali pada fold 3. Nilai rata-rata dan standar deviasi MAPE maupun RMSE pada model PSO-ELM lebih kecil dibandingkan dengan model sebelumnya, hal ini menunjukan bahwa model PSO-ELM memiliki performa yang baik sekaligus memiliki stabilitas yang lebih baik dibandingkan dengan model-model sebelumnya.

F. Perbandingan Algoritma Data Mining

TABLE VIII. PERFORMA SETIAP ALGORITMA DATA MINING

Model	**MAPE**	**RMSE**
SVR	4.76	9.17
MLR	4.81	8.63
ELM	4.80	8.62
PSO-ELM	4.14	7.37

Menurut tabel di atas dapat dilihat bahwa ELM secara general memiliki performa yang lebih baik dibandingkan SVR dan MLR apabila ditinjau dari nilai RMSE-nya. Selain itu, penambahan PSO untuk mencari bobot awal dan bias yang optimal pada ELM juga dapat meningkatkan performa pada model ELM itu sendiri. Dibuktikan dengan improvisasi MAPE 4.80% menjadi 4.14% dan improvisasi RMSE 8.62 menjadi 7.37. Algoritma *data mining* terbaik pada penelitian ini adalah PSO-ELM yang memiliki nilai rata-rata MAPE 4.14% dan nilai rata-rata RMSE 7.37.

IV. KESIMPULAN

Algoritma data *mining* yang digunakan untuk merekomendasikan suhu dan durasi *roasting* pada proses *roasting* biji kakao dapat dirancang. Adapun algoritma *data mining* yang digunakan adalah SVM, MLR, ELM, dan PSO-ELM. Algoritma tersebut diujikan menggunakan skema *3-fold*. Berdasarkan MAPE dan RMSE, ELM dan PSO-ELM memiliki performa terbaik, diikuti SVM, kemudian MLR. Penambahan PSO untuk mencari bobot awal dan bias pada ELM juga dapat meningkatkan performa dari algoritma ELM. Algoritma terbaik yang dapat digunakan untuk merekomendasikan suhu yang optimal berdasarkan durasi *roasting* tertentu pada proses *roasting* biji kakao adalah PSO-ELM yang memiliki nilai rata-rata MAPE 4.14% dan nilai rata-rata RMSE 7.37. Fitur atau variabel yang memiliki pengaruh besar terhadap suhu adalah nibs_capacity, moist_N, durasi_roasting, dan beans_source.

Penelitian lanjutan dapat dilakukan untuk melakukan uji algoritma dan beberapa perlakuan pada data seperti misalnya melakukan *oversampling* dan menambah metode *hyperparameter tuning* atau bobot. Kemudian, penelitian lanjutan juga dapat dilakukan untuk melakukan eksplorasi terhadap metode seleksi fitur yang lain seperti *wrapped*, *embedded*, atau metode filter yang lainnya. Selain itu, penelitian lanjutan juga dapat dilakukan untuk melakukan eksplorasi terhadap algoritma yang belum diujikan pada penelitian ini terutama pengujian pada *tree-based algorithm*.

REFERENCES

[1] V. A. Dihni. "5 Negara Penghasil Kakao Terbesar, Indonesia Urutan Berapa?," *Databoks*, 04 October 2021. [Online]. Available: https://databoks.katadata.co.id/datapublish/2021/10/04/5-negara-penghasil-kakao-terbesar-indonesia-urutan-berapa. [Accessed: 26 March 2023].

[2] A. Mahmudan. "Ekspor Kakao Indonesia Turun 2,92% pada 2021," *DataIndonesia.id*, 20 May 2022. [Online]. Available: https://dataindonesia.id/sektor-riil/detail/ekspor-kakao-indonesia-turun-292-pada-2021 . [Accessed: 26 March 2023].

[3] Direktorat Jendral Perkebunan, "Luas Areal Kakao Menurut Provinsi di Indonesia, 2017 - 2021," *Direktorat Jendral Perkebunan*, 2021. [Online]. Available: https://www.pertanian.go.id/home/index.php?show=repo&fileNum=224. [Accessed: 26 March 2023].

[4] Ditpui. "Proses Pengolahan Cokelat di Tingkat UGM Cocoa Teaching and Learning Industry," *Direktorat Pengembangan Bisnis dan Inkubasi UGM*, 31 October 2020. . [Online]. Available: https://ditpui.ugm.ac.id/proses-pengolahan-cokelat-di-tingkat-ugm-cocoa-teaching-and-learning-industry/ . [Accessed: 26 March 2023].

[5] S. Wijanarti, A. M. Rahmatika, and R. Hardiyanti, "Pengaruh lama penyangraian Manual Terhadap Karakteristik Kakao Bubuk," *Jurnal Nasional Teknologi Terapan (JNTT)*, vol. 2, no. 2, p. 212, 2019.

[6] K. Potdar, T. S., and C. D., "A comparative study of categorical variable encoding techniques for neural network classifiers," *International Journal of Computer Applications*, vol. 175, no. 4, pp. 7–9, 2017.

[7] O. A. Akanbi, I. S. Amiri, and E. Fazeldehkordi, *A machine learning approach to phishing detection and Defense*. Amsterdam: Elsevier, 2015.

[8] F. F. Firdaus, H. A. Nugroho, and I. Soesanti, "A review of feature selection and classification approaches for heart disease prediction," *IJITEE (International Journal of Information Technology and Electrical Engineering)*, vol. 4, no. 3, p. 75, 2021.

[9] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection – A comparative study," *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, pp. 178–187, 2007.

[10] J. Brownlee, "How to choose a feature selection method for machine learning," *MachineLearningMastery.com*, 20 August 2020. [Online]. Available: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/. [Accessed: 27 March 2023].

[11] A. Shirzad, M. Tabesh, and R. Farmani, "Performance Comparison between Support Vector Regression and Artificial Neural Network for Prediction of Oil Palm Production," *Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)*, vol. 9, no. 1, pp. 1-8, 2016.

[12] D. M. Belete and M. D. Huchaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results," *International Journal of Computers and Applications*, vol. 44, no. 9, pp. 875–886, 2021.

[13] Sulistiana and M. A. Muslim, "Support Vector Machine (SVM) optimization using grid search and UNIGRAM to improve e-commerce review accuracy," *Journal of Soft Computing Exploration*, vol. 1, no. 1, 2020.

[14] G. K. Uyanık and N. Güler, "A study on multiple linear regression analysis," *Procedia - Social and Behavioral Sciences*, vol. 106, pp. 234–240, 2013.

[15] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[16] P. Godbole and Dr. M. Pathak, "Particle Swarm Optimization (PSO) Model and Its Application in ANN Controller", I*nternational Journal for Modern Trends in Science and Technology*, vol. 8, no.1, pp. 153-157, 2022.

[17] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," *Encyclopedia of Database Systems*, pp. 532–538, 2009.

[18] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *International Journal of Forecasting*, vol. 32, no. 3, pp. 669–679, 2016.

[19] D. Christie and S. P. Neill, "Measuring and observing the Ocean Renewable Energy Resource," *Comprehensive Renewable Energy*, pp. 149–175, 2022.