Jurnal Rekayasa Elektrika Vol. 18, No. 3, September 2022, hal. 175-184
ISSN. 1412-4785; e-ISSN. 2252-620X, Terakreditasi Dikbudristekdikti No. 105/KPT/2022
DOI: 10.17529/jre.v18i3.25903

Handling Missing Value dengan Pendekatan Regresi pada Dataset Akuakultur Berukuran Kecil

Ricky Afiful Maula[1], Agus Indra Gunawan[1], Bima Sena Bayu Dewantara[1], M. Udin Harun Al Rasyid[1], Setiawardhana[1], Ferry Astika Saputra[1], dan Junaedi Ispianto[2]
[1]Politeknik Elektronika Negeri Surabaya
Kampus PENS, Jalan Raya ITS Sukolilo, Surabaya 60111
[2]Asosiasi Tambak Intensif, Indonesia
Perumahan Puri Indah, Sepanjang, Sidoarjo 61257
e-mail: ricky@pasca.student.pens.ac.id

***Abstrak*—Budidaya udang sangat dipengaruhi kondisi kualitas air tambak. Petambak harus mengetahui tindakan tepat dalam mengatur kualitas air yang sesuai untuk kelangsungan hidup udang. Kondisi kualitas air dapat diketahui dengan pengukuran parameter-parameter tambak menggunakan berbagai sensor. Pemasangan sensor yang dilengkapi dengan modul *artificial intelligence* untuk memberitahu kondisi kualitas air merupakan tindakan tepat. Namun sensor tidak terlepas dari *error* sehingga berakibat tidak bisa mendapatkan data atau terjadi *missing* data. Pada kasus ini dilakukan pendekatan 5 parameter kualitas air tambak dari 13 parameter yang tersedia. Paper ini mengusulkan teknik untuk mendapatkan data hilang yang disebabkan oleh kesalahan sensor dan mencari model terbaik. Pendekatan sederhana bisa dilakukan seperti *Handling Missing Value* (HMV) yang umum digunakan yaitu *mean*, dengan *classifier K-Nearest Neighbors* (KNN) yang dioptimasi menggunakan *grid search*. Namun nilai akurasi teknik ini masih rendah, yaitu mencapai 0.739 pada *20-fold cross-validation*. Untuk lebih meningkatkan akurasi prediksi, dilakukan perhitungan dengan teknik-teknik lain, dan didapatkan bahwa *Linear Regression* (LR) dapat meningkatkan akurasi hingga 0.757 yang mengungguli pendekatan lain seperti pendekatan statistik mean 0.739, modus 0.716, median 0.734 serta pendekatan regresi KNN 0.742, *Lasso* 0.751, *Passive Aggressive Regressor* (PAR) 0.737, *Support Vector Regression* (SVR) 0.739, *Kernel Ridge* (KR) 0.731, dan *Stochastic Gradient Descent* (SGD) 0.734.**

Kata kunci: ***handling missing value, iterative imputation, algoritma regresi, akuakultur***

***Abstract*—Shrimp cultivation is strongly influenced by pond water quality conditions. Farmers must know the appropriate action in regulating water quality that is suitable for shrimp survival. The state of water quality can be understood by measuring pond parameters using various sensors. Installing sensors equipped with artificial intelligence modules to inform water quality conditions is the right action. However, the sensor cannot be separated from errors, so it results in not being able to get data or missing data. In this case, the approach of 5 parameters of pond water quality from 13 available parameters is carried out. This paper proposes a technique to obtain lost data caused by sensor error and looks for the best model. A simple approach can be taken, such as the Handling Missing Value (HMV) which is commonly used, namely the mean, with the K-Nearest Neighbors (KNN) classifier optimized using a grid search. However, the accuracy of this technique is still low, reaching 0.739 at 20-fold cross-validation. Calculations were carried out with other methods to further improve the prediction accuracy. It was found that Linear Regression (LR) can increase accuracy up to 0.757, which outperforms different approaches such as the statistical approach to mean 0.739, mode 0.716, median 0.734, and regression approach KNN 0.742, Lasso 0.751, Passive Aggressive Regressor (PAR) 0.737, Support Vector Regression (SVR) 0.739, Kernel Ridge (KR) 0.731, and Stochastic Gradient Descent (SGD) 0.734.**

Keywords: ***handling missing value, iterative imputation, regression algorithm, aquaculture***

I. PENDAHULUAN

Budidaya udang di Indonesia merupakan salah satu prioritas pengembangan budidaya perikanan di Indonesia untuk meningkatkan perekonomian nasional. Keberhasilan budidaya sangat dipengaruhi oleh teknik pembudidaya dalam memelihara tambaknya. Petambak perlu memahami kondisi kualitas air tambak yang mempengaruhi kelangsungan hidup udang. Kondisi tambak yang tidak tepat akan menyebabkan udang mudah terserang penyakit dan mati. Dengan memahami kualitas air tambak, tindakan atau penanganan yang tepat dapat diberikan untuk mencegah hal-hal yang tidak diinginkan terjadi [1].

Penerapan teknologi monitoring di sektor akuakultur telah meningkatkan pemantauan dalam memperoleh data yang beragam untuk membantu pembudidaya dalam menganalisa kondisi tambak [2]-[3]. Permasalahannya, mengetahui kondisi kualitas air tambak tidaklah mudah karena banyaknya parameter air. Kisaran nilai parameter

Received 13 May 2022; Revised 08 July 2022; Accepted 12 July 2022

air yang cenderung sama antara kondisi baik maupun buruk juga menjadi kendala dalam menentukan pengelolaan yang tepat. Penerapan algoritma *Machine Learning* (ML) juga sudah dilakukan dengan tujuan untuk identifikasi kondisi tambak yang tidak akurat karena kesalahan asumsi dalam menganalisa data kondisi baik dan buruk yang cenderung sama [4]. Proses kerja ML yaitu dengan memanfaatkan dataset yang sudah dicatat sebelumnya dalam melakukan pembelajaran dan membuat model klasifikasi akuakultur. Sementara itu, alat monitoring tidak selalu bisa mendapatkan semua data parameter air yang menimbulkan adanya data yang hilang atau bisa disebut data kosong. ML yang sangat bergantung pada dataset akan mengembangkan model yang salah apabila data kosong ini dibiarkan untuk melatih ML. Dengan ukuran dataset yang kecil, pembuangan sampel yang memiliki data kosong bukanlah tindakan tepat karena hanya sedikit data untuk melatih ML dan berpengaruh terhadap keakuratan analisis data.

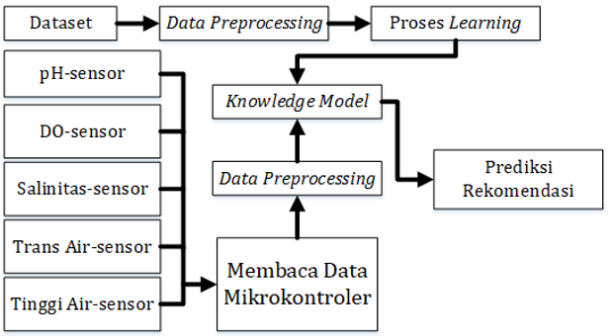
Untuk mengatasi masalah ini, teknik *Handling Missing Value* (HMV) dapat digunakan untuk mengisi data kosong dengan nilai yang sesuai dengan kondisi sampel tersebut. Terisinya data kosong membuat sampel dapat digunakan untuk melatih ML untuk mendapatkan model terlatih dengan performa terbaik. Model terlatih memprediksi keadaan tambak dari data parameter kualitas air yang didapatkan oleh alat monitoring. Jenis keputusan pembudidaya dalam memberikan perlakuan yang tepat bergantung pada hasil prediksi kualitas air tambak dari model terlatih.

Dalam penelitian ini, data yang digunakan untuk proses pemodelan adalah dataset budidaya yang dikumpulkan dari beberapa tambak udang di Bulukumba, Sulawesi Selatan, Indonesia. Data berisi beberapa parameter air yang diukur dan memiliki contoh yang cukup.

Beberapa parameter dalam dataset tidak digunakan dalam mengisi data kosong dan melatih ML. Hal ini karena menyesuaikan alat monitoring yang hanya dapat mengukur 5 parameter dalam dataset yaitu pH, *dissolved oxygen* (DO), salinitas, tinggi air, dan transparansi air. Algoritma ML yang digunakan adalah algoritma *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), *Naïve Bayes* (NB), *Gaussian Progress Classifier* (GPC) sebagai algoritma klasifikasi untuk memprediksi kualitas air tambak udang berdasarkan beberapa parameter air. Kemudian, algoritma optimasi *grid search* digunakan untuk *tuning* parameter ML yang berfungsi untuk menemukan nilai parameter terbaik dalam memberikan performa tertinggi. Sedangkan untuk teknik HMV menggunakan pendekatan algoritma regresi yaitu KNN, *Linear Regression* (LR), *Lasso*, *Passive Aggressive Regressor* (PAR), *Support Vector Regression* (SVR), *Kernel Ridge* (KR), *Stochastic Gradient Descent* (SGD) dan dibandingkan dengan teknik HMV statistik (konvensional) yang sering digunakan yaitu rata-rata (mean), nilai tengah (median), nilai yang sering muncul (modus).

II. Studi Pustaka

Dalam beberapa dekade terakhir, ML memiliki dampak yang luar biasa pada beberapa penelitian terkait pemodelan karakteristik air. Model hibrid *Long Short-Term Memory Network* (LSTM) dan *Gradient Boosting Decision Tree* (GBDT) telah dilakukan untuk memprediksi konsentrasi DO dalam budidaya yang dikembangkan oleh Huan dkk. [5]. Huan dkk. memprediksi DO dengan menggunakan temperatur air, temperatur udara, kelembapan, pH, kecepatan angin, dan tekanan udara. Model ini menunjukkan kemampuan prediksi yang lebih baik dari model kompleks lain seperti *Extreme Learning Machine*, *Back Propagation* NN, *Particle Swarm* LSSVM, dan LSTM dengan RMSE, MAE, dan MAPE secara urut bernilai 0.197, 0.299, dan 0.092. Dikatakan bahwa *tuning* GBDT dalam kasus prediksi DO dapat menghindari kelalaian *tuning parameter* dan dapat mengoptimalkan model. Sinshaw dkk. menjelaskan bahwa solusi lain untuk memantau kualitas air adalah prediksi menggunakan ML dan *Deep Learning* [6]. Dibandingkan dengan pengukuran manual secara metode statistik, penggunaan ML ini memiliki beberapa keunggulan dari biaya lebih rendah, efisien dalam hal waktu yang diperlukan untuk pengiriman dan pengumpulan data, memungkinkan prediksi dalam berbagai jenis sistem, dan memprediksi nilai yang diinginkan saat terdapat data menggangu. Pada kasus prediksi banyaknya Fosfor dan Nitrogen di danau US, Sinshaw dkk. menggunakan *Artificial Neural Network* (ANN) dalam membangun model dimana model pembelajarannya ditanamkan pada alat pengukuran danau regional. Performa prediksi model dibandingkan dengan *linear regression*. Dikatakan bahwa model ANN lebih unggul daripada model pembandingnya yaitu *linear regression*. Lebih lanjut, Samsudin dkk. juga mengembangkan ANN dan *Multiple Linear Regression* (MLR) pada kasus prediksi model indeks kualitas air laut di Muara Bakau. Dataset yang digunakan memiliki 13 parameter dari 6 pos pengawasan Malaysia Barat. Sebelum dataset dimodelkan, dataset diproses terlebih dahulu menggunakan *Spatially Discriminant Analysis* (SDA) dan terpilih 7 parameter yang akan dipelajari oleh model. Model terbaik yang terpilih adalah ANN karena nilai R^2 (0.9044) dan nilai validasi (0.7113) hasil *training* lebih tinggi daripada model MLR. Untuk studi ini, diketahui bahwa proses SDA sangat membantu model dalam meningkatkan akurasi. SDA memilih parameter kualitas air yang paling berpengaruh terhadap pembelajaran ML [7]. Dilain sisi Ahmed dkk. mengatakan bahwa pengukuran secara tradisional sangat bergantung pada kumpulan data dan sulit dalam mengelompokkan zona tiap kelas data. Berbeda dengan *artificial intelligence* yang mengarah pada struktur matematika fleksibel dan memiliki kemampuan untuk mengidentifikasi hubungan nonlinear kompleks sehingga mereka menggunakan beberapa ML jenis *supervised* dalam memprediksi kualitas air [8]. Mereka menerapkan model mereka pada empat parameter kualitas air. Mereka menemukan bahwa dengan



Gambar 1. Rancangan sistem alat monitoring

menggunakan *Gradient Boosting* dan *Regresi Polinomial*, prediksi kualitas air lebih tinggi dimana *perceptron* dengan banyak *layer* mengklasifikasikan kualitas air secara lebih efektif. Studi ini dilakukan pada kasus kualitas air dengan parameter dataset yang sedikit, serta performa model klasifikasi yang diusulkan akurasinya tidak lebih dari 75%.

Parameter kualitas air sulit untuk diidentifikasi apabila jumlah sampelnya sedikit dan hanya jika datasetnya tidak memiliki data kosong. Jika terdapat data kosong pada dataset, akan sangat buruk apabila sampel tersebut dibuang karena dapat terjadi bias informasi. Mayoritas penelitian-penelitian yang sudah dilakukan, dataset yang digunakan dalam kondisi utuh dan ini berbeda dengan keadaan di lapangan dimana dapat terjadi kemungkinan sensor tidak mengukur parameter karena *error*. Untuk itu dalam makalah ini, kami mengusulkan metode *Handling Missing Value* (HMV) regresi. HMV regresi mengganti nilai kosong dengan cara memprediksi parameter yang terdapat nilai kosong tersebut menggunakan algoritma regresi. Disini juga akan dilakukan pencarian model terbaik untuk dipilih dalam kasus kualitas air pada budidaya udang.

III. METODE/DESAIN

Penelitian ini bertujuan untuk mendapatkan model terbaik untuk klasifikasi kualitas air tambak budidaya udang pada dataset berukuran kecil yang memiliki data kosong dengan menggunakan teknik HMV pendekatan regresi yang tepat dalam mengisi data kosong tersebut. Oleh karena itu, dilakukan evaluasi performa pada berbagai macam tahap dan algoritma dimulai dari proses pengambilan data, pengolahan data hingga proses pemodelan data. Untuk mengetahui seberapa bagus teknik HMV pendekatan regresi, dilakukan perbandingan dengan teknik HMV statistik. Pada bab ini dijelaskan bagaimana penelitian dilakukan dan penjelasan secara detail pada setiap tahapan penelitian. Adapun proses akuisisi hingga pemodelan data dan validasi model terlatih dapat diringkas sesuai dengan blok diagram pada Gambar 1.

A. Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset budidaya udang yang diperoleh dari beberapa tambak udang di Bulukumba, Sulawesi Selatan, Indonesia.

Tabel 1. Dataset Akuakultur

No	Spesifikasi	
1.	Tugas	Klasifikasi
2.	Jumlah Sampel	174
3.	Tipe Data	Numerik dan kategorikal
4.	Missing Value	Ya
5.	Bidang	Akuakultur

Tabel 2. Data kosong pada dataset

No	Parameter	Jumlah Data Kosong
1.	pH	24
2.	Salinitas	48
3.	Transparansi Air	13
	Total	85

Dataset ini terdiri dari parameter yang mewakili kondisi tambak udang. Ada 174 sampel data dengan 14 fitur (13 *input* dan 1 *output*). Pencatatan kumpulan data bertujuan untuk mengklasifikasikan keadaan tambak berdasarkan 13 nilai *input*. Informasi detail dari dataset ini dapat dilihat pada Tabel 1.

Output dataset terdiri dari 2 kelas mewakili kondisi kualitas air (0 untuk buruk dan 1 untuk baik). Kelas 0 terdiri dari 84 sampel data, sedangkan jumlah sampel data untuk kelas 1 adalah 90. Semua parameter *input* diukur menggunakan alat ukur standar dan berdasarkan uji laboratorium. Diketahui terdapat data yang hilang/ data kosong (*missing value*) di sebagian sampel. Uraian data kosong di dataset serta rincian parameter air beserta deskripsinya ditunjukkan pada Tabel 2 dan Tabel 3.

B. Data Pre-processing

Data pre-processing dilakukan jika dataset memiliki data kosong dan *range* nilai data yang bervariasi antar parameter. Tujuan dari *data pre-processing* adalah untuk mengolah suatu dataset sebelum dataset tersebut dimodelkan. Proses ini akan meningkatkan akurasi dan mengurangi persentase *error* komputasi pemodelan. Metode yang dilakukan pada tahap *pre-processing* yaitu seleksi fitur, HMV, dan normalisasi.

Seleksi Fitur

Beberapa parameter dataset tidak diperlukan dalam menentukan prediksi karena tidak konsistensinya nilai pada parameter atau adanya nilai yang berulang. Seleksi fitur merupakan proses pemilihan parameter yang optimal menurut kriteria tertentu guna untuk meningkatkan akurasi prediksi model terlatih [9]. Sesuai yang telah dijelaskan sebelumnya, bahwa dari 13 parameter dataset digunakan 5 parameter untuk mengembangkan model ML klasifikasi dalam memprediksi kondisi kolam. Hal ini menyesuaikan

Tabel 3. Parameter kondisi air

No	Parameter	Penjelasan
		Atribut Input
1.	pH	Kadar pH dalam air tambak
2.	Alkalinitas	Jumlah HCO3 (Bikarbonat) dan CO3 (Karbonat) (ppm)
3.	DO	Jumlah Oksigen terlarut dalam air (ppm)
4.	TOM	*Total organic matter* dalam air (ppm)
5.	NH_4	Jumlah ammonium dalam air (ppm)
6.	NH_3	Jumlah ammonia dalam air (ppm)
7.	NO_2	Jumlah nitrogen dioksida dalam air (ppm)
8.	NO_3	Jumlah nitrat dalam air (ppm)
9.	PO_4	Jumlah ortofosfat dalam air (ppm)
10.	NP Ratio	Rasio nitrat dan fosfat dalam air
11.	Salinitas	Tingkat ke-asin-an air (ppt)
12.	Transparansi Air	Jarak pandang air dari permukaan tambak (cm)
13.	Tinggi Air	Ketinggian air tambak (cm)
		Atribut *Output*
14.	State	Menggambarkan status tambak (baik dan buruk)

ketersediaan sensor pada alat monitoring dan hasil kriteria dari ANOVA-F. ANOVA-F bekerja dengan cara memeriksa rata-rata dua atau lebih parameter yang berbeda secara signifikan [10]. Dengan seleksi fitur ini terpilih 5 parameter yaitu pH, DO, salinitas, transparansi air, dan tinggi air.

HMV

Dari sub-bab sebelumnya menyatakan bahwa fokusan dari penelitian ini adalah pada teknik HMV dan dataset yang digunakan memiliki beberapa data kosong. HMV merupakan proses dalam mengisi nilai kosong dengan nilai lain sesuai dengan kondisi data sampel atau data pada parameter tersebut. Data kosong apabila dibiarkan akan menyebabkan masalah ketika proses pelatihan ML berlangsung. Dalam penelitian ini, teknik yang digunakan adalah pendekatan dengan regresi dan dibandingkan dengan pendekatan statistik. Teknik pendekatan regresi terdiri dari beberapa tahap. Dari dataset yang digunakan, tahap pertama menghilangkan atribut *output* dataset terlebih dahulu untuk tidak diperhitungkan dalam mencari nilai data kosong. Proses selanjutnya menghitung nilai rata-rata (mean) tiap parameter yang sementara digunakan untuk mengisi nilai kosong pada masing-masing parameter. Rumus perhitungan mean dapat dilihat pada Persamaan (1).

$$\overline{X_i} = \frac{\sum_{n=1}^{N_i} x_{in}}{N_i} \quad (1)$$

Dimana $\overline{X_i}$ adalah nilai rata-rata pada parameter i, x_{in} adalah data sampel yang bukan data kosong pada parameter i dan N_i merupakan jumlah sampel data yang tidak terdapat data kosong pada parameter i.

Setelah perhitungan mean selesai, didapatkan nilai rata-rata setiap parameter: $\{\overline{X_1}, \overline{X_2}, \overline{X_3}, \ldots, \overline{X_i}\}$. Dimana i merupakan nilai urutan parameter pada dataset. Nilai rata-rata ini sebagai permulaan disebut sebagai *data zeroth* yang bisa disimbolkan sebagai berikut: $\{z_{11}, z_{12}, z_{13}, \ldots, z_{in}\}$. Dimana i adalah nilai urutan parameter dan merupakan urutan angka sampel pada dataset. Selanjutnya memulai prediksi nilai data kosong menggunakan algoritma regresi dimulai dari parameter pertama yang dirubah menjadi atribut *output* dan dari sampel yang tercatat paling awal. Apabila data kosong yang digantikan nilai mean berada pada sampel n, maka nilai mean dikembalikan menjadi data kosong dan sampel tersebut dijadikan data yang diprediksi, sedangkan data yang lain menjadi data pelatihan bagi algoritma regresi. Dari sini algoritma regresi memprediksi nilai data kosong dan nilai prediksi digunakan untuk mengisi data kosong. Proses ini dilakukan secara bergantian pada sampel lain yang memiliki data kosong hingga berpindah ke parameter lain.

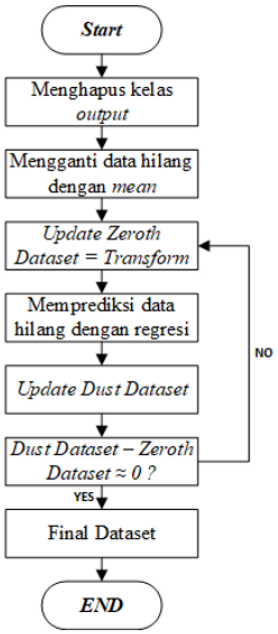
Sesudah memprediksi semua data kosong maka proses ini selesai dengan satu iterasi dan hasil prediksi pertama algoritma regresi disebut dengan *data dust* yang dapat disimbolkan sebagai berikut: $\{d_{11}, d_{12}, d_{13}, \ldots, d_{in}\}$. *Data dust* belum bisa menjadi nilai akhir untuk menggantikan data kosong dikarenakan data ini masih data yang tidak sesuai dengan keadaan sampel. Tahap berikutnya, mencari *data transform* dengan cara mengurangi *data dust* dengan *data zeroth*. Rumus perhitungan *data transform* dapat dilihat pada Persamaan (2).

$$T_{in} = d_{in} - z_{in} \quad (2)$$

Dimana T_{in} adalah nilai *data transform* pada parameter i dan sampel n, d_{in} adalah nilai *data dust* pada parameter i dan sampel n serta z_{in} adalah nilai *data zeroth* pada parameter i dan sampel n.

Data transform biasanya sangat jarang untuk mendapatkan nilai 0 pada proses awal, dan kebanyakan kasus nilainya berjarak cukup jauh dari 0. Maka dari itu dilakukan iterasi selanjutnya yang memiliki tahap seperti sebelumnya dengan *data transform* dijadikan sebagai *data zeroth*. Iterasi ini terjadi berulang-ulang dan berakhir apabila semua nilai *transform* data mendekati 0 atau berada pada *range* yang telah ditentukan. Apabila semua nilai *transform* sudah mendekati 0, maka *data dust* pada iterasi tersebut menjadi nilai final dalam mengisi data kosong. Blok diagram proses HMV dapat dilihat pada Gambar 2.

Seluruh proses dari HMV pendekatan regresi dapat diringkas menjadi langkah pertama melibatkan penggantian setiap data kosong dengan nilai rata-rata pada parameter yang diamati dan bertindak sebagai pengganti. Langkah kedua melibatkan pengaturan imputasi nilai rata-rata dikembalikan ke 'kosong'. Pada langkah ketiga, nilai yang diamati dari suatu parameter (misalnya, 'x_{in}') diregresikan pada variabel lain sehingga 'x_{in}' adalah data



Gambar 2. Blok diagram alur kerja HMV

yang diprediksi dan sisanya adalah *train data*. Dalam teknik imputasi ini, banyak algoritma regresi dijalankan dari KNN, LR, *Lasso*, PAR, SVR, KR, SGD. Langkah keempat melibatkan penggantian nilai data kosong dengan prediksi yang diturunkan dari algoritma regresi. Nilai yang diperhitungkan ini kemudian menjadi bagian dari variabel *input* dalam proses penggantian data kosong selanjutnya pada sampel lain maupun parameter lain. Langkah '2' hingga '4' dilakukan kembali untuk setiap variabel yang memiliki data kosong hingga semuanya tergantikan dan ini masih dalam satu iterasi. Setelah satu iterasi, semua nilai data kosong digantikan oleh prediksi regresi yang terkait dengan data yang diamati. Nilai prediksi yang diperhitungkan terus menerus digantikan oleh prediksi regresi terbaru untuk setiap penambahan iterasi, dan jumlah iterasi dapat bervariasi tergantung apakah data sudah jauh dari bias atau masih bisa disebut sebagai *data dus*t. Sejumlah iterasi idealnya menghasilkan konvergensi koefisien regresi [11].

Imputasi dilakukan dengan menjaga nilai yang diamati dari semua variabel tetap konstan dan hanya data kosong yang berubah ke prediksi imputasi masing-masing. Hasil ini didapat dari pembentukan beberapa set data yang bergantung pada jumlah imputasi. Jumlah imputasi tergantung pada nilai-nilai yang hilang [11]. Pada penelitian White dkk, jumlah imputasi yang ideal berjumlah sama dengan proporsi data kosong [12].

Normalisasi

Normalisasi adalah proses penskalaan nilai atribut dari dataset sehingga dapat berada pada *range* tertentu. Salah satu teknik normalisasi data yang digunakan pada penelitian ini adalah *Min-Max Scaling*. Pada proses normalisasi, data di transformasi nilainya kedalam nilai antara 0 – 1. Hal ini ditujukan supaya masing – masing nilai dalam parameter yang berbeda memberikan pengaruh

		Actual State	
		Positive (1)	Negative (0)
Predicted State	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

Gambar 3. *Confusion* matriks

yang setara. Secara matematis *Min-Max Scaling* ini dapat dituliskan seperti pada Persamaan (3).

$$x_{in} = \frac{x_{in} - x_{i(min)}}{x_{i(\max)} - x_{i(\min)}} \quad (3)$$

Dimana x_{in} adalah variable data pada parameter i dan sampel n, $x_{i(min)}$ adalah nilai yang paling kecil pada parameter i dan $x_{i(max)}$ adalah nilai yang paling besar pada parameter i.

C. *Performance Analysis*

Metrik yang digunakan untuk mengevaluasi kinerja model klasifikasi diberikan pada Gambar 3.

Banyak peneliti menggunakan akurasi sebagai pengukuran performa model untuk masalah klasifikasi dimana rumus metodenya dapat dilihat pada Persamaan (4).

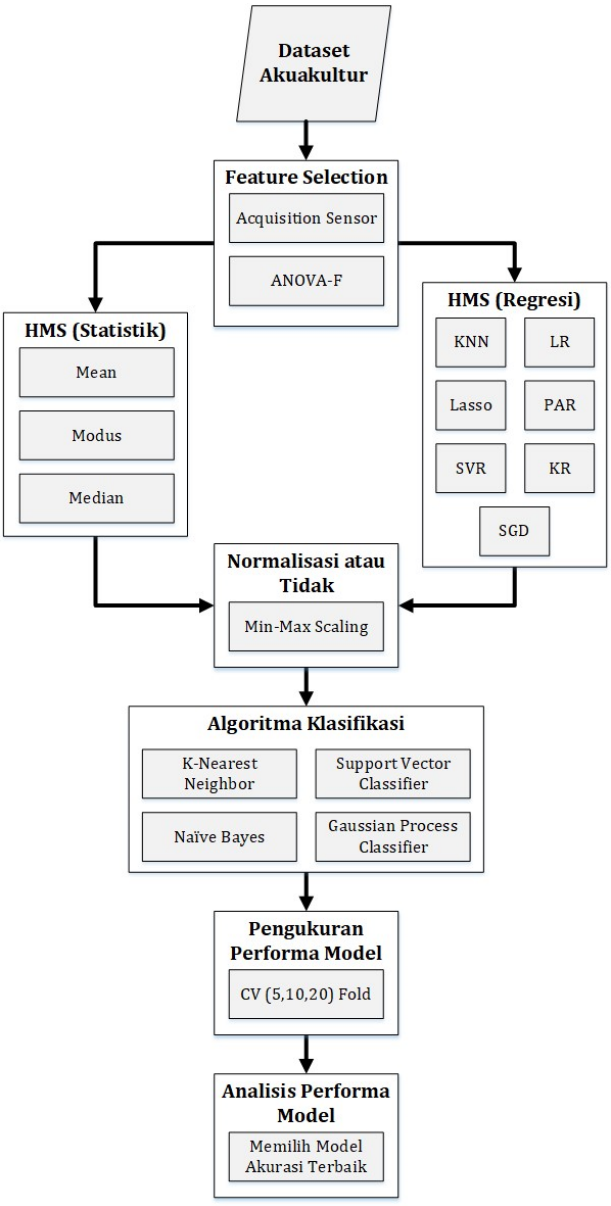
$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

True Positive (TP) yang menunjukkan banyaknya prediksi benar kondisi 1, *True Neagtive* (TN) yang menunjukkan banyaknya prediksi benar kondisi 0, *False Positive* (FP) yang menunjukkan banyaknya prediksi salah kondisi 1, dan *False Negative* (FN) yang menunjukkan banyaknya prediksi salah kondisi 0. Ilustrasi TP, TN, FP, dan FN dapat dilihat dengan *confusion matriks* pada Gambar 3.

IV. Hasil dan Pembahasan

Pada bab ini mengutarakan tentang hasil perbandingan teknik HMV dengan pendekatan algoritma yang berbeda-beda pada kasus budidaya udang. Performa yang digunakan untuk mengetahui penilaian antar teknik HMV pada model terlatih adalah analisis akurasi. Akurasi menjelaskan seberapa bagus model dalam memprediksi kondisi tambak air. Semakin tinggi akurasi, maka semakin baik model yang dibuat. Terdapat empat algoritma model yang digunakan pada penelitian ini yaitu KNN + *grid search*, SVM + *grid search*, GPC, dan NB.

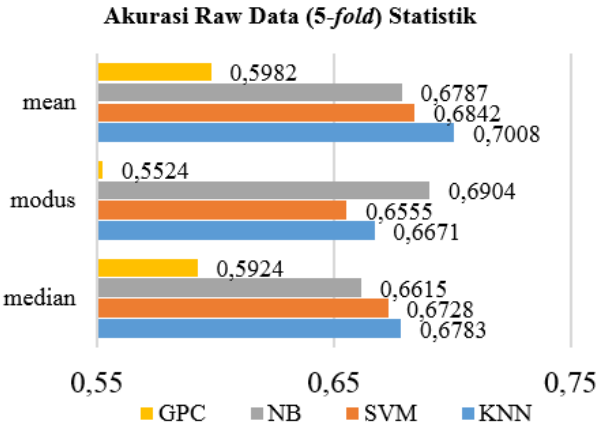
Dari Gambar 4 menunjukkan bahwa dataset yang digunakan dibuat menjadi dua skema dalam memodelkan algoritma. Dataset yang pertama yaitu tanpa melalui tahap normalisasi (*raw* data) sedangkan dataset kedua dinormalisasi terlebih dahulu sebelum dimodelkan (normalisasi data). Dari berbagai macam uji coba, didapatkan akurasi tiap-tiap teknik HMV terhadap masing-masing model dan kemudian dibandingkan. Terdapat tiga validasi yang digunakan dalam melihat performa tiap model. Keputusan akhirnya memilih proses dan model yang memiliki akurasi tertinggi.



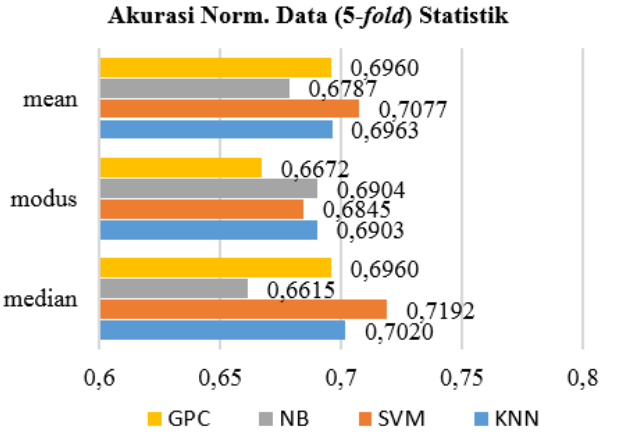
Gambar 4. Diagram proses dari *knowledge modelling*

A. 5-Fold Cross Validation Dataset

Cross Validation dengan 5-*fold* mengartikan bahwa data dibagi menjadi lima bagian dan terdapat lima iterasi. Pada iterasi pertama, data potongan pertama menjadi *test data* dan sisanya menjadi *train data*. *Train data* adalah potongan dataset yang digunakan oleh model klasifikasi untuk mempelajari data dalam mengembangkan model algoritma. Model yang telah mempelajari *train data* diuji coba untuk memprediksi *test data*. Performa model diketahui dengan melihat hasil prediksi dari model sesuai dengan data sebenarnya atau tidak. Semakin banyak prediksi model yang sama dengan data sebenarnya maka semakin bagus performanya. Tahap selanjutnya memulai iterasi kedua, dengan menjadikan potongan kedua sebagai *test data* dan sisanya menjadi *train data*. Skema ini dibuat hingga setiap bagian menjadi *test data*. Dengan demikian didapatkan lima nilai akurasi dari masing-masing iterasi.



Gambar 5. Hasil perbandingan *5-fold cross validation* dengan menggunakan raw data HMV statistik



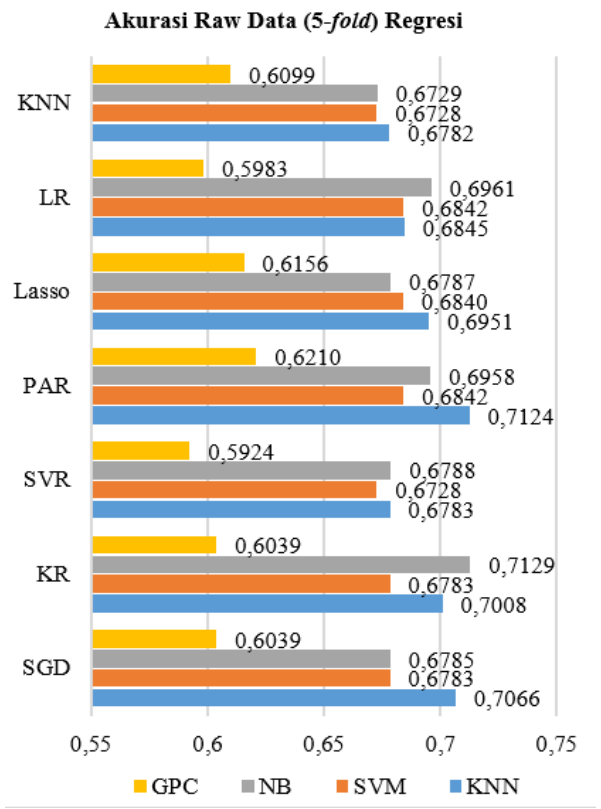
Gambar 6. Hasil perbandingan *5-fold cross validation* dengan menggunakan normalisasi data HMV statistik

Akurasi model klasifikasi secara keseluruhan didapatkan dari rata-rata akurasi tiap iterasi.

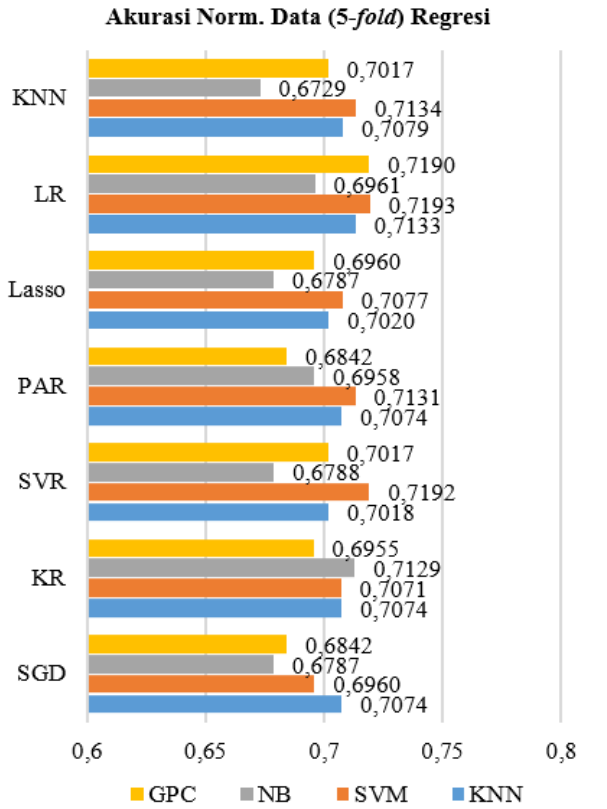
Hasil perbandingan teknik HMV terhadap beberapa algoritma klasifikasi dalam mengembangkan model dengan menggunakan *raw* data dapat dilihat pada Gambar 5 dan Gambar 7 sedangkan ketika menggunakan normalisasi data sebagai dataset yang dipelajari dapat dilihat pada Gambar 6 dan Gambar 8.

Seperti yang ditunjukkan pada Gambar 7, teknik HMV dengan pendekatan regresi menggunakan algoritma KR pada model klasifikasi NB memberikan akurasi tertinggi. Selanjutnya disusul dengan algoritma PAR dengan model klasifikasi KNN. Akurasi dari kedua algoritma tersebut secara berurutan yaitu 0.7129 dan 0.7124. Terlihat pada Gambar 5 bahwa pendekatan regresi lebih unggul daripada pendekatan statistik (konvensional) dengan akurasi tertinggi yang dapat dicapai ada pada mean yaitu 0.7008.

Pada Gambar 6 dan Gambar 8 menunjukkan terjadinya peningkatan akurasi ketika dataset dinormalisasi terlebih dahulu sebelum dipelajari oleh model. Hasil akurasi tertinggi pada pendekatan regresi ada pada penggunaan algoritma LR yaitu 0.7193 disusul dengan SVR yang bernilai 0.7192. Pada eksperimen ini pendekatan statistik



Gambar 7. Hasil perbandingan *5-fold cross validation* dengan menggunakan raw data HMV regresi



Gambar 8. Hasil perbandingan *5-fold cross validation* dengan menggunakan normalisasi data HMV regresi

juga tidak kalah tinggi yaitu 0.7192 dengan teknik median.

B. 10-Fold Cross Validation Dataset

Pada 10-*fold cross validation*, dataset dipecah menjadi 10 bagian. Skema bagian yang menjadi train dan *test data* dilakukan secara bergantian sama seperti proses sebelumnya sehingga mendapatkan 10 akurasi. Grafik perbandingan teknik HMV yang diteliti terhadap *raw* data ada pada Gambar 9 dan Gambar 11 serta normalisasi data dapat dilihat pada Gambar 10 dan Gambar 12.

Dari Gambar 11 dapat diketahui bahwa HMV dengan pendekatan regresi memiliki performa yang lebih tinggi dibandingkan dengan HMV yang menggunakan pendekatan statistik. Akurasi paling tinggi pada pendekatan regresi dicapai oleh algoritma KR sebesar 0.7258 menggunakan model klasifikasi NB. Di sisi lain pada Gambar 9 akurasi pendekatan statistik tertinggi ada pada mean yang bernilai 0.7082 dengan model klasifikasi KNN.

Berdasarkan Gambar 10 ditemukan bahwa akurasi tertinggi yang dapat dicapai dengan pendekatan statistik masih belum bisa melebihi pendekatan regresi. Penggunaan metode mean yang merupakan akurasi tertinggi pada pendekatan statistik mendapatkan 0.7379 sedangkan terlihat pada Gambar 12 pendekatan regresi menggunakan algoritma LR bisa mencapai 0.7552.

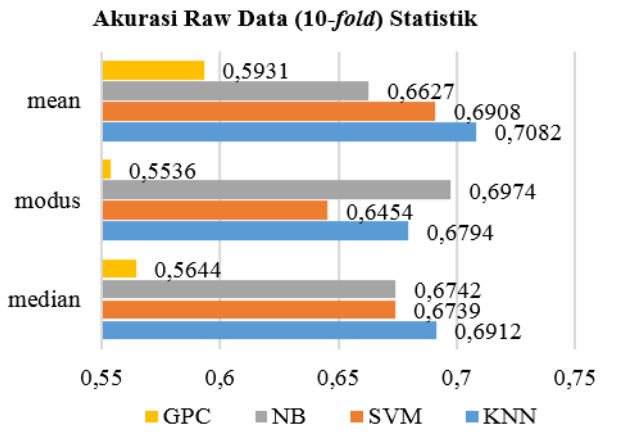
C. 20-Fold Cross Validation Dataset

Pada 20-*fold cross validation*, ada 20 bagian dari dataset akuakultur untuk membangun model klasifikasi dan menganalisa performa model. Proses pengujian dan skemanya sama dengan sebelumnya sehingga didapatkan 20 nilai akurasi yang dirata-ratakan untuk mengetahui performa akurasi akhirnya. Hasil perbandingan teknik HMV terhadap pemakaian *raw* data dengan normalisasi data secara urut dapat dilihat pada Gambar 13, Gambar 14, Gambar 15, dan Gambar 16.

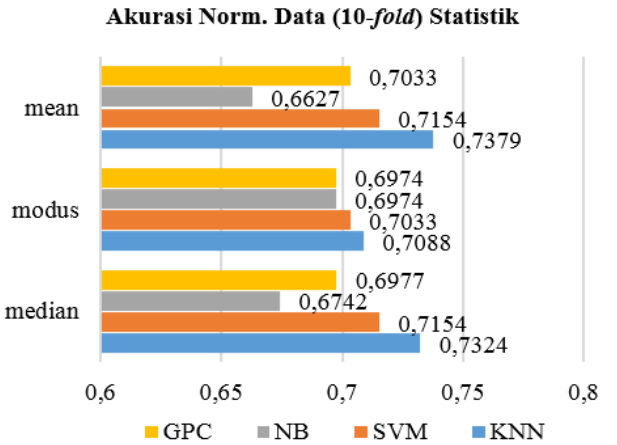
Dalam grafik yang ditunjukkan oleh Gambar 13 diketahui bahwa pendekatan statistik dengan teknik mean bisa mencapai akurasi sebesar 0.7090. Namun pada Gambar 14 pendekatan regresi dapat mencapai akurasi yang lebih tinggi lagi terutama pada algoritma PAR. Algoritma PAR dengan model klasifikasi KNN memiliki akurasi sebesar 0.7264 dan ini merupakan akurasi tertinggi pada pendekatan regresi.

Nilai akurasi tertinggi yang pernah dicapai dalam uji coba penelitian ini terdapat pada Gambar 16. Teknik HMV yang digunakan dan bisa dibilang performa terbaiknya yaitu ketika menggunakan pendekatan regresi LR pada persoalan 20-*fold cross validation*. Akurasi yang dicapai oleh algoritma LR bernilai 0.7576. Sementara itu di Gambar 15, performa terbaik yang berhasil diperoleh dengan pendekatan statistik mean yaitu 0.7396.

Pada makalah ini setelah dilakukan uji coba validasi dari 5-*fold*, 10-*fold*, dan 20-*fold,* dengan dua skema yaitu dinormalisasi dan tidak dinormalisasi, diketahui bahwa HMV regresi memiliki akurasi tertinggi pada setiap validasi. Terlepas dari hal itu, diketahui bahwa akurasi ketika validasi menggunakan 5-*fold* dengan 10-*fold*



Gambar 9. Hasil perbandingan *10-fold cross validation* dengan menggunakan raw data HMV statistik

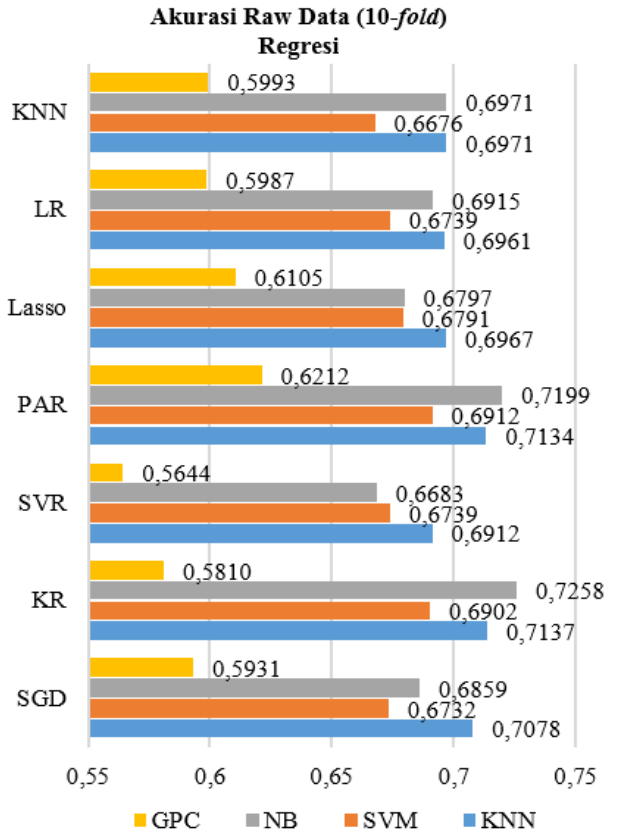


Gambar 10. Hasil perbandingan *10-fold cross validation* dengan menggunakan normalisasi data HMV statistik

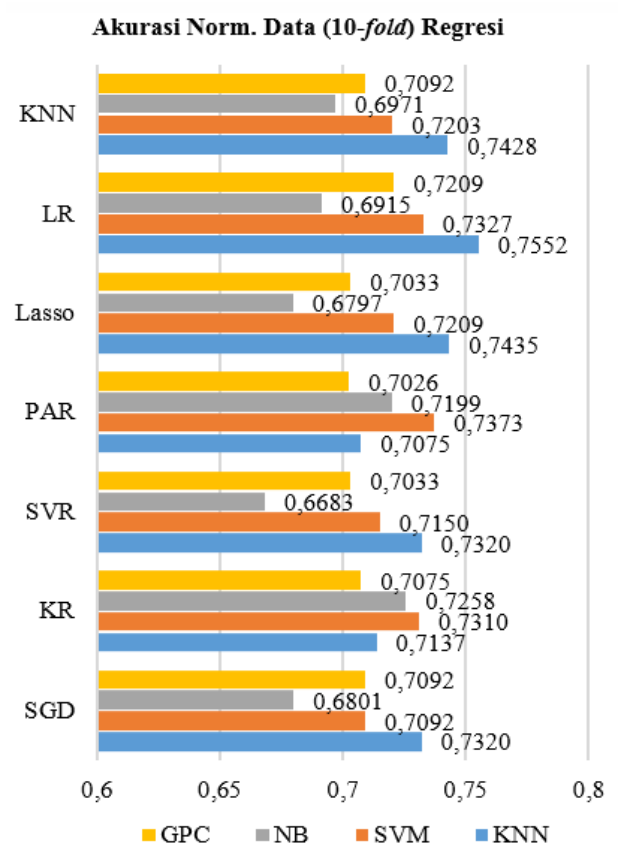
memiliki selisih antara 1-4%. Hal ini dapat terjadi karena penggunaan 5-*fold* membuat dataset yang dipelajari model lebih sedikit daripada 10-*fold* yang berakibat model belum bisa mengelompokkan data dengan baik. Namun, selisih untuk akurasi validasi 10-*fold* dengan 20-*fold* terpaut sangat sedikit yang berarti kemampuan model klasifikasi sudah tergeneralisasi dalam mengelompokkan data dengan validasi 10-*fold* pada kasus dataset penelitian ini. Di lain sisi, penggunaan skema data dinormalisasi terlebih dahulu lebih baik daripada tanpa normalisasi (*raw*). Hal ini karena tanpa normalisasi, *range* tiap parameter berbeda dan bisa saja parameter yang memiliki *range* lebar lebih dominan dalam model. Akibat yang ditimbulkan adanya parameter yang dominan adalah hasil prediksi akan sangat bergantung pada perubahan parameter tersebut dan perubahan parameter lain tidak memberikan dampak yang cukup besar.

V. KESIMPULAN

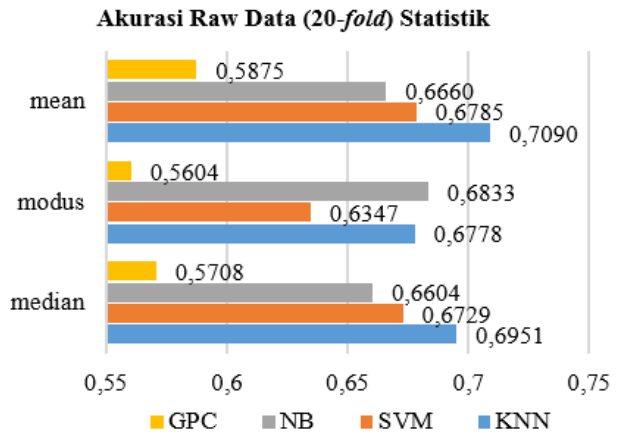
Berdasarkan hasil pengujian teknik HMV yang telah dilakukan, dapat disimpulkan bahwa penggunaan teknik HMV dengan pendekatan regresi pada dataset berukuran kecil dapat meningkatkan akurasi lebih tinggi daripada menggunakan pendekatan statistik (konvensional).



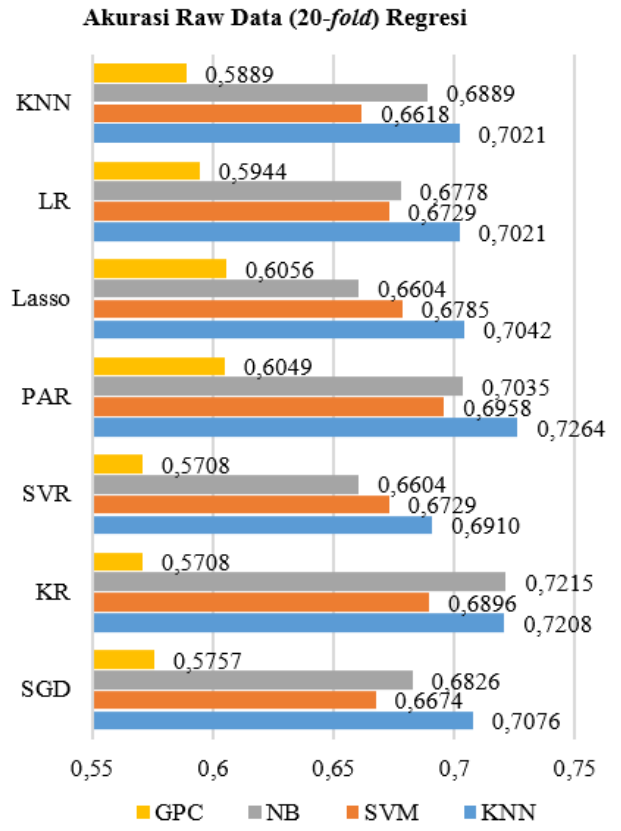
Gambar 11. Hasil perbandingan *10-fold cross validation* dengan menggunakan *raw* data HMV regresi



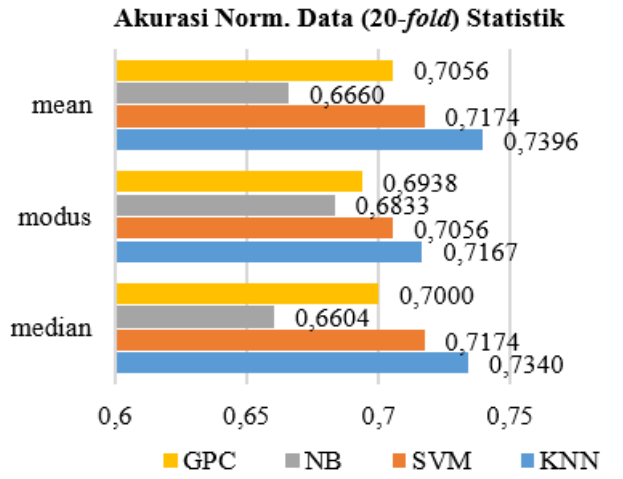
Gambar 12. Hasil perbandingan *10-fold cross validation* dengan menggunakan normalisasi data HMV regresi



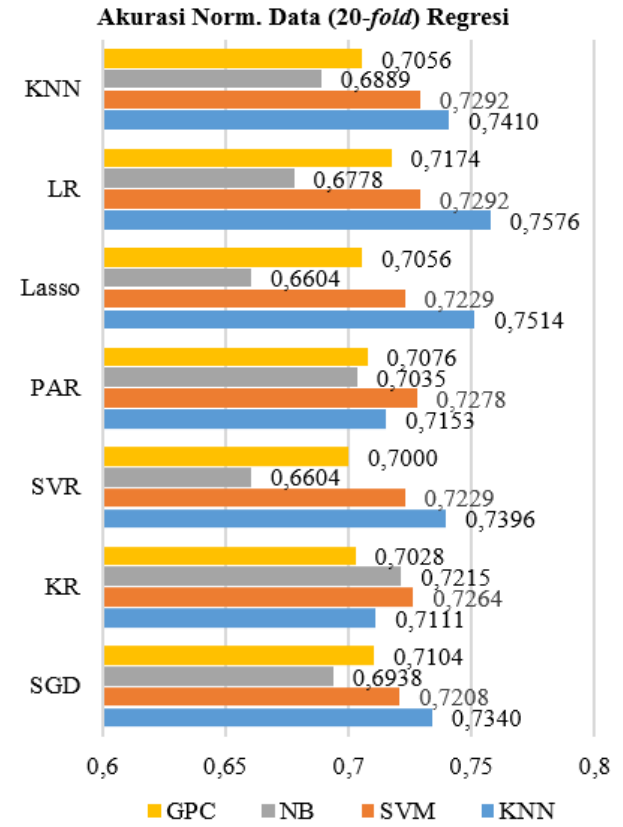
Gambar 13. Hasil perbandingan *20-fold cross validation* dengan menggunakan *raw* data HMV statistik



Gambar 14. Hasil perbandingan *20-fold cross validation* dengan menggunakan *raw* data HMV regresi



Gambar 15. Hasil perbandingan *20-fold cross validation* dengan menggunakan normalisasi data HMV statistik



Gambar 16. Hasil perbandingan *20-fold cross validation* dengan menggunakan normalisasi data HMV regresi

Teknik HMV dengan pendekatan regresi memberikan performa terbaik dari semua validasi yang diajukan yaitu 5-*fold*, 10-*fold*, dan 20-*fold*. Ada satu kondisi yaitu pada pengolahan *raw* data 5-*fold cross validation* disaat pendekatan secara statistik median memberikan kinerja terbaik menyamai dengan pendekatan regresi LR dan SVR sebesar 0.7192. Meskipun demikian, tidak dapat dipungkiri bahwa pendekatan secara statistik yang terdiri dari mean, modus, median tidak dapat mengungguli akurasi tertinggi yang dapat dicapai pendekatan regresi pada 5-*fold*, 10-*fold*, 20-*fold* raw data serta 10-*fold*, 20-*fold* normalisasi data dengan nilai akurasinya secara urut bernilai 0.7129, 0.7258, 0.7264, 0.7552, dan 0.7576. Hal ini dikarenakan HMV pendekatan statistik menggantikan semua nilai data kosong dengan nilai yang sama secara paksa sehingga dapat memunculkan 2 sampel kondisi air (*output*) yang berbeda padahal nilai parameternya sama (data tidak konsisten). Komputasi mengalami kesalahan dalam membuat model yang berakibat model memiliki akurasi rendah untuk memprediksi kondisi baik atau kondisi buruk. Berbeda dengan pendekatan statistik, pendekatan regresi bekerja dengan memperhitungkan nilai pada parameter lain sehingga memungkinkan terjadinya

perbedaan nilai dalam mengganti data kosong dan model terlatih lebih stabil. Setelah menganalisis kinerja beberapa teknik HMV, pendekatan regresi algoritma LR dan ML KNN pada 20-*fold* memiliki akurasi tertinggi sehingga menjadi solusi yang efektif untuk dataset budidaya udang dengan akurasi sebesar 0.7576.

Terlepas dari pencapaian yang telah diuraikan, beberapa perbaikan masih mungkin dilakukan karena algoritma regresi belum di optimasi. Pada penelitian mendatang, akan dilakukan studi dan eksperimen terhadap algoritma metaheuristik dalam pengoptimalan algoritma regresi yang digunakan pada teknik HMV.

Ucapan Terima Kasih

Kegiatan riset ini didukung oleh Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi dan Lembaga Pengelola Dana Pendidikan melalui Program Pendanaan Program Riset Keilmuan Terapan Tahun 2021 berdasarkan kontrak nomor: 0761/D6/KU.04.00/2021 antara Direktorat Jenderal Pendidikan Vokasi, Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi Republik Indonesia dan Politeknik Elektronika Negeri Surabaya.

Referensi

[1] Djumanto, Ustadi, Rustadi, and B. Triyatno, "Utilization of wastewater from vannamei shrimp pond for rearing milkfish in keburuhan coast purworejo sub-district," *Aquacultura Indonesiana*, vol. 19 (1), pp. 38-46, 2018.

[2] K. Li and L. Liu, "Preliminary research on modeling and simulation technology of artificial intelligence system," IOP Conf. Series: Materials Science and Engineering, vol. 452, 2018.

[3] L. Yang, Y. Liu, H. Yu, X. Fang, L. Song, D. Li, and Y. Chen, "Computer Vision models in intelligent aquaculture with emphasis on fish detection and behavior analysis: A Review," *Archives of Computational Methods in Engineering*, vol. 28, no. 4, pp. 2785–2816, 2020.

[4] S. Spänig, A. Emberger-Klein, J.-P. Sowa, A. Canbay, K. Menrad, and D. Heider, "The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes," *Artificial Intelligence in Medicine*, vol. 100, p. 101706, 2019.

[5] J. Huan, H. Li, M. Li, and B. Chen, "Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long short-term memory network: A study of chang zhou fishery demonstration base, China," *Computers and Electronics in Agriculture*, vol. 175, p. 105530, 2020.

[6] T. A. Sinshaw, C. Q. Surbeck, H. Yasarer, and Y. Najjar, "Artificial Neural Network for prediction of total nitrogen and phosphorus in US lakes," *Journal of Environmental Engineering*, vol. 145, no. 6, p. 04019032, 2019.

[7] M. S. Samsudin, A. Azid, S. I. Khalit, M. S. Sani, and F. Lananan, "Comparison of prediction model using spatial discriminant analysis for marine water quality index in mangrove estuarine zones," *Marine Pollution Bulletin*, vol. 141, pp. 472–481, 2019.

[8] A. Najah Ahmed, F. Binti Othman, H. Abdulmohsin Afan, R. Khaleel Ibrahim, C. Ming Fai, M. Shabbir Hossain, M. Ehteram, and A. Elshafie, "Machine learning methods for better water quality prediction," *Journal of Hydrology*, vol. 578, p. 124084, 2019.

[9] Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Approaches to multi-objective feature selection: A systematic literature review," *IEEE Access*, vol. 8, pp. 125076–125096, 2020.

[10] U. Moorthy and U. D. Gandhi, "A novel optimal feature selection technique for medical data classification using ANOVA based whale optimization," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 3527–3538, 2020.

[11] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?," *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40–49, 2011.

[12] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for Practice," *Statistics in Medicine*, vol. 30, no. 4, pp. 377–399, 2010.