Deteksi Dini Status Keanggotaan Industri Kebugaran Menggunakan Pendekatan *Supervised Learning*

http://dx.doi.org/10.28932/jutisi.v6i2.2675

Julio Narabel [#1], Setia Budi ✉ [#2]

[#]*Magister Ilmu Komputer,Universitas Kristen Maranatha Bandung*

[1]mi1879003@student.it.maranatha.edu

[2]setia.budi@it.maranatha.edu

Abstract — **In the fitness industry, the number of members is a major factor for the sustainability of its business. The ability of managers and trainers to detect members who represent traits to quit membership is critical. Four supervised learning classification methods like Support Vector Machine, Random Forest, K-Nearest Neighbor, and Artificial Neural Network were used to generate early detection using two variants of datasets that have different amounts of data. Classification results are separated into three different zones, which are Green Zone, Yellow Zone, and Red Zone. Artificial Neural Network methods using backpropagation training give 99.90% of accuracy on a dataset which has more amount of data. The evaluation has been done using the confusion matrix and AUC-ROC curves.**

Keywords— **Artificial Neural Network; K-Nearest Neighbor; Random Forest; supervised learning; Support Vector Machine.**

I. PENDAHULUAN

Industri kebugaran menjadi salah satu industri yang perkembangannya pesat dalam enam tahun terakhir [1]. Kepedulian masyarakat terhadap kesehatan menjadi salah satu faktor pendorong munculnya banyak industri kebugaran. Dalam menjalankan bisnis kebugaran, faktor utama yang harus diperhatikan adalah keanggotaan. Operasional dan keberlangsungan bisnis kebugaran sangat bergantung pada jumlah anggota (*member*) yang dimilikinya. Banyak faktor yang mempengaruhi anggota untuk tetap berlatih / menjadi anggota di satu klub (*loyal*), tetapi tidak sedikit juga yang memilih untuk berpindah-pindah klub sebab pelayanan, perhatian pelatih (*trainer*), frekuensi kehadiran di klub, biaya yang dibebankan, serta pencapaian yang diperoleh (*progress, achievement*) dirasa kurang memuaskan [2].

Masalah utama yang diangkat dalam penelitian ini yaitu deteksi dini (*early detection*) terhadap prediksi / kemungkinan apakah seorang anggota akan menetap di klub atau mengakhiri keanggotaannya. Dari dataset yang berhasil dikumpulkan, setidaknya ada tiga komponen yang dapat digunakan untuk melakukan deteksi dini, antara lain:

1. Data log latihan setiap anggota.
2. Data pencapaian anggota dilihat dari hasil penimbangan dan pengukuran (*progress, achievement*).
3. Data durasi keanggotaan.

Berdasarkan permasalahan yang telah diuraikan, terdapat dua pertanyaan penelitian yang dapat dirumuskan, yaitu bagaimana cara mendeteksi anggota yang menunjukkan ciri-ciri akan mengundurkan diri pada sebuah pusat kebugaran, dan bagaimana klasifikasi anggota yang menunjukkan ciri-ciri akan mengundurkan diri pada sebuah pusat kebugaran.

Adapun yang menjadi tujuan dari penelitian antara lain membantu manajer dan pelatih sebuah pusat kebugaran agar dapat menentukan langkah preventif yang harus dilakukan terhadap setiap anggota klubnya (terutama yang menunjukkan ciri-ciri akan mengundurkan diri) melalui laporan hasil deteksi dini yang diperoleh dengan memanfaatkan teknologi *machine learning.* Membuat klasifikasi anggota yang menunjukkan ciri-ciri akan mengundurkan diri, untuk memudahkan manajer dan pelatih sebuah pusat kebugaran dalam mengambil tindakan / keputusan.

Hipotesis yang hendak dibuktikan dalam penelitian ini yaitu frekuensi kehadiran, pencapaian anggota, serta biaya keanggotaan yang merupakan satu kesatuan (operasi AND) memiliki pengaruh terhadap berakhirnya status keanggotaan setiap anggota pusat kebugaran.

Penelitian dilakukan dengan menggunakan data riil yang diperoleh dari sebuah industri kebugaran di Kota Bandung periode Desember 2017 sampai dengan Februari 2020. Durasi keanggotaan yang berakhir melebihi bulan Februari 2020 diasumsikan berakhir pada akhir Februari 2020 untuk menunjang penentuan zona klasifikasi.

Dataset yang digunakan dibagi menjadi dua bagian, yaitu:

1. Dataset yang berisikan data anggota yang dicatat dengan frekuensi mingguan.
2. Dataset yang berisikan data anggota yang dicatat dengan frekuensi bulanan.

Pengujian menggunakan dua dataset yang memiliki frekuensi (mingguan dan bulanan) berbeda bertujuan untuk memperoleh model dan akurasi di atas 90%, dengan bias dan variansi di bawah 1%.

Dataset yang terdiri atas 28 ("mingguan") dan 29 ("bulanan") fitur ini akan melalui proses seleksi fitur dengan menggunakan tiga metode, yaitu *Pearson Correlation, Chi-Squared,* serta *Recursive Feature Elimination* (RFE) hingga diperoleh 21 fitur untuk masing-masing dataset.

Deteksi dini dilakukan dengan menggunakan algoritma klasifikasi *supervised learning*, yaitu *Support Vector Machin*e (SVM), *Random Forest* (RF), *K-Nearest Neighbor* (KNN) dan *Artificial Neural Network* (ANN). Hasil deteksi dini berupa klasifikasi yang terdiri dari tiga zona, yaitu Zona Hijau, Zona Kuning dan Zona Merah.

II. KAJIAN LITERATUR.

A. Paradoks Niat-untuk-Melanjutkan-Keanggotaan: Industri Kebugaran [2]

Berikut ini merupakan lima faktor yang mempengaruhi niat setiap anggota pada industri kebugaran untuk melanjutkan keanggotaannya, yaitu:

1. Kepuasan pelanggan.
2. Atribut-atribut pelayanan, seperti aktivitas, properti fisik, orang, dan elemen-elemen pendukung lainnya yang ditawarkan oleh sebuah organisasi untuk dirasakan oleh pelanggan dalam bentuk layanan.
3. Merek, salah satu faktor yang paling penting sebab dapat mempengaruhi kepercayaan dan pengakuan masyarakat.
4. Perilaku pembelian sebelumnya.
5. Persepsi harga.

Pelayanan yang ditawarkan oleh klub, keamanan, serta promosi gambar klub memberikan dampak positif bagi kepuasan pelanggan. Hubungan pelanggan pada sebuah klub, komitmen pelanggan, dan frekuensi kehadiran mingguan pelanggan di klub memberikan dampak positif secara langsung pada niat pelanggan untuk melanjutkan keanggotaannya. Sedangkan biaya yang dibebankan pada pelanggan memberikan sebuah dampak langsung yang negatif.

B. Tren Industri Kebugaran [1]

Melalui survei yang dilakukan pada tahun 2018, *wearable technology* menempati posisi pertama, diikuti oleh *group training* pada tempat kedua dan *High-Intensity Interval Training* (HIIT) pada tempat ketiga.

Wearable technology merupakan komponen digital pendukung aktivitas yang berupa *fitness tracker, smart watches, heart rate monitor* dan GPS *tracking device*.

Group training merupakan sebuah grup dalam pusat kebugaran yang berisikan *coach* dan *lead*, biasanya merupakan grup *personal training*. Grup ini dirasakan efektif dalam memberikan motivasi dan porsi latihan.

High-Intensity Interval Training (HIIT) merupakan latihan berdurasi pendek dengan intensitas tinggi yang disertai waktu istirahat singkat.

C. Survei

Sebelum menentukan sumber data dan metode *machine learning* yang akan digunakan, langkah awal yang harus dilakukan yaitu mendapatkan pemahaman yang lebih mendalam mengenai permasalahan utama yang menjadi perhatian pemangku kepentingan. Pemahaman bisa diperoleh melalui kajian literatur yang dilakukan serta *interview* terhadap pihak-pihak terkait. Pemahaman tersebut menghasilkan kategori-kategori / komponen-komponen penting yang menjadi fokus utama penelitian.

Langkah berikutnya yaitu membuat kuesioner yang terdiri dari tiga bagian utama, yaitu informasi umum, informasi spesifik, dan bobot kategori. Kuesioner ini menjadi bahan utama survei untuk menentukan kategori yang paling penting bagi penelitian. Dalam kasus ini, survei dijadikan tahapan dalam melakukan seleksi fitur.

D. Koleksi dan Persiapan Data

Setelah kategori yang akan digunakan telah ditentukan, langkah selanjutnya yaitu melakukan *data collection and preparation* [4]. Setelah pengumpulan dan persiapan data selesai, maka dilanjutkan dengan *data pre-processing* menggunakan tahapan sebagai berikut [3]:

1. Normalisasi data.
2. Pemilihan *setting*. Menggunakan teknik sampling bertingkat, diambil 10% grup homogen dari data untuk dieksekusi dalam tes algoritma.
3. *Cross-validation.* Data dibagi untuk *five fold cross-validation.*
4. Pembagian *learning set* menjadi dua, 70% untuk *training* dan 30% untuk *testing* / validasi.

E. Melakukan Training pada Machine Learning

Setelah dataset tersedia, langkah berikutnya yaitu melakukan *training* dengan menggunakan algoritma klasifikasi *supervised learning* [4][5]. Adapun algoritma *supervised learning* yang digunakan antara lain:

Support Vector Machine (SVM) [3][6][7] memiliki beberapa fungsi kernel tradisional antara lain linear, polinomial, *Gauss, Sigmoid,* dan *Fourier*. Dalam hal kernel, *Gauss* dipilih karena memiliki efisiensi tinggi, tidak membutuhkan banyak komputasi, serta mudah diakses.

Random Forest (RF) [4][8] dengan jumlah *Decision Tree* yang digunakan untuk tahap *learn* dari model sebanyak 100 buah.

Artificial Neural Network (ANN) yang menggunakan model *Supervised and Feed-Forward Neural Network* dengan dua *neuron* output dan tiga *hidden layer* [3].

F. Menganalisis Hasil

Hasil dari SVM, RF dan ANN dibandingkan dengan menggunakan *Confusion Matrix* yang diakses dengan menggunakan KNIME [3]. Dalam menggabungkan hasil prediksi antara dua model atau lebih digunakan fitur *prediction function* pada KNIME.

Selain *Confusion Matrix*, kurva *Receiver Operating Characteristic* (ROC) digunakan untuk melihat akurasi model *machine learning*. Dalam penggunaan ROC, *Area Under the Curve* (AUC) digunakan untuk membantu mengurangi kurva ROC menjadi satu nilai saja, merepresentasikan kinerja yang diharapkan dari *classifier* [3].

III. METODOLOGI

Alur metodologi dimulai dari koleksi dan persiapan data, seleksi fitur, pembagian set *training* dan *testing*, *cross validation*, pengujian model *machine learning*, pengecekan bias dan variansi, dan evaluasi hasil ditunjukkan oleh Gambar 1.

A. Koleksi dan Persiapan Data

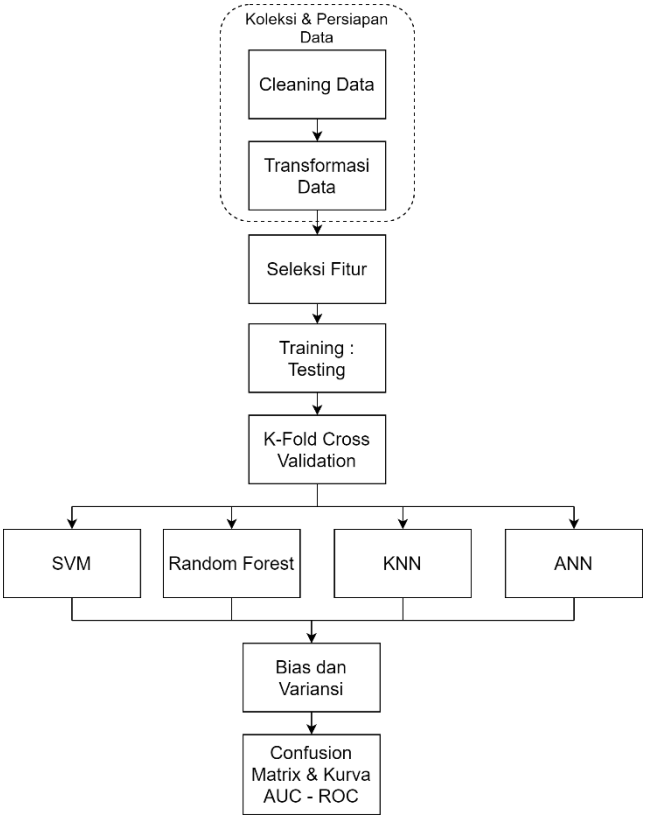
Ketiga data *raw* (yaitu, frekuensi latihan, pencapaian anggota, durasi keanggotaan) digabung menjadi satu kesatuan dataset. Ada dua macam dataset yang dibuat, yaitu dataset "mingguan" yang mengacu pada frekuensi kehadiran anggota setiap minggunya, serta dataset "bulanan" yang mengacu pada pencapaian anggota yang dicatat setiap bulan.

Rentang klasifikasi yang dibagi menjadi tiga zona (yaitu, Zona Hijau, Zona Kuning, Zona Merah) ditentukan berdasarkan frekuensi kehadiran anggota. Pemilihan faktor frekuensi kehadiran anggota sebagai patokan untuk menentukan zona mengacu pada tiga hal, yakni:

1. Memiliki jumlah baris data paling banyak dibandingkan dua faktor lainnya.
2. Seluruh anggota setidaknya mencatatkan data kehadirannya minimal satu kali, sehingga terhindar dari *missing value*.
3. Pencapaian anggota sangat bergantung pada frekuensi latihan di pusat kebugaran.

Berikut merupakan rincian dari tiga zona klasifikasi:

1. Zona Hijau. Anggota tidak memerlukan penanganan / berada dalam kondisi ideal (frekuensi kehadiran baik).
2. Zona Kuning. Anggota memerlukan sedikit penanganan (frekuensi kehadiran cukup / di bawah standar).
3. Zona Merah. Anggota sangat membutuhkan pendekatan, perhatian, dan penanganan ekstra dari manajer / pelatih pusat kebugaran (frekuensi kehadiran buruk).



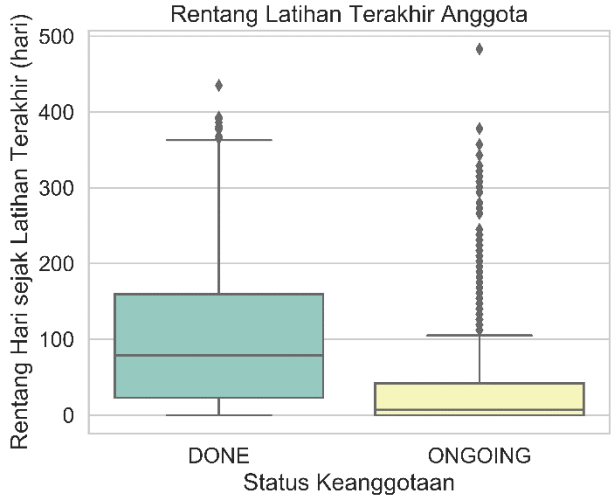
Gambar 1. Metodologi

Proses menentukan klasifikasi dibagi menjadi dua bagian, yaitu klasifikasi pada dataset "mingguan" dan dataset "bulanan". Pada dataset "mingguan", klasifikasi dilakukan dengan menggunakan data frekuensi latihan anggota per minggu yang digabungkan dengan data durasi keanggotaan. Pada dataset "bulanan" klasifikasi dilakukan dengan menggunakan data frekuensi latihan anggota per bulan yang digabungkan dengan data durasi keanggotaan. Hasil penggabungan tersebut menghasilkan rentang (jumlah hari) antara latihan terakhir dengan masa berakhirnya keanggotaan.

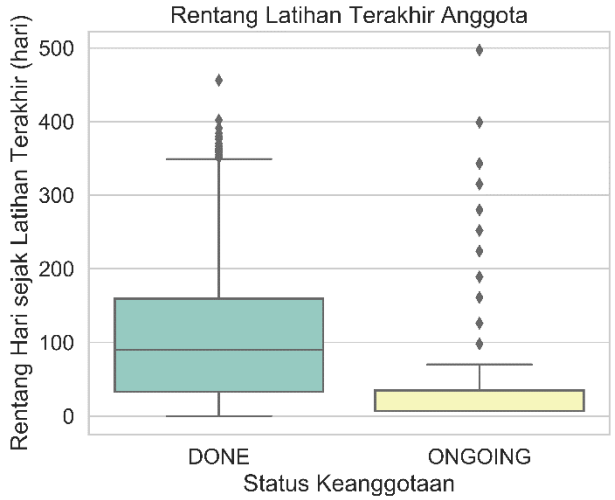
Gambar 2 menunjukkan data sebaran "rentang waktu" (dalam satuan hari) antara latihan terakhir dengan masa berakhirnya keanggotaan pada data "mingguan". Gambar 3 menunjukkan data sebaran "rentang waktu" (dalam satuan hari) antara latihan terakhir dengan masa berakhirnya keanggotaan pada data "bulanan". "DONE" mewakili "rentang waktu" untuk anggota yang masa keanggotaannya telah berakhir, sedangkan "ONGOING" mewakili "rentang waktu" untuk anggota yang masa keanggotaannya masih berjalan.

Gambar 2 dan Gambar 3 menunjukkan sejumlah *outlier* pada *boxplot* "DONE" dan *boxplot* "ONGOING". *Outlier* ini merupakan rentang sejak latihan terakhir bagi anggota yang masa berakhir keanggotaannya melebihi Februari 2020, serta anggota yang sedang dalam status "*on hold*" atau "cuti" (status keanggotaannya aktif kembali setelah Februari 2020).

Oleh karena itu jumlah *outlier* pada status keanggotaan "ONGOING" lebih banyak daripada *outlier* pada status keanggotaan "DONE". Keberadaan *outlier* ini tidak mempengaruhi penentuan zona klasifikasi sehingga dapat diabaikan.



Gambar 2. Rentang antara latihan terakhir dengan masa berakhirnya keanggotaan per minggu



Gambar 3. Rentang antara latihan terakhir dengan masa berakhirnya keanggotaan per bulan

Tabel I menunjukkan rangkuman klasifikasi zona yang didapatkan dari hasil *plotting* yang ditunjukkan pada Gambar 2 dan Gambar 3.

Setelah klasifikasi ditentukan, tahap selanjutnya yaitu membuat dataset "mingguan" dan dataset "bulanan". Dataset ini dibentuk dengan menggabungkan data frekuensi kehadiran anggota (per minggu / per bulan) dengan data pencapaian anggota.

TABEL I
RINCIAN KLASIFIKASI ZONA

Zona	Data "Mingguan" (Hari)	Data "Bulanan" (Hari)
Hijau	Rentang ≤ 7	Rentang ≤ 7
Kuning	7 < Rentang < 23	7 < Rentang < 33
Merah	Rentang ≥ 23	Rentang ≥ 33

Penggabungan yang dilakukan untuk membuat dataset "mingguan" terkendala *missing value* sejumlah 75% dari data pencapaian. Hal ini disebabkan oleh data pencapaian anggota yang dicatat dalam frekuensi bulanan, sehingga dalam empat baris data frekuensi kehadiran hanya terdapat 1 baris data pencapaian anggota, seperti diperlihatkan pada Tabel II.

TABEL II
MISSING VALUE PADA PENGGABUNGAN DATASET "MINGGUAN"

Durasi	Latihan	Pencapaian
Minggu 1	Latihan 1	Pencapaian 1
Minggu 2	Latihan 2	Tidak ada data
Minggu 3	Latihan 3	Tidak ada data
Minggu 4	Latihan 4	Tidak ada data
Minggu 5	Latihan 5	Pencapaian 2
Minggu 6	Latihan 6	Tidak ada data

Pada penggabungan yang dilakukan untuk membuat dataset "bulanan" masih didapatkan *missing value*, tetapi tidak sebanyak pada dataset "mingguan". Pada dataset "bulanan", fitur latihan dibagi menjadi dua yaitu latihan total (dalam satu bulan) dan rata-rata latihan per minggu dalam satu bulan. Struktur datanya dapat dilihat pada Tabel III.

TABEL III
STRUKTUR DATA DAN *MISSING VALUE* PADA DATASET "BULANAN"

Durasi	Latihan (Total)	Latihan (Rata-Rata)	Pencapaian
Bulan 1	Latihan 1 – 4	Latihan 1 – 4	Pencapaian 1
Bulan 2	Latihan 5 – 8	Latihan 5 – 8	Pencapaian 2
Bulan 3	Latihan 9 – 12	Latihan 9 – 12	Pencapaian 3
Bulan 4	Latihan 13 – 16	Latihan 13 – 16	Tidak ada data
Bulan 5	Latihan 17 – 20	Latihan 17 – 20	Pencapaian 4
Bulan 6	Latihan 21 - 24	Latihan 21 - 24	Pencapaian 5

Teknik interpolasi linier merupakan alternatif yang dipilih untuk mengisi kekosongan pada *missing value* [9]. Fitur "UserId" dan "Durasi" menjadi acuan dalam melakukan interpolasi linier. Teknik interpolasi linier dipilih karena fitur "Durasi" yang menjadi acuan berisikan data yang nilainya terus bertambah dengan kenaikan konstan (bertambah satu untuk setiap sampel data, baik mingguan maupun bulanan). Tidak seluruh *missing value* dapat diisi

dengan cara interpolasi, melainkan hanya data (berdasarkan "UserId" dan "Durasi") yang memiliki nilai awal dan akhir fitur "Pencapaian" yang dapat menerapkan teknik interpolasi linier.

Setelah dilakukan interpolasi, masih terdapat *missing value* pada fitur "Pencapaian" untuk beberapa baris data. Oleh karena itu dilakukan *cleaning* data dengan cara menghapus data-data tersebut. Pada dataset "mingguan" terdapat 40 baris data yang dihapus sehingga menyisakan 35813 baris data yang akan digunakan. Pada dataset "bulanan" terdapat 211 baris data yang dihapus, menyisakan 9806 baris data yang dapat digunakan.

B. Seleksi Fitur

Seleksi fitur dilakukan terhadap kedua dataset ("mingguan" dengan 28 fitur dan "bulanan" dengan 29 fitur). Masing-masing dataset akan diseleksi dengan menggunakan 3 metode, yaitu *Pearson Correlation, Chi-Squared*, dan *Recursive Feature Elimination* (RFE) [10][11][12][13].

Seleksi fitur dengan menggunakan tiga metode dilakukan berdasarkan pertimbangan berikut:

1. *Pearson Correlation* hanya dapat melakukan cek korelasi antar dua variabel linier.
2. *Chi-Squared* melakukan cek korelasi antar dua variabel berdasarkan frekuensi kemunculan, sehingga dapat menutupi ketidakmampuan *Pearson Correlation* dalam melakukan cek korelasi antar dua variabel nonlinier.
3. RFE merupakan metode dengan teknik *wrapper-based* yang mengecek korelasi tidak berdasarkan *score*. Estimator yang digunakan adalah *Logistic Regression*. *Logistic Regression* dipilih sebab dapat diandalkan dalam klasifikasi data nominal [4].

Dalam melakukan seleksi fitur, klasifikasi zona yang diwakili oleh tiga data numerik yaitu 0 (Zona Merah), 1 (Zona Kuning) dan 2 (Zona Hijau) menjadi patokan. Fitur-fitur yang melalui seleksi hanya fitur yang terdiri dari data numerik. Fitur lainnya yang memiliki tipe data *string* dan *timestamps* tidak disertakan dalam proses seleksi fitur.

Seleksi fitur yang dilakukan menghasilkan masing-masing 21 fitur dari dataset "mingguan" dan dataset "bulanan". Jika ada fitur yang tidak memberikan satupun nilai "True" dari ketiga metode seleksi, maka fitur tersebut akan dieliminasi.

C. Pembagian Set Training dan Testing

Set *training* dan *testing* menggunakan komposisi 70:30, yaitu 70% untuk *training* dan 30% untuk *testing*. Pada dataset "mingguan", dari total baris data yang digunakan (35813 baris), 25069 baris data digunakan untuk *training*, dan 10744 baris data digunakan untuk *testing*. Pada dataset "bulanan", dari total baris data yang digunakan (9806 baris), 6864 baris data digunakan untuk *training*, dan 2942 baris data digunakan untuk *testing*.

D. Cross Validation

Data dibagi menjadi lima paket set *training* dan *testing* secara acak melalui teknik validasi hasil *5-fold cross validation* dengan komposisi seperti yang ditunjukkan Tabel IV [14].

TABEL IV
FIVEFOLD CROSS VALIDATION

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
Pembagian 1	Testing	Training	Training	Training	Training
Pembagian 2	Training	Testing	Training	Training	Training
Pembagian 3	Training	Training	Testing	Training	Training
Pembagian 4	Training	Training	Training	Testing	Training
Pembagian 5	Training	Training	Training	Training	Testing

E. Pengujian Support Vector Machine (SVM)

Pengujian dilakukan menggunakan SVM dengan kernel RBF. Nilai C yang digunakan adalah 1, dengan gamma 1/21 (1/jumlah fitur). Nilai C menunjukkan margin *hyperplane*. Semakin besar nilai C, maka akan semakin kecil marginnya. Gamma menunjukkan pengaruh dari satu sampel *training*. Semakin kecil gamma, semakin kecil juga pengaruh dari satu sampel *training*-nya. Persamaan dari SVM dengan kernel RBF adalah sebagai berikut:

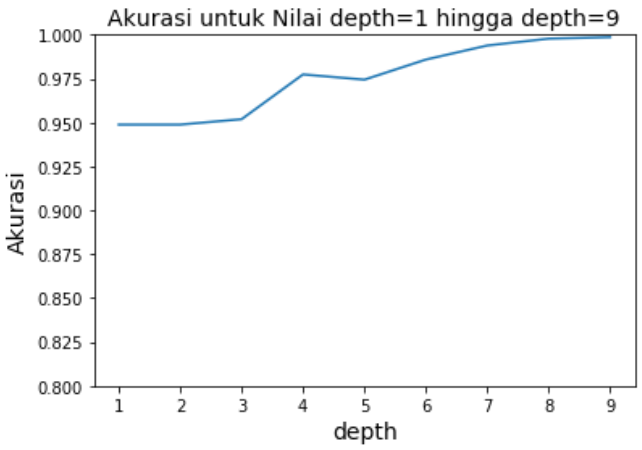
$$k_{(x_i,x_j)} = \exp\left(-\gamma \left\|x_i, x_j\right\|^2\right) \quad (1)$$

Dari pengujian dengan model SVM yang telah dilakukan, dataset "mingguan" memiliki akurasi 96.95% dan dataset "bulanan" memiliki akurasi 94.83%.

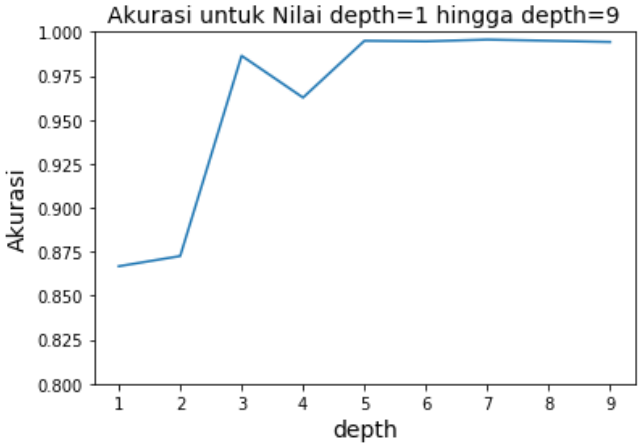
F. Pengujian Random Forest (RF)

Dalam pengujian RF, hal pertama yang dilakukan yaitu menentukan "*depth*" yang akan digunakan, dimulai dari *depth* = 1 hingga *depth* = 9. Berdasarkan hasil yang ditunjukkan pada Gambar 4 dan Gambar 5, "*depth*" yang dipilih adalah 4 untuk dataset "mingguan" dan 5 untuk dataset "bulanan", dengan pertimbangan hasil akurasi menunjukkan nilai paling baik sebelum mencapai 100% tanpa memerlukan kedalaman yang besar. Nilai "*depth*" dengan akurasi 100% tidak dipilih karena memiliki kemungkinan menjadikan "model" *overfitting*. Jumlah *tree* yang digunakan adalah 100.

Dari pengujian dengan model RF yang telah dilakukan, dataset "mingguan" memiliki akurasi 96.95% dan dataset "bulanan" memiliki akurasi 99.26%.



Gambar 4. Menentukan "*depth*" *Random Forest* pada dataset "mingguan"



Gambar 5. Menentukan "*depth*" *Random Forest* pada dataset "bulanan"

G. Pengujian K-Nearest Neighbor (KNN)

Dalam pengujian KNN [15][16], hal pertama yang dilakukan yaitu menentukan nilai k (tetangga / *neighbor*) yang akan digunakan. Pencarian nilai k dimulai dari k = 1 hingga k = 25. Berdasarkan hasil yang ditunjukkan pada Gambar 6 dan Gambar 7, nilai k yang dipilih adalah 4 untuk dataset "mingguan" dan 6 untuk dataset "bulanan", dengan pertimbangan hasil akurasi menunjukkan nilai tertinggi sebelum kemudian semakin berkurang dengan "curam".

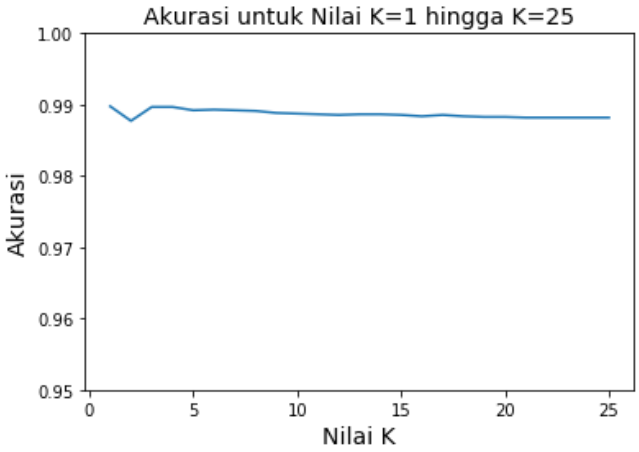
Pada nilai k = 2 (Gambar 6 dan Gambar 7), akurasi cenderung menurun sebelum kemudian naik kembali pada nilai k = 3. Hal ini disebabkan oleh jumlah "tetangga" yang dijadikan acuan hanya dua buah, padahal klasifikasi yang merupakan output terdiri atas tiga buah. Oleh karena itu, jumlah minimal "tetangga" (k) yang digunakan dalam KNN harus sama dengan jumlah zona klasifikasinya

Dari pengujian dengan model KNN yang telah dilakukan, dataset "mingguan" memiliki akurasi 98.89% dan dataset "bulanan" memiliki akurasi 98.25%.

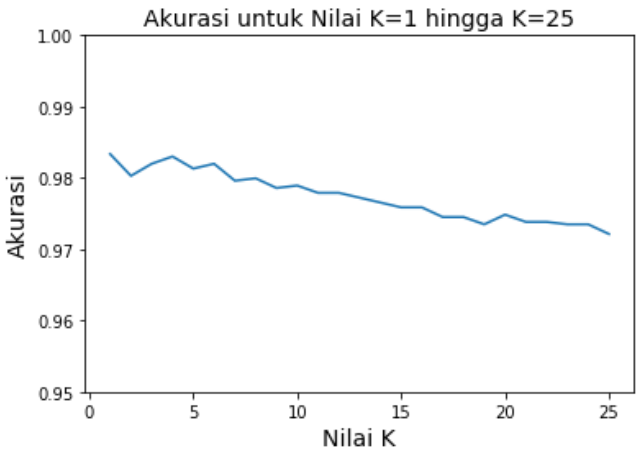
H. Pengujian Artificial Neural Network (ANN)

Pengujian dilakukan menggunakan ANN [17][18][19] dengan *solver adam*, lima *hidden layer* dan tiga *neuron* output (Zona Hijau, Zona Kuning, dan Zona Merah), serta algoritma *training back propagation*. Dipilihnya *adam* sebagai *solver* didasari oleh ukuran dataset dengan 21 fitur dan 35813 baris (dataset "mingguan") dan 9806 baris (dataset "bulanan"), agar memberikan waktu *training* yang singkat dan skor validasi di atas 85%.

Dari pengujian dengan model ANN yang telah dilakukan, dataset "mingguan" memiliki akurasi 99.76% dan dataset "bulanan" memiliki akurasi 99.01%.



Gambar 6. Menentukan nilai k KNN pada dataset "mingguan"



Gambar 7. Menentukan nilai k KNN pada dataset "bulanan"

IV. EVALUASI HASIL DAN PEMBAHASAN

Evaluasi hasil dan pembahasan dibagi menjadi tiga bagian, antara lain evaluasi hasil pengujian SVM, RF, KNN dan ANN dalam *confusion matrix* [20][21], kurva ROC [22][23], serta evaluasi bias dan variansi [24].

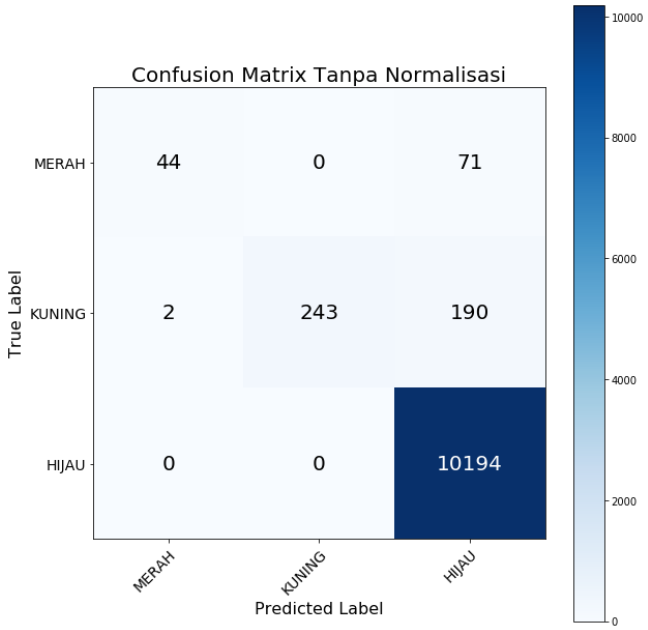
p-ISSN : 2443-2210
e-ISSN : 2443-2229

A. Evaluasi Hasil Confusion Matrix Dataset "Mingguan"

Confusion matrix dibuat dengan memanfaatkan *library confusion_matrix* pada *scikit-learn* dengan menggunakan bahasa pemrograman *Python.*

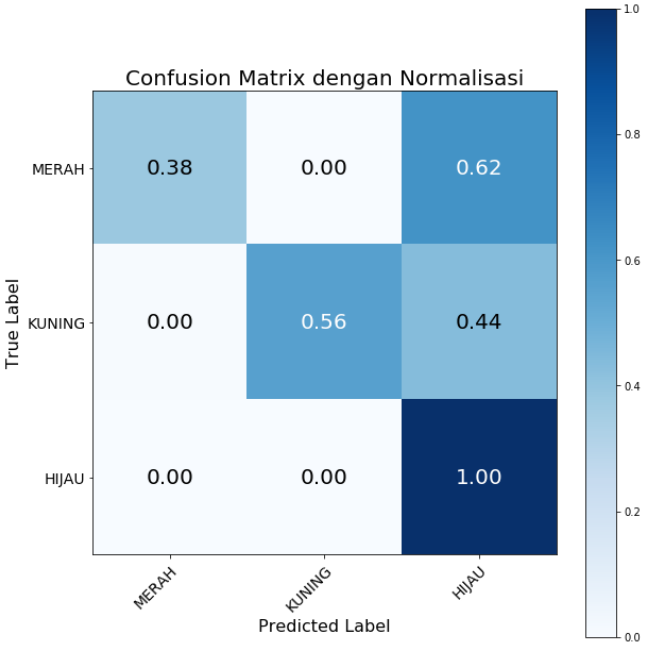
Gambar 9 menunjukkan *confusion matrix* SVM dengan normalisasi. Normalisasi dilakukan untuk menunjukkan interpretasi yang lebih visual pada *confusion matrix* yang memiliki ketidakseimbangan kelas. Normalisasi dilakukan per kelompok klasifikasi (MERAH, KUNING, HIJAU) pada wilayah *true positive*. Satu baris klasifikasi pada wilayah *true positive* berjumlah 1.00 [25].

Berdasarkan hasil yang ditunjukkan *confusion matrix* SVM (Gambar 8 dan Gambar 9), klasifikasi Zona Hijau seluruhnya berada pada area *true positive*, yang berarti bahwa 100% hasil pengujian sesuai dengan prediksi. Klasifikasi Zona Kuning (56%) dan Zona Merah (38%) memberikan hasil pengujian yang sesuai dengan prediksi. Hasil *false positive* yang berasal dari Zona Kuning sebesar 44%, dan Zona Merah sebesar 62%, yang seluruhnya diprediksikan sebagai Zona Hijau. Model SVM menunjukkan error yang besar (62%) dalam memprediksi Zona Merah.

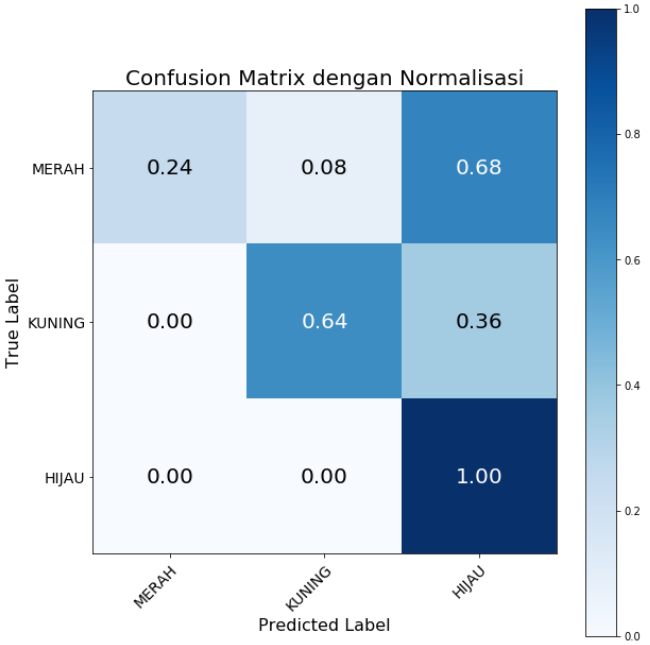


Gambar 8. *Confusion matrix* SVM dataset "mingguan" tanpa normalisasi

Confusion matrix RF (Gambar 10) menunjukkan klasifikasi Zona Hijau seluruhnya berada pada area *true positive*, yang berarti bahwa 100% hasil pengujian sesuai dengan prediksi. Klasifikasi Zona Kuning (64%) dan Zona Merah (24%) memberikan hasil pengujian yang sesuai dengan prediksi. Hasil *false positive* yang berasal dari Zona Kuning sebesar 36% (diprediksikan sebagai Zona Hijau). Pada Zona Merah, hasil *false positive* sebesar 68% (Zona Hijau), serta hasil *true negative* sebesar 8% (Zona Kuning). Model RF menunjukkan *error* yang besar (76%) dalam memprediksi Zona Merah.



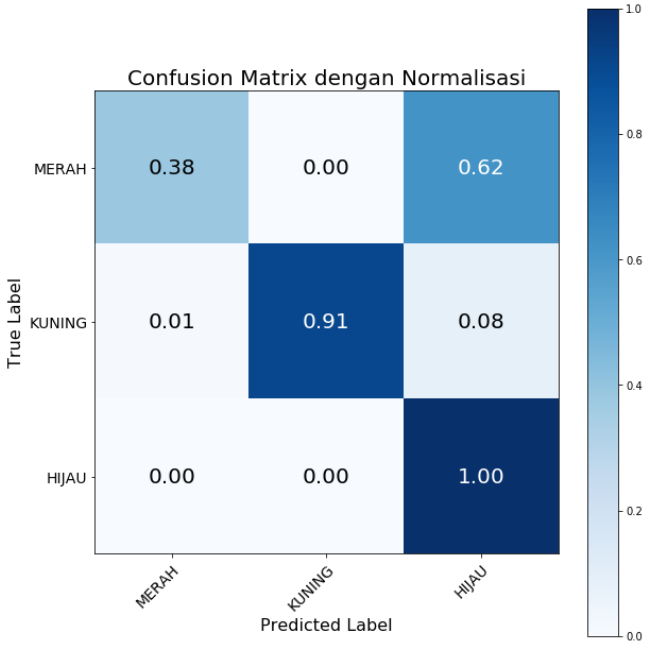
Gambar 9. *Confusion matrix* SVM dataset "mingguan" ternormalisasi



Gambar 10. *Confusion matrix* RF dataset "mingguan" ternormalisasi

Confusion matrix KNN (Gambar 11) menunjukkan klasifikasi Zona Hijau seluruhnya berada pada area *true positive*, yang berarti bahwa 100% hasil pengujian sesuai dengan prediksi. Klasifikasi Zona Kuning memberikan 91% hasil pengujian yang sesuai dengan prediksi. Sedangkan Zona Merah memberikan hasil pengujian sesuai prediksi sebesar 38%. Hasil *false positive* yang berasal dari Zona

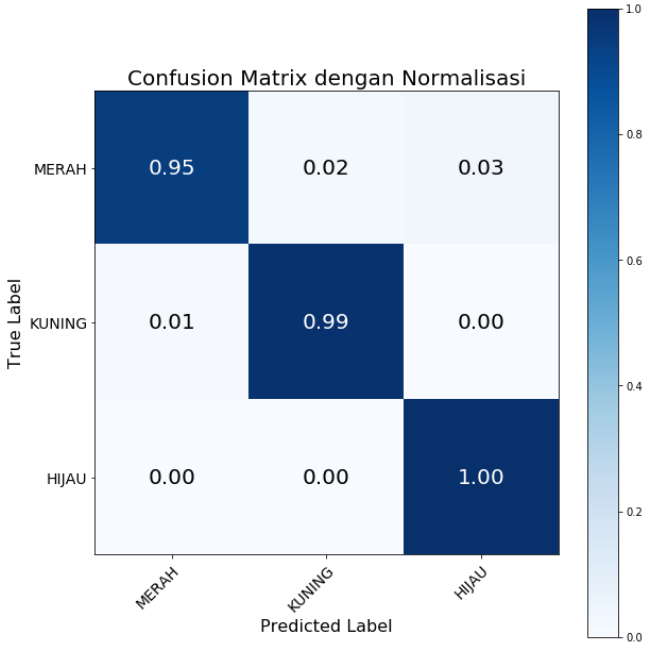
Kuning (8%) dan Zona Merah (62%) seluruhnya diprediksikan sebagai Zona Hijau. Model KNN menunjukkan *error* yang besar (62%) dalam memprediksi Zona Merah.



Gambar 11. *Confusion matrix* KNN dataset "mingguan" ternormalisasi

Confusion matrix ANN (Gambar 12) menunjukkan klasifikasi Zona Hijau seluruhnya berada pada *area true positive*, yang berarti bahwa 100% hasil pengujian sesuai dengan prediksi. Klasifikasi Zona Kuning memberikan 99% hasil pengujian yang sesuai dengan prediksi. Sedangkan Zona Merah memberikan hasil pengujian sesuai prediksi sebesar 95%. Hasil *true negative* Zona Merah yang berasal dari Zona Kuning sebesar 1%, serta hasil *true negative* Zona Kuning yang berasal dari Zona Merah sebesar 2%. *False positive* Zona Merah (diprediksikan sebagai Zona Hijau) sebesar 3%. Model ANN memberikan hasil *true positif* di atas 95% untuk masing-masing zona.

Berdasarkan *confusion matrix* dari hasil pengujian SVM, RF, KNN dan ANN menggunakan dataset "mingguan", seluruh model 100% berhasil dalam memprediksi Zona Hijau, tetapi memiliki *error* yang besar (di atas 62%) dalam memprediksi Zona Merah, kecuali ANN (error 5%). Secara umum, ANN memberikan hasil prediksi terbaik (99.90%) jika dibandingkan dengan SVM, RF dan KNN. Kesalahan yang ditimbulkan dalam memprediksi Zona Merah pada SVM (error 62%), RF (error 76%) dan KNN (error 62%) disebabkan oleh jumlah data (klasifikasi Zona Merah) yang sangat kecil (1.3%) sehingga tidak mampu ditangani ketiga model tersebut. Hasil akurasi dan *error* dari setiap pengujian *machine learning* dengan dataset "mingguan" ditunjukkan pada Tabel V.



Gambar 12. *Confusion matrix* ANN dataset "mingguan" ternormalisasi

TABEL V
HASIL *CONFUSION MATRIX* DATASET "MINGGUAN"

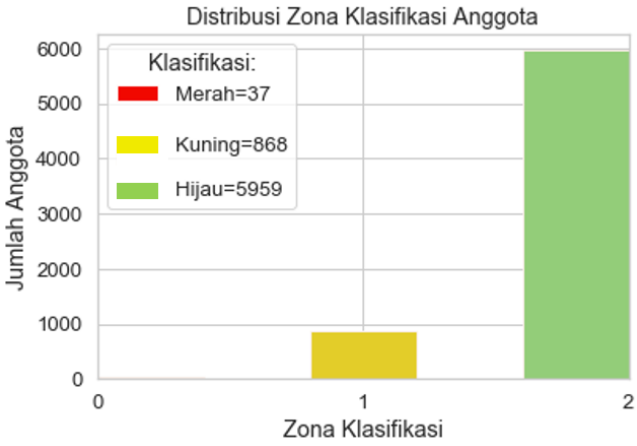
	SVM	**RF**	**KNN**	**ANN**
Klasifikasi Tepat	10481.00	10501.00	10633.00	10734.00
Klasifikasi Gagal	263.00	243.00	111.00	10.00
Akurasi %	97.55	97.73	98.96	99.90
Error %	2.45	2.27	1.04	0.10

B. *Evaluasi Hasil Confusion Matrix Dataset "Bulanan"*

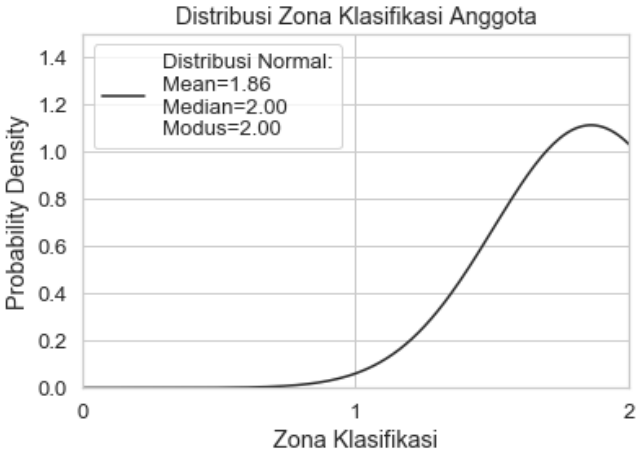
Evaluasi hasil pengujian *machine learning* pada dataset "bulanan" dilakukan dengan teknik yang sama dengan evaluasi hasil pengujian pada dataset "mingguan".

Berdasarkan *confusion matrix* dari hasil pengujian SVM, RF, KNN dan ANN menggunakan dataset "bulanan", seluruh model memiliki akurasi 100% dalam memprediksi Zona Hijau (*true positive*), tetapi memiliki akurasi 0% dalam memprediksi Zona Merah (100% *false positive*). Kesalahan 100% yang ditimbulkan dalam memprediksi Zona Merah disebabkan oleh jumlah data *training* (klasifikasi Zona Merah) yang sangat kecil (0.5%), yaitu sebanyak 37 data (Gambar 13), sehingga menyebabkan distribusi miring negatif. Pada Gambar 14 (Zona Merah diwakili 0, Zona Kuning diwakili 1, Zona Hijau diwakili 2) dapat dilihat bahwa distribusi miring negatif dengan *Mean* lebih kecil daripada *Median* dan *Modus* [26]. Oleh sebab itu, algoritma SVM, RF, KNN dan ANN tidak berhasil dalam mengklasifikasikan Zona Merah pada dataset "bulanan".

p-ISSN : 2443-2210
e-ISSN : 2443-2229



Gambar 13. Sebaran jumlah anggota pada setiap zona klasifikasi dataset "bulanan"



Gambar 14. Distribusi miring negatif dataset "bulanan"

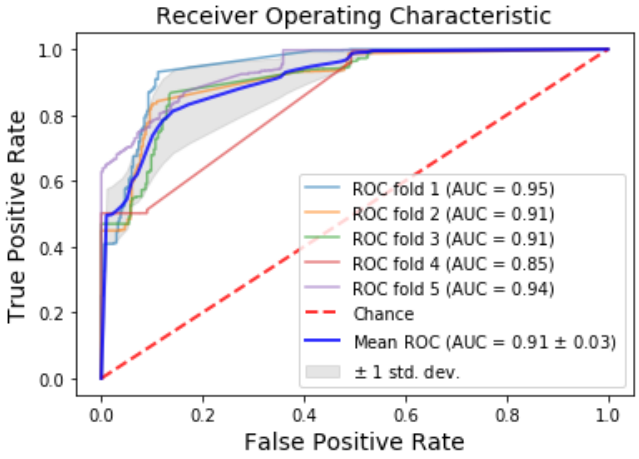
Hasil akurasi dan *error* dari setiap pengujian *machine learning* dengan dataset "bulanan" ditunjukkan pada Tabel VI.

TABEL VI
HASIL *CONFUSION MATRIX* DATASET "BULANAN"

	SVM	**RF**	**KNN**	**ANN**
Klasifikasi Tepat	2767.00	2927.00	2889.00	2921.00
Klasifikasi Gagal	175.00	15.00	53.00	21.00
Akurasi %	94.05	99.49	98.19	99.28
Error %	5.95	0.51	1.81	0.72

C. Evaluasi Hasil Kurva ROC pada Dataset "Mingguan"

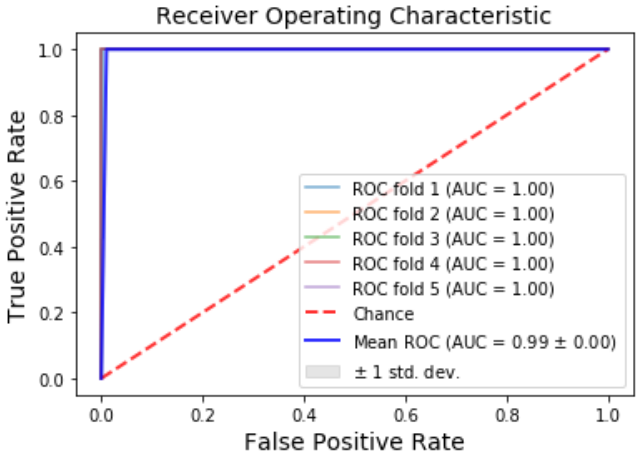
Kurva ROC pada Gambar 15 menunjukkan bahwa nilai AUC berkisar antara 0.85 hingga 0.95 dari hasil *5-fold cross validation*. Nilai AUC yang berbeda-beda pada setiap *fold* menghasilkan simpangan sebesar ± 0.03. Model SVM dengan nilai rata-rata AUC 0.91 ± 0.03 yang ditunjukkan Gambar 15 akan memberikan performa baik sebab memiliki nilai *true positive* yang tinggi.



Gambar 15. Kurva ROC pengujian SVM pada dataset "mingguan"

Dari kurva ROC, dapat dilihat bahwa AUC menyederhanakan ROC menjadi satu nilai. Sumbu x kurva ROC menunjukkan *false positive rate*, sedangkan sumbu y menunjukkan *true positive rate*. AUC yang dibentuk merupakan rata-rata (*mean*) dari 5 AUC hasil *5-fold cross validation*. AUC menunjukkan hasil yang berbeda dengan akurasi *confusion matrix*.

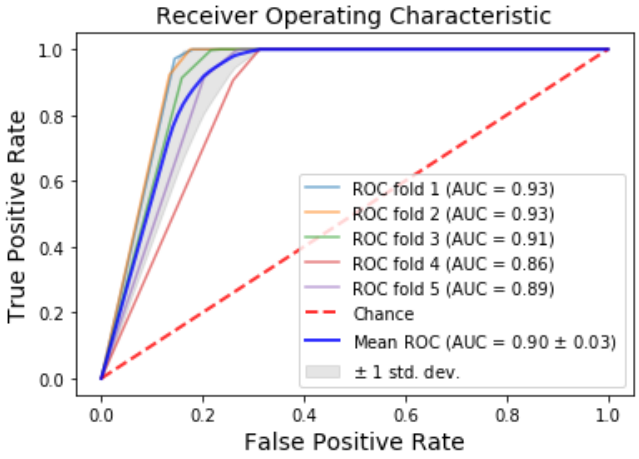
Kurva ROC pada pengujian model RF yang ditunjukkan oleh Gambar 16 memberikan nilai rata-rata AUC yang sangat tinggi (0.99). Model dengan nilai AUC mendekati 1.00 akan memberikan performa maksimal sebab memiliki nilai *true positive* yang tinggi.



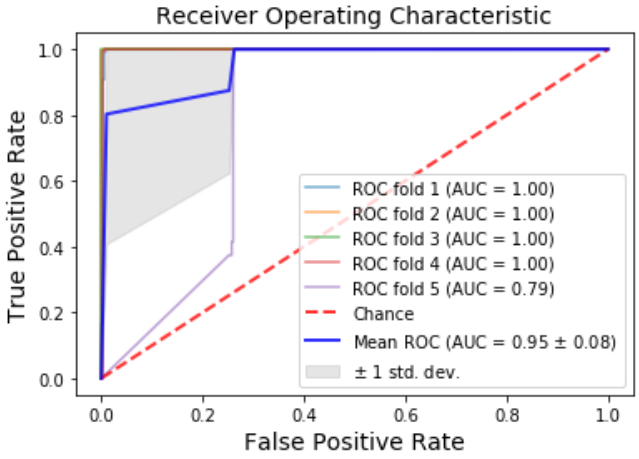
Gambar 16. Kurva ROC pengujian RF pada dataset "mingguan"

Kurva ROC pada pengujian model KNN (Gambar 17) menunjukkan kurva yang landai. Hal ini menandakan bahwa terdapat hasil *false positive* yang besar pada model. Walaupun memiliki nilai rata-rata AUC 0.9, tetapi model ini memiliki simpangan ± 0.03. Model KNN tidak akan memberikan performa maksimal sebab memiliki nilai *false*

positive besar yang menyebabkan kurva ROC jauh dari sumbu *true positive*.



Gambar 17. Kurva ROC pengujian KNN pada dataset "mingguan"



Gambar 18. Kurva ROC pengujian ANN pada dataset "mingguan"

Gambar 18 memperlihatkan kurva ROC pada pengujian model ANN yang menunjukkan nilai rata-rata AUC 0.95. Empat pengujian (*fold*) menunjukkan nilai AUC 1.00, yang menandakan bahwa model berfungsi secara maksimal karena memiliki nilai *true positif* yang tinggi. *Fold* kelima yang menunjukkan nilai AUC 0.79 menyebabkan nilai simpangan (standar deviasi) ± 0.08. Perhitungan standar deviasi dilakukan dengan menggunakan persamaan berikut [27]:

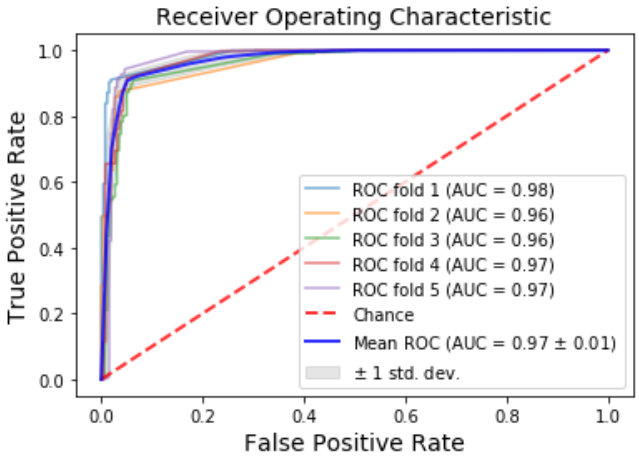
$$Standard\ Deviation = \sqrt{mean(abs(x - \bar{x})^2)} \qquad (2)$$

Standar Deviasi pada kurva ROC ANN adalah simpangan nilai AUC dari hasil pengujian yang dilakukan sebanyak lima *fold.* Walaupun simpangannya ± 0.08 pada nilai rata-rata AUC 0.95, tetapi nilai maksimal AUC adalah 1.

Kurva ROC menunjukkan bahwa dalam pengujian dataset "mingguan", RF (0.99) memiliki nilai AUC paling baik dibandingkan dengan SVM (0.91 ± 0.03), KNN (0.90 ± 0.03) dan ANN (0.95 ± 0.08).

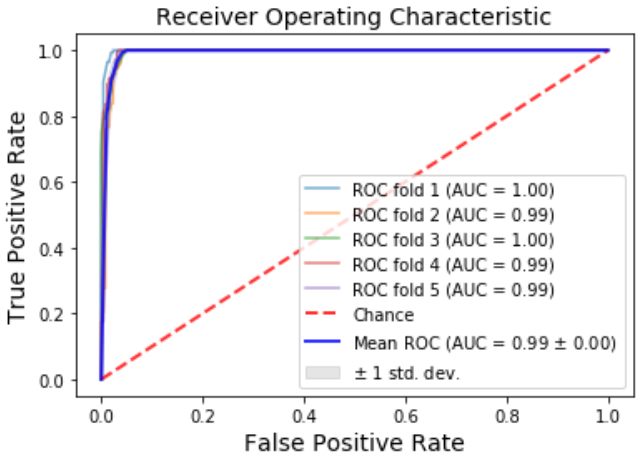
D. Evaluasi Hasil Kurva ROC pada Dataset "Bulanan"

Kurva ROC pada pengujian SVM menunjukkan nilai rata-rata AUC 0.97 dengan simpangan ± 0.01. Pada Gambar 19 dapat dilihat bahwa model memiliki nilai *true positive* yang tinggi sehingga akan memberikan performa baik.



Gambar 19. Kurva ROC pengujian SVM pada dataset "bulanan"

Pada Gambar 20, diperlihatkan kurva ROC pada pengujian RF memiliki nilai rata-rata AUC 0.99. Model dengan nilai AUC mendekati 1.00 akan memberikan performa maksimal sebab memiliki nilai *true positive* yang tinggi.

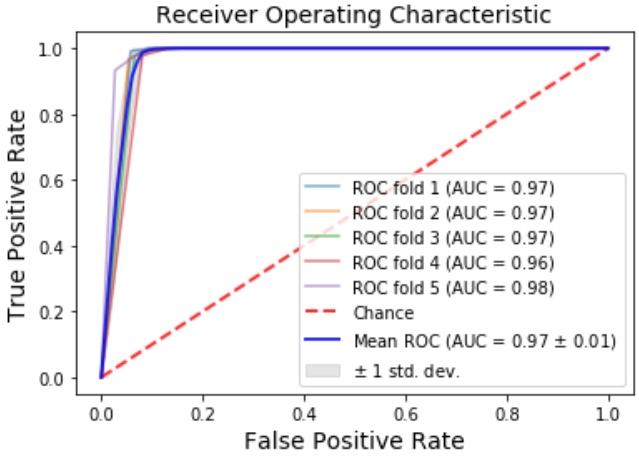


Gambar 20. Kurva ROC pengujian RF pada dataset "bulanan"

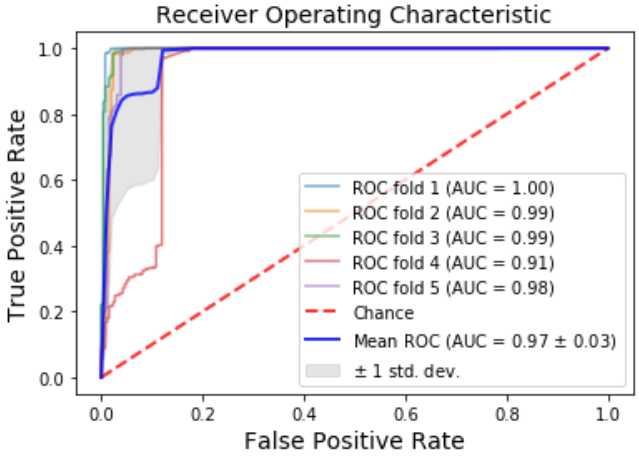
Kurva ROC pada pengujian KNN dengan dataset "bulanan" seperti yang diperlihatkan pada Gambar 21, menunjukkan hasil yang lebih baik daripada kurva ROC pengujian KNN pada dataset "migguan" (Gambar 17). Hal ini ditunjukkan dengan nilai rata-rata AUC 0.97 dengan simpangan ± 0.01 dari hasil *5-fold cross validation.* Model

p-ISSN : 2443-2210
e-ISSN : 2443-2229

KNN dengan kurva ROC pada Gambar 21 akan memberikan performa yang baik karena memiliki nilai *true positive* yang tinggi



Gambar 21. Kurva ROC pengujian KNN pada dataset "bulanan"



Gambar 22. Kurva ROC pengujian ANN pada dataset "bulanan"

Pada pengujian ANN, kurva ROC yang dihasilkan menunjukkan nilai rata-rata AUC 0.97 dengan simpangan sebesar ± 0.03 (Gambar 22). Pada *fold* keempat, AUC menunjukkan nilai terendah (0.91) yang mengakibatkan simpangan (standar deviasi) nilai rata-rata AUC menjadi ± 0.03. Nilai AUC yang rendah pada *fold* keempat disebabkan oleh nilai *false positive* yang tinggi, akibat *error* dalam mengklasifikasikan Zona Merah. Walaupun memiliki simpangan ± 0.03, model ANN akan memberikan performa yang baik sebab memiliki nilai *true positive* yang tinggi.

Kurva ROC dalam pengujian dataset "bulanan" menunjukkan bahwa RF (0.99) memiliki nilai AUC paling baik dibandingkan dengan SVM (0.97 ± 0.01), KNN (0.97 ± 0.01) dan ANN (0.97 ± 0.03). Hal unik terjadi pada kurva ROC ANN dengan simpangan AUC sebesar ± 0.03, yang menyebabkan nilai rata-rata AUC dari kurva ROC ANN tidak lebih baik daripada SVM dan KNN.

E. Evaluasi Bias dan Variansi

Evaluasi bias dan variansi dilakukan terhadap pengujian dataset "mingguan" dan dataset "bulanan". Rincian bias dan variansi ditampilkan pada Tabel VII dan Tabel VIII.

TABEL VII
BIAS DAN VARIANSI *MACHINE LEARNING* DATASET "MINGGUAN"

	SVM	**RF**	**KNN**	**ANN**
Akurasi *Training* %	96.95	96.95	98.89	99.76
Akurasi *Testing* %	97.55	97.73	98.96	99.90
Bias %	3.05	3.05	1.11	0.24
Variansi %	0.6	0.08	0.07	0.14
Hasil	*undefitting*	*undefitting*	*low bias low variance*	*low bias low variance*

TABEL VIII
BIAS DAN VARIANSI *MACHINE LEARNING* DATASET "BULANAN"

	SVM	**RF**	**KNN**	**ANN**
Akurasi *Training* %	94.83	99.26	98.25	99.01
Akurasi *Testing* %	94.05	99.49	98.19	99.28
Bias %	5.17	0.74	1.75	0.99
Variansi %	0.78	0.23	0.06	0.27
Hasil	*undefitting*	*low bias low variance*	*low bias low variance*	*low bias low variance*

Berdasarkan hasil evaluasi bias dan variansi, KNN dan ANN memiliki performa yang baik (*low bias low variance*) pada dataset "mingguan" dan dataset "bulanan". ANN menunjukkan hasil bias dan variansi di bawah 0.25% pada dataset "mingguan". Hal ini sejalan dengan hasil akurasi di atas 99% (tertinggi dibandingkan SVM, RF dan KNN) yang ditunjukkannya pada saat *training* dan *testing*. RF memiliki hasil bias dan variansi di bawah 0.75% pada dataset "bulanan". Hal ini sejalan dengan hasil akurasi di atas 99% (terbaik dibandingkan SVM, KNN dan ANN) dan nilai AUC 0.99 yang ditunjukkannya.

V. KESIMPULAN

Deteksi dini terhadap anggota yang menunjukkan ciri-ciri akan mengundurkan diri pada sebuah pusat kebugaran berhasil dilakukan menggunakan metode *Support Vector Machine* (SVM), *Random Forest* (RF), *K-Nearest Neighbor* (KNN) dan *Artificial Neural Network* (ANN) pada dataset "mingguan" dan "bulanan", dengan hasil berupa klasifikasi yang terbagi ke dalam tiga zona yaitu Zona Hijau, Zona Kuning dan Zona Merah. Klasifikasi ditentukan berdasarkan frekuensi kehadiran, pencapaian anggota, serta biaya yang dibebankan sebagai deteksi dini untuk menentukan langkah yang harus diambil manajer dan pelatih industri kebugaran

terhadap anggota pusat kebugaran. Jika dibandingkan dengan ketiga metode lainnya, hasil pengujian metode ANN pada dataset "mingguan" menunjukkan tingkat akurasi tinggi (99.90%), memiliki nilai *true positive* yang tinggi, serta memiliki nilai bias (0.24%) dan variansi (0.11%) yang kecil, sehingga anggota klub kebugaran berhasil diklasifikasikan ke dalam tiga zona dengan *error rate* yang rendah (*error rate* tertinggi terdapat pada klasifikasi Zona Merah sebesar 5.20%).

DAFTAR PUSTAKA

[1] W. R. Thompson, "WORLDWIDE SURVEY OF FITNESS TRENDS FOR 2019," *ACSM's Health & Fitness Journal*, vol. 22, no. 6, pp. 10-17, 2018.

[2] A. Ferrand, L. Robinson, P. Valette-Florence, "The Intention-to-Repurchase Paradox: A Case of the Health and Fitness Industry," *Journal of Sport Management*, vol. 24,no. 1, pp. 83-105, 2010.

[3] Y. Nieto, V. Garcia-Diaz, C. Montenegro, R. G. Crespo, "Supporting academic decision making at higher educational institutions using machine learning-based algorithms," *Soft Computing*, vol. 23, no. 12, pp. 4145-4153, 2019.

[4] Y. Nieto, V. Garcia-Diaz, C. Montenegro, C. C. Gonzalez, R. G. Crespo, "Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions," *IEEE Access*, vol. 7, pp. 75007-75017, 2019.

[5] R. Mehta, *Big Data Analytics with Java*, Packt Publishing, 2017.

[6] Samsudiney. (2019) Penjelasan Sederhana tentang Apa Itu SVM. [Online]. Tersedia: https://medium.com/@samsudiney/penjelasan-sederhana-tentang-apa-itu-svm-149fec72bd02/

[7] A.S. Nugroho, A. B. Witarto, D. Handoko. (2003) Support Vector Machine Teori dan aplikasinya dalam Bioinformatika. [Online]. Tersedia: http://asnugroho.net/papers/ikcsvm.pdf/

[8] T. You. (2019) Understanding Random Forest. [Online]. Tersedia: https://towardsdatascience.com/understanding-random-forest-58381e0602d2/

[9] J. Needham, *Science and Civilisation in China: Volume 3, Mathematics and the Sciences of the Heavens and the Earth*, Cambridge University Press, 1959.

[10] M. Dash & H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1,no. 1-4, pp. 131-156, 1997.

[11] R. Agarwal. (2019) The 5 Feature Selection Algorithms every Data Scientists should know. [Online]. Tersedia: https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2/

[12] R. G. van den Berg. (2020) Pearson Correlations - Quick Introduction. [Online]. Tersedia: https://www.spss-tutorials.com/pearson-correlation-coefficient/

[13] R. D. Mason & A. Lind. Douglas, *Teknik Statistik untuk Bisnis dan Ekonomi*, Jilid 2, Jakarta, Penerbit Erlangga, 1999.

[14] J. Brownlee. (2018) A Gentle Introduction to k-fold Cross-Validation. [Online]. Tersedia: https://machinelearningmastery.com/k-fold-cross-validation/

[15] K. Q. Weinberger, L. K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *Journal of Machine Learning Research*, vol. 10, no. 9, pp. 207-244, 2009.

[16] A. M. Ismail. (2018) Cara Kerja Algoritma k-Nearest Neighbor (k-NN). [Online]. Tersedia: https://medium.com/bee-solution-partners/cara-kerja-algoritma-k-nearest-neighbor-k-nn-389297de543e/

[17] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, M. Ettaouil, "Multilayer perceptron: architecture optimization and training," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4,no. 1, pp. 26, 2016.

[18] S. Robinson. (2018) Introduction to Neural Networks with Scikit-Learn. [Online]. Tersedia: https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/

[19] J. Heaton, *Introduction to neural networks with Java*, vol. 99, 2nd ed., Heaton Research Inc, St. Louis, 2008.

[20] A. M. Hay, "The derivation of global estimates from a confusion matrix," *International Journal of Remote Sensing*, vol. 9, no. 8, pp. 1395-1398, 1988.

[21] K. S. Nugroho. (2019) Confusion Matrix untuk Evaluasi Model pada Supervised Learning. [Online]. Tersedia: https://medium.com/@ksnugroho/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f/

[22] J. Fan, S. Upadhye, A. Worster, "Understanding receiver operating characteristic (ROC) curves," *CJEM*, vol. 8, no. 1, pp. 19-20, 2006.

[23] S. Narkhede. (2018) Understanding AUC - ROC Curve. [Online]. Tersedia: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5/

[24] A. Ng. (2018) Machine Learning Yearning. [Online]. Tersedia: https://www.deeplearning.ai/machine-learning-yearning/

[25] P. Philipp, L. Schreiter, J. Giehl, Y. Fischer, J. Raczkowsky, M. Schwarz, H. Wörn, J. Beyerer, "Situation Detection for an Interactive Assistance in Surgical Interventions Based on Dynamic Bayesian Networks," *Prosiding 6th Joint Workshop on New Technologies for Computer/Robot Assisted Surgery*, 2016, paper 45, p. 117.

[26] J. K. Patel & C. B. Read, *Handbook of the Normal Distribution*, 2nd ed., CRC Press, 1996.

[27] M. Gupta. (2020) numpy.std() in Python. [Online]. Tersedia: https://www.geeksforgeeks.org/numpy-std-in-python/