p-ISSN: 1979-5661
e-ISSN: 2622-321X

Perbandingan Metode *Artificial Neural Network* (ANN) dan *Support Vector Machine* (SVM) untuk Klasifikasi Kinerja Perusahaan Daerah Air Minum (PDAM) di Indonesia

[1]Pardomuan Robinson Sihombing, [2] Oki Prasetia Hendarsin

[1]Badan Pusat Statistik,Indonesia,
robinson@bps.go.id,

[2]Universitas Padjadjaran, Indonesia
prasetiaoki@gmail.com

Abstrak

Indikator kinerja Perusahan Daerah Air Minum (PDAM) dapat diartikan sebagai suatu ukuran yang dapat digunakan untuk memberikan gambaran tingkat keberhasilan kegiatan pengelolaan PDAM. Tingkat keberhasilan pengelolaan PDAM ini diukur melalui proses penilaian terhadap kinerja PDAM yang didasarkan pada indikator kinerja penyelenggaraan pengembangan SPAM meliputi: aspek keuangan, operasional, pelayanan pelanggan dan sumber daya manusia sesuai dengan ketentuan di dalam Pasal 59 Permen PU No. 18/PRT/M/2007. Berdasarkan aspek dan indikator diatas, PDAM dapat memprediksikan kinerja mereka untuk tahun berjalan. Namun butuh banyak indikator perhitungannya cenderung rumit dan memerlukan audit internal PDAM terlebih dahulu yang mana membutuhkan waktu banyak. Dalam penelitian ini bertujuan membuat model dari algoritma *Artificial Neural Network* (ANN) dan *Support Vector Machine* (SVM) untuk mengklasifikasikan kinerja PDAM berdasarkan indikator terpilih. Hasil penelitian menunjukan perusahaan dapat menggunakan model klasifikasi ANN untuk memprediksi kinerja perusahaan di tahun berjalan dengan menggunakan 3 atribut yaitu Rasio Operasi, Jam Operasi Layanan/hari, dan Rasio Jumlah Pegawai/1000 pelanggan dengan rata-rata akurasi 83.93% dan tingkat presisi prediksi untuk kinerja Tidak Sehat sebesar 86.36%. Hal ini lebih baik apabila dibandingkan dengan algoritma SVM model terpilih yaitu memiliki rata-rata akurasi 82.14% dan tingkat presisi untuk kinerja Tidak Sehat sebesar 80%.
kata kunci: kinerja, *vector machine*, *neural network*

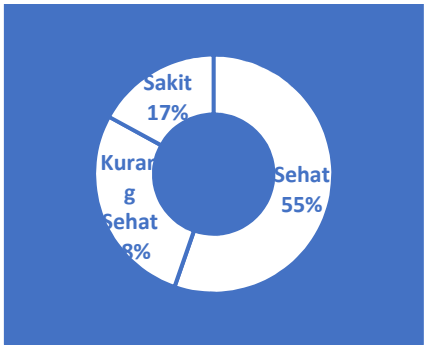
Abstract

The performance indicators of the Regional Water Supply Company (RWSC/PDAM) can be interpreted as a measure that can be used to provide a picture of the success rate of RWSC management activities. The level of success of RWSC management is measured through a process of evaluating RWSC performance based on performance indicators for the implementation of SPAM development including: financial, operational, customer service and human resource aspects in accordance with the provisions in Article 59 of PU Permen No. 18 / PRT / M / 2007. Based on the above aspects and indicators, RWSCs can predict their performance for the current year. But it takes a lot of indicators the calculation tends to be complicated and requires an internal audit of theRWSC in advance which requires a lot of time. In this study aims to create a model of Artificial Neural Network (ANN) and Support Vector Machine (SVM) algorithms to classify RWSC performance based on selected indicators. The results shows that company can use the ANN classification model to predict company performance in the current year by using 3 attributes, namely Operational Ratio, Operating Hours of Service / day, and Ratio of Number of Employees / 1000 customers with an average accuracy level of 83.93% and the level of prediction precision for Unhealthy performance amounted to 86.36%. This is better than the SVM algorithm of the chosen model which has an average accuracy of 82.14% and a precision level for Unhealthy performance of 80%.
keywords: performance, vector machine, neural network

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Pendahuluan

Berdasarkan data Badan Pendukung Pengembangan Sistem Penyediaan Air Minum (BPPSPAM), pada Tahun 2017 dari 376 PDAM yang dinilai terdapat 208 berkinerja "Sehat", 104 "Kurang Sehat", dan 64 "Sakit"



Sumber : Buku Kinerja PDAM 2017, BPPSPAM

Gambar 1 Kondisi Kinerja PDAM di Indonesia Tahun 2017

Penilaian dilakukan menggunakan kriteria BPPSPAM yang merupakan 4 aspek dengan 18 indikator yang berbeda

Tabel 1. Aspek dan Indikator Penilaian Kinerja PDAM

Aspek Keuangan	Aspek Pelayanan	Aspek Operasional	Aspek SDM
Return on Equity	Cakupan Pelayanan Teknis	Efisiensi Produksi	Rasio Pegawai terhadap Pelanggan
Rasio Operasi	Pertumbuhan Pelanggan	Air Tak Berekening	Rasio Diklat Pegawai
Rasio Kas	Tingkat Penyelesaian Pengaduan	Jam Operasi Layanan	Rasio Beban Diklat terhadap Beban Pegawai
Efektifitas Penagihan	Kualitas Air Pelanggan	Tekanan Air pada Sambungan Pelanggan	
Solvabilitas	Konsumsi Air Domestik	Penggantian Meter Air Pelanggan	

BPPSPAM menggunakan hasil audit Badan Pengawasan Keuangan dan Pembangunan (BPKP) untuk menilai kinerja PDAM. Berdasarkan aspek dan indikator diatas, PDAM dapat memprediksikan kinerja mereka untuk tahun berjalan. Namun butuh banyak indikator perhitungannya cenderung rumit dan memerlukan audit internal PDAM terlebih dahulu yang mana membutuhkan waktu banyak. Salah satunya bila ingin menghitung Air Tak Berekening, maka PDAM membutuhkan pengukuran yang dapat memakan waktu dan biaya yang tinggi.

Oleh karena itu dibutuhkan model dengan variabel indikator yang sedikit namun dapat memprediksi kinerja perusahaan dengan tingkat akurasi yang tinggi. Algoritma klasifikasi dapat digunakan untuk membentuk model prediksi. Beberapa algoritma klasifikasi yang dapat digunakan adalah *Artificial Neural Network* (ANN) dan *Support Vector Machine* (SVM).

2. Kajian Teoritis

Indikator kinerja PDAM dapat diartikan sebagai suatu ukuran yang dapat digunakan untuk memberikan gambaran tingkat keberhasilan kegiatan pengelolaan PDAM. Tingkat keberhasilan pengelolaan PDAM ini diukur melalui proses penilaian terhadap kinerja PDAM yang didasarkan pada indikator kinerja penyelenggaraan pengembangan SPAM meliputi: aspek keuangan, operasional, pelayanan pelanggan dan sumber daya manusia sesuai dengan ketentuan di dalam Pasal 59 Permen PU No. 18/PRT/M/2007. Masing-masing aspek dirinci ke dalam beberapa indikator penilaian melalui pendekatan *balanced score card* (Petunjuk Teknis Penilaian Kinerja PDAM, BPPSPAM). Penilaian Kinerja PDAM terdiri dari 3 kategori yaitu Sehat, Kurang Sehat, dan Sakit.

p-ISSN: 1979-5661
e-ISSN: 2622-321X

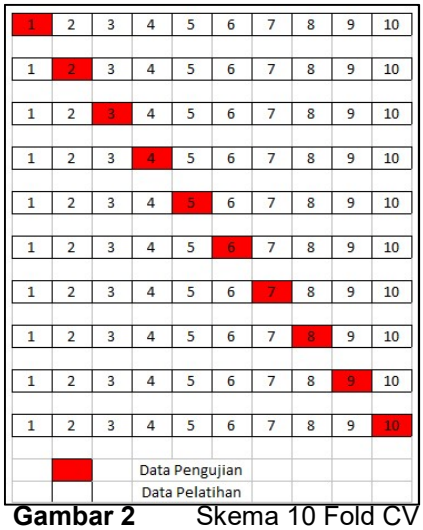
2.1. Algoritma Klasifikasi

Klasifikasi merupakan operasi untuk memisahkan beragam entitas kedalam beberapa kelas (*Nisbet, 2009*: 236). Pengklasifikasian merupakan pelatihan atau pembelajaran terhadap fungsi *f* (target/label) yang memetakan setiap atribut *x* ke satu dari jumlah label kelas *y* yang tersedia.

Apabila target kelas sudah diketahui maka proses klasifikasi terasuk dalam *supervised* dan apabila dataset belum memiliki target kelas maka termasuk *unsupervised* contohnya adalah proses klaster. Penelitian ini menggunakan teknik klasifikasi *supervised* yaitu dengan target kelas kinerja Sehat dan Tidak Sehat dengan mengkomparasi 2 metode yaitu *Artificial Neural Network* (ANN) dan *Support Vector Machine* (SVM). SVM dan ANN merupakan algorithma klasifikasi yang digunakan untuk menilai objek data dan memasukkan data kedalam kelas tertentu. Dalam klasifikasi dilakukan pembangunan model dan menggunakan model tersebut untuk melakukan klasifikasi/prediksi pada suatu obyek agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpan.

2.2.1. *Cross Validation* (CV)

Cross-validation (CV) adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dipisahkan menjadi dua subset yaitu data proses pembelajaran dan data validasi / evaluasi. Model atau algoritma dilatih oleh subset pembelajaran dan divalidasi oleh subset validasi. Selanjutnya pemilihan jenis CV dapat didasarkan pada ukuran dataset. Biasanya CV K-fold digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi.



Gambar 2 Skema 10 Fold CV

10 fold CV adalah salah satu K fold CV yang direkomendasikan untuk pemilihan model terbaik karena cenderung memberikan estimasi akurasi yang kurang bias dibandingkan dengan CV biasa, leave-one-out CV dan bootstrap. Dalam 10 fold CV, data dibagi menjadi 10 fold berukuran kira-kira sama, sehingga kita memiliki 10 subset data untuk mengevaluasi kinerja model atau algoritma. Untuk masing-masing dari 10 subset data tersebut, CV akan menggunakan 9 fold untuk pelatihan dan 1 fold untuk pengujian seperti diilustrasikan pada Gambar 1.

Misalkan t_l^kdan y_l^k (k=1,2,3,..,10 dan l=1,2,3,...,Ncv) menjadi nilai tragte dan nilai prediksi pengamatan data l dalam subset k. Maka, kinerja model atau algoritma dapat diperkirakan dengan *Mean Squared Error Cross-Validation* (MSECV) sebagai berikut:

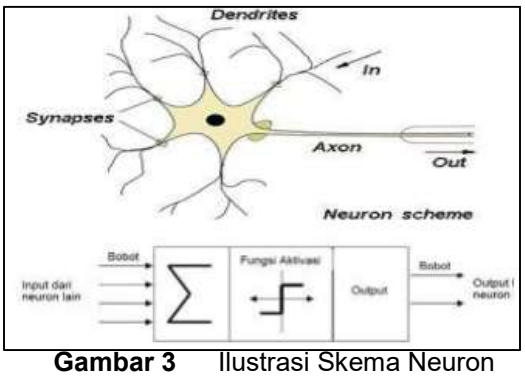
$$MSECV = \frac{1}{10}\frac{1}{N_{cv}}\sum_{k=1}^{10}\sum_{l=1}^{N_{cv}}\left(t_l^k - y_l^k\right)^2 \quad (2.1)$$

Kita menggunakan CV untuk memilih model yang sesuai dengan membandingkan nilai mean square error dari coss-validation (MSECV). Model atau algoritma terbaik dipilih jika memiliki nilai MSECV terendah disbanding yang lain. Selanjutnya, misalkan $\mathrm{t_m^{val}}$ dan $\mathrm{y_m^{val}}$ adalah nilai target dan prediksi nilai untuk validasi, dan M adalah data validasi (m=1,2,3,..,M). Selanjutnya, kita menggunakan *mean square error* (MSE): $MSE_{val} = \frac{1}{M}\sum_{m=1}^{M}(t_m^{val} - y_m^{val})^2$ untuk melakukan validasi model terbaik.

p-ISSN: 1979-5661
e-ISSN: 2622-321X

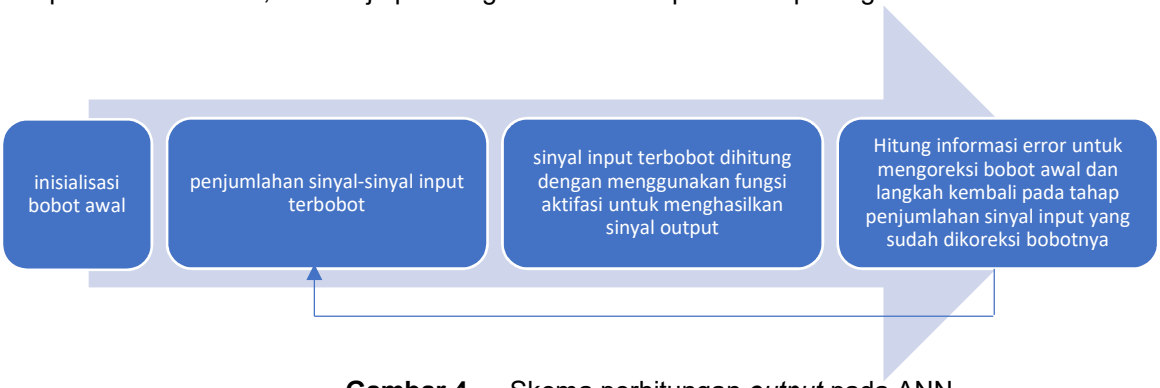
2.2.2. *Artificial Neural Network* (ANN)

Artificial Neural Network atau yang biasa disebut Jaringan Saraf Tiruan (JST) merupakan sistem pengolahan data yang terinspirasi dari konfigurasi otak manusia. pada dasarnya terbuat dari neuron buatan yang diidentifikasi sebagai konstituen pemrosesan yang saling berhubungan bertindak sama sekali untuk mencapai masalah tertentu (Khademi et al., 2016).



Gambar 3 Ilustrasi Skema Neuron

Cara kerja ANN seperti cara kerja manusia, yaitu belajar melalui contoh. Lapisan-lapisan penyusun ANN dibagi menjadi 3, yaitu lapisan input (input layer), lapisan tersembunyi (hidden layer), dan lapisan output (ouput layer). Hidden layer terhubung ke layer lain dengan bobot, bias dan fungsi transfer. Fungsi error ditentukan oleh perbedaan antara output jaringan dan target. Error disebarkan kembali serta bobot dan bias disesuaikan dengan menggunakan beberapa teknik optimasi yang meminimalkan error. Seluruh proses yang disebut training diulang sejumlah epoch sampai keakuratan yang diinginkan dalam output tercapai. Setelah jaringan di-training, jaringan tersebut dapat digunakan untuk memvalidasi data yang tidak terlihat dengan menggunakan bobot dan bias. Adapun secara umum, alur kerja pada algoritma ANN dapat dilihat pada gambar berikut



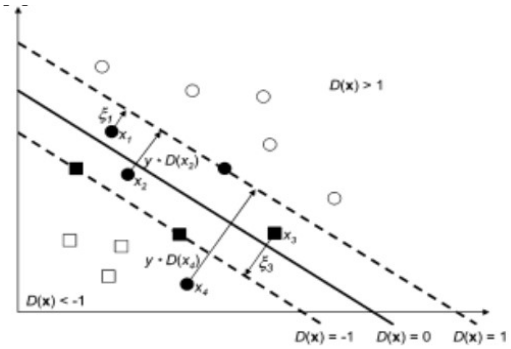
Gambar 4 Skema perhitungan *output* pada ANN

2.2.3. *Support Vector Machine (SVM)*

Support Vector Machines (SVM) adalah algoritma supervised learning yang sangat kuat yang digunakan untuk klasifikasi atau regresi (Burges, 1998). SVM dapat digunakan untuk prediksi numerik serta klasifikasi (Han, Kamber,& Pei, 2012). SVM adalah algoritma yang berfungsi untuk pemetaan nonlinier dan mengubah data pelatihan asli ke dimensi yang lebih tinggi. Di dalam dimensi baru ini, LR mencari linear optimal dengan memisahkan hyperplane. Dengan pemetaan nonlinier yang sesuai ke dalam dimensi yang cukup tinggi, data dari dua kelas selalu dapat dipisahkan oleh hyperplane. SVM menemukan hyperplane tersebut dengan menggunakan vektor dukungan dan margin. Dua sifat khusus dari SVM yaitu (1) mencapai generalisasi yang tinggi dengan memaksimalkan margin, dan (2) mendukung pembelajaran yang efisien dari fungsi nonlinier pada trik kernel sehingga membuat kinerja generalisasinya baik dalam menyelesaikan masalah pengenalan pola (Gorunescu, 2011). Pada permasalahan klasifikasi SVM mencoba untuk mencari garis pemisah yang optimal yang diekspresikan sebagai kombinasi linier dari subset data training dengan menyelesaikan masalah keterbatasan linier pemrograman quadrat (QP) dengan margin maksimum antara dua kelas.

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Pengklasifikasi SVM dilakukan dengan prosedur dua langkah: Pertama, vektor data sampel dipetakan ("diproyeksikan") ke ruang dimensi. Dimensi ruang ini secara signifikan lebih besar dari dimensi ruang data asli. Kemudian, algoritma digunakan untuk menemukan hyperplane di ruang ini dengan margin terbesar yang memisahkan kelas data. Hal itu menunjukkan bahwa keakuratan klasifikasi biasanya hanya akan turun tipis pada proyeksi tertentu. Ide dasar SVM adalah memaksimalkan batas hyperplan (maximal margin hyperplane). Konsep klasifikasi dengan SVM dapat dijelaskan secara sederhana sebagai usaha untuk mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah kelas data pada ruang input.



Gambar 5 *Decision Boundary* dengan Marginal Maksimum

Gambar diatas memperlihatkan beberapa pola yang merupakan anggota dari dua buah kelas data : +1 dan -1 . Data yang tergabung pada kelas -1 disimbolkan dengan bentuk kotak bujur sangkar dan data pada kelas +1 disimbolkan dengan bentuk lingkaran. Hyperplane (batas keputusan) pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur margin hyperplane tersebut dan mencari titik maksimalnya. Margin adalah jarak antara hyperplane dengan data terdekat dari masing-masing kelas. Data yang paling dekat inilah disebut support vector. Garis solid pada gambar 1 sebelah kanan menunjukkan hyperplane yang terbaik yaitu yang terletak tepat pada tengah-tengah kedua kelas, sedangkan data lingkaran dan bujur sangkar yang dilewati garis batas margin (garis putus-putus) adalah support vector. Usaha untuk mencari lokasi hyperplane ini merupakan proses pembelajaran. Konsep Klasifikasi dengan SVM adalah mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua kelas data. SVM mampu bekerja pada dataset yang berdimensi tinggi dengan menggunakan kernel trik. SVM hanya menggunakan beberapa titik data terpilih yang berkontribusi (Support Vector) untuk membentuk model yang akan digunakan dalam proses klasifikasi.

- Titik data : $x_i = \{x_1, x_2, \ldots, x_n\} \in R^n$
- Kelas data : $y_i \in \{-1, +1\}$
- Pasangan data dan kelas : $\{(x_i, y_i)\}_{i=1}^{N}$
- Maksimalkan fungsi berikut :

$$Ld = \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \;\; syarat \;\; : 0 \le \alpha_i \le C \;\; dan \;\; \sum_{i=1}^{N} \alpha_i y_i = 0$$

- Hitung nilai w dan b :

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \;\; b = -\frac{1}{2}\left(w.x^{+} + w.x^{-}\right)$$

- Fungsi keputusan klasifikasi sign(f(x)) :

$$f(x) = w.x + b \quad atau \quad f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b$$

Keterangan :

N = banyaknya data
n = dimensi data atau banyaknya fitur
Ld = Dualitas Lagrange Multipier
αi = nilai bobot setiap titik data
C = nilai konstanta
m = jumlah support vector/titik data yangmemiliki αi > 0)
K (x,xi) = fungsi kernel

p-ISSN: 1979-5661
e-ISSN: 2622-321X

3. Metodologi

Data yang digunakan merupakan data Kinerja PDAM Tahun 2016, dengan 376 instansi PDAM seluruh Indonesia sebagai observasi beserta 1 variabel kinerja sebagai label dan 18 sub-aspek sebagai variabel atribut. Untuk variabel kinerja dibentuk menjadi 2 kelas yaitu kelas kinerja Sehat untuk PDAM berkinerja Sehat dan kelas kinerja Tidak Sehat untuk PDAM berkinerja Kurang Sehat dan Sakit.

3.1 *Data Preparation*

Tahapan *data preparation* merupakan tahapan untuk membuat *dataset* dalam format yang tepat untuk digunakan dalam analisis pembuatan model. (*Nisbet, 2009*: 40). Tahapan ini juga disebut tahap *preprocessing* sebelum masuk dalam tahap permodelan. Pada tahapan ini dapat mencakup kegiatan transformasi data, penghapusan objek yang mengandung *missing value*, dan reduksi data untuk mengurangi jumlah data yang digunakan (*Nisbet, 2009*: 40). Pada penelitian ini, tahap *preprocessing* akan melakukan reduksi data dalam hal ini reduksi variabel untuk mendapatkan model yang efektif yang dapat digunakan oleh PDAM. Reduksi variabel atribut menggunakan metode *Principal Component Analysis* (PCA). PCA merupakan suatu teknik statistik untuk mengunbah dari sebagian besar variabel asli yang digunakan yang saling berkorelasi satu dengan lainnya menjadi satu set variabel baru yang lebih kecil dan saling bebas (Susilawati, 2011:32). Salah satu metode yang digunakan untuk menentukan komponen utama adalah berdasarkan proporsi kumulatif total varians yang mampu dijelaskan. Patokan baku batas minimum proporsi kumulatif varians berkisar antara 70%-90% (Susilawati, 2011:33).

3.2 *Modelling*

Pada tahap *modelling* akan dilakukan pembentukan model yang dapat membedakan kelas data. Singkatnya, pada tahapan *modelling* penerapan algoritma klasifikasi akan dilakukan. Untuk melakukan *modelling*, pada tahap ini dibutuhkan dua jenis data set yang berasal dari data hasil *preprocessing* yaitu *training data* dan *testing data*.

Training data merupakan dataset yang digunakan untuk membangun model sementara *Testing data* digunakan untuk menghitung *performance* dari model yang terbentuk dengan membandingkan label data sebenarnya dan label data hasil klasifikasi model. Untuk membentuk *Training data* dan *Testing data* penelitian menggunakan dua buah metode yaitu *V-Fold Cross Validation* dan secara deterministik.

a. K-Fold Cross-Validation (*K-Fold CV*)

Validasi *testing data* dilakukan terhadap dataset yang tidak digunakan pada pembentukan model. *Cross-Validation* merupakan bentuk dari *resampling* yang mengambil beberapa sampel dari keseluruhan observasi dan menjadikannya sebagai *training data* untuk model (*Nisbet, 2009*:140).

b. Deterministik

Pembagian data set menjadi *training data* dan *testing data* dapat menggunakan cara deterministik, yaitu dengan menentukan sendiri rasio pembagian dari kedua dataset tersebut. Contohnya rasio dataset dapat menggunakan 7:3, artinya 0.7 dari keseluruhan data digunakan untuk *training data* dan 0.3 sisanya digunakan untuk *testing data* yang dapat menghasilkan *performances*.

Tahap selanjutnya setelah terbentuk dua jenis dataset adalah tahap klasifikasi. Klasifikasi dalam peneltiian ini adalah menggunakan algoritma *Artificial Neural Network* dan *Support Vector Machine*

3.3 *Evaluation*

Evaluasi dilakukan untuk memilih metode pembagian data set dan klasifikasi yang mana yang dapat menghasilkan tingkat akurasi. Evaluasi dalam penelitian adalah dengan memperhatikan *Confussion Matrix*. *Confussion Matrix* merupakan sebuah alat untuk mengetahui sejauh mana pengklasifikasian dapat mengenal atau memprediksi kelas data (Han, 2012:366). *Confussion Matrix* merupakan tabel berukuran *mxm* dengan *m*=jumlah kelas. Bagian kolom diisi oleh label aktual untuk tiap kelas, sementara bagian baris diisi oleh label kelas hasil prediksi

Tabel 2. *Confusion Matrix* (Han, 2012:366)

Confusion Matrix		*Actual Class*		*Total*
		Yes	*No*	
Predicted Class	*Yes*	*TP*	*FP*	*P'*
	No	*FN*	*TN*	*N'*
Total		*P*	*N*	

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Berdasarkan tabel diatas, perhitungan akurasi dan tingkat kekeliruan klasifikasi untuk keseluruhan pengklasifikasian adalah

$$Akurasi = \frac{TP + TN}{P + N} \quad (3.1)$$

$$Error\ Rate = \frac{FP + FN}{P + N} \quad (3.2)$$

Perhitungan di atas diterapkan untuk kelas dengan masing-masing jumlah anggota relative seimbang. Apabila terjadi masalah ketidakseimbangan kelas (*class imbalance*), dimana kelas yang menjadi perhatian berjumlah sedikit, contohnya seperti contoh kasus penyakit yang sangat jarang, maka dapat memperhatikan perhitungan *sensitivity* dan *specificity* (Han, 2012:367). *Sensitivity* merupakan proporsi kelas yang menjadi perhatian/diinginkan terprediksi dengan benar. *Specificity* merupakan proporsi kelas yang tidak menjadi perhatian/tidak diinginkan terprediksi dengan benar.

$$Sensitivity = \frac{TP}{P} \quad (3.3)$$

$$Specificity = \frac{TN}{N} \quad (3.4)$$

Apabila tingkat akurasi tinggi, namun *sensitivity* rendah, maka pengklasifikasian dapat dikatakan tidak baik (Han, 2012:368)

Ukuran yang dapat digunakan dalam mengevaluasi pengklasifikasian adalah *precision* dan *recall* (Han, 2012:368). *Precisions* merupakan ukuran ketepatan dari proses pengklasifikasian, atau juga proporsi klasifikasi positif hasil prediksi yang benar terhadap seluruh hasil prediksi positif *Recall* merupakan ukuran *completeness* dari proses klasifikasi, atau juga proporsi klasifikasi posiif hasil prediksi yang benar terhadap seluruh positif aktual.

$$Precision = \frac{TP}{TP + FP} \quad (3.5)$$

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{P} \quad (3.6)$$

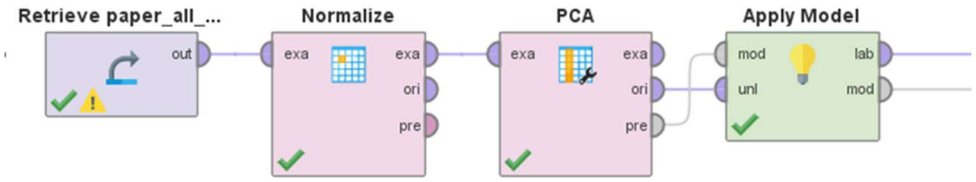
Untuk kasus class imbalance dapat menggunakan *F1-Score* (Han, 2012:369) yang merupakan rata-rata harmonik dari *precision* dan *recall* sebagai tingkat akurasi

$$F1 = \frac{2\ x\ precision\ x\ recall}{precision + recall} \quad (3.7)$$

4. Hasil dan Pembahasan

4.1 Reduksi Variabel dengan PCA

Reduksi variabel dengan metode PCA menggunakan bantuan *software* RapidMiner Studio dengan alur kerja seperti Gambar 4.1 dan menghasilkan Tabel 4.2 yang merupakan proporsi varians dari komponen utama



Gambar 6 *Alur* Kerja PCA dengan RapidMiner

Tabel 3. Proporsi Varians Komponen Utama

Component	*Std Deviation*	*Proportion of Variance*	*Cummulative variance*
PC 1	1.948	0.211	0.211
PC 2	1.264	0.089	0.299
PC 3	1.2	0.08	0.379
PC 4	1.196	0.079	0.459
PC 5	1.12	0.07	0.529
PC 6	1.023	0.058	0.587

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Component	*Std Deviation*	*Proportion of Variance*	*Cummulative variance*
PC 7	0.984	0.054	0.641
PC 8	0.904	0.045	0.686
PC 9	0.898	0.045	0.731
PC 10	0.866	0.042	0.773
PC 11	0.827	0.038	0.811
PC 12	0.822	0.038	0.848
PC 13	0.776	0.033	0.882
PC 14	0.75	0.031	0.913
PC 15	0.718	0.029	0.942
PC 16	0.629	0.022	0.964
PC 17	0.602	0.02	0.984
PC 18	0.543	0.016	1

Berdasarkan tabel diatas variabel atribut terpilih akan diambil dari komponen utama PC 1 karena memiliki proporsi varians terbesar dibanding 17 komponen utama lainnya. Untuk memeilih variabel atribut akan memperhatikan tabel *Eigen Vector* hasil *output* Rapidminer. Selanjutnya berdasarkan tabel *Eigen Vector* di bawah ini ditentukan variabel atribut yang digunakan dalam *modelling* adalah yang memiliki absolut nilai eigen diatas 0.3, Maka terpilih 3 variabel atribut untuk tahap *modelling* yaitu Rasio Operasi, Jam Operasional Layanan, dan Rasio Jumlah Pegawai/1000 Pelanggan.

Tabel 4. *Eigen Vectors* Komponen Utama PC 1

Attribute	PC 1
R O E	0.153
Ratio Operasi	-0.314
Ratio Kas	0.004
Efektivitas Penagihan	0.259
Solvabilitas	-0.067
Cakupan Pelayanan	0.246
Pertumbuhan Pelanggan	-0.116
Tingkat Penyelesaian Pengaduan	0.218
Kualitas Air Pelanggan	0.274
Konsumsi Air Domestik	0.29
Effisiensi Produksi	0.285
Tingkat Kehilangan air	-0.276
Jam Operasi Layanan / hari	0.321
Tekanan Sambungan Pelanggan	0.179
Penggantian Meter Air	0.173
Rasio juml peg /1000 plg	-0.332
Ratio diklat pegawai/peningkatan kompetensi	0.274
Biaya Diklat terhadap Biaya Pegawai	0.135

4.2.1 Data Olah

Proses reduksi variabel menghasilkan 3 variabel atribut untuk masuk pada tahap *modelling* yaitu variabel Rasio operasi, Jam Operasional Layanan, dan Rasio Jumlah Pegawai
Data hasil reduksi variabel menggunakan PCA digunakan untuk masuk dalam tahap *Modelling*

4.3 *Modelling*

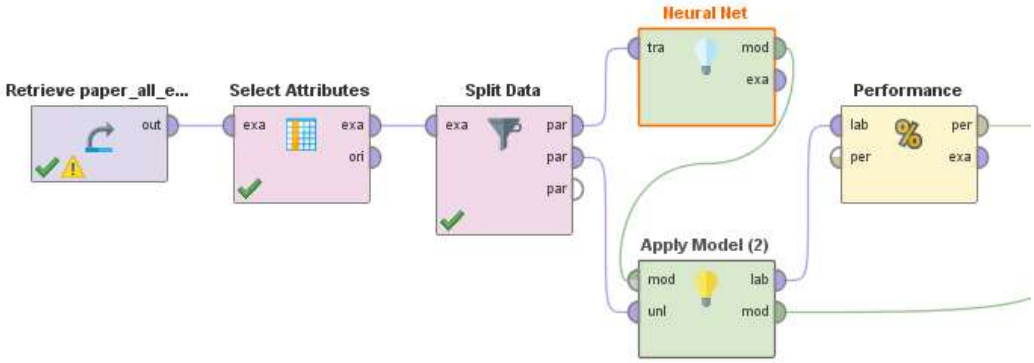
Pada tahap *modelling* akan dilakukan klasifikasi terhadap data. Sebelum dilakukan klasifikasi, data terlebih dahulu proses *split* data yaitu membagi data menjadi subset *data training* dan *data testing* Adapun tujuan dari pembagian subset data tersebut adalah menggunakan *data training* sebagai dasar pembuatan model dan menggunakan *data testing* digunakan sebagai dasar evaluasi atas diterapkannya model pada data eksisting.

4.3.1 *Split Data*

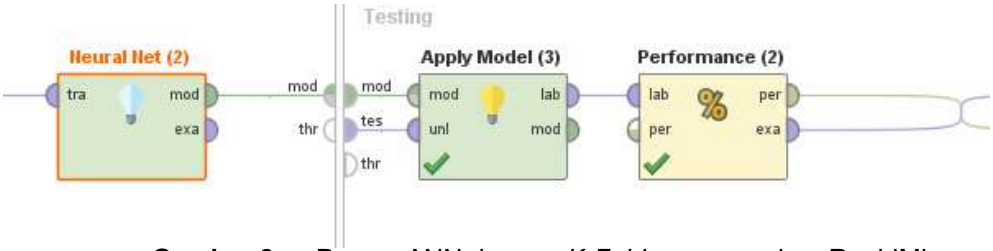
Proses *split data* dalam penelitian ini menggunakan dua Teknik yaitu secara deterministik dan *K-fold Cross Validation*. Proses *split data* secara deterministik adalah dengan menentukan proporsi data yang akan dijadikan *data training* dan *data testing* yang dalam penelitian ini ditentukan 0.7 bagian data digunakan untuk *data training* dan 0.3 bagian data digunakan untuk *data testing*. Proses *split data* menggunakan *K-fold Cross Validation* adalah dengan menjadikan (1-(1/K)) bagian data menjadi *data training* dan (1/K) bagian data sisanya menjadi *data testing* yang dalam penelitian ini menggunakan *K=5, K=7,* dan *K=10*

4.3.2 Klasifikasi

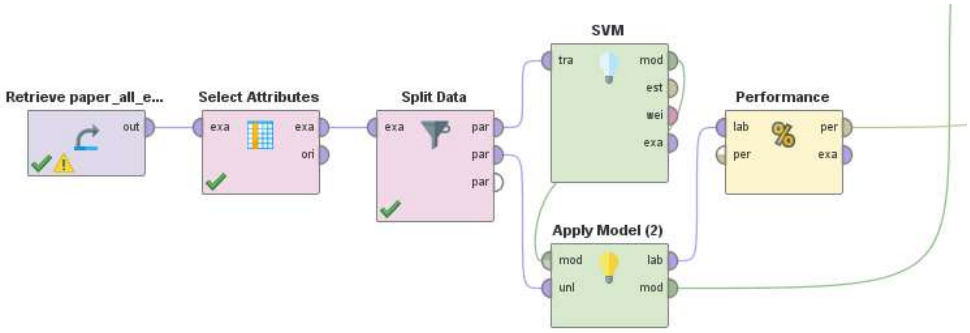
Proses klasifikasi dalam penelitian ini menggunakan software RapidMiner Studio sebagai alat bantu. Arsitektur proses *Artificial Neural Network (ANN)* dalam penelitian ini ditetapkan menggunakan 1 *Hidden Layer* dan 4 *Neuron Hidden*. Parameter ANN yang ditetapkan adalah *training cycle=200, momentum=0.9,* dan *error epsilon=1.0E-4.* Adapun untuk parameter yang dijadikan parameter pembanding adalah *learning rate* yang ditetapkan sebesar 0.1, 0.3, dan 0.5. Sementara untuk proses *Support Vector* Machine (SVM) parameter yang ditetapkan adalah menggunakan fungsi kernel *dot*, *kernel cache=200, convergence epsilon=0.001,* dan iterasi maksimal 100.000 Adapun untuk parameter yang dijadikan parameter pembanding adalah nilai *C* konstan yang merupakan toleransi dari kekeliruan misklasifikasi. Nilai C yang tinggi menunjukan semakin *soft* margin yang digunakan. Resiko dari menggunakan nilai *C* yang tinggi bias menimbulkan *overfitting* pada model, sementara apabila nilai *C* terlalu rendah membuat model terlalu menggeneralisir dalam klasifikasi. Nilai *C* yang digunakan dalam penelitian ini adalah 0.0, 0.5, dan 1.0. Adapun tahap alur kerja dari proses *ANN* dan *SVM* menggunakan RapidMiner Studio adalah sebagai berikut



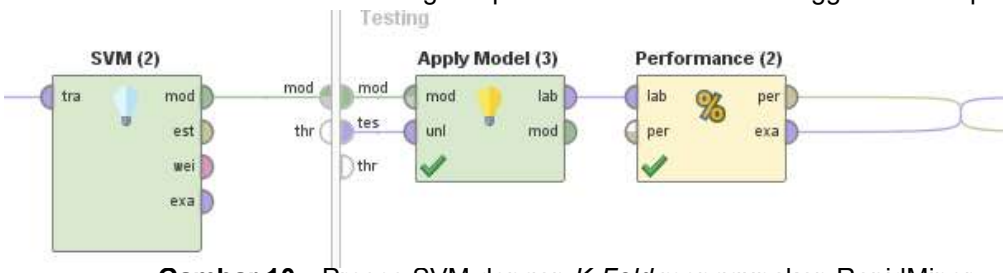
Gambar 7 Proses ANN dengan Split Data Deterministik menggunakan RapidMiner



Gambar 8 Proses ANN dengan *K-Fold* menggunakan RapidMiner



Gambar 9 Proses SVM dengan Split Data Deterministik menggunakan RapidMiner



Gambar 10 Proses SVM dengan *K-Fold* menggunakan RapidMiner

Proses klasifikasi menghasilkan ukuran-ukuran *performance* yang dapat dijadikan acuan evaluasi pemilihan model klasifikasi yang baik untuk data.

4.4 *Evaluasi*

Pada tahap evaluasi akan dilakukan dua tahap peninjauan ukuran-ukuran *performances*. Tahap pertama adalah memperhatikan tingkat akurasi secara keseluruhan yang dihasilkan oleh model-model klasifikasi serta nilai RMSE. Untuk model terpilih diinginkan model yang memiliki nilai akurasi relatif tinggi dengan nilai RMSE relatif rendah. Setelah terpilih model pada Tahap pertama, maka proses maju pada Tahap Kedua, yaitu mengevaluasi ukuran *precision* dan *recall* dari model terpilih. *Precision* merupakan ukuran ketepatan model klasifikasi dalam mengklasifikasikan data yang diinginkan/relevan atau dapat dikatakan merupakan ukuran dari kualitas model klasifikasi. *Recall* merupakan ukuran seberapa lengkap model klasifikasi dapat mendeteksi data yang dinginkan/relevan atau dapat dikatakan merupakan ukuran dari kuantitas model klasifikasi.

Tahap 1

Berdasarkan Tabel 4.5, dapat dilihat berdasarkan tingkat akurasi tertinggi dan RMSE terendah terdapat pada proses ANN dengan spit data deterministik dan *Learning Rate=0.3*. Pada Proses SVM, rata-rata tingkat akurasi tertinggi terdapat pada proses split deterministik dan untuk setiap *C* yang digunakan bernilai sama yaitu 82.14%, artinya dari 100% data, 82.14% terklasifikasikan dengan benar oleh model. Untuk nilai RMSE terendah ada pada SVM dengan nilai *C*=1.0 dengan proses split deterministik. Oleh karena itu model klasifikasi yang masuk dalam tahap kedua evaluasi adalah model ANN dengan LR=0.3 dan SVM dengan C=1.0

Tabel 5. Nilai Akurasi dan RMSE Model-Model Klasifikasi

Accuracy (%)		*Split*	*K-Fold Cross Validation*		
		7:3	*K=5*	*K=7*	*K=10*
Artificial Neural Network	*LR = 0.1*	*83.04*	*81.07% +/- 5.11%*	*81.37% +/- 7.11%*	*79.45% +/- 6.34%*
	LR = 0.3	*83.93*	*80.00% +/- 5.73%*	*80.29% +/- 6.87%*	*78.65% +/- 8.20%*
	LR = 0.5	*83.93*	*77.87% +/- 5.55%*	*81.10% +/- 7.39%*	*79.74% +/- 7.55%*
SVM (Kernel : Dot)	*C = 0.0*	*82.14*	*82.13% +/- 6.37%*	*81.34% +/- 4.53%*	*82.40% +/- 7.03%*
	C = 0.5	*82.14*	*82.13% +/- 5.63%*	*80.81% +/- 4.48%*	*82.13% +/- 7.01%*
	C = 1.0	*82.14*	*81.87% +/- 5.70%*	*81.07% +/- 4.19%*	*81.60% +/- 6.83%*

RMSE		Split	K-Fold Cross Validation		
		7:3	K=5	K=7	K=10
Artificial Neural Network	LR = 0.1	0.341 +/- 0.000	0.377 +/- 0.049	0.366 +/- 0.047	0.376 +/- 0.048
	LR = 0.3	0.346 +/- 0.000	0.387 +/- 0.059	0.378 +/- 0.053	0.400 +/- 0.066
	LR = 0.5	0.369 +/- 0.000	0.412 +/- 0.059	0.395 +/- 0.065	0.400 +/- 0.081
SVM (Kernel : Dot)	C = 0.0	0.340 +/- 0.000	0.366 +/- 0.022	0.363 +/- 0.023	0.363 +/- 0.027
	C = 0.5	0.339 +/- 0.000	0.364 +/- 0.022	0.363 +/- 0.023	0.363 +/- 0.028
	C = 1.0	0.337 +/- 0.000	0.363 +/- 0.022	0.362 +/- 0.024	0.361 +/- 0.028

Tahap 2

Pada tahap kedua akan *Confusion Matrix* dari masing-masing hasil klasifikasi yang terpilih pada tahap pertama. Berdasarkan *Confusion Matrix* yang terbentuk akan diperhatikan kelas data yang relevan, yaitu kinerja Tidak Sehat. Kinerja Tidak Sehat dipilih sebagai kelas perhatian dikarenakan apabila perusahaan diprediksi akan berkinerja Sehat padahal pada kenyataannya berkinerja Tidak Sehat, maka hal ini akan dapat merugikan perusahaan tersebut karena terjadi kekeliruan prediksi. Apabila terjadi kekeliruan perusahaan diprediksi berkinerja Tidak Sehat padahal pada kenyataanya berkinerja Sehat, maka perusahaan akan berusaha lebih untuk mendapatkan hasil prediksi berkinerja Sehat dengan mendorong atribut yang digunakan dalam model agar bernilai baik.

Tabel 6. *Confusion Matrix* Model terpilih

ANN dengan LR=0.3

ANN ; LR = 0.3 ; Split	true TIDAK SEHAT	true SEHAT	class precision
pred. TIDAK SEHAT	38	6	86.36%
pred. SEHAT	12	56	82.35%
class recall	76.00%	90.32%	

SVM dengan C=1.0

SVM ; C = 1 ; Split	true TIDAK SEHAT	true SEHAT	class precision
pred. TIDAK SEHAT	40	10	80.00%
pred. SEHAT	10	52	83.87%
class recall	80.00%	83.87%	

Berdasarkan tabel *Confusion Matrix* diatas dapat dikatakan bahwa untuk model ANN dengan LR=0.3 memiliki nilai *precision* 86%, artinya model dapat memprediksi perusahaan berkinerja Tidak Sehat dengan tepat sebesar 86%, sementara untuk nilai *recall* 76% memiliki arti model dapat menghasilkan prediksi sejumlah 76% perusahaan yang berkinerja Tidak Sehat dari seluruh perusahaan yang berkinerja Tidak Sehat pada data aktual.

Model SVM dengan C=1.0 memiliki nilai *precision* 80%, artinya model dapat memprediksi perusahaan berkinerja Tidak Sehat dengan tepat sebesar 80%, sementara untuk nilai *recall* 80% memiliki arti model dapat menghasilkan prediksi sejumlah 80% perusahaan yang berkinerja Tidak Sehat dari seluruh perusahaan yang berkinerja Tidak Sehat pada data actual.

Berdasarkan informasi diatas dapat dikatakan bahwa untuk data yang digunakan, apabila ingin menghasilkan model klasifikasi yang memperhatikan kualitas hasil maka dapat digunakan model ANN dengan LR=0.3 sementara apabilia diinginkan model yang dapat memprediksi lebih lengkap secara kuantitas maka dapat menggunakan model SVM dengan C=1.0. Kembali pada rumusan masalah pada peelitian ini, maka model klasifikasi yang digunakan adalah model dengan kualitas hasil prediksi tertinggi yaitu model ANN dengan 1 *Hidden Layer* dan 4 *Neuron Hidden* dengan LR=0.3 dan metode validasi *split data* secara deterministik.

5.Kesimpulan

Tujuan dari penelitian ini adalah membentuk model klasifikasi yang dapat memprediksi kinerja PDAM dengan indikator yang sedikit namun memiliki tingkat akurasi yang tinggi. Berdasarkan bab sebelumnya yaitu hasil dan pembahasan, maka dapat diambil kesimpulan sebagai berikut :

1. Model Klasifikasi yang terebntuk dalam penelitian ini adalah *Neural Network* dengan *1 Hidden Layer* dan 4 *Neuron Hidden* dengan LR=0.3 dan metode validasi *split data* dengan 70% untuk *training data* dan 30% untuk *testing data.*
2. Penelitian ini menghasilkan metode validasi yang memiliki tingkat rata-rata akurasi yang paling tinggi adalah secara deterministik dibandingkan dengan metode *k-fold cross validation.*
3. Perusahaan dapat menggunakan model klasifikasi *Artificial Neural Network* untuk memprediksi kinerja perusahaan di tahun berjalan dengan menggunakan 3 atribut yaitu Rasio Operasi, Jam Operasi Layanan/hari, dan Rasio Jumlah Pegawai/1000 pelanggan dengan rata-rata akurasi 83.93% dan tingkat presisi prediksi untuk kinerja Tidak Sehat sebesar 86.36%. Hal ini sedikit lebih baik apabila dibandingkan dengan algoritma SVM model terpilih yaitu memiliki rata-rata akurasi 82.14% dan tingkat presisi untuk kinerja Tidak Sehat sebesar 80%.

DAFTAR PUSTAKA

Badan Peningkatan Penyelenggaraan Sistem Penyediaan Air Minum Kementerian Pekerjaan Umum dan Perumahan Rakyat, Buku Petunjuk Teknis Penilaian Kinerja PDAM.

Burges, C. 1998. *A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2, 2.*

Gorunescu, F. 2011. *Data Mining: Concepts, models and techniques (2011th ed.).* Springer.

Han, Jiawei, *Micheline Kamber dan Jian Pei. 2012. Data Mining: Concepts and Techniques 3rd Edition*. Massachusetts: Elsevier Inc

Khademi, F., Jamal, S. M., Deshpande, N., & Londhe,S. (2016). *Predicting strength of recycled aggregate concrete using Artificial Neural Network, Adaptive Neuro-Fuzzy Inference System and Multiple Linear Regression. International Journal of Sustainable Built Environment*, 5(2). https://doi.org/10.1016/j.ijsbe.2016.09.003

Nisbet, Robert, John Elder, dan Gary Miner. 2009. *Handbook of Statistical Analysis and Data Mining Applications*. California: Elsevier Inc.

Susilawati, Sumarni. 2011. *Analisis Komponen Utama dalam memonitor Pengendalian Kualitas Produksi Karet*. Makassar: Skripsi Universitas Islam Negeri Alauddin

Sutojo, T., et al, 2010, Kecerdasan Buatan, Yogyakarta: Andi Offset.