**Klasifikasi kanker payudara menggunakan model CNN dan Support Vector Machine**

Indra Cahya Ramdani 1\* , Sugiyarto 2

1 Jurusan Matematika, Fakultas Terapan dan Teknologi, Universitas A hmad Dahlan, Yogyakarta , Indonesia

2 1 Jurusan Matematika, Fakultas Terapan dan Teknologi, Universitas A hmad Dahlan, Yogyakarta , Indonesia

Email: indra1800015053 @webmail.uad.ac.id 1 , Sugiyarto@math.uad.ac.id 2

**Abstrak**

Jaringan Syaraf Tiruan merupakan sistem komputasi yang terinspirasi dari cara kerja sistem saraf pada manusia dan terus berkembang pesat hingga sekarang. Sama seperti sistem saraf pada manusia, jaringan saraf tiruan bekerja melalui proses mempelajari data yang ada untuk merumuskan keluaran data baru. Jaringan syaraf tiruan dengan menggunakan metode Recurrent Neural Network (RNN) merupakan salah satu model yang populer digunakan saat ini terutama pada kasus peramalan. Secara sederhana, alur peramalan dengan menggunakan metode RNN dimulai dengan membagi data uji dan data latih, proses perhitungan maju, proses perhitungan mundur, perhitungan optimasi, dan perhitungan evaluasi model peramalan. Kendala utama dari metode RNN adalah adanya vanishing gradient yang dapat menyebabkan hasil peramalan yang kurang baik. Dalam penelitian ini, penulis mengusulkan metode reduksi dimensi Principal Component Analysis (PCA) untuk mendapatkan variabel yang paling berpengaruh dan menjadi masukan untuk model prediksi yang dibangun untuk meminimalkan kesalahan yang ada. Penulis juga menggunakan metode K-means clustering untuk membagi data dengan variasi trend yang sama. Untuk meningkatkan efek clustering, peneliti menggunakan perhitungan kemiripan berdasarkan jarak Euclidean. Sehingga dalam upaya membangun hasil prediksi yang optimal akan dipilih data first time series dengan variabel yang paling berpengaruh menggunakan metode PCA. Selanjutnya data tersebut dikelompokkan menggunakan metode K-means dan akan dimasukkan ke dalam model prediksi yang dibangun. Pada model prediksi RNN, data akan dilatih menggunakan metode Backpropagation Through Time (BPTT) dan metode optimasi yang digunakan adalah Stochastic Gradient Descent (SGD). Peramalan dengan metode RNN dengan PCA menghasilkan akurasi sebesar 93%, sedangkan peramalan dengan menggunakan metode RNN tanpa PCA menghasilkan akurasi sebesar 82%. Hasil eksperimen menunjukkan bahwa metode RNN dengan PCA mencapai akurasi prediksi dan fleksibilitas yang lebih tinggi daripada RNN tanpa PCA.

*Kata kunci* : PCA , K -means Clustering , RNN , BPTT, SGD

**pengantar**

Peramalan adalah bidang ilmu untuk memprediksi peristiwa yang akan terjadi di masa depan dengan menggunakan data masa lalu dan saat ini dan memproyeksikannya ke dalam model matematika (Ariyanto et al., 2017) . Secara sederhana ada 2 metode umum peramalan, yaitu metode kualitatif dan metode kuantitatif. Metode kualitatif adalah metode yang digunakan ketika tidak ada data historis dan bersifat intuitif, sehingga tidak mungkin dilakukan perhitungan matematis. Umumnya metode kualitatif diperoleh dari pendapat seorang ahli sebagai bahan pertimbangan dalam pengambilan keputusan. Sedangkan metode kuantitatif adalah metode yang digunakan apabila terdapat data historis, sehingga perhitungan matematis dapat dilakukan secara outetode (Maricar, 2019) .

Metode yang sering digunakan untuk peramalan adalah metode kuantitatif dengan menggunakan data time series. Data runtun waktu adalah data yang dikumpulkan berdasarkan urutan waktu tertentu untuk menggambarkan perkembangan suatu keadaan (Desmonda et al., 2018) . Beberapa metode peramalan yang menggunakan data time series adalah *Fuzzy* (Sumartini et al., 2017) , *Markov Chain* (Rukhansah et al., 2016) , ARIMA (Nur Hadiansyah, 2017) , *Monte Carlo* (Moh. Jufriyanto, 2020) , dan *Neural Jaringan* (SAWITRI dkk., 2018) . Belakangan ini, *Recurrent Neural Network* (RNN) telah menjadi salah satu metode peramalan yang menjanjikan karena aplikasinya yang luas untuk peramalan dan kinerja generalisasi yang sangat tinggi.

RNN pertama kali dikembangkan oleh Jeff Elman pada tahun 1990 dan merupakan salah satu arsitektur dari Neural Network (Novita, 2016) . RNN termasuk dalam kategori Deep Learning karena data diproses melalui banyak lapisan (Firmansyah et al., 2020) . Keunikan metode RNN adalah memiliki deskripsi yang sangat baik dan dapat mengatasi kelemahan feedforward. Keunikan lain dari RNN adalah terdapat koneksi umpan balik yang membawa informasi noise pada saat input sebelumnya yang akan ditampung untuk input berikutnya (Informasi, 2017) .

Dalam penelitian yang dilakukan (Cao et al., 2020) , Cao et al menggunakan model GRU dalam peramalan untuk mendapatkan nilai akurasi yang lebih baik. Pengembangan yang disarankan adalah penggunaan optimasi yang berbeda untuk mendapatkan akurasi yang lebih baik dari penelitian. Maka dalam penelitian ini penulis menggunakan model prediksi RNN dengan data training menggunakan *Backpropagation Through Time* (BPTT). Pengembangan lain yang dilakukan pada penelitian ini adalah penggunaan fungsi aktivasi *Stochastic Gradient Descent* (SGD) pada saat pelatihan data. Yang mana sebelum data diolah menjadi model RNN maka data time series akan diolah menjadi 2 tahap sebagai berikut. Pertama, data time series akan menyeleksi faktor yang paling berpengaruh menggunakan metode *Principal Component Analysis* (PCA). Banyaknya faktor yang digunakan dalam peramalan seringkali disertai dengan kompleksitas yang tinggi yang akan menyebabkan beban komputasi yang besar. Selain itu, pengaruh faktor utama dan faktor pendukung memiliki peranan penting dalam peramalan sehingga hanya faktor-faktor penting yang akan digunakan. Oleh karena itu perlu dilakukan pemilihan fitur, salah satunya dengan menggunakan PCA. Tujuan seleksi fitur adalah untuk mereduksi faktor atau variabel agar menghasilkan lebih sedikit kompleksitas dalam analisis peramalan.

Kedua, membagi data deret waktu dan mengelompokkan data yang memiliki tren serupa menggunakan metode K-means Clustering. Jika data time series diprediksi secara langsung, maka akurasi prediksi yang diperoleh cenderung rendah karena karakteristik dan tren variasi data yang berbeda, sehingga perlu dilakukan pengelompokan data untuk mendapatkan hasil prediksi yang optimal. Untuk meningkatkan efek clustering, peneliti menggunakan perhitungan kemiripan berdasarkan jarak Euclidean dan jarak Manhattan. Selanjutnya data tersebut akan diolah menjadi model RNN.

Metode ini akan diterapkan pada kasus peramalan kualitas air sungai dan iklim harian. Tingkat kualitas air dan iklim sehari-hari saat ini menjadi salah satu topik pembicaraan terutama di kota-kota besar. Manusia sebagai makhluk hidup tidak mungkin dapat hidup tanpa air, oleh karena itu sungai sebagai sumber air perlu dijaga agar dapat berfungsi secara lestari. Beberapa penelitian terkait peramalan kualitas air sungai menggunakan metode Neural Network (Wu et al., 2021) , Analisis Regresi (Shakhari et al., 2019) , Fuzzy Neural Network (Sun & He, 2018) , hingga Adaptive Neuro-Fuzzy Inference Sistem (ANFIS) (Evangelista et al., 2020) .

**Metodologi**

1. Matriks Kovarian

Misalkan .....,adalah pengukuran n di kolom pertama data. Pengukuran rata-rata sampel adalah

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

Secara umum untuk variabel ke- *j* , jika terdapat m variabel dan n adalah jumlah data, maka rata-rata sampel adalah sebagai berikut.

|  |  |  |
| --- | --- | --- |
|  |  | ( 2 ) |

Varians sampel untuk variabel ke - *j*

|  |  |  |
| --- | --- | --- |
|  | = | ( 3 ) |

Kovarians sampel untuk variabel ke -j dan ke- *h*

|  |  |  |
| --- | --- | --- |
|  | = | ( 4 ) |

1. Nilai Eigen dan Vektor Eigen

Definisi 1 (sLuis & Moncayo, nd) Jika adalah suatu matriks maka vektor tak nol x di disebut vektor eigen dari jika x adalah kelipatan skalar dari x yaitu

|  |  |  |
| --- | --- | --- |
|  |  | ( 5 ) |

Untuk skalar disebut nilai eigen dari dan x disebut vektor eigen yang sesuai dengan .

ukuran matriks maka dapat dituliskan sebagai berikut :

|  |  |  |
| --- | --- | --- |
|  |  | ( 6 ) |

Persamaan (6) setara dengan

|  |  |  |
| --- | --- | --- |
|  |  | ( 7 ) |

Dimana . Persamaan (7) akan memiliki solusi bukan-nol jika

|  |  |  |
| --- | --- | --- |
|  |  | ( 8 ) |

Persamaan (8) disebut persamaan karakteristik dengan skalar yang memenuhi persamaan ini adalah nilai eigen dari . Jika diekspansikan, maka *det* adalah polinomial yang disebut polinomial karakteristik If adalah matriks ukuran , maka polinomial karakteristik tersebut memiliki derajat n dan adalah 1. Polinomial karakteristik dari matriks memiliki bentuk sebagai berikut.

|  |  |  |
| --- | --- | --- |
|  |  | ( 9 ) |

1. Analisis Komponen Utama (PCA)

Definisi 2 (Luis & Moncayo, nd) Suatu vektor dikatakan kombinasi linier dari vektor-vektor jika skalarnya sedemikian sehingga

|  |  |  |
| --- | --- | --- |
|  |  | (1 0 ) |

Komponen utama hanya bergantung pada matriks kovarians ( matriks korelasi ) dari , maka komponen utama yang dibentuk berdasarkan kombinasi linier adalah sebagai berikut.

|  |  |  |
| --- | --- | --- |
|  |  | (1 1 ) |

Dapat juga dituliskan dalam persamaan matriks sebagai berikut.

|  |  |  |
| --- | --- | --- |
|  |  | (1 2 ) |

Dengan adalah transpos matriks vektor eigen, PC dan merupakan matriks variabel baru (komponen utama) dan matriks data. Komponen utama adalah kombinasi linier yang tidak berkorelasi dan memiliki varians maksimum. Menurut Breezin (Brézin, 2002) jika PC= dan diketahui , maka dari persamaan tersebut diperoleh varians dari masing-masing komponen utama yaitu

|  |  |  |
| --- | --- | --- |
|  |  | (1 3 ) |

Untuk menentukan vektor koefisien pembobotan komponen utama pertama dapat dilakukan dengan menggunakan fungsi Lagrange sebagai berikut:

|  |  |  |
| --- | --- | --- |
|  |  | (1 4 ) |

Memaksimalkan varians dengan limit atau Untuk mendapatkan nilai maksimum maka digunakan limit , dengan menggunakan metode pengali Lagrange diperoleh :

|  |  |  |
| --- | --- | --- |
|  |  | (1 5 ) |

Jika persamaan dikalikan dengan dan , maka diperoleh hasil

|  |  |  |
| --- | --- | --- |
|  |  | (1 6 ) |

Karena dan begitu

|  |  |  |
| --- | --- | --- |
|  |  | (1 7 ) |

adalah komponen utama pertama yang merupakan kombinasi linier dengan tujuan memaksimumkan dengan suatu konstanta dan diperoleh bahwa

|  |  |  |
| --- | --- | --- |
|  |  | (1 8 ) |

Dimana adalah nilai eigen terbesar dari matriks ***S*** .Komponen utama pertama dapat ditulis sebagai berikut.

|  |  |  |
| --- | --- | --- |
|  |  | ( 19 ) |

Bentuk komponen utama dari- *p* yaitu merupakan kombinasi linier yang bertujuan untuk memaksimalkan varians ( andi s tidak berkorelasi dengan komponen utama lainnya tetapi ortogonal terhadap komponen utama lainnya. Batasan yang harus dipenuhi oleh ( are dan Cov( untuk *l <m* , sehingga komponen utama Cmp adalah sebagai berikut.

|  |  |  |
| --- | --- | --- |
|  |  | ( 20 ) |

|  |  |  |
| --- | --- | --- |
|  |  | ( 2 1) |

1. K-means Clustering

Algoritma K-means adalah algoritma clustering yang cukup sederhana yang mempartisi data terbaik menjadi beberapa k cluster. k-means bertujuan untuk mengelompokkan data dengan memaksimalkan kesamaan data dalam satu cluster dan meminimalkan kesamaan data antar cluster. Ukuran kemiripan yang digunakan dalam cluster adalah fungsi jarak. Sehingga memaksimalkan kesamaan data diperoleh berdasarkan jarak terpendek antara data dan titik centroid. Algoritma k-means dimulai dengan pembentukan partisi cluster di awal, kemudian secara iteratif memperbaiki partisi cluster hingga tidak ada perubahan yang signifikan pada partisi cluster (Sibuea & Safta, 2017) .

Misal diberikan matriks data X= yang berukuran i=1,2,...,n, j=1,2,...,p, dan asumsikan jumlah cluster awal adalah K. Selanjutnya, menghitung centroid cluster ke - *i* berikutnya , rumus yang digunakan

|  |  |  |
| --- | --- | --- |
|  |  | ( 23 ) |

Jarak Euclidean

Jarak Euclidean adalah perhitungan untuk mengukur jarak dua titik dalam ruang Euclidean yang mempelajari hubungan antara sudut dan jarak (Mustofa & Suasana, 2018) .

|  |  |  |
| --- | --- | --- |
|  |  | ( 24 ) |

1. Jaringan Saraf Berulang (RNN)

RNN merupakan salah satu arsitektur Neural Network untuk mengolah data sekuensial. Perbedaan utama dengan Neural Network adalah sinyal dapat mengalir maju dan mundur secara berulang (Al Kindhi et al., 2019) . Proses pelatihan untuk RNN sama dengan proses pelatihan untuk jaringan saraf pada umumnya. Ada tiga langkah utama dari proses pelatihan di RNN. Pertama, melakukan proses forward pass dan membuat prediksi (Babaee et al., 2018) . Pada proses ini dilakukan perhitungan untuk setiap hidden state ( ) berdasarkan setiap input ( ) dan bobot yang ditentukan. Setelah menemukan nilai hidden state, maka langkah selanjutnya adalah menghitung output atau hasil prediksi ( ) . Langkah kedua, bandingkan hasil prediksi ( ) dengan nilai keluaran yang sebenarnya atau disebut juga *target* , menggunakan *Loss Function* . *Loss Function* dengan nilai keluaran aktual atau disebut juga dengan target, menggunakan Loss Function. Loss Function menghasilkan nilai error yang dapat menunjukkan apakah hasil prediksi tepat sasaran atau bahkan jauh dari target sehingga dapat disimpulkan seberapa baik atau buruk kinerja RNN tersebut. Langkah terakhir, dari nilai error yang dihasilkan oleh Loss Function, kemudian dilakukan proses Backpropagation Through Time (BPTT) untuk menghitung gradien untuk setiap langkah waktu dalam jaringan. Proses BPTT dilakukan untuk mencari bobot dan bias yang lebih baik dari proses sebelumnya. Setelah proses BPTT selesai dilakukan pemutakhiran bobot dan bias menggunakan metode Stochastic Gradient Descent (SGD). Untuk menghitung nilai keadaan tersembunyi untuk waktu *t,* digunakan rumus :

|  |  |  |
| --- | --- | --- |
|  |  | ( 25 ) |

Untuk menghitung keluaran sebagai prediksi, digunakan rumus

|  |  |  |
| --- | --- | --- |
|  |  | ( 26 ) |

1. Waktu Pikir Backpropagation (BPTT)

BPTT merupakan algoritma yang digunakan untuk mengubah nilai bobot pada RNN. Algoritma pelatihan BPTT biasanya digunakan untuk data deret waktu. Konsep utama BPTT adalah memperluas jaringan pada setiap langkah waktu dengan meletakkan salinan RNN yang sama, dan mengatur ulang koneksi jaringan untuk mendapatkan koneksi antara salinan berikutnya (Juanda et al., 2018) . Untuk menghasilkan prediksi yang akurat, parameter dalam RNN seperti learning rate, jumlah neuron, dan jumlah data akan diuji.

Gradien adalah nilai yang digunakan untuk mengatur parameter atau bobot yang sesuai dalam suatu jaringan, sehingga jaringan tersebut dapat belajar. Dalam proses BPTT inilah proses perhitungan gradien terjadi. Semakin besar gradien, semakin besar penyesuaian, dan sebaliknya. BPTT menggunakan konsep aturan rantai, yaitu hubungan antara beberapa turunan (seperti rantai). Berikut ini adalah aturan rantai BPTT.

Rumus (28) dan (29) digunakan dalam aturan rantai (27), yaitu menghitung gradien bobot

|  |  |  |
| --- | --- | --- |
|  |  | ( 27 ) |

|  |  |  |
| --- | --- | --- |
|  |  | ( 28 ) |

|  |  |  |
| --- | --- | --- |
|  |  | ( 29 ) |

Rumus (3 1), (32), (33) digunakan dalam aturan rantai (30), yaitu menghitung gradien bobot

|  |  |  |
| --- | --- | --- |
|  |  | ( 30 ) |

|  |  |  |
| --- | --- | --- |
|  |  | ( 3 1) |

|  |  |  |
| --- | --- | --- |
|  |  | ( 32 ) |

|  |  |  |
| --- | --- | --- |
|  |  | ( 33 ) |

Rumus (35) digunakan dalam aturan rantai (34), yaitu menghitung gradien bobot

|  |  |  |
| --- | --- | --- |
|  |  | ( 34 ) |

|  |  |  |
| --- | --- | --- |
|  |  | ( 35 ) |

Rumus (37) digunakan dalam aturan rantai (36), yaitu untuk menghitung bias gradien

|  |  |  |
| --- | --- | --- |
|  |  | ( 36 ) |

|  |  |  |
| --- | --- | --- |
|  |  | ( 37 ) |

Rumus (39) digunakan dalam aturan rantai (38), yaitu untuk menghitung gradien bobot

|  |  |  |
| --- | --- | --- |
|  |  | ( 38 ) |

|  |  |  |
| --- | --- | --- |
|  |  | ( 39 ) |

1. Stochastic Gradient Descent (SGD)

Penggunaan Stochastic Gradient Descent pada neural network dilatarbelakangi oleh cost atau loss value yang tinggi dan membutuhkan running backpropagation setelah training (Haqqi & Kusumoputro, 2022) . SGD dapat mengatasi nilai loss yang tinggi dengan memperbarui parameter atau bobot dan bias setelah backpropagation, atau di RNN-BPTT.

|  |  |  |
| --- | --- | --- |
|  |  | ( 40 ) |
|  |  |  |

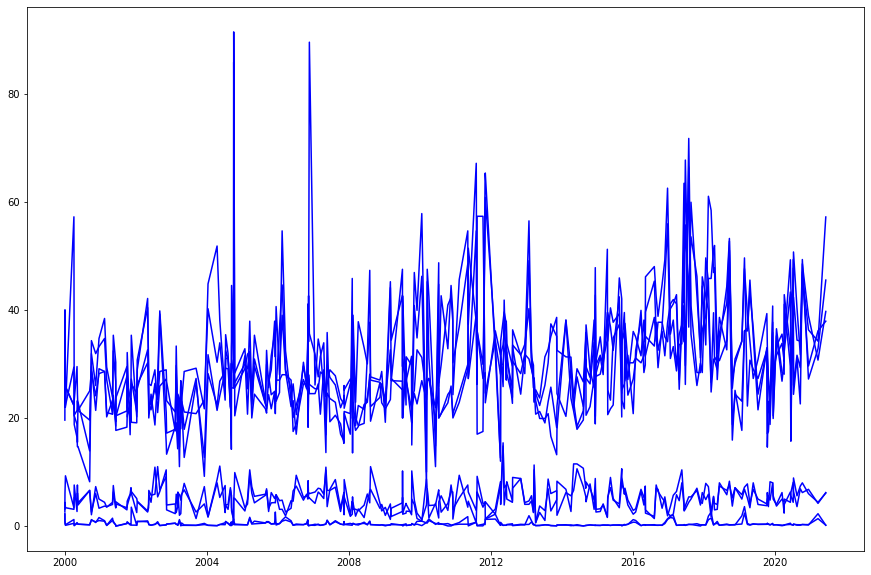
**Hasil dan Diskusi**

Objek yang digunakan dalam penelitian ini adalah data sekunder yaitu data kualitas air sungai yang diperoleh dari website resmi Kaggle. Dataset diperoleh dari situs resmi Kaggle di laman <https://www.kaggle.com>. Dataset terdiri dari variabel yang diambil dari 2 stasiun. Variabel data kualitas air dapat dilihat pada Tabel 1.

Tabel 1. Dataset Kualitas Air Sungai

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 15\_BSK5 | 15\_SO4 | 15\_CL | 15\_NH4 | 16\_BSK5 | | 16\_SO4 | 16\_CL | 16\_NH4 |
| 4,4 | 6,0 | 26,8 | 25,0 | | 0,87 | 0,54 | 40,0 | 32,0 |
| 3,4 | 9,3 | 25,8 | 25,8 | | 0,25 | 0,14 | 22,30 | 22,0 |
| 5,3 | 7,6 | 23,2 | 20,50 | | 0,07 | 0,14 | 20,40 | 18,8 |
|  |  |  |  | |  |  |  |  |
| 6,9 | 7,4 | 34,2 | 31,70 | | 0,25 | 0,16 | 29,30 | 22,6 |
| 6,8 | 6,0 | 38,9 | 29,61 | | 0,28 | 0,20 | 36,30 | 27,2 |
| 6,2 | 6,1 | 39,7 | 37,90 | | 0,17 | 0,16 | 57,17 | 45,5 |

Sebelum data diolah menggunakan metode PCA, terlebih dahulu dilakukan EDA pada dataset. Gambar 1 menunjukkan dataset yang telah diproses sebelumnya.



Gambar 1. Data kualitas air sungai

Langkah pertama dalam mengolah data ini adalah memilih faktor yang paling berpengaruh dengan menggunakan metode PCA. Proses perhitungan PCA diawali dengan normalisasi data sehingga data berada pada range tertentu. Selanjutnya dilakukan perhitungan nilai kovarians antar variabel yang dibentuk ke dalam matriks. Kemudian nilai vektor eigen dihitung, dan komponen utama ditentukan.

Hasil normalisasi data menggunakan softmax ditulis dalam bentuk matriks sebagai berikut.

Matriks kovarians menunjukkan hubungan antar variabel yang akan dianalisis menggunakan PCA.

Hasil perhitungan matriks kovarians disajikan pada matriks di bawah ini .

Berdasarkan matriks di atas, terlihat bahwa setiap variabel saling berhubungan.

Selanjutnya, cari nilai dan vektor eigen dari matriks kovarians. Hasil perhitungan eigenvector disajikan pada matriks di bawah ini.

Sedangkan hasil perhitungan matriks eigenvalue adalah: [2.500 1.8665 1.1125 0,4132 0,2337 0,1793 0,1545].

Setelah didapatkan eigenvalues, dihitung nilai Cumulative Proportion of Variance (PKV) untuk menentukan jumlah komponen utama yang akan dipilih. Dapat dilihat bahwa proporsi kumulatif eigenvalues terdapat pada tabel berikut.

Tabel 2.Nilai Eigen dan Varians Total

|  |  |  |  |
| --- | --- | --- | --- |
| Komponen | Nilai eigen ( | Jumlah varians (%) | Kumulatif (%) |
| 1 | 2.5700 | 31,95 | 31,95 |
| 2 | 1,8665 | 23,20 | 55,16 |
| 3 | 1,5129 | 18,81 | 73,97 |
| 4 | 1.1125 | 13,83 | 87,80 |
| 5 | 0,4132 | 5,13 | 92,94 |
| 6 | 0,2337 | 2,90 | 95,84 |
| 7 | 0,1793 | 2,22 | 98,07 |
| 8 | 0,1545 | 1,92 | 100 |

Pada Tabel 2, diperoleh 3 variabel baru atau komponen utama yang telah mewakili variabel yang dianalisis. Kemampuan setiap komponen untuk merepresentasikan variabel yang dianalisis ditunjukkan dengan besarnya varians yang dijelaskan, yang disebut nilai eigen. Besarnya nilai eigen menunjukkan kontribusi komponen utama terhadap varians semua variabel asli yang dianalisis.

Setelah melakukan reduksi dimensi dan mendapatkan indikator yang paling berpengaruh menggunakan PCA, langkah selanjutnya adalah melakukan clustering menggunakan K-means. Hasil pengelompokan pada k-means disajikan pada Tabel 3 di bawah ini.

Tabel 3. Dataset Kualitas Air Setelah Pengelompokan

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X1 | X2 | X3 | Label | Jarak Gugus 1 | | Jarak Cluster 2 | Jarak Cluster 3 | |
| -0,5956 | 2,3740 | 1,4649 | 0 | | 0,800 | 3,66 | 3,54 |
| -0,7177 | -1,3372 | 0,2241 | 2 | | 3.500 | 2,72 | 1,04 |
| -1,3524 | -2,1054 | -0,4967 | 2 | | 4,419 | 3,71 | 1,58 |
|  |  |  |  | |  |  |  |
| 0,7676 | -1,2681 | 1,1475 | 1 | | 3.837 | 1,78 | 2,43 |
| 1.5042 | -0,5197 | -0,9615 | 1 | | 3.707 | 0,92 | 2,81 |
| 4,4425 | 0,1223 | 0,6564 | 1 | | 5.951 | 2,77 | 5,52 |

Teknik clustering ini juga digunakan untuk mengubah tipe data unsupervised menjadi data supervised. Selanjutnya data yang diperoleh dari hasil pengelompokan k-means akan masuk ke proses peramalan RNN. Data tersebut terlebih dahulu akan dibagi menjadi dua untuk proses training dan testing. Distribusi data dibagi menjadi 80% data latih dan 20% data uji. Langkah selanjutnya adalah proses pelatihan untuk model RNN dengan variabel input dan output yang telah ditentukan. Kemudian melakukan prediksi pada data uji dan langkah terakhir adalah proses evaluasi model menggunakan mean squared error, root mean squared error, dan mean absolute deviation yang telah dibahas sebelumnya. Analisis hasil prediksi dari model RNN dapat dilihat dengan bantuan grafik sehingga dapat dengan mudah dipahami oleh orang awam sekalipun. Label data aktual dan hasil prediksi dibandingkan menggunakan plot garis. Jika plot garis hasil prediksi mendekati plot garis label data yang sebenarnya, maka model prediksi yang dibangun akan lebih baik. Selain itu, analisis model yang dikategorikan baik juga dapat dilihat dengan membandingkan hasil error. Jika tingkat kesalahan semakin kecil, maka akan semakin baik model yang telah diuji.

Hasil peramalan menggunakan RNN dengan PCA dan RNN tanpa PCA disajikan pada Gambar di bawah ini.

Gambar 2 . RNN dengan PCA

Gambar 3 . RNN tanpa PCA

Gambar 2 merupakan hasil prediksi menggunakan RNN dengan metode PCA, menunjukkan sampel prediksi selama 12 hari. Gambar 3 merupakan hasil prediksi RNN tanpa metode PCA, menunjukkan sampel prediksi selama 30 hari. Selanjutnya, kami akan menganalisis metrik kesalahan metode ini, yang ditunjukkan pada Tabel 4.

Tabel 4. Nilai Akurasi MSE, RMSE, dan MA D

|  |  |  |
| --- | --- | --- |
|  | RNN dengan PCA | RNN tanpa PCA |
| UMK | 0,014921 | 21,00179 |
| RMSE | 0,12215 | 4.582772 |
| GILA | 0,098413 | 1,141026 |
| Ketepatan | 93% | 82% |

Hasil pengukuran matriks error MSE, RMSE, dan MAD diperoleh dari model prediksi yang telah dibangun dan disajikan pada tabel di atas. Berdasarkan analisis Gambar 1, Gambar 2, dan Tabel 4 diperoleh hasil bahwa model peramalan menggunakan metode RNN dengan reduksi dimensi PCA sangat baik untuk mengolah data ini.

**Kesimpulan**

Makalah ini telah menggambarkan model untuk memprediksi kualitas air sungai. Pertama, faktor yang paling besar pengaruhnya terhadap data kualitas air sungai dipilih menggunakan PCA. Data time series kemudian dikelompokkan dengan variasi trend yang sama menggunakan k-means clustering. Setelah mendapatkan data tersupervisi, selanjutnya dibangun model prediksi berdasarkan metode RNN. Model yang dibangun menghasilkan error metric pada data kualitas air sungai dengan MSE 0,014921, RMSE 0,12215, MAD 0,098413, dan hasil akurasi yang didapat adalah 93%. Model yang diusulkan menghasilkan hasil prediksi yang lebih akurat dibandingkan model prediksi RNN tanpa pemilihan fitur PCA. Untuk model prediksi RNN tanpa PCA pada data kualitas air sungai didapatkan nilai error metric MSE sebesar 21.00179, RMSE 4.582772, MAD 1.141026, dan hasil akurasi yang diperoleh sebesar 82%.

Dalam pekerjaan di masa depan, gunakan algoritma optimasi yang berbeda untuk mengoptimalkan pemilihan parameter model RNN dalam upaya untuk meningkatkan akurasi eksperimental.

**ucapan terima kasih**

Jurusan Matematika, Fakultas Sains dan Teknologi Terapan, Universitas Ahmad Dahlan, Indonesia

**Referensi**

Al Kindhi, B., Sardjono, TA, & Hery Purnomo, M. (2019). Prediksi DNA Virus Hepatitis C Berdasarkan Recurrent Neural Network-Back Propagation Through Time (RNN-BPTT). *Konferensi Internasional 2019 tentang Mekatronika Tingkat Lanjut, Manufaktur Cerdas, dan Otomasi Industri, ICAMIMIA 2019 - Prosiding* , 208–214. https://doi.org/10.1109/ICAMIMIA47173.2019.9223395

Ariyanto, R., Puspitasari, D., & Ericawati, F. (2017). Penerapan Metode Double Exponential Smoothing Pada Peramalan Produksi Tanaman Pangan. *Jurnal Informatika Polinema* , *4* (1), 57. https://doi.org/10.33795/jip.v4i1.145

Babaee, M., Li, Z., & Rigoll, G. (2018). Penanganan Oklusi dalam Melacak Banyak Orang Menggunakan RNN. *Prosiding - Konferensi Internasional tentang Pengolahan Gambar, ICIP* , 2715-2719. https://doi.org/10.1109/ICIP.2018.8451140

Brezin, E. (2002). Pengantar Model Matriks. *Kombinatorika Asimtotik dengan Aplikasi pada Fisika Matematika* , 23–50. https://doi.org/10.1007/978-94-010-0575-3\_2

Cao, X., Liu, Y., Wang, J., Liu, C., & Duan, Q. (2020). Prediksi oksigen terlarut dalam air budidaya tambak berdasarkan K-means clustering dan jaringan saraf unit berulang yang terjaga keamanannya. *Teknik Budidaya Perairan* , *91* (17), 102122. https://doi.org/10.1016/j.aquaeng.2020.102122

Desmonda, D., Tursina, T., & Irwansyah, MA (2018). Prediksi Besaran Curah Hujan Menggunakan Metode Fuzzy Time Series. *Jurnal Sistem Dan Teknologi Informasi (JUSTIN)* , *6* (4), 141. https://doi.org/10.26418/justin.v6i4.27036

Evangelista, DGD, Bedruz, RAR, Vicerra, RRP, & Bandala, AA (2020). Perancangan Model Adaptive Neuro-Fuzzy Inference System (ANFIS) pada Penilaian Efluen pada Perairan Permukaan Tawar Kelas C. *Konferensi Internasional ke-12 IEEE 2020 tentang Humanoid, Nanoteknologi, Teknologi Informasi, Komunikasi dan Kontrol, Lingkungan, dan Manajemen, HNICEM 2020* . https://doi.org/10.1109/HNICEM51456.2020.9399993

Firmansyah, MR, Ilyas, R., & Kasyidi, F. (2020). Klasifikasi Kalimat Ilmiah Menggunakan Recurrent Neural Network. *Prosiding Lokakarya Penelitian Industri dan Seminar Nasional ke -11* , *11* (1), 488–495.

Haqqi, MS, & Kusumoputro, B. (2022). *Komparasi Metode Optimasi Adam dan SGD dalam Skema Direct Inverse Control untuk Sistem Kendali Data Sikap dan Ketinggian Quadcopter* . *10* (2), 458–469.

Informasi, FT (2017). *Peramalan Laju Inflasi Di Indonesia Menggunakan Back Propagation Neural Network Peramalan Tingkat Inflasi di Indonesia Menggunakan Back Propagation Neural Network Menggunakan Back Propagation Neural* .

Juanda, RA, Jondri, & Rohmawati, AA (2018). Prediksi Harga Bitcoin Dengan Menggunakan Recurrent Neural Network. *E-Prosiding Teknik* , *5* (2), 3682–3690.

Luis, F., & Moncayo, G. (nd). *Aljabar Linier Dasar: Versi Aplikasi, Edisi ke-11* .

Maricar, MA (2019). Analisa Perbandingan Nilai Akurasi Moving Average Dan Exponential Smoothing Untuk Sistem Peramalan Pendapatan Pada Perusahaan XYZ. *Jurnal Sistem Dan Informatika* , *13* (2), 36–45.

Moh. Jufriyanto. (2020). Peramalan Buka Keripik Singkong dengan Simulasi Monte Carlo Peramalan Permintaan Keripik Singkong dengan Simulasi Monte Carlo. *Jurnal Teknik Industri* , *6* (2), 107-113.

Mustofa, Z., & Suasana, IS (2018). Algoritma Clustering K-Medoids Pada E-Government Bidang Informasi Dan Komunikasi. *Jurnal Teknologi Dan Komunikasi* , *9* , 1–10.

Novita, A. (2016). Prediksi Pergerakan Harga Saham Pada Bank Terbesar Di Indonesia Dengan Metode Backpropagation Neural Network. *Jutisi* , *05* (01), 965–972.

Nur Hadiansyah, F. (2017). Prediksi Harga Cabai dengan Menggunakan pemodelan Time Series ARIMA. *Jurnal Komputasi Indonesia (Indo-JC)* , *2* (1), 71. https://doi.org/10.21108/indojc.2017.2.1.144

Seri (Studi Kasus: PT. Telekomunikasi Indonesia, Tbk Kandatel Sukabumi). *Jurnal Ilmiah SANTIKA* , *8* (2), 1–17.

Rukhansah, N., Muslim, MA, & Arifudin, R. (2016). Peramalan Harga Emas Menggunakan Fuzzy Time Series Markov Chain Model. *Komputaki* , *1* (1), 56–74. https://www.unaki.ac.id/ejournal/index.php/komputaki/article/view/113

S awitri , MND, S umarjaya , IW, & T astrawati , NKT (2018). Peramalan Menggunakan Metode Backpropagation Neural Network. *E-Jurnal Matematika* , *7* (3), 264. https://doi.org/10.24843/mtk.2018.v07.i03.p213

Shakhari, S., Verma, AK, & Banerjee, I. (2019). Prediksi kualitas air lokasi terpencil ganga sungai India: Analisis regresi dan kesalahan. *Konferensi Internasional tentang ICT dan Rekayasa Pengetahuan* , *2019* - *Novem* . https://doi.org/10.1109/ICTKE47035.2019.8966796

Sibuea, ML, & Safta, A. (2017). Pemetaan Siswa Berprestasi Menggunakan Metode K-Means Clustring. *Jurteksi* , *4* (1), 85–92. https://doi.org/10.33330/jurteksi.v4i1.28

Sumartini, Hayati, MN, & Wahyuningsih, S. (2017). Peramalan Menggunakan Metode Fuzzy Time Series Cheng. *Jurnal EKSPONENSIAL* , *8* , 51–56.

Sun, H., & Dia, Y. (2018). Penelitian dan Penerapan Evaluasi Kualitas Air Pada Bagian Tertentu Sungai Yangtze Berdasarkan Fuzzy Neural Network. *Prosiding - Konferensi Internasional 2017 tentang Informatika Industri - Teknologi Komputasi, Teknologi Cerdas, Integrasi Informasi Industri, ICIICII 2017 ,* 2017 *-* Desember , 301–304. https://doi.org/10.1109/ICIICII.2017.39

Wu, Y., Ling, R., Zhou, J., Zhang, M., & Gao, W. (2021). Prediksi Kualitas Air Berbasis Jaringan Syaraf Tiruan. *Jurnal Fisika: Seri Konferensi* , *1738* (1), 1-5. https://doi.org/10.1088/1742-6596/1738/1/012066