e-ISSN: 2460-870X | p-ISSN: 2085-1367

Model Prediksi Data Besar Distribusi Produk Farmasi: Analisis Kinerja Model Deep Learning

Big Data Prediction Model of Pharmaceutical Product Distrubution: Deep Learning Model Performance Analysis

Faisal Fadli[a,1], Saib Suwilo[b,2], Muhammad Zarlis[c,3]
[a,b,c]Universitas Sumatera Utara; Jl. Dr. T. Mansur No. 9, Padang Bulan, Kec. Medan Baru, Kota Medan, Sumatera Utara 20222, Indonesia
e-mail: faisalfadli23@gmail.com[1], saib@usu.ac.id[2], m.zarlis@usu.ac.id[3]

ABSTRAK

Seiring dengan berjalannya bisnis perusahaan, masalah dalam penyimpanan dan pengolahan data besar pun akan semakin kompleks. data yang tidak terorganisir dapat menyebabkan perusahaan gagal dalam memaksimalkan strategi penjualan. Salah satu pendekatan untuk memaksimalkan strategi penjualan tersebut adalah dengan peramalan.penelitian ini bertujuan untuk mengurangi tingkat persediaan pelanggan jangka pendek dan membantu dalam menentukan target penjualan yang realistis di masa depan dengan mengusulkan metode pembelajaran mendalam berdasarkan segmentasi pelanggan. Kerangka analisa diusulkan menggunakan teknik Robust Principal Component Analysis (RPCA) untuk mengurangi dimensi kumpulan dataset, kemudian algoritma K-Means Clustering diterapkan untuk mengidentifikasi kelompok populasi guna melihat beberapa kluster yang dapat sangat mewakili karakteristik basis pelanggan perusahaan yang ada. Terakhir lapisan CNN dan LSTM digabungkan untuk memperkirakan penjualan masa depan. Hasil peramalan dievaluasi menggunakan Mean Absolute Error (MAE) dan Root Mean Square Error (RMSE). Pendekatan yang diusulkan guna mengisi celah masalah yang terjadi karena kurangnya informasi mengenai kurangnya informasi tentang kinerja bisnis dalam hal kategorisasi produk. Deep Learning dan Machine Learning secara keseluruhan menghasilkan tingkat akuasi yang baik, akan tetapi kombinasi CNN-LSTM belum dapat menghasilkan akurasi seperti yang diharapkan. Pendekatan Gradient Boost membutuhkan waktu lebih sedikit dibandingkan dengan Random Forest. Sehingga model Gradient Boost dapat dikatakan sebagai algoritma terbaik dari algoritma lainnya. Gradient Boost memberikan RMSE 0,625 pada data pengujian.

Kata Kunci : Robust Principal Component Analysis (RPCA); *K-Means Clustering*; CNN; LSTM; *Mean Absolute Error* (MAE); *Root Mean Square Error* (RMSE)

ABSTRACT

As the company's business goes on, the problems in storing and processing big data will become more complex. Unorganized data can cause companies to fail in maximizing sales strategies. One approach to maximize the sales strategy is forecasting. This study aims to reduce short-term customer inventory levels and assist in determining realistic sales targets in the future by proposing a deep learning method based on customer segmentation. The analytical framework is proposed using the Robust Principal Component Analysis (RPCA) technique to reduce the dimensions of the dataset, then the K-Means Clustering algorithm is applied to identify population groups in order to see several clusters that can best represent the characteristics of the company's existing customer base. Finally, the CNN and LSTM layers are combined to estimate future sales. Forecasting results were evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The proposed approach is to fill the gaps in problems that occur due to lack of information regarding the lack of information about business performance in terms of product categorization. Deep Learning and Machine Learning as a whole produce a good level of accuracy, but the combination of CNN-LSTM has not been able to produce accuracy as expected. The Gradient Boost approach takes less time than Random Forest. So that the Gradient Boost model can be said to be the best algorithm from other algorithms. Gradient Boost gives RMSE 0.625 on test data.

Keywords : Robust Principal Component Analysis (RPCA); *K-Means Clustering*; CNN; LSTM; *Mean Absolute Error* (MAE); *Root Mean Square Error* (RMSE)

Info Artikel :
Disubmit: 11 November 2021 *Direview: 20 Januari 2021* *Diterima : 27 Januari 2022*

DOI: https://www.doi.org/10.22303/csrid.14.1.2022.68-80

1. PENDAHULUAN

Peramalan merupakan pendekatan penting untuk merencanakan masa depan secara efektif dan efisien, sebagai dasar dari setiap tahap perencanaan manajemen yang memiliki pengaruh sangat tinggi terhadap kinerja bisnis perusahaan. Perkiraan yang akurat dapat menghasilkan keputusan yang lebih baik dalam bisnis dan meningkatkan kinerja perusahaan terkait dengan manajemen persediaan, pengadaan perusahaan, dan manajemen penjualan, sehingga meningkatkan keuntungan perusahaan dan menurunkan biaya (Chen & Lu, 2017). Secara umum, sistem prakiraan penjualan menghadapi masalah dalam menghasilkan prakiraan mingguan dari unit penjualan untuk item ritel. Penjualan barang eceran pada minggu tertentu dipengaruhi oleh banyak faktor, seperti faktor musiman, diskon, harga dan lainnya. Peramalan deret waktu dengan pendekatan model matematis seperti *Auto-Regressive Integrated Moving Average* (ARIMA) menghasilkan akurasi prediksi yang baik untuk kasus peramalan harga saham (Choi, 2018), perkiraan nilai tukar uang (Reddy SK, 2015), keuangan (Siami-Namini, Tavakoli, & Namin, 2019), penjualan e-commerce (M. Li et al., 2018), penjualan eceran konsumen (Ramos et al., 2015) dan lainnya. Akan tetapi, ARIMA sangat tergantung pada plot *Autocorrelation Function* (ACF), *Partial Autocorrelation* (PACF) dan validasi parameter (p,d,q) untuk menghasilkan akurasi yang baik, selain itu juga terbatas pada struktur linier dan kemampuan memasukkan variabel eksternal (Karb et al., 2020).

Dalam beberapa tahun terakhir, penerapan arsitektur pembelajaran mendalam seperti *Long-Short Term Memory* (LSTM) (Shams et al., 2020), *Recurrent Neural Network* (RNN) (Hewamalage et al., 2021), *Convolutional Neural Network* (CNN) (Yang et al., 2020) dan lainnya dilaporkan menghasilkan tingkat akurasi yang menjanjikan untuk peramalan deret waktu. RNN adalah salah satu metode pembelajaran mendalam yang digunakan untuk menemukan korelasi temporal dalam prediksi deret waktu (Sherstinsky, 2020) berkinerja baik terhadap informasi terbaru, tetapi sulit untuk memodelkan ketergantungan jangka panjang (Sherstinsky, 2020). Model LSTM dalam meramalkan deret waktu dan khususnya untuk masalah prediksi jangka panjang lebih unggul dari ARIMA (Siami-Namini, Tavakoli, & Siami Namin, 2019), namun memiliki kelemahan pada masalah linear, non-temporal dan beberapa asumsi saat memodelkan jaringan (Chimmula & Zhang, 2020).

Berapa karya penelitian mengusulkan metode hybrid untuk mengoptimalkan fitur akurasi peramalan deret waktu (Khashei & Hajirahimi, 2019). Lu dan Wang (Lu et al., 2020) mengusulkan kombinasi CNN-LSTM untuk memprediksi harga saham. Hasil prediksi dibandingkan dengan MLP, CNN, RNN, LSTM, CNN-RNN dimana model CNN-LSTM lebih unggul dari model lainnya. Selanjutnya karya (T. Li et al., 2020) juga menerapkan CNN-LSTM untuk Peramalan Particulate Matter (PM2.5) dan (Kuo & Huang, 2018) untuk peramalan harga listrik. Namun, dalam industri e-commerce produk memiliki banyak jenis. Artinya, ada lebih dari satu deret waktu yang harus diprediksi (Zhao & Wang, 2017). Selain itu, prediksi e-commerce membutuhkan ketepatan waktu (Bandara et al., 2019). Oleh karena itu, mekanisme dan pendekatan untuk memprediksi penjualan sangat penting untuk menangkap perubahan yang dapat mempengaruhi penjualan produk dalam waktu dekat dari sumber yang paling dekat dengan pelanggan akhir.

Pada penelitian ini tujuan utamanya adalah untuk mengurangi tingkat persediaan pelanggan jangka pendek dan membantu dalam menentukan target penjualan yang realistis di masa depan dengan mengusulkan metode pembelajaran mendalam berdasarkan segmentasi pelanggan. Kerangka analisa diusulkan teknik *Robust Principal Component Analysis* (RPCA) untuk mengurangi dimensi kumpulan dataset, kemudian algoritma *K-Means Clustering* diterapkan untuk mengidentifikasi kelompok populasi guna melihat beberapa kluster dapat sangat mewakili karakteristik basis pelanggan perusahaan yang ada. Terakhir lapisan CNN dan LSTM digabungkan untuk memperkirakan penjualan masa depan. Hasil peramalan dievaluasi menggunakan M*ean Absolute Error* (MAE) dan *Root Mean Square Error* (RMSE). Pendekatan yang diusulkan guna mengisi celah masalah yang terjadi karena kurangnya informasi mengenai kurangnya informasi tentang kinerja bisnis dalam hal kategorisasi produk.

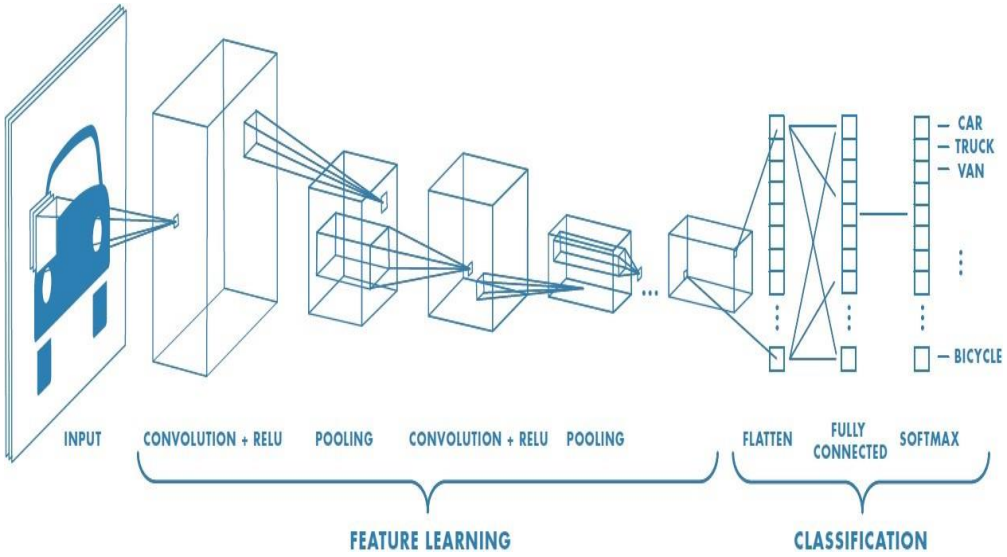
e-ISSN: 2460-870X | p-ISSN: 2085-1367

2. METODE

Pada bagian ini akan diuraikan penjelasan tentang model yang diusulkan serta pemahaman tentang cara kerja metode *Convolutional Neural Network* (CNN), *Recurrent Neural Networks* (RNN) dan *Long Short-Term Memory* (LSTM) secara singkat.

A. Convolutional Neural Network (CNN)

Jaringan Neural Konvolusional atau *Convolutional Neural Network* (CNN) sangat mirip dengan Jaringan Neural biasa yang terdiri dari dari neuron yang memiliki bobot dan bias. Setiap neuron akan menerima beberapa masukan, kemudian melakukan perkalian titik dan secara opsional mengikutinya dengan non-linearitas. Seluruh jaringan masih mengekspresikan satu fungsi skor yang dapat dibedakan dari piksel gambar mentah di satu sisi hingga skor kelas di sisi lain. Selain itu, memiliki fungsi kerugian (mis. SVM / Softmax) pada lapisan terakhir (*Fully Connected*) (Stanford University Course cs231n, 2018). CNN memiliki dua metode; yakni klasifikasi menggunakan *feedforward* dan tahap pembelajaran menggunakan *backpropagation*. CNN memiliki kesamaan cara kerja dengan *Multilayer Perceptron* (MLP), namun pada CNN setiap neuron dipresentasikan dalam bentuk dua dimensi, sedangkan MLP setiap neuron hanya berukuran satu dimensi. Pada gambar 1 dapat dilihat Arsitektur CNN



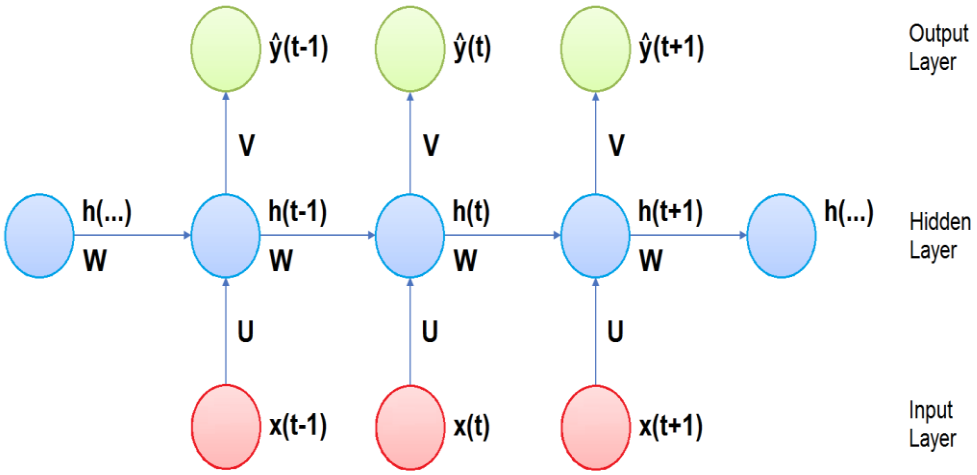
Gambar 1. Arsitektur Convolutional Neural Network

B. Recurrent Neural Networks (RNN)

Jaringan saraf berulang (RNN) merupakan jenis arsitektur jaringan saraf tiruan yang pemrosesannya dilakukan secara berulang-ulang. Pada umumnya model ini dipakai buat memproses input data sekuensial. RNN merupakan bagian dari metode *deep learning* lantaran data diproses terdiri berdasarkan beberapa lapis (layer). RNN sudah mengalami kemajuan yangg pesat dengan merevolusi bidang-bidang penelitian misalnya pemrosesan bahasa alami (NLP), sosialisasi suara, sintesa musik, peramalan, analisa deret DNA, analisa video, & sebagainya (Karpathy, 2015). RNN cocok buat data deret waktu lantaran mempunyai perluasan kemampuan buat memanfaatkan liputan temporal berdasarkan data memakai tautan berulang antara neuron.

RNN memakai jaringan node mirip neuron yang diatur pada lapisan berurutan, dimana setiap neuron dibagi pada lapisan masukan, lapisan tersembunyi, dan lapisan keluaran. Setiap koneksi antar neuron memiliki bobot latihan yang sesuai (Pra, n.d.). RNN melakukan tugas yang sama untuk setiap elemen urutan, dengan keluaran yang bergantung pada perhitungan sebelumnya (Neural & Tutorial, 2000). Arsitektur jaringan RNN bisa dilihat pada Gambar 2.

DOI: https://www.doi.org/10.22303/csrid.14.1.2022.68-80

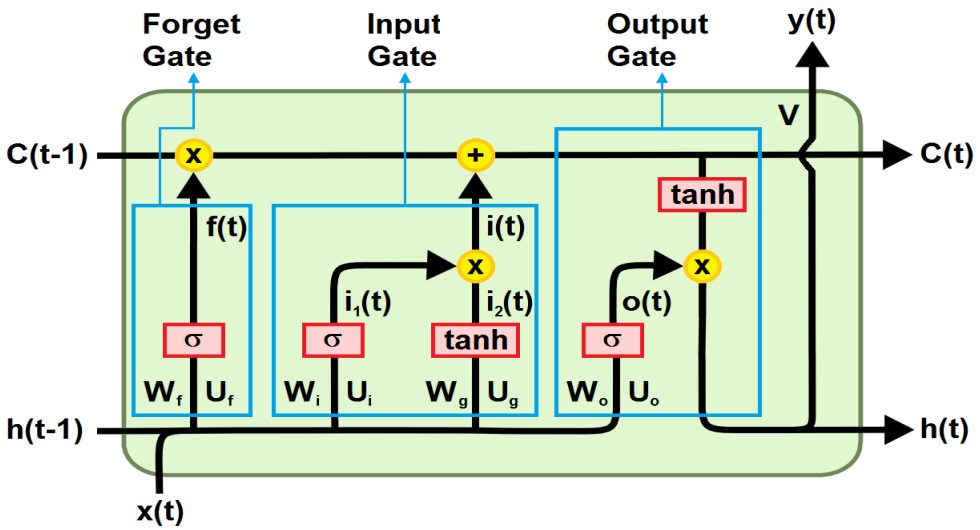


Gambar 2. Arsitektur Recurrent Neural Network

C. Long Short-Term Memory (LSTM)

Long Short-Term Memory Networks (LSTM) adalah salah satu algoritma untuk mengatasi masalah gradien yang hilang dalam RNN standar dengan meningkatkan aliran gradien dalam jaringan. Model ini menggunakan unit LSTM sebagai pengganti lapisan tersembunyi. LSTM dapat mempelajari dependensi jangka panjang karena penggunaan mekanisme gerbang dan sel memori. LSTM terdiri dari *Sel State, Forget Gate, Input Gate,* dan *Output Gat*e. Sel memori dapat mengingat informasi selama periode waktu yang sewenang-wenang dan aliran informasi ke dan dari sel diatur oleh mekanisme gerbang. Seperti yang ditunjukkan pada Gambar 2.3, dimana unit LSTM terdiri dari:

a. Sel State, yang membawa informasi sepanjang seluruh urutan dan mewakili memori jaringan;
b. Forget Gate, yang memutuskan apa yang relevan untuk menjaga dari langkah waktu sebelumnya;
c. Input Gate, yang memutuskan informasi apa yang relevan untuk menambahkan dari waktu saat ini langkah;
d. Output Gate, yang menentukan nilai output pada waktu saat ini langkah.

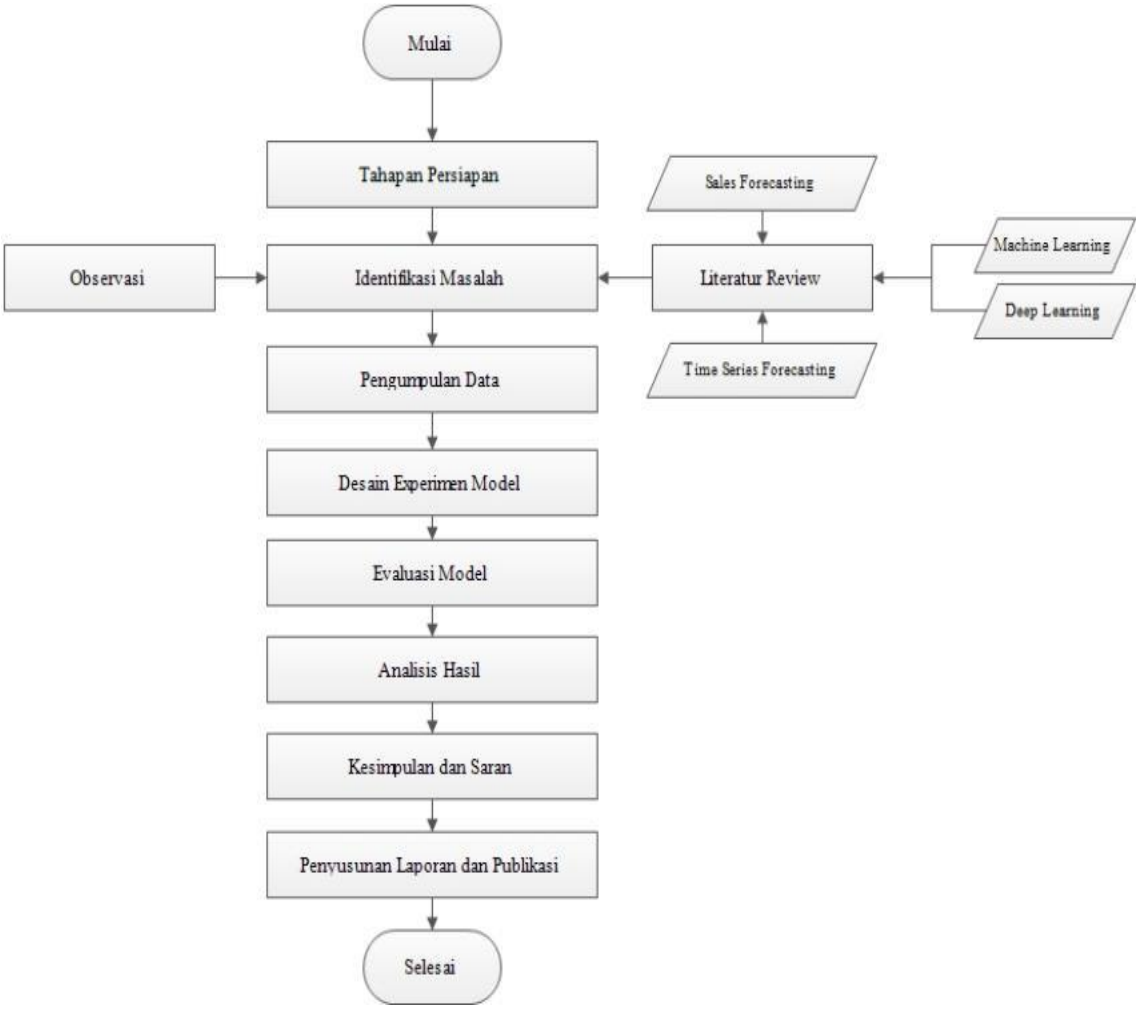


Gambar 3 Arsitektur Long Short-Term Memory Network

e-ISSN: 2460-870X | p-ISSN: 2085-1367

D. Desain Alur Penelitian

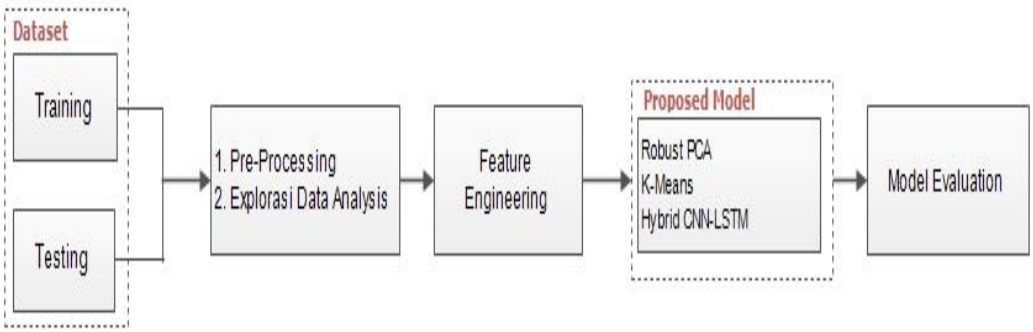
Desain alur penelitian yang dilakukan diilustrasikan pada gambar 2.1 berikut ini.



Gambar 4. Desain Alur Penelitian

E. Metode Yang Diusulkan

Metode yang diusulkan pada penelitian ini menggunakan teknik *Robust Principal Component Analysis* (RPCA) untuk mengurangi dimensi kumpulan dataset, kemudian algoritma *K-Means Clustering* diterapkan untuk mengidentifikasi kelompok populasi basis pelanggan perusahaan yang ada dan analisis prediktif menggunakan pendekatan *hybrid* CNN-LSTM untuk memperkirakan penjualan masa depan. Ilustrasi metode yang diusulkan ditunjukkan pada gambar 2.2



Gambar 5. Model yang diusulkan

DOI: https://www.doi.org/10.22303/csrid.14.1.2022.68-80

F. Experimen dan Pengujian Metode

Tahap ini data diolah dengan metode metode *Convolutional Neural Network* (CNN), *Long Short-Term Memory Networks* (LSTM), *Recurrent Neural Network* (RNN) dan *hybrid* CNN-LSTM, hasil peramalan pada metode yang diusulkan akan dibandingkan untuk mendapatkan metode yang paling akurat.

G. Evaluasi dan Validasi Hasil

Tahap ini dilakukan pengujian kesalahan atau *error* untuk mengetahui metode yang terbaik dalam melakukan peramalan penjualan produk farmasi menggunakan *Mean Absolute Error* (MAE) dan *Root Mean Square Error* (RMSE).

3. HASIL DAN PEMBAHASAN

A. Hasil

Pada bagian ini akan diuraiakan hasil penelitian dengan menguji dataset yang bersumber dari PT. Anugrah Argon Medica yang merupakan salah satu perusahaan distribusi farmasi di Indonesia.

B. Persiapan Data

Kumpulan dataset memuat data informasi transaksi penjualan tahun 2019 sampai tahun 2020 yang terdiri dari 3 (tiga) bagian yang di pisahkan, yaitu Data Customer, Data Produk dan Data Sales. Pada dataset sales terdiri dari dua dataset yaitu dataset sales tahun 2019 dan tahun 2020. Langkah awal akan dilakukan penggabungan dataset menjadi satu seperti terlihat pada gambar 3.1 berikut:

	Invoice_No	Full_Date	Order_No	customer_id	Customer_Name	Item_Code	Product_Desc	Item_Price	Qty	Gross_Value	...	CHANNEL_C
0	121000000000.0	01/02/2020	12200007465	120	12 SENTOSA. APT./MEDAN	165	LIPITOR 20 MG	575989.0	2.0	1151978.0	...	
1	121000000000.0	01/02/2020	12200007445	37	12 ROYAL PRIMA MEDAN. RS	121	RHINOS NEO DROP 10mL	45000.0	150.0	6750000.0	...	
2	121000000000.0	01/02/2020	12200007429	97	12 MITRA MEDIKA. RSU	122	RHINOS SR @50	250000.0	1.0	250000.0	...	
3	121000000000.0	01/02/2020	12200007433	97	12 MITRA MEDIKA. RSU	157	ALPRAZOLAM 0.5MG@100TAB	61000.0	3.0	192150.0	...	
4	121000000000.0	01/02/2020	12200007486	259	12 CENDANA. APT.	135	FOLAMIL GENIO Soft Cap(Btl/30's)	114000.0	2.0	228000.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
95	121000000000.0	03/02/2020	12200007514	92	12 TEO FARMA. APT	421	LEVOFLOXACIN FC 500MG(Box/20)	20976.0	3.0	66075.0	...	
96	121000000000.0	03/02/2020	12200007592	8	12 SANTA ELIZABETH. RSU.	399	GLUCOVANCE 500/5MG(Box/100)	434253.0	2.0	868506.0	...	
97	121000000000.0	03/02/2020	12200007514	92	12 TEO FARMA. APT	83	HERBAKOF Syrup 60ml	11000.0	5.0	55000.0	...	
98	121000000000.0	03/02/2020	12200007588	407	12 ANDI. APT.	49	CANDERIN 16MG(Box/30's)	270000.0	1.0	270000.0	...	
99	121000000000.0	03/02/2020	12200007576	1	12 YAKINI. APT.	267	COLERGIS SYR 60ML	38000.0	2.0	76000.0	...	

Gambar 6. Dataset Penjualan

Pada gambar 3.1 merupakan kumpulan dataset yang digunakan dalam penelitian ini, secara keseluruhan dataset terdiri dari 311.101 baris dan 21 kolom. Selanjutnya dilakukan eksplorasi data untuk mengumpulkan informasi data yang hilang dan dilakukan penghapusan data dengan mencari kolom dan baris yang akan di hapus sesuai dengan proporsinya. Beberapa kolom data juga dilakukan perubahan nama pada kolom untuk memudahkan pemrosesan. Untuk baris data yang hilang sekitar 800 baris data sehingga secara keseluruhan dataset terdiri dari 303.279 baris dan 13 kolom. Hasil pembersihan data dapat dilihat pada gambar 3.2 berikut:

e-ISSN: 2460-870X | p-ISSN: 2085-1367

	Order_No	Invoice_Date	Customer_ID	Customer_Name	Type_Customer	Group_Customer	Item_Code	Product_Desc	Item_Price	Qty	Sale
0	12200007465	2020-02-01	120	12 SENTOSA. APT./MEDAN	PHARMACY	-	165	LIPITOR 20 MG	575989.0	2.0	1151978.
1	12200007445	2020-02-01	37	12 ROYAL PRIMA MEDAN. RS	HOSPITAL	B	121	RHINOS NEO DROP 10mL	45000.0	150.0	6750000.
2	12200007429	2020-02-01	97	12 MITRA MEDIKA. RSU	HOSPITAL	C	122	RHINOS SR @50	250000.0	1.0	250000.
3	12200007433	2020-02-01	97	12 MITRA MEDIKA. RSU	HOSPITAL	C	157	ALPRAZOLAM 0.5MG@100TAB	61000.0	3.0	192150.
4	12200007486	2020-02-01	259	12 CENDANA. APT.	PHARMACY	-	135	FOLAMIL GENIO Soft Cap(Btl/30's)	114000.0	2.0	228000.
...	...	...	...	...	...	...	...	...	...	...	.
95	12200007514	2020-02-03	92	12 TEO FARMA. APT	PHARMACY	-	421	LEVOFLOXACIN FC 500MG(Box/20)	20976.0	3.0	66075.
96	12200007592	2020-02-03	8	12 SANTA ELIZABETH. RSU.	HOSPITAL	B	399	GLUCOVANCE 500/5MG(Box/100)	434253.0	2.0	868506.
97	12200007514	2020-02-03	92	12 TEO FARMA. APT	PHARMACY	-	83	HERBAKOF Syrup 60ml	11000.0	5.0	55000.
98	12200007588	2020-02-03	407	12 ANDI. APT.	PHARMACY	-	49	CANDERIN 16MG(Box/30's)	270000.0	1.0	270000.
99	12200007576	2020-02-03	1	12 YAKINI. APT.	PHARMACY	-	267	COLERGIS SYR 60ML	38000.0	2.0	76000.

Gambar 7. Hasil Pembersihan Dataset

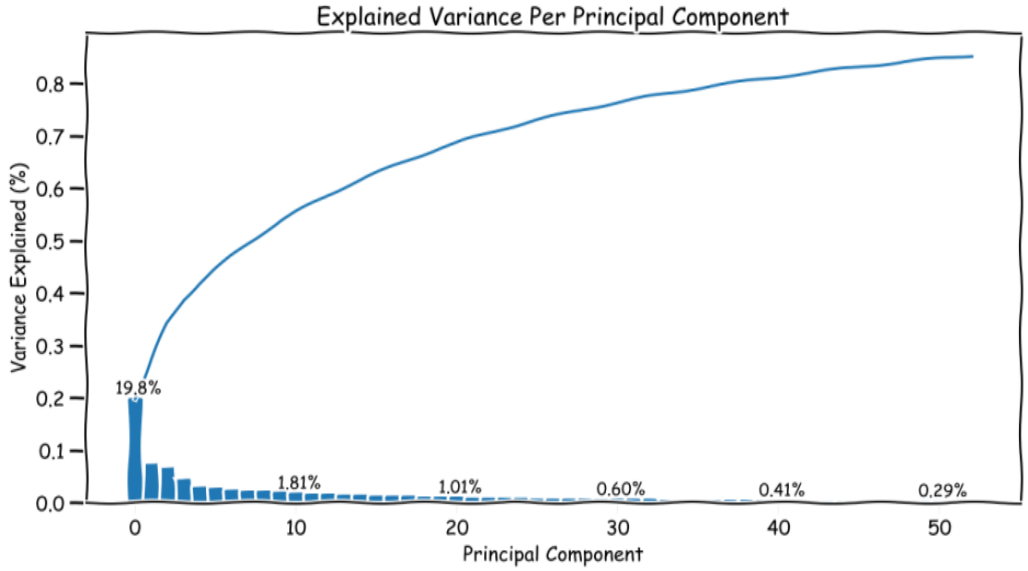
C. Segmentasi Pelanggan

Proses segmentasi pelanggan akan menggunakan teknik pembelajaran tanpa pengawasan *Robust Principal Component Analysis* (RPCA) dan *K-Means Clustering* untuk mengidentifikasi segmen pelanggan dari populasi data transaksi penjualan produk farmasi di PT. Anugrah Argon Medica.

a. Pengurangan Dimensi – *Robust Principal Component Analysis* (RPCA)

Langkah pertama adalah menemukan jumlah komponen dan cluster yang optimal untuk analisis segmentasi dengan mengurangi dimensi kumpulan data yang terdiri dari banyak variabel yang berkorelasi satu sama lain dengan tetap mempertahankan variasi yang ada dalam dataset.

Berikut akan ditampilkan visualisasi data pengurangan dimensi:



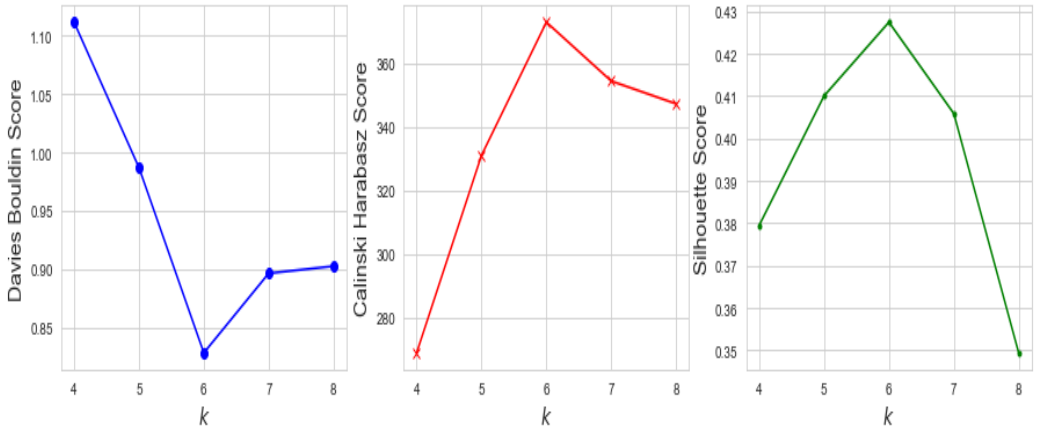
Gambar 8. Pengurangan Dimensi

Hasil RPCA menunjukkan bahwa komponen pada balok biru pertama mendapat penjelasan varian tertinggi sebesar 19,8%. Total ada 50 komponen yang menangkap semua varians dalam data dan mempertahankan pada sekitar 80% dari total varian data.

DOI: https://www.doi.org/10.22303/csrid.14.1.2022.68-80

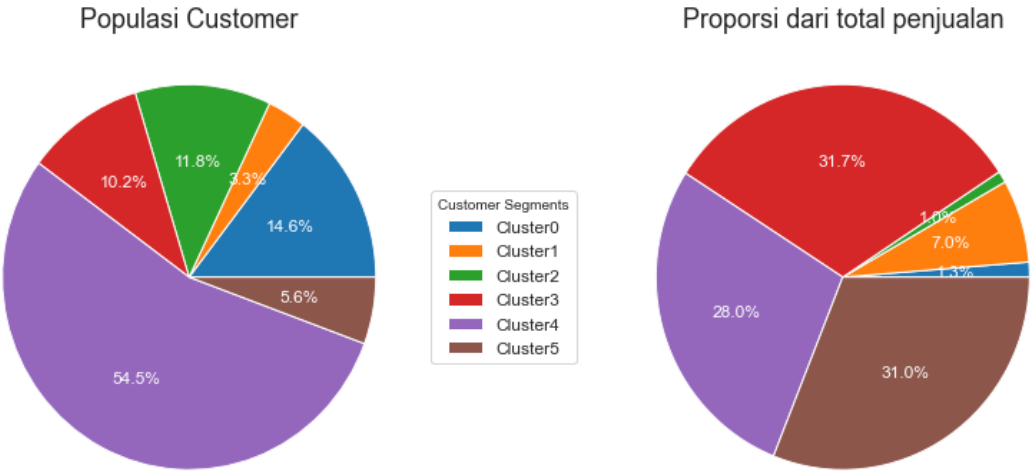
b. Analisis *K-Means Clustering*

Pada penelitian ini, penulis menerapkan pengelompokan *k-means* ke kumpulan data dan menggunakan rata-rata jarak dalam cluster dari setiap titik ke pusat cluster yang ditugaskan untuk memutuskan jumlah cluster mana yang akan dipertahankan. Untuk menentukan pusat cluster yang optimal, penulis menggunakan 3 (tiga) model berbeda yaitu model *Davies Bouldin*, *alinski Harabasz* dan *Silhouette*. Ketiga model tersebut akan dievaluasi dan kemudian model yang paling optimal akan di terapkan pada algorima *K-Means Clustering*. Hasil perbandingan ketiga model dalam menetukan pusat cluster yang optimal tersebut dapat dilihat pada gambar dibawah ini:



Gambar 9. Penentuan Pusat Cluster

Berdasarkan hasil pengelompokan data pada gambar 3.4, hasil dari tiga model yang digunakan untuk menentukan pusat cluster pada algoritma *k-means* tersebut menunjukkan bahwa pusat cluster menghasilkan nilai yang sama yaitu [6,6,6]. Maka dapat ditentukan nilai K=6. Kemudian nilai tersebut akan digunakan untuk mengukur kinerja model dalam hal akuisisi pelanggan dan membangun loyalitas pelanggan menggunakan metode *Analisys Recency Frequency* dan *Monetary* berdasarkan *K-Means Clustering* untuk memisahkan pelanggan berdasarkan perilaku. Semakin baru pembelian dan juga semakin responsif pelanggan terhadap promosi, maka akan semakin sering pula pelanggan tersebut membeli dan juga semakin terlibat serta puas mereka terhadap nilai moneter yang membedakan antara pembelanja berat dari pembeli bernilai rendah. Hasil analisis RFM berdasarkan algoritma K-Means Clustering dapat dilihat pada gambar 3.5 berikut:



Gambar 10. Hasil *K-Means Clustering*

e-ISSN: 2460-870X | p-ISSN: 2085-1367

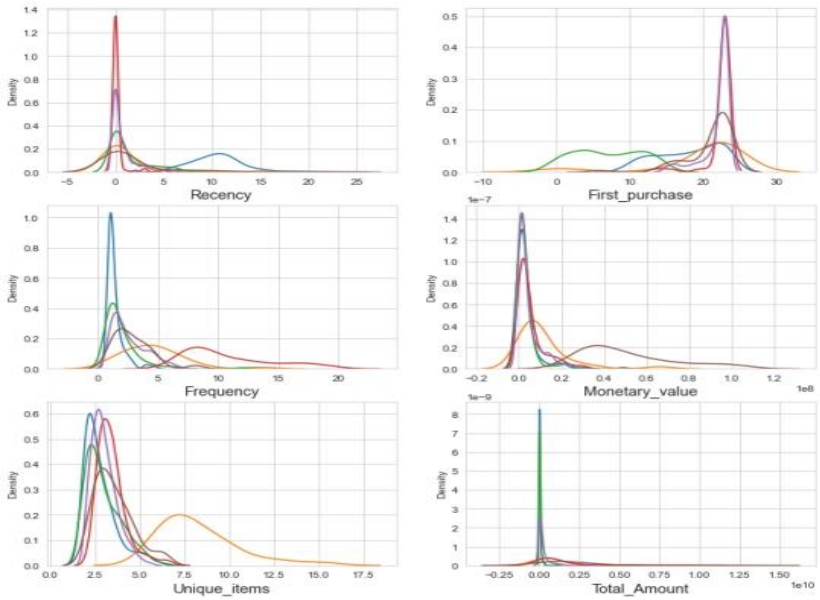
	#Customers	%customers	#Purchases	%transactions	Total_Amount	%sales_amount
Cluster0	135	14.56	3717	1.96	5.00e+09	1.26
Cluster1	31	3.34	16491	8.70	2.78e+10	6.98
Cluster2	109	11.76	3624	1.91	3.87e+09	0.97
Cluster3	95	10.25	79688	42.02	1.26e+11	31.72
Cluster4	505	54.48	77775	41.02	1.12e+11	28.03
Cluster5	52	5.61	8328	4.39	1.24e+11	31.05

Gambar 11. Hasil *K-Means Clustering*

Dapat dijelaskan dari gambar diatas, Cluster 0 yang berisi 135 pelanggan merupakan 14.56% dari seluruh populasi pelanggan dan menyumbang 1.96% dari total penjualan. Segmen ini memiliki profitabilitas yang cukup tinggi dengan nilai moneter rata-rata 1.26% per transaksi dan frekuensi sedang dengan rata-rata 1,96% transaksi per bulan. Yang menarik dari cluster ini adalah rata-rata jumlah *unique* item memiliki nilai yang banyak dalam setiap transaksi. Hal ini menunjukkan bahwa sebagian besar pelanggan di segmen ini sebenarnya adalah pelanggan organisasi bukan individu.

Selanjutnya, Pelanggan di Cluster 1 merupakan pelanggan dengan status lebih baru daripada yang ada di Cluster 2 dan Cluster 5. Segmen ini adalah yang terkecil dengan 3,34% dari seluruh populasi dan menyumbang 6.98% dari total penjualan. Segmen ini mencakup pelanggan setia yang mulai berbelanja dengan pengecer online pada kuartal pertama tahun ini dengan rata-rata pembelian pertama sebesar 8.7%.

Akhirnya, Cluster 4 berisi sekitar 54.48% dari seluruh populasi dan menyumbang 28% dari total jumlah penjualan. Pelanggan di segmen ini sering berbelanja dengan frekuensi rata-rata 42,2 transaksi per bulan. Mereka juga memiliki nilai moneter moderat, £308,7 per transaksi. Segmen ini dapat dianggap sebagai segmen pertama yang paling menguntungkan. Untuk lebih detailnya, hasil cluster pada masing-masing segmen pelanggan dapat dilihat pada gambar 3.7 berikut:



Gambar 12. Analisis RFM berdasarkan K-Means Clustering

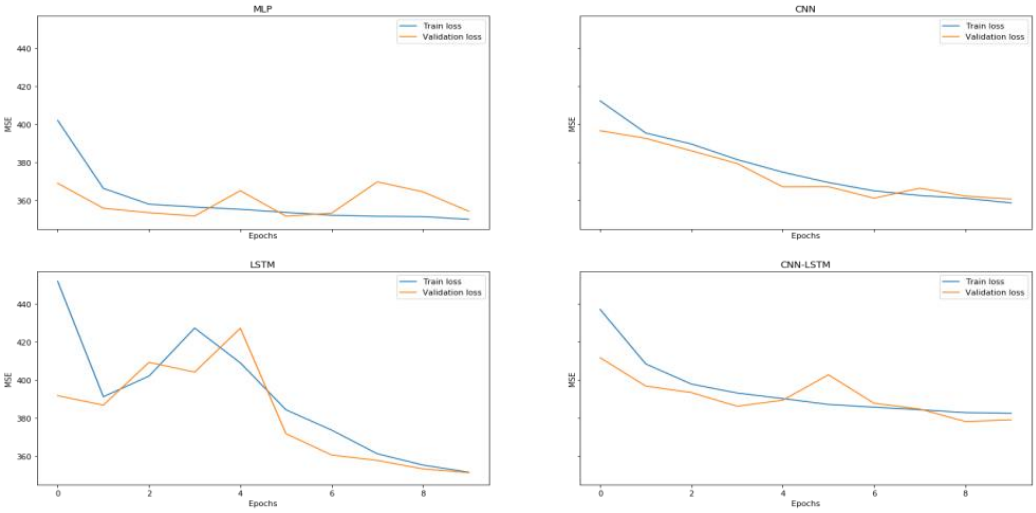
DOI: https://www.doi.org/10.22303/csrid.14.1.2022.68-80

D. Peramalan

Berdasarkan hasil dari analisis RFM, maka dapat disimpulkan bahwa segmentasi pelanggan yang paling tinggi berada pada Cluster 4 dengan 54.48% dari seluruh populasi dan total produk terjual setiap bulannya sebanyak 77.775 produk. Dari hasil ini maka tahapan selanjutnya adalah meramalkan jumlah persediaan produk yang optimal sehingga mengurangi tingkat persediaan pelanggan jangka pendek dan membantu dalam menentukan target penjualan yang realistis di masa depan. Pada penelitian ini, metode peramalan yang digunakan adalah model MLP, CNN, LSTM dan CNN-LSTM kemudian sebagai pembanding juga disajikan hasil peramalan dengan pendekatan model *Machine Learning*.

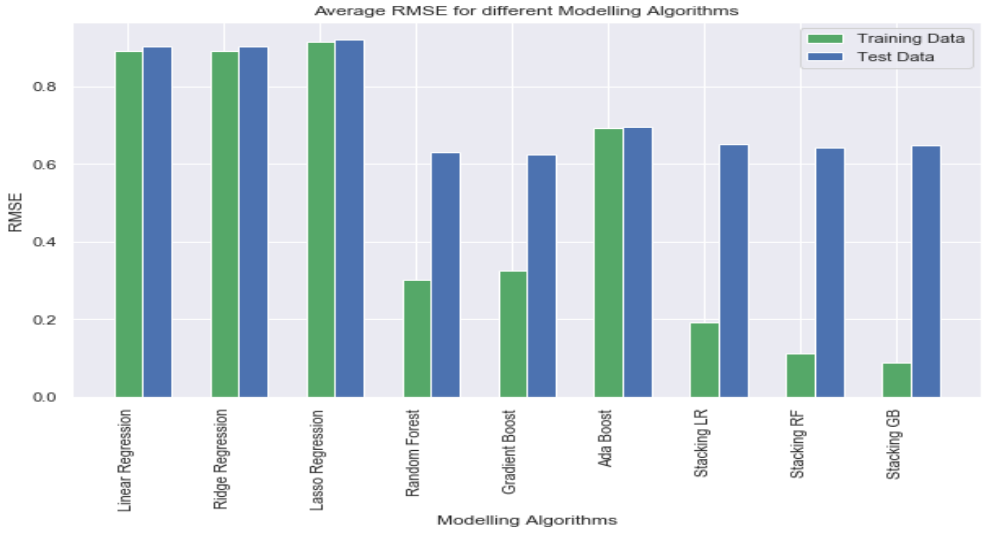
a. Analisis Peralaman *Deep Learning*

Pada bagian ini disajikan hasil peramalan menggunakan pendekatan *Deep Learning* yaitu MLP, CNN, LSTM dan CNN-LSTM. Tahapan pertama dilakukan adalah proses pelatihan yang dievaluasi menggunakan metode RMSE yang dapat dilihat pada gambar dibawah ini:



Gambar 13. Hasil Pelatihan model MLP, CNN, LSTM dan CNN-LSTM

b. Analisis Peramalan *Machine Learning*

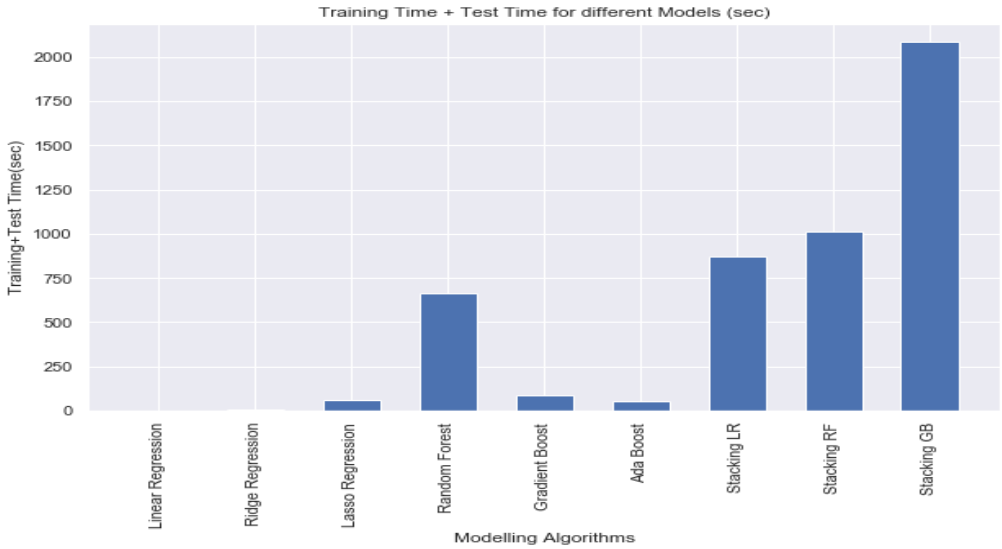


Gambar 14. Hasil pelatihan dan pengujian data model *Machine Learning*

e-ISSN: 2460-870X | p-ISSN: 2085-1367

E. Pembahasan

Berdasarkan hasil pengujian data transaksi penjualan produk farmasi yang bersumber dari PT. Anugrah Argon Medica yang merupakan salah satu perusahaan distribusi farmasi di Indonesia, maka dapat dianalisis hasilnya dimana penerapan metode RPCA sebagai model untuk pengurangan dimensi fitur pada dataset memiliki hasil yang signifikan terhadap waktu komputasi data. Selain itu analisis RFM berdasarkan algoritma K-Means Clustering dapat menghasilkan analisis segmentasi pelanggan dimana Cluster 4 merupakan segmen pelanggan yang paling menjanjikan untuk dipertahankan. Pada cluster ini, terdapat 505 customer setia dengan nilai rata-rata transaksi per tahunnya sebesar 41.02 % dan 28.03% keuntungan bersumber dari customer ini. Selanjutnya hasil peramalan dengan menggunakan dua pendekatan berbeda yaitu *Deep Learning* dan *Machine Learning* secara keseluruhan menghasilkan tingkat akuasi yang baik, akan tetapi kombinasi CNN-LSTM belum dapat menghasilkan akurasi seperti yang diharapkan. Pendekatan Gradient Boost membutuhkan waktu lebih sedikit dibandingkan dengan Random Forest. Sehingga model Gradient Boost dapat dikatakan sebagai algoritma terbaik dari algoritma lainnya. Gradient Boost memberikan RMSE 0,625 pada data pengujian.



Gambar 15. Hasil Waktu Pelatihan dan pengujian data

4. KESIMPULAN

Berdasarkan hasil pengujian yang dilakukan untuk peramalan data penjualan dengan kombinasi RPCA dan *K-Means Clustering*, maka dapat diambil kesimpulan sebagai berikut:

a. Penerapan algoritma RPCA secara signifikan dapat berguna untuk mengurangi dimensi fitur datataset. Selain itu juga dapat mengoptimalkan waktu komputasi yang digunakan.
b. Analisis RFM berdasarkan *K-Means Clustering* untuk segmentasi pelanggan dapat menggambarkan individu mana yang paling mungkin menjadi pelanggan perusahaan yang paling loyal dan berpotensi, dimana Cluster 4 merupakan segmen pelanggan yang paling menjanjikan untuk dipertahankan. Pada cluster ini, terdapat 505 customer setia dengan nilai rata-rata transaksi per tahunnya sebesar 41.02 % dan 28.03% keuntungan bersumber dari customer ini.
c. Peramalan dengan pendekatan *Deep Learning* MLP, CNN, LSTM dan CNN-LSTM secara keseluruhan menghasilkan tingkat akuasi yang baik, akan tetapi kombinasi CNN-LSTM belum dapat menghasilkan akurasi seperti yang diharapkan. Pendekatan Gradient Boost membutuhkan waktu lebih sedikit dibandingkan dengan Random Forest. Sehingga model Gradient Boost dapat dikatakan sebagai algoritma terbaik dari algoritma lainnya. Gradient Boost memberikan RMSE 0,625 pada data pengujian.

DOI: https://www.doi.org/10.22303/csrid.14.1.2022.68-80

UCAPAN TERIMA KASIH

Terima kasih saya sampaikan kepada Prof. Dr. Saib Suwilo dan Prof. Dr. Muhammad Zalis beserta keluarga dan rekan-rekan saya yang telah membimbing dan mendukung saya dalam terlaksananya penelitian ini.

REFERENSI

[1] Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019). Sales demand forecast in E-commerce using a long short-term memory neural network methodology. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11955 LNCS*, 462–474. https://doi.org/10.1007/978-3-030-36718-3_39

[2] Chen, I. F., & Lu, C. J. (2017). Sales forecasting by combining clustering and machine-learning techniques for computer retailing. *Neural Computing and Applications*, *28*(9), 2633–2647. https://doi.org/10.1007/s00521-016-2215-x

[3] Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons and Fractals*, *135*(January), 109864. https://doi.org/10.1016/j.chaos.2020.109864

[4] Choi, H. K. (2018). Stock Price Correlation Coefficient Prediction with ARIMA-LSTM Hybrid Model. *ArXiv*.

[5] Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent Neural Networks for Time Series Forecasting: Current status and future directions. *International Journal of Forecasting*, *37*(1), 388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008

[6] Karb, T., Kühl, N., Hirt, R., & Glivici-Cotruţă, V. (2020). A network-based transfer learning approach to improve sales forecasting of new products. In *arXiv* (Issue Ml, pp. 1–17).

[7] Karpathy, A. (2015). The Unreasonable Effectiveness of Recurrent Neural Networks. *Web Page*, 1–28.

[8] Khashei, M., & Hajirahimi, Z. (2019). A comparative study of series arima/mlp hybrid models for stock price forecasting. *Communications in Statistics: Simulation and Computation*, *48*(9), 2625–2640. https://doi.org/10.1080/03610918.2018.1458138

[9] Kuo, P. H., & Huang, C. J. (2018). An electricity price forecasting model by hybrid structured deep neural networks. *Sustainability (Switzerland)*, *10*(4), 1–17. https://doi.org/10.3390/su10041280

[10] Li, M., Ji, S., & Liu, G. (2018). Forecasting of Chinese E-Commerce Sales: An Empirical Comparison of ARIMA, Nonlinear Autoregressive Neural Network, and a Combined ARIMA-NARNN Model. *Mathematical Problems in Engineering*, *2018*, 1–12. https://doi.org/10.1155/2018/6924960

[11] Li, T., Hua, M., & Wu, X. (2020). A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5). *IEEE Access*, *8*, 26933–26940. https://doi.org/10.1109/ACCESS.2020.2971348

[12] Lu, W., Li, J., Li, Y., Sun, A., & Wang, J. (2020). A CNN-LSTM-Based Model to Forecast Stock Prices. *Complexity*, *2020*, 1–10. https://doi.org/10.1155/2020/6622927

[13] Neural, R., & Tutorial, N. (2000). *Recurrent Neural Networks Tutorial, Part 1 – Introduction to RNNs – WildML*. 1–14.

[14] Pra, M. Del. (n.d.). *Time Series Forecasting with Deep Learning and Attention Mechanism*. Medium Towards Data Science.

[15] Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and ARIMA models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing*, *34*, 151–163.

e-ISSN: 2460-870X | p-ISSN: 2085-1367

https://doi.org/10.1016/j.rcim.2014.12.015

[16] Reddy SK, B. A. (2015). Exchange Rate Forecasting using ARIMA, Neural Network and Fuzzy Neuron. *Journal of Stock & Forex Trading*, *04*(03). https://doi.org/10.4172/2168-9458.1000155

[17] Shams, M. Bin, Hossain, M. J., & Noori, S. R. H. (2020). A Time Series Analysis of Trends With Twitter Hashtags Using LSTM. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/ICCCNT49239.2020.9225349

[18] Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. https://doi.org/10.1016/j.physd.2019.132306

[19] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). A comparative analysis of forecasting financial time series using ARIMA, LSTM, and BiLSTM. In *arXiv*. arXiv.

[20] Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2019). A Comparison of ARIMA and LSTM in Forecasting Time Series. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 1394–1401. https://doi.org/10.1109/ICMLA.2018.00227

[21] Stanford University Course cs231n. (2018). CS231n Convolutional Neural Networks for Visual Recognition. *Stanford University Course Cs231n*, 30.

[22] Yang, C., Zhai, J., Tao, G., & Haajek, P. (2020). Deep Learning for Price Movement Prediction Using Convolutional Neural Network and Long Short-Term Memory. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/2746845

[23] Zhao, K., & Wang, C. (2017). Sales forecast in E-commerce using convolutional neural network. In *arXiv*.