PREDIKSI PERSEDIAAN DAN PERMINTAAN PASAR DENGAN MENGGUNAKAN METODE RANDOM FOREST DAN PRINCIPAL COMPONENT ANALYSIS

(MARKET SUPPLY AND DEMAND PREDICTION USING RANDOM FOREST AND PRINCIPAL COMPONENT ANALYSIS METHOD)

Fransiscus Aditya Wibowo[1)], dan Mardiani[2)]
[1, 2)]Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data palembang
Jl. Rajawali No. 14, Palembang 30113
e-mail: fransiscusadityawibowo_2426311003@mhs.mdp.ac.id [1)], mardiani@mdp.ac.id[2)]

ABSTRAK

Ketidaktepatan dalam memprediksi kebutuhan stok dan permintaan pasar merupakan permasalahan utama yang dihadapi oleh PT Bintang Jaya, yang berpotensi menimbulkan kelebihan stok (overstock) maupun kekurangan stok (stockout). Kondisi ini dapat berdampak pada peningkatan biaya penyimpanan, terganggunya arus kas, serta menurunnya tingkat kepuasan pelanggan. Permasalahan tersebut terjadi karena proses perencanaan stok masih mengandalkan pendekatan manual berbasis pengalaman historis tanpa pemanfaatan analisis data secara sistematis. Oleh karena itu, penelitian ini bertujuan menerapkan teknik prediksi berbasis machine learning guna meningkatkan akurasi peramalan kebutuhan stok dan permintaan pasar. Metode yang digunakan adalah algoritma Random Forest sebagai model baseline serta Random Forest berbasis Principal Component Analysis (PCA) untuk mengevaluasi pengaruh reduksi dimensi terhadap kinerja prediksi. Data yang digunakan merupakan data historis transaksi penjualan PT Bintang Jaya periode 2022–2024 yang telah melalui tahapan pra-pemrosesan, agregasi bulanan, dan pembentukan fitur time series. Dataset dibagi menggunakan skema 70:30, yaitu 70% sebagai data latih dan 30% sebagai data uji untuk mengukur kemampuan generalisasi model. Hasil penelitian menunjukkan bahwa Random Forest menghasilkan prediksi yang lebih stabil dan mendekati nilai aktual dibandingkan model hybrid RF+PCA. Nilai koefisien determinasi (R^2) sebesar 0,69 menunjukkan bahwa model mampu menjelaskan sekitar 69% variasi permintaan, dengan nilai MAE dan RMSE yang konsisten. Temuan ini mengindikasikan bahwa penggunaan Random Forest efektif dalam mendukung pengambilan keputusan berbasis data pada perencanaan stok perusahaan.

***Kata Kunci**: Machine Learning, PCA, Permintaan Pasar, Prediksi Stok, Random Forest*

ABSTRACT

Inaccuracies in predicting stock requirements and market demand are major problems faced by PT Bintang Jaya, which have the potential to cause overstocking and stockouts. These conditions can lead to increased storage costs, disrupted cash flow, and decreased customer satisfaction. This problem occurs because the inventory planning process still relies on a manual approach based on historical experience without the systematic use of data analysis. Therefore, this study aims to apply machine learning-based prediction techniques to improve the accuracy of forecasting inventory requirements and market demand. The methods used are the Random Forest algorithm as a baseline model and Random Forest based on Principal Component Analysis (PCA) to evaluate the effect of dimension reduction on prediction performance. The data used is historical sales transaction data from PT Bintang Jaya for the 2022–2024 period, which has undergone pre-processing, monthly aggregation, and time series feature formation. The dataset is divided using a 70:30 scheme, with 70% as training data and 30% as test data to measure the model's generalization ability. The results showed that Random Forest produced more stable predictions that were closer to the actual values than the RF+PCA hybrid model. The coefficient of determination (R^2) value of 0.69 indicated that the model was able to explain approximately 69% of the demand variation, with consistent MAE and RMSE values. These findings indicated that the use of Random Forest was effective in supporting data-driven decision making in corporate inventory planning.

***Keywords**: Inventory Prediction, Machine Learning, Market Demand, PCA, Random Forest*

I. PENDAHULUAN

Pada era persaingan global dan kebutuhan konsumen yang dinamis, pengelolaan persediaan menjadi salah satu tantangan utama dalam manajemen rantai pasokan. Ketidakseimbangan antara jumlah stok dalam angka dengan permintaan pasar yang juga ditunjukkan dalam angka dapat mengakibatkan pemborosan biaya atau kerugian pendapatan.

Banyak perusahaan mengalami kelebihan maupun kekurangan stok akibat minimnya akurasi dalam prediksi permintaan. Kondisi ini tidak hanya berdampak pada efisiensi operasional, tetapi juga menurunkan tingkat kepuasan pelanggan serta mempengaruhi profitabilitas perusahaan secara keseluruhan.

Seiring dengan perkembangan teknologi informasi dan ketersediaan data dalam jumlah besar (*big data*), banyak perusahaan mulai memanfaatkan pendekatan berbasis data (*data-driven decision making*) untuk mengatasi berbagai permasalahan operasional [1], termasuk pengelolaan persediaan. Salah satu pendekatan yang banyak digunakan adalah konsep data mining penggunaannya dibuktikan secara empiris [2] yang menunjukkan bahwa data mining mampu meningkatkan akurasi prediksi penjualan bahan bangunan. [3] dan pemanfaatan algoritma kecerdasan buatan (*Artificial Intelligence*) dan *machine learning* untuk memodelkan dan memprediksi permintaan pasar. Pendekatan data mining juga digunakan [4] yang membandingkan efektivitas algoritma *Random Forest* dengan *K-Nearest Neighbor* dalam klasifikasi produk. Penelitian tersebut menekankan bahwa *Random Forest* unggul dalam stabilitas hasil pada data klasifikasi berskala kecil, [5] mengembangkan sistem prediksi kebutuhan stok yang lebih stabil dengan mengintegrasikan *Random Forest* dan PCA [6], sementara [7] menunjukkan efektivitas *Random Forest* dalam mengidentifikasi pola musiman penjualan pada *platform e-commerce*. Temuan-temuan tersebut menjadi relevan dalam konteks penelitian ini. Dalam konteks ini, PCA (*Principal Component Analysis*) sering digunakan untuk menyaring informasi terpenting dari data yang kompleks, sehingga membantu dalam proses ekstraksi pengetahuan secara efisien. Misalnya, [8] memanfaatkan PCA untuk menyederhanakan kompleksitas data spasial, sedangkan [9] menerapkan PCA pada dataset penjualan untuk merangkum fitur-fitur paling signifikan

Dengan adanya model prediksi yang akurat, perusahaan dapat melakukan perencanaan produksi, distribusi, dan pengadaan barang yang lebih efisien dan tepat sasaran. Salah satu metode langkah dalam prediksi permintaan dan stok adalah algoritma *Random Forest*, sebuah metode machine learning berbasis ensemble yang memiliki kemampuan klasifikasi dan regresi dengan tingkat akurasi tinggi. *Random Forest* bekerja dengan membangun banyak decision tree dan menggabungkan hasilnya untuk memperoleh prediksi yang lebih akurat dan stabil. Keunggulan utama dari algoritma ini adalah kemampuannya dalam menangani data dalam jumlah besar, kompleks, serta resistensinya terhadap *overfitting* yang kerap terjadi pada model prediksi lain [10]. Untuk meningkatkan efektivitas dari algoritma ini, metode *Principal Component Analysis* (PCA) sering digunakan sebagai teknik reduksi dimensi. PCA memungkinkan pengurangan jumlah fitur tanpa menghilangkan informasi penting, sehingga mempercepat proses perhitungan dan meningkatkan akurasi model. Dengan kata lain, PCA membantu menyaring informasi yang paling relevan dari dataset besar yang kompleks, sehingga mempermudah proses pembelajaran mesin dan pengambilan keputusan berbasis data [11]. Untuk meningkatkan efektivitas pemodelan prediksi, Principal Component Analysis (PCA) sering digunakan sebagai teknik reduksi dimensi karena mampu mentransformasikan sejumlah fitur yang saling berkorelasi menjadi komponen utama yang ortogonal (tidak berkorelasi). Pada dataset prediksi permintaan berbasis time series, fitur yang umum dibentuk—misalnya *lag* penjualan, *moving average*, tren musiman, dan fitur turunan lain—cenderung menghasilkan multikolinearitas dan redundansi informasi. Dalam kondisi seperti ini, PCA relevan karena dapat merangkum variasi terbesar data ke dalam beberapa komponen yang mewakili pola utama (tren dan musiman), sehingga berpotensi mengurangi noise, menekan risiko *overfitting*, dan mempercepat proses pelatihan model tanpa menghilangkan struktur informasi yang dominan. Selain itu, studi demand forecasting menunjukkan bahwa kombinasi PCA + Random Forest dapat digunakan untuk mereduksi variabel dan menjaga akurasi prediksi pada beberapa konteks peramalan permintaan [12]. Di sisi lain, literatur juga menegaskan bahwa PCA merupakan pendekatan reduksi dimensi yang luas digunakan dalam *forecasting* untuk mengekstraksi faktor/komponen dari prediktor berdimensi tinggi [13]. Namun, dampak PCA dapat bervariasi tergantung karakteristik data dan algoritma, sehingga evaluasi empiris tetap diperlukan pada dataset penelitian ini [14].

Oleh karena itu, dibutuhkan upaya untuk menerapkan pendekatan yang lebih modern dan adaptif dalam memprediksi kebutuhan stok. Dengan memanfaatkan teknologi machine learning dan pendekatan statistika modern seperti *Random Forest* dan PCA, perusahaan diharapkan mampu

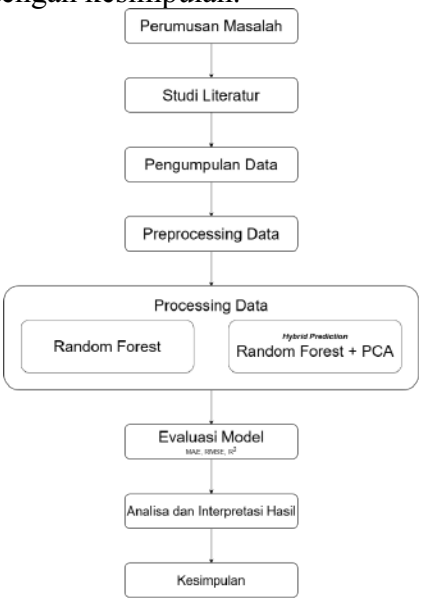
membuat strategi pengelolaan persediaan yang lebih akurat dan efisien. Dengan menggabungkan kedua pendekatan tersebut, penelitian ini bertujuan untuk menerapkan metode *Random Forest* dan PCA secara bersamaan dalam rangka memprediksi kebutuhan stok dan permintaan pasar secara lebih akurat di PT Bintang Jaya.

II. STUDI PUSTAKA

Beberapa penelitian sebelumnya juga telah mengkaji penerapan metode *Random Forest* [15] dan PCA dalam konteks prediksi penjualan, manajemen stok, maupun analisis data yang kompleks [16] menunjukkan bahwa *Random Forest* mampu menghasilkan prediksi yang lebih stabil dibandingkan metode konvensional seperti regresi linear [17] membuktikan bahwa PCA efektif dalam menyederhanakan struktur data tanpa mengurangi kualitas informasi, sehingga model prediksi menjadi lebih ringan dan efisien. Sementara itu, [18] berhasil mengintegrasikan PCA dengan *Random Forest* [19] untuk menstabilkan prediksi kebutuhan stok barang dalam sistem distribusi juga menerapkan kombinasi PCA dan *Random Forest* pada kasus prediksi penjualan bahan bangunan dan memperoleh peningkatan akurasi yang signifikan. Selanjutnya, [1] menyimpulkan bahwa integrasi kedua metode tersebut mampu mengoptimalkan proses pengambilan keputusan dalam manajemen rantai pasok. Namun demikian, penelitian sebelumnya umumnya berfokus pada dataset berdimensi tinggi atau konteks industri tertentu, sehingga belum banyak yang mengevaluasi efektivitas integrasi PCA dan Random Forest pada dataset penjualan regional dengan karakteristik fitur time series yang relatif terbatas. Oleh karena itu, novelty penelitian ini terletak pada evaluasi empiris terhadap kontribusi PCA dalam meningkatkan kinerja Random Forest pada kasus distribusi bahan bangunan di wilayah Sumatera Selatan dan Bengkulu. Masalah ketidaksesuaian antara stok dan permintaan yang dialami PT Bintang Jaya menunjukkan pentingnya pendekatan berbasis data untuk meningkatkan akurasi prediksi serta efisiensi distribusi.

III. METODE PENELITIAN

Alur metodologi penelitian yang digunakan secara sistematis mulai dari tahap awal hingga menghasilkan kesimpulan yang digambarkan pada Gambar 1. Proses diawali dengan perumusan masalah sebagai dasar dalam menentukan tujuan dan fokus penelitian, kemudian dilanjutkan dengan studi literatur untuk memperkuat landasan teori serta menemukan pendekatan yang relevan dengan topik penelitian. Setelah itu dilakukan pengumpulan data yang menjadi sumber utama dalam proses analisis, lalu data tersebut dipersiapkan melalui tahap *preprocessing* untuk memastikan kualitas data layak digunakan, seperti pembersihan, seleksi, dan transformasi data. Pada tahap *processing* data, penelitian menerapkan dua pendekatan pemodelan, yaitu *Random Forest* sebagai model *baseline* dan *Hybrid Prediction Random Forest* + PCA sebagai metode kombinasi untuk meningkatkan performa melalui reduksi dimensi dan penguatan pola fitur. Selanjutnya dilakukan evaluasi model menggunakan metrik MAE, RMSE, dan R^2 untuk membandingkan tingkat akurasi dan kestabilan hasil prediksi. Tahap akhir meliputi analisis dan interpretasi hasil guna menjelaskan makna temuan penelitian, serta ditutup dengan kesimpulan.



Gambar 1. Metode Penelitian

Langkah-langkah penelitian yang dilakukan dalam studi ini bertujuan untuk menerapkan metode *Principal Component Analysis* (PCA) dan *Random Forest* dalam melakukan prediksi permintaan stok barang pada perusahaan manufaktur. Penelitian ini dilaksanakan dengan melalui tahapan-tahapan yang saling berkesinambungan, dimulai dari identifikasi masalah, studi literatur, pengumpulan dan *preprocessing* data, transformasi data menggunakan PCA, pembangunan model prediktif dengan algoritma *Random Forest*, evaluasi performa model, hingga penarikan kesimpulan dari hasil yang diperoleh. Secara umum, tahapan

penelitian yang dilakukan dapat diuraikan sebagai berikut:

A. Perumusan Masalah

Pada Langkah pertama ini dilakukan penentukan isu utama yang menjadi fokus penelitian, yaitu bagaimana melakukan prediksi permintaan stok secara lebih akurat menggunakan pendekatan data mining.

B. Studi Literatur

Melakukan kajian terhadap jurnal dan penelitian terdahulu dengan rentang waktu 5 tahun terakhir yang relevan dengan topik PCA dan *Random Forest* dalam konteks prediksi.

C. Pengumpulan data

Mengambil data historis penjualan dan persediaan barang dari sistem Penjualan perusahaan pada PT Bintang Jaya. Data yang dikumpulkan mencakup informasi tanggal transaksi, nama produk, jumlah penjualan, harga, stok, dan atribut terkait lainnya.

D. Preprocessing Data

Melakukan tahap pembersihan data (*missing value* dan *outlier*), normalisasi variabel numerik, konversi format, serta encoding variabel kategorikal. Data kemudian disiapkan dalam bentuk dataset terstruktur yang dapat diproses oleh algoritma.

E. Processing Data

Tahap ini membangun dua model prediktif secara paralel untuk keperluan evaluasi dan komparasi :

- Model A – *Random Forest*: Menggunakan semua fitur asli tanpa reduksi dimensi.
- Model B – PCA + *Random Forest*: Menggunakan komponen utama hasil PCA sebagai input model, untuk mengurangi kompleksitas dan potensi multikolinearitas.

F. Evaluasi Model

Melakukan pengukuran performa model menggunakan tiga metrik utama: MAE (*Mean Absolute Error*), RMSE (*Root Mean Square Error*), dan R-*Squared* (R^2). Evaluasi dilakukan untuk kedua model guna mengetahui perbedaan akurasi dan efisiensi.

G. Analisis dan Interpretasi Hasil

Menafsirkan hasil evaluasi dan mengidentifikasi faktor-faktor yang paling berpengaruh dalam prediksi permintaan berdasarkan *feature importance* dari *Random Forest*. Hasil dari kedua model dibandingkan untuk melihat apakah penggunaan PCA memberikan dampak positif terhadap kinerja model.

H. Kesimpulan

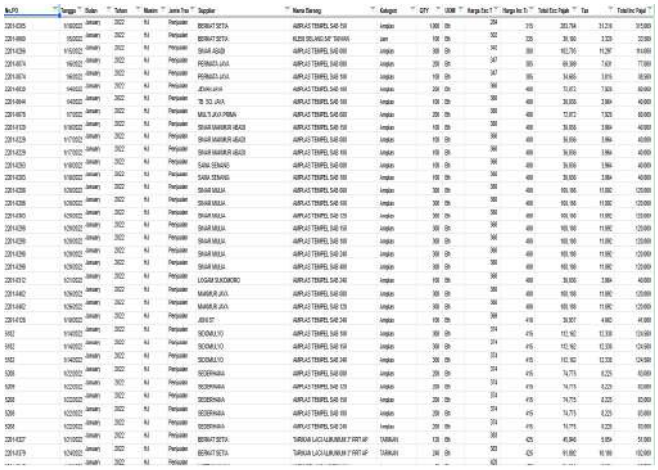
Menarik simpulan berdasarkan hasil eksperimen dan analisis. Peneliti memberikan rekomendasi terhadap metode terbaik yang dapat diterapkan oleh PT Bintang Jaya dalam mengoptimalkan pengadaan dan distribusi stok secara *data-driven*.

IV. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil penelitian dan pembahasan terkait penerapan metode *Random Forest* serta pendekatan *hybrid Random Forest* + PCA dalam membangun model prediksi. Pembahasan difokuskan pada hasil pengolahan data setelah tahap preprocessing, performa model yang dihasilkan, serta perbandingan kedua pendekatan berdasarkan metrik evaluasi MAE, RMSE, dan R^2. Hasil yang diperoleh selanjutnya dianalisis untuk menilai efektivitas penggunaan PCA dalam meningkatkan akurasi dan stabilitas prediksi, sekaligus memberikan interpretasi terhadap pola dan karakteristik data yang memengaruhi kinerja model.

A. Hasil Eksperimen

Pada pengolahan data, dataset yang digunakan dalam penelitian diperoleh dari sistem *Enterprise Resource Planning* (ERP) perusahaan, yaitu sistem HexPos, yang mencatat transaksi permintaan barang selama periode Januari 2022 hingga Desember 2024, dengan total 59.388 baris data transaksi penjualan yang diperoleh dari departemen *Production Planning and Inventory Control* (PPIC) serta gudang pusat. Dataset ini bersifat sekunder dan diambil secara langsung dari bagian perencanaan produksi (PPIC) dan gudang, dengan izin dan pengawasan dari pemilik perusahaan. Dataset ini dipilih karena mewakili variasi kondisi ekonomi dan permintaan pasar yang berfluktuasi, termasuk masa pandemi COVID-19 dan periode pemulihan ekonomi setelahnya. Kondisi tersebut memberikan gambaran dinamis mengenai pola distribusi dan permintaan stok di wilayah Sumatera Selatan dan Bengkulu. Berikut dataset yang digunakan pada gambar 2.



Gambar 2. Dataset Transaksi Penjualan

Selanjutnya, tahap preprocessing data dilakukan secara bertahap untuk memastikan dataset siap digunakan dalam proses pemodelan dan menghasilkan performa prediksi yang optimal. Tahapan ini diawali dengan proses import data ke dalam lingkungan pemrograman untuk dilakukan pemeriksaan awal terhadap struktur dataset, termasuk jumlah baris, jumlah atribut, tipe data, serta distribusi nilai pada masing-masing variabel. Pemeriksaan ini bertujuan untuk mengidentifikasi potensi permasalahan seperti inkonsistensi format data, duplikasi, maupun ketidaksesuaian tipe data.

Tahap berikutnya adalah pemeriksaan dan penanganan nilai kosong (missing values) guna mencegah distorsi dalam proses pelatihan model. Setelah itu dilakukan proses pembersihan data (cleansing), seperti penghapusan duplikasi, koreksi kesalahan input, serta penyaringan data yang tidak relevan. Selanjutnya dilakukan transformasi dan encoding terhadap variabel kategorikal agar dapat diproses oleh algoritma machine learning, serta normalisasi atau standarisasi pada variabel numerik untuk menyamakan skala nilai antar fitur.

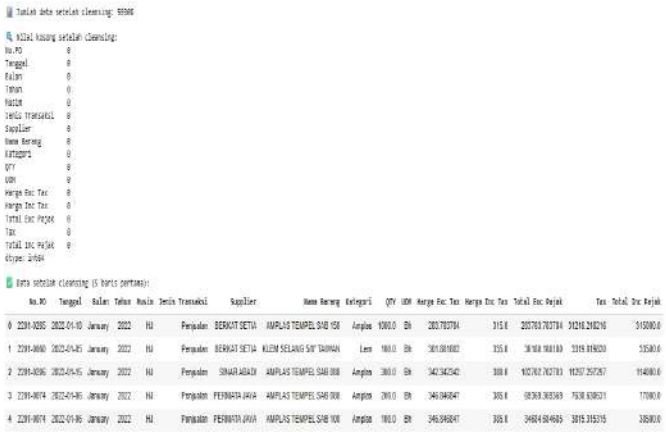
Pada tahap akhir, dataset dibagi menjadi data latih dan data uji dengan proporsi tertentu untuk memastikan evaluasi model dilakukan secara objektif. Gambar yang disajikan memperlihatkan hasil pemeriksaan struktur data, hasil proses cleansing, transformasi dan encoding, serta pembagian data latih dan data uji sebagai bagian dari tahapan preprocessing yang sistematis.

```
🔍 Informasi struktur dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59309 entries, 0 to 59308
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   No.PO            59309 non-null  object
 1   Tanggal          59309 non-null  datetime64[ns]
 2   Bulan            59309 non-null  object
 3   Tahun            59309 non-null  int64
 4   Musim            59309 non-null  object
 5   Jenis Transaksi  59309 non-null  object
 6   Supplier         59309 non-null  object
 7   Nama Barang      59309 non-null  object
 8   Kategori         59309 non-null  object
 9   QTY              59309 non-null  float64
 10  UOM              59309 non-null  object
 11  Harga Exc Tax    59309 non-null  float64
 12  Harga Inc Tax    59309 non-null  float64
 13  Total Exc Pajak  59308 non-null  float64
 14  Tax              59308 non-null  float64
 15  Total Inc Pajak  59308 non-null  float64
dtypes: datetime64[ns](1), float64(6), int64(1), object(8)
memory usage: 7.2+ MB
None

📊 Jumlah nilai kosong per kolom:
No.PO              0
Tanggal            0
Bulan              0
Tahun              0
Musim              0
Jenis Transaksi    0
Supplier           0
Nama Barang        0
Kategori           0
QTY                0
UOM                0
Harga Exc Tax      0
Harga Inc Tax      0
Total Exc Pajak    1
Tax                1
Total Inc Pajak    1
dtype: int64
```

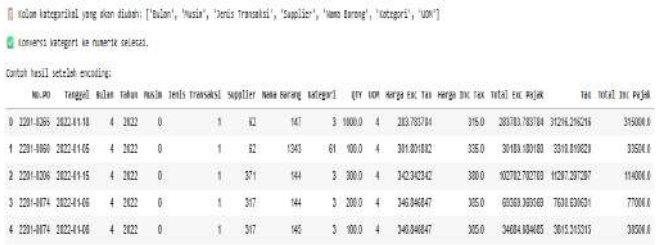
Gambar 3. Hasil Pemeriksaan Struktur

Langkah awal dalam proses pengolahan data adalah melakukan pemeriksaan terhadap struktur dataset dan kelengkapan nilai data pada Gambar 3. Hasil pemeriksaan menunjukkan bahwa terdapat beberapa nilai kosong, namun jumlahnya sangat kecil (kurang dari 3 baris) dan baris-baris tersebut dihapus agar data menjadi konsisten dan tidak mengganggu proses pembelajaran model. Gambar 2 memperlihatkan bahwa dataset memiliki total 59.309 baris (records) dan 16 kolom (features). Ukuran memori dataset tercatat sekitar 7,2 MB, yang berarti data berukuran ringan dan dapat diolah secara efisien dalam lingkungan komputasi berbasis *cloud* seperti Google Colab.



Gambar 4. Hasil Data Cleansing

Setelah dilakukan pemeriksaan terhadap struktur dan kelengkapan dataset, tahap selanjutnya adalah pembersihan data (data *cleansing*) dan *preprocessing* yang terdiri dari menangani nilai kosong (*missing value)*, menyesuaikan format data kolom tanggal, dan pemeriksaan ulang jumlah data, Gambar 4 menunjukan hasil dari data *cleansing*. Tahapan ini bertujuan untuk memastikan bahwa data yang digunakan dalam proses analisis dan pemodelan memiliki kualitas yang baik, tidak mengandung nilai kosong, dan memiliki format yang sesuai dengan kebutuhan algoritma machine learning.


Kolom kategorikal yang akan diubah: ['Bulan', 'Musim', 'Jenis Transaksi', 'Supplier', 'Nama Barang', 'Kategori', 'UOM']

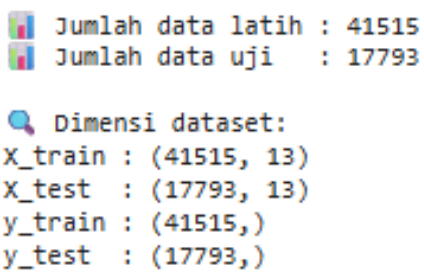
Konversi kategori ke numerik selesai.

Contoh hasil setelah encoding:

	No.PO	Tanggal	Bulan	Tahun	Musim	Jenis Transaksi	Supplier	Nama Barang	Kategori	QTY	UOM	Harga Exc Tax	Harga Inc Tax	Total Exc Pajak	Pajak	Total Inc Pajak
0	2201-0265	2022-01-10	4	2022	0	1	62	147	3	1000.0	4	283783784	315.0	283783.783784	31216.216216	315000.0
1	2201-0060	2022-01-05	4	2022	0	1	62	1345	61	100.0	4	301.801802	335.0	30180.180180	3318.819820	33500.0
2	2201-0206	2022-01-15	4	2022	0	1	371	144	3	300.0	4	342.342342	380.0	102702.702703	11297.297297	114000.0
3	2201-0174	2022-01-06	4	2022	0	1	317	144	3	200.0	4	340.846847	385.0	69369.369369	7630.630631	77000.0
4	2201-0174	2022-01-06	4	2022	0	1	317	145	3	100.0	4	340.846847	385.0	34684.684685	3815.315315	38500.0

Gambar 5. Hasil Transformasi dan Encoding Data

Tahapan transformasi data dan encoding dilakukan untuk menyiapkan dataset agar dapat diproses secara optimal oleh algoritma machine learning. Dalam penelitian ini, dataset terdiri atas kombinasi variabel kategorikal (bertipe teks) dan numerik, sehingga diperlukan proses konversi nilai teks menjadi angka serta normalisasi skala nilai numerik agar memiliki bobot yang seimbang dalam proses pelatihan model. Proses transformasi ini dilakukan melalui dua variable kategorikal utama, yaitu encoding variabel kategorikal dan normalisasi fitur numerik, sebagaimana hasilnya ditunjukkan pada Gambar 5.


Jumlah data latih : 41515
Jumlah data uji : 17793

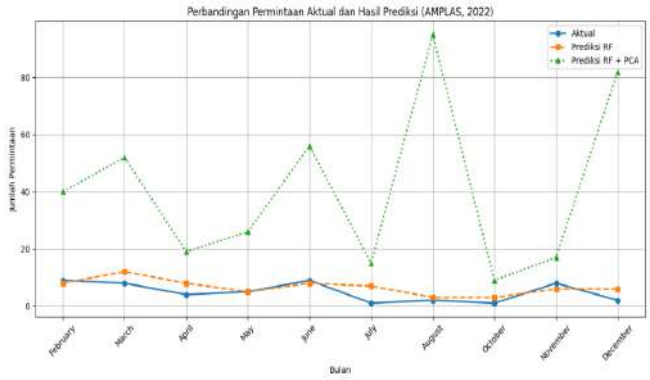
Dimensi dataset:
X_train : (41515, 13)
X_test : (17793, 13)
y_train : (41515,)
y_test : (17793,)

Gambar 6. Data Split Uji dan Data Train

Setelah data selesai melalui tahap transformasi dan encoding, Langkah selanjutnya adalah melakukan pemisahan data (*data splitting*) menjadi dua bagian utama, yaitu data latih (*training data*) dan data uji (*testing data*). Tujuan dari proses ini adalah untuk memisahkan bagian data yang digunakan untuk membangun model dan bagian lainnya untuk mengukur performa serta kemampuan generalisasi model terhadap data baru yang belum pernah dilihat sebelumnya. Proses pemisahan dilakukan menggunakan fungsi train_test_split() dari *library Scikit-Learn* dengan parameter test_size=0.3, yang berarti 70% dari total data digunakan untuk pelatihan model, dan 30% sisanya untuk pengujian. Proporsi ini dipilih karena dianggap ideal dalam menjaga keseimbangan antara kebutuhan pelatihan dan validasi model pada dataset berukuran menengah, yang ditunjukan pada gambar 6.

Selanjutnya dilakukan pembangunan model prediksi untuk mengestimasi jumlah penjualan (permintaan) barang pada periode berikutnya berdasarkan data historis transaksi PT Bintang Jaya. Model prediksi yang dikembangkan bertujuan untuk membantu perusahaan dalam melakukan perencanaan persediaan secara lebih akurat, sehingga risiko terjadinya kelebihan stok (overstock) maupun kekurangan stok (stockout) dapat diminimalkan. Penelitian ini menggunakan pendekatan machine learning berbasis supervised learning, di mana model dilatih menggunakan data historis yang telah diketahui nilai targetnya, kemudian diuji kemampuannya dalam memprediksi data baru. Variabel target (output) dalam penelitian ini adalah jumlah penjualan barang, sedangkan variabel input (fitur) mencakup kombinasi atribut waktu, kategori produk, serta variabel numerik transaksi yang telah melalui proses pra-pemrosesan pada tahap sebelumnya. Dua pendekatan model prediksi digunakan dalam penelitian ini, yaitu Model Random Forest dan Model Hybird Random Forest + Principal Component Analysis (PCA). Model Random Forest sebagai model baseline, yang digunakan untuk membangun prediksi langsung dari data hasil pra-pemrosesan tanpa reduksi dimensi. Model didefinisikan dengan parameter utama n_estimators=100, yang berarti terdapat 100 pohon keputusan yang digunakan untuk menghasilkan hasil akhir prediksi. Parameter n_jobs=-1 digunakan agar seluruh prosesor komputasi dapat dimanfaatkan, sehingga mempercepat proses komputasi. Gambar 7 menunjukan visualisasi dari hasil model prediksi dengan random forest yang dibandingkan dengan pendekatan random forest berbasis PCA yang menjelaskan bahwa model mampu menangkap pola permintaan historis dengan cukup baik, khususnya pada kategori dengan volume permintaan yang besar dan konsisten, meskipun terdapat selisih, nilai prediksi

tersebut masih berada dalam rentang yang wajar dan mencerminkan tren permintaan tinggi pada kategori tersebut.



Gambar 7. Hasil Prediksi Stok Dengan Random Forest

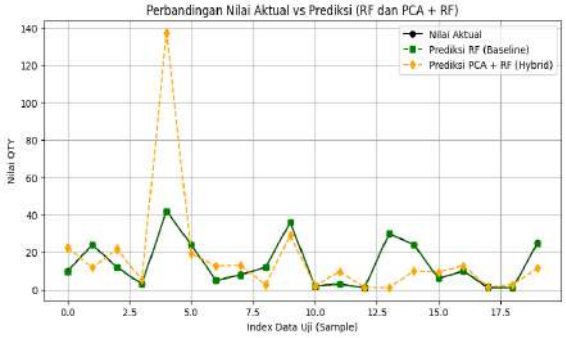
Model Hybrid Random Forest dan Principal Component Analysis (PCA) dimana PCA digunakan sebagai tahap reduksi dimensi setelah data diproses oleh algoritma Random Forest. Setelah model baseline Random Forest berhasil dibangun dan dievaluasi, tahap berikutnya adalah melakukan penggabungan antara metode Principal Component Analysis (PCA) dengan algoritma Random Forest untuk membentuk model hybrid. Tujuan dari pendekatan ini adalah untuk mengurangi dimensi data (reduksi fitur) sehingga model dapat bekerja lebih efisien dengan fokus pada komponen utama yang paling berpengaruh terhadap variabel target (QTY). Dataset hasil normalisasi (X_train dan X_test) ditransformasikan menggunakan PCA dengan parameter n_components=0.95, yang berarti komponen utama dipilih hingga mampu menjaga 95% dari total variansi data asli. Berdasarkan hasil proses PCA, diperoleh 1 komponen utama (principal component) yang mewakili keseluruhan fitur input dengan proporsi variansi terbesar. Komponen utama hasil PCA kemudian dijadikan input bagi model Random Forest Regressor dengan parameter yang sama seperti model baseline (n_estimators=100 dan random_state=42). Proses pelatihan dan pengujian dilakukan menggunakan data latih (X_train_pca, y_train) dan data uji (X_test_pca, y_test). Evaluasi lanjutan terhadap model prediksi dengan mengintegrasikan Principal Component Analysis (PCA) ke dalam algoritma Random Forest. Model ini merupakan pendekatan hybrid yang bertujuan untuk menguji pengaruh reduksi dimensi terhadap kinerja model baseline Random Forest yang telah dibahas pada bagian sebelumnya. PCA diterapkan sebagai tahap pra-pemrosesan data dengan mereduksi variabel input ke dalam sejumlah komponen utama yang memiliki varian terbesar. Selanjutnya, komponen hasil PCA digunakan sebagai input bagi model Random Forest untuk menghasilkan prediksi permintaan stok. Pendekatan ini diharapkan dapat mengurangi redundansi antar fitur serta meningkatkan efisiensi komputasi model. Hasil prediksi permintaan stok menggunakan model Random Forest berbasis PCA untuk tahun 2022 ditunjukkan pada gambar 8.

Hasil Perbandingan Prediksi Model Random Forest dan PCA+RF

	Nilai Aktual (y_test)	Prediksi_RF	Prediksi_PCA_RF	Selisih_RF	Selisih_PCA_RF
0	10.0	9.76	22.481984	0.24	12.481984
1	24.0	24.05	11.856667	0.05	12.143333
2	12.0	12.00	21.620000	0.00	9.620000
3	3.0	3.00	5.217500	0.00	2.217500
4	42.0	41.91	137.010000	0.09	95.010000
5	24.0	24.00	19.320000	0.00	4.680000
6	5.0	5.00	12.694857	0.00	7.694857
7	8.0	7.86	13.000000	0.14	5.000000
8	12.0	12.00	2.770000	0.00	9.230000
9	36.0	36.10	29.040000	0.10	6.960000
10	2.0	1.99	2.030000	0.01	0.030000
11	3.0	3.00	9.820000	0.00	6.820000
12	1.0	1.00	1.180000	0.00	0.180000
13	30.0	30.27	1.030000	0.27	28.970000
14	24.0	24.00	10.140000	0.00	13.860000

Gambar 8. Perbandingan Hasil Prediksi Random Forest vs Model Hybird Random Forest+PCA

Gambar tersebut menampilkan perbandingan antara permintaan aktual dan hasil prediksi RF+PCA pada beberapa kategori produk utama. Berdasarkan hasil yang diperoleh, terlihat bahwa prediksi RF+PCA cenderung memiliki deviasi yang lebih besar terhadap nilai aktual dibandingkan model Random Forest tanpa PCA. Pada beberapa kategori dengan permintaan tinggi, seperti PAKU, SEAL, dan Pipa, hasil prediksi RF+PCA masih mampu mengikuti pola umum permintaan, namun tingkat akurasinya relatif lebih rendah dibandingkan model baseline. Sementara itu, pada kategori dengan permintaan rendah hingga menengah, selisih antara nilai aktual dan hasil prediksi RF+PCA terlihat lebih signifikan.



Gambar 8. Visualisasi Perbandingan Model

Penggunaan dua pendekatan ini bertujuan untuk mengevaluasi pengaruh penerapan reduksi dimensi terhadap kinerja prediksi, serta membandingkan tingkat akurasi antara model baseline dan model hybrid. Hasil dari kedua model selanjutnya akan dievaluasi menggunakan metrik Mean Absolute Error (MAE), Root Mean Square Error (RMSE), dan Coefficient of Determination (R^2) untuk menentukan model yang paling sesuai diterapkan pada kasus prediksi permintaan di PT Bintang Jaya.

Terakhir dilakukan evaluasi model, secara kuantitatif, perbandingan metrik evaluasi dari kedua model disajikan pada tabel 1 berikut:

Tabel 1.
Perbandingan Hasil Evaluasi Model Random Forest vs Model Random Forest+PCA

Metrik Evaluasi	*Random Forest (Baseline)*	PCA + *Random Forest*
Mean Absolute Error (MAE)	2.3259	23.8795
Root Mean Squared Error (RMSE)	97.9378	170.3655
Coefficient of Determination (R^2)	0.6931	0.0714

Berdasarkan hasil pada Tabel 1, dapat dijelaskan bahwa nilai MAE dan RMSE pada model Random Forest (Baseline) lebih kecil dibandingkan dengan model PCA + Random Forest. Hal ini menunjukkan bahwa tingkat kesalahan rata-rata serta deviasi prediksi terhadap data aktual lebih rendah pada model Random Forest tanpa reduksi dimensi, sehingga menghasilkan akurasi prediksi yang lebih baik. Nilai R^2 sebesar 0,6931 pada model Random Forest mengindikasikan bahwa sekitar 69,31% variasi permintaan dapat dijelaskan oleh model. Sebaliknya, nilai R^2 pada model PCA + Random Forest yang jauh lebih rendah menunjukkan bahwa proses reduksi dimensi melalui PCA menyebabkan berkurangnya informasi penting yang dibutuhkan untuk menjelaskan variasi data permintaan secara optimal.

Untuk memastikan bahwa perbedaan performa antar model tidak disebabkan oleh variasi pembagian data, secara metodologis analisis signifikansi statistik seperti k-fold cross-validation disertai uji paired t-test atau Wilcoxon signed-rank test sangat direkomendasikan. Namun, penelitian ini menggunakan skema hold-out 70:30, sehingga evaluasi dilakukan pada data uji independen yang tidak terlibat dalam proses pelatihan. Meskipun demikian, perbedaan nilai MAE, RMSE, dan R^2 antara kedua model menunjukkan selisih yang cukup besar, sehingga secara empiris Random Forest menunjukkan keunggulan performa yang konsisten dibandingkan model PCA + Random Forest. Analisis signifikansi statistik yang lebih komprehensif dapat menjadi pengembangan pada penelitian selanjutnya untuk memperkuat validitas inferensial hasil penelitian.

B. Pembahasan Hasil

Fenomena ini menunjukkan bahwa penerapan PCA pada data penelitian ini tidak memberikan peningkatan performa prediksi, bahkan dalam beberapa kasus justru menurunkan akurasi model. Hal ini dapat dijelaskan oleh karakteristik algoritma Random Forest yang pada dasarnya tidak sensitif terhadap multikolinearitas dan mampu menangani fitur dalam jumlah besar secara efektif. Oleh karena itu, proses reduksi dimensi menggunakan PCA berpotensi menghilangkan informasi penting yang bersifat relevan terhadap variabel target, terutama pada data permintaan yang memiliki pola non-linear dan fluktuatif. Selain itu, PCA merupakan metode unsupervised yang melakukan reduksi dimensi berdasarkan variasi data tanpa mempertimbangkan hubungan langsung terhadap variabel target. Akibatnya, komponen utama yang terbentuk tidak selalu merepresentasikan fitur-fitur yang paling berpengaruh terhadap permintaan stok, sehingga performa prediksi model hybrid menjadi kurang optimal.

Model Random Forest berbasis PCA tetap dipertahankan sebagai model pembanding untuk menunjukkan bahwa penerapan reduksi dimensi tidak selalu meningkatkan kinerja prediksi, khususnya pada data dengan dimensi rendah dan karakteristik non-linear seperti pada penelitian ini.

Secara umum, dari hasil evaluasi menunjukkan bahwa Random Forest tanpa PCA merupakan model yang paling efektif untuk prediksi permintaan dan stok pada PT Bintang Jaya. Sementara itu, PCA diposisikan sebagai pendekatan metodologis alternatif yang dapat dipertimbangkan pada kondisi dataset dengan jumlah fitur yang sangat besar atau tingkat multikolinearitas yang tinggi, namun tidak direkomendasikan sebagai komponen utama dalam model prediksi pada penelitian ini.

V. KESIMPULAN

Penelitian ini membuktikan bahwa penerapan model prediksi berbasis machine learning dapat dilakukan secara efektif untuk mengolah data historis dari ERP HexPos, sehingga mampu menghasilkan estimasi permintaan yang lebih akurat dibandingkan pendekatan manual. Model Random Forest sebagai baseline menunjukkan performa yang baik dan stabil, ditandai dengan nilai MAE dan RMSE yang konsisten serta koefisien determinasi (R^2) sekitar 0,69, yang berarti model mampu menjelaskan kurang lebih 69% variasi permintaan berdasarkan fitur historis. Selain itu, hasil analisis menunjukkan bahwa faktor yang paling memengaruhi permintaan meliputi kategori produk, jenis barang, periode waktu (bulan/musim), serta karakteristik harga, di mana pola permintaan juga cenderung bersifat musiman pada periode tertentu. Penerapan Principal Component Analysis (PCA) digunakan sebagai pendekatan hybrid evaluatif untuk menguji pengaruh reduksi dimensi terhadap performa model, khususnya terkait kompleksitas fitur dan potensi multikolinearitas. Namun, hasil penelitian menunjukkan bahwa kombinasi Random Forest + PCA tidak memberikan peningkatan akurasi yang signifikan, bahkan pada beberapa kondisi meningkatkan deviasi prediksi, karena karakteristik data yang relatif berdimensi rendah dan bersifat non-linear sudah dapat ditangani dengan baik oleh Random Forest tanpa reduksi dimensi tambahan. Dengan demikian, Random Forest tanpa PCA dinilai lebih layak digunakan sebagai model utama untuk mendukung perencanaan stok karena mampu menghasilkan prediksi yang lebih mendekati permintaan aktual dan lebih stabil, serta dapat menjadi dasar pengembangan sistem prediksi stok berkelanjutan, termasuk integrasi ke dashboard analitik pada ERP perusahaan.

UCAPAN TERIMA KASIH

Penulis menyampaikan apresiasi dan terima kasih kepada Universitas Multi Data Palembang atas dukungan institusional serta sarana dan prasarana yang telah disediakan selama proses penelitian, sehingga penelitian ini dapat berjalan dengan lancar hingga tahap penyelesaian.

REFERENSI

[1] J. Lim and K. Lee, "Supply Chain Optimization Using Random Forest and PCA," *International Journal of Supply Chain Management*, vol. 9, no. 4, pp. 300–309, 2020.

[2] Z. Huo and X. Zhan, "Revolutionizing Inventory Management: The Role of Data Mining in Industry 4.0," *Advances in Economics, Management and Political Sciences*, vol. 109, no. 1, pp. 67–72, Oct. 2024, doi: 10.54254/2754-1169/109/2024bj0121.

[3] A. Dharma, "Penerapan Algoritma Random Forest untuk Prediksi Churn Pelanggan di Sektor Ritel," *Jurnal Teknologi Informasi*, vol. 8, no. 2, pp. 55–66, 2023.

[4] D. I. Purnamasari, V. A. Permadi, A. Saepudin, and R. P. Agusdin, "Demand Forecasting for Improved Inventory Management in Small and Medium-Sized Businesses," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 12, no. 1, pp. 56–66, Mar. 2023, doi: 10.23887/janapati.v12i1.57144.

[5] S. Polo-Triana, J. C. Gutierrez, and J. Leon-Becerra, "Integration of Machine Learning in the Supply Chain for Decision Making: A Systematic Literature Review," *Journal of Industrial Engineering and Management*, vol. 17, no. 2, pp. 344–372, 2024, doi: 10.3926/jiem.6403.

[6] R. F. Tarigan, J. T. Informatika, S. Tinggi, T. Harapan, and S. M. Honda, "SISTEM PENDUKUNG KEPUTUSAN UNTUK PEMBELIAN SEPEDA MOTOR MEREK HONDA MENGGUNAKAN METODE ANALYTIC HIERARCHY PROCESS," pp. 1–9.

[7] S. Li, " Sales Forecasting Model of E-commerce Activities Based on Improved Random Forest Algorithm ," in *2022 2nd International Conference on Computer Graphics, Image and Virtualization (ICCGIV)* , Los Alamitos, CA, USA: IEEE Computer Society, Sep. 2022, pp. 195–198. doi: 10.1109/ICCGIV57403.2022.00045.

[8] R. Akbar, "Pemanfaatan Principal Component Analysis dalam Visualisasi Data Spasial," *Jurnal Geoinformatika*, vol. 10, no. 3, pp. 112–119, 2022.

[9] D. Farizal and I. Setiawan, "Reduksi Dimensi Fitur Menggunakan PCA pada Dataset Penjualan," *Jurnal Informatika*, vol. 7, no. 3, pp. 210–218, 2020.

[10] R. Koushiki, "Retail Sales Forecasting: Random Forest vs LSTM," *Journal of Retail Analytics*, vol. 6, no. 2, pp. 101–110, 2021.

[11] A. Moudy Fajria, A. Faqih, and G. Dwilestari, "The Impact of Principal Component Analysis Dimensionality Reduction on Sentiment Classification Performance Using Support Vector Machine," 2025. [Online]. Available: https://ioinformatic.org/

[12] G. Spanos, A. Lalas, K. Votis, and D. Tzovaras, "Principal Component Random Forest for Passenger Demand Forecasting in Cooperative, Connected, and Automated Mobility," *Sustainability (Switzerland)*, vol. 17, no. 6, Mar. 2025, doi: 10.3390/su17062632.

[13] P. Fang, Z. Gao, and R. S. Tsay, "Supervised kernel principal component analysis for forecasting," *Financ. Res. Lett.*, vol. 58, Dec. 2023, doi: 10.1016/j.frl.2023.104292.

[14] I. Maulana, A. M. Siregar, R. Rahmat, and A. Fauzi, "Optimization Of Machine Learning Model Accuracy For Brain Tumor Classification With Principal Component Analysis," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 3, pp. 903–915, Jun. 2024, doi: 10.52436/1.jutif.2024.5.3.2058.

[15] W. Sari and R. Taufik, "Prediksi Permintaan Produk Digital Menggunakan Algoritma Random Forest," *Jurnal Informatika*, vol. 9, no. 1, pp. 23–31, 2021.

[16] Y. Huang, T. Yang, and J. Zhang, “Forecasting Demand with Random Forest Algorithm in Logistics Industry,” *International Journal of Forecasting and Planning*, vol. 11, no. 2, pp. 97–108, 2022.
[17] C. Wu and Y. Zhang, “Seasonal Demand Forecasting Using Optimized Random Forest,” *Expert Syst. Appl.*, vol. 210, pp. 118–130, 2025.
[18] R. Emha and S. Totok, “Stabilisasi Prediksi Kebutuhan Stok Barang Menggunakan Kombinasi PCA dan Random Forest,” *Jurnal Teknologi Informasi*, vol. 9, no. 1, pp. 45–55, 2021.
[19] L. Wu and X. Zhang, “Random Forest Parameter Tuning for Seasonal Demand Forecasting,” *Journal of Applied AI and Data Science*, vol. 10, no. 1, pp. 71–84, 2025.