P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

Analisis Perbandingan Decision Tree dan Random Forest dalam Klasifikasi Penjualan Produk pada Supermarket

Putri Ayu Firnanda[1*], Litasya Shofwatillah[2], Fauziah Rahma[3], Fatkhurokhman Fauzi[4]

[1,2,3,4] Program Studi Statistika, Universitas Muhammadiyah Semarang, Jalan Kedungmundu No.18, Kecamatan Tembalang, Semarang, 50273, Jawa Tengah, Indonesia

*Corresponding author: firnanda.putria@gmail.com

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Riwayat Artikel

Dikirim: 30 Juli 2024
Direvisi: 29 November 2024
Diterima: 23 Desember 2024

ABSTRAK

Penelitian ini bertujuan untuk mendapatkan model klasifikasi terbaik antara model algoritma Decision Tree dan Random Forest untuk melihat apakah sebuah produk laris atau tidak laris berdasarkan data dari Supermarket ASDA. Kedua metode tersebut menggunakan teknik klasifikasi pohon keputusan dengan pendekatan top-down untuk memecah masalah menjadi keputusan yang sederhana. Metode Random Forest merupakan pengembagan dari Decision Tree dengan menggunakan ensemble untuk meningkatkan akurasi dan mengurangi resiko overfitting. Sebagai sebuah bisnis retail, Supermarket ASDA memerlukan informasi-informasi tersebut untuk melihat pola konsumsi pelanggan sehingga dapat digunakan dalam membuat strategi dan keputusan yang tepat. Sumber data penelitian meliputi data penjualan dari Supermarket ASDA yang terdiri dari variabel harga, harga per unit, nama produk, tanggal, kategori dan kepemilikan merek untuk melatih model klasifikasi. Penelitian ini akan melibatkan pengujian kedua model pada dataset untuk mengukur kemampuan prediksi model dalam memprediksi produk yang laris dan tidak laris.Berdasarkan hasil penelitian, model algoritma Random Forest memiliki performa lebih baik dari model algoritma Decision Tree baik pada pemodelan dasar maupun setelah dilakukan hyperparameter tuning dengan presentasi akurasi sebesar 99%. Hasil dari matrik evaluasi (precision, recall, F1-Score) model algoritma Random Forest juga menunjukkan nilai yang lebih tinggi sehingga menjadikan model algoritma Random Forest lebih akurat dalam mengklasifikasikan penjualan laris dan tidak laris di Supermarket ASDA.

Kata Kunci: Akurasi, Decision Tree, Klasifikasi, Penjualan, Random Forest.

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

ABSTRACT

This research aims to obtain the best classification model between the Decision Tree and Random Forest algorithms to determine whether a product is popular or not based on data from ASDA Supermarket. Both methods use decision tree classification techniques with a top-down approach to break down problems into simpler decisions. The Random Forest method is an extension of the Decision Tree, utilizing ensemble techniques to enhance accuracy and reduce the risk of overfitting. As a retail business, ASDA Supermarket needs this information to understand customer consumption patterns, enabling it to make strategic and informed decisions. The dataset for this research consists of sales data from ASDA Supermarket, including variables such as price, unit price, product name, date, category, and brand ownership used to train the classification models. This research will involve testing both models on a dataset to measure their prediction capabilities in classifying which products are popular and which are not. The results show that the Random Forest algorithm performs better than the Decision Tree algorithm, both in its base model and after hyperparameter tuning, with an accuracy of 99%. The evaluation metrics (precision, recall, F1-Score) for the Random Forest algorithm also demonstrate higher values, making it a more accurate model for classifying popular and unpopular products at ASDA Supermarket.

Keywords: Accuracy, Decision Tree, Classification, Sales, Random Forest.

1. Pendahuluan

Supermarket merupakan sebuah bisnis ritel yang memainkan peran krusial dalam menyediakan berbagai produk kebutuhan sehari-hari dan barang konsumsi lainnya. Menurut Theodore Cohn, seorang ekonom dan penulis buku tentang industri supermarket, supermarket adalah "sebuah bisnis ritel yang menjual berbagai macam produk kebutuhan sehari-hari dan barang-barang konsumen lainnya di satu tempat, biasanya dengan ukuran yang besar" [1].

Dunia bisnis ritel, khususnya supermarket, memahami pola penjualan dan perilaku konsumen merupakan hal yang sangat penting sebagai strategi dalam dunia persaingan bisnis [2]. Melakukan klasifikasi produk-produk menjadi dua kategori yaitu produk yang laris dan produk yang tidak laris, merupakan salah satu cara yang dapat dilakukan oleh Supermarket ASDA dalam membantu strategi pemahaman tersebut. Memahami dan mengklasifikasikan produk yang laris dan tidak laris pada suatu supermarket dapat meningkatkan efisiensi operasional, meningkatkan kepuasan pelanggan, sehingga mampu meningkatkan profitabilitas bisnis supermarket.

Produk yang laris biasanya adalah produk yang memiliki tingkat permintaan konsumen yang tinggi, berputar dengan cepat, dan sering menjadi favorit di kalangan konsumen. Produk-produk ini biasanya memiliki kontribusi yang cukup signifikan terhadap pendapatan supermarket dan ketersediaannya sangat penting untuk menjaga kepuasan pelanggan. Sebaliknya, produk yang tidak laris adalah produk yang memiliki permintaan konsumen yang rendah, berputar lambat, dan memungkinan menjadi penyebab

penumpukan stok yang berlebihan [3]. Melakukan klasifikasi penjualan, berbagai metode analisis data dan teknik machine learning dapat digunakan dengan menggunakan data historis penjualan dalam setiap hari yang telah tersedia. Serta, produk-produk penjualan dapat dianalisis untuk mengidentifikasi sebuah pola penjualan.

Pada penelitian-penelitian sebelumnya, yang pertama oleh Wicaksono, S. A., Kadek, I., & Nuryana, D. (2024) yang berjudul "Perbandingan Klasifikasi Hasil Indeks Kepuasan Masyarakat Terhadap UPT Balai Latihan Kerja Surabaya Menggunakan Algoritma Decision Tree, Random Forest, dan K-Nearest Neighbor". Penelitian ini dilakukan dengan membandingkan algoritma Decision Tree, Random Forest, dan K-Nearest Neighbor untuk melakukan analisis tingkat kepuasan masyarakat terhadap sarana prasarana UPT Balai Latihan Kerja Surabaya. Kesimpulan yang diperoleh dari penelitian ini bahwa algoritma Random Forest menjadi algoritma yang paling cocok untuk digunakan klasifikasi pada kasus ini dengan hasil akurasi sebesar 100%, sedangkan algoritma Decision Tree sebesar 91% dan algoritma K-Nearest Neighbor sebesar 94% [4].

Penelitian oleh Azis Rahmat, W., Madinah Ladjamuddin, S., & Teruna Awaludin, D. (2023) dengan judul yaitu "Perbandingan Algoritma Decision Tree, Random Forest, dan Naive Bayes pada Prediksi Penilaian Kepuasan Penumpang Maskapai Pesawat Menggunakan Dataset Kaggle". Pada penelitian ini menggunakan tiga model yaitu Decision Tree, Random Forest, dan Naive Bayes yang dibandingkan untuk memperoleh algoritma yang paling relevan untuk melakukan prediksi pada penilaian kepuasan penumpang dan diperoleh kesimpulan bahwa algoritma Random Forest memiliki nilai akurasi yang paling tinggi yaitu sebesar 95.44%, sedangkan algoritma Decision Tree sebesar 93.41%, dan algoritma Naive Bayes memiliki akurasi paling rendah yaitu sebesar 82.51%. Sehingga pada kasus dalam penelitian ini algoritma yang paling relevan dalam melakukan prediksi pada penilaian kepuasan penumpang yaitu menggunakan algoritma Random Forest [5].

Pada penelitian lainnya dengan judul "Perbandingan Algoritma Random Forest, Naive Bayes, dan Decision Tree dengan Oversampling Untuk Klasifikasi Bakteri E. Coli" oleh Kusumarini, A. I., Hogantara, P. A., & Chamidah, N. (2021). Penelitian ini memiliki tujuan untuk mengetahui dari ketiga algoritma yaitu Random Forest, Naive Bayes, dan Decision Tree, manakah yang menghasilkan akurasi paling tinggi dan paling baik untuk melakukan klasifikasi bakteri E. Coli. Pada penelitian ini diperoleh hasil akurasi dengan menggunakan algoritma Naive Bayes sebesar 78%, algoritma Decision Tree sebesar 76%, dan algoritma Random Forest menghasilkan nilai akurasi sebesar 84%. Sehingga dapat disimpulkan bahwa algoritma Random Forest merupakan algoritma yang menghasilkan nilai akurasi tertinggi dan paling tepat untuk melakukan klasifikasi bakteri E. Coli [6].

Penelitian yang lainnya yaitu oleh Lalo, A. K., Batarius, P., & Siki, Y. C. H. (2021) yang berjudul "Implementasi Algoritma C4.5 Untuk Klasifikasi Penjualan Barang di Swalayan Dutalia". Pada penelitian tersebut, algoritma C4.5 digunakan dalam melakukan klasifikasi penjualan sehingga pihak swalayan dapat memperoleh informasi berupa pola penjualan produk atau barang berdasarkan riwayat penjualan barang di masa lampau yang akan turut mendukung dalam pengambilan kebijakan perusahaan di masa yang akan datang, dan diperoleh hasil akurasi sebesar 100% [2].

Pada penelitian dengan judul "Klasifikasi Data Penjualan pada Supermarket dengan Metode Decision Tree" yang merupakan penelitian dari Wardhana, A. W., Patimah, E., Shafarindu, A. I., Siahaan, Y. M., Haekal, B. V., & Prasvita, D. S. (2021). Pada penelitian ini, dilakukan klasifikasi menggunakan algoritma Decision tree untuk memperoleh informasi tingkat penjualan yang lebih baik dengan akurasi terbaik dengan menggunakan dua jenis tipe kelas yaitu kelas branch dan kelas customer type (tipe pelanggan). Pada penelitian ini, diperoleh hasil akurasi tertinggi pada kelas branch adalah 1.0 dan pada kelas customer type ada pada akurasi decision tree dengan menggunakan metode Hold Out yang sebesar 0.5 [1].

P-ISSN: 2986-4178
E-ISSN: 2988-4004
Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

Berdasarkan latar belakang dan penelitian-penelitian sebelumnya, pada penelitian ini berupaya untuk memperkenalkan pendekatan komparatif atau perbandingan. Pada penelitian ini, algoritma klasifikasi Decision Tree dan Random Forest menjadi algoritma yang diterapkan sebagai perbandingan untuk menentukan algoritma yang paling baik dalam mengelompokkan produk berdasarkan kinerja penjualannya. Penelitian ini menggunakan teknik data mining seperti hyperparameter dengan grid search, serta penerapan feature engineering untuk dapat meningkatkan akurasi dalam klasifikasi. Selain itu, penelitian ini juga melakukan evaluasi performa dengan menggunakan matrik evaluasi, seperti precision, recall, F1-score, dan akurasi yang berguna untuk memberikan analisis yang lebih mendalam dan relevan dalam melakukan klasifikasi atau pengelompokkan produk, sehingga memperoleh hasil klasifikasi yang terbaik.

2. Tinjauan Pustaka

2.1. Data Mining

Data Mining secara umum terdiri dari dua komponen:

a. Data: Sekumpulan fakta yang tercatat, atau entitas yang tidak memiliki makna dan sering diabaikan.
b. Mining: Proses penambangan, sehingga Data Mining dapat diartikan sebagai proses menambang data untuk menghasilkan pengetahuan.

Data mining adalah proses otomatis untuk menemukan informasi yang berguna dari penyimpanan data berukuran besar. Teknik ini digunakan untuk memeriksa basis data besar guna menemukan pola-pola baru yang bermanfaat. Namun, tidak semua kegiatan pencarian informasi dapat dikategorikan sebagai data mining. Berikut adalah contoh yang termasuk dan tidak termasuk dalam data mining:

a. Bukan Data Mining: Mencari informasi tertentu di internet dengan menggunakan kata kunci di mesin pencari. Data Mining: Mengelompokkan informasi yang serupa dalam hasil pencarian mesin pencari berdasarkan konteks tertentu, seperti mengelompokkan universitas berdasarkan akreditasi, jumlah mahasiswa, kualitas alumni, dan sebagainya.
b. Bukan Data Mining: Membandingkan data laporan keuangan bulan Januari dan Februari untuk menentukan peningkatan permintaan. Data Mining: Menggunakan laporan keuangan saat ini untuk memprediksi pengeluaran dan pembelian di masa mendatang.

2.2 Decision Tree

Decision tree adalah metode yang mengubah data menjadi struktur pohon keputusan (decision tree) dan aturan-aturan keputusan. Dalam pohon keputusan, terdapat tiga jenis node [7], yaitu:

1. Root node: Node ini berada di bagian paling atas dari pohon dan tidak memiliki input. Root node bisa tidak memiliki output atau memiliki lebih dari satu output.
2. Internal node: Node ini adalah titik percabangan dalam pohon. Setiap internal node memiliki satu input dan minimal dua output.
3. Leaf node atau terminal node: Node ini merupakan node terakhir dalam pohon, di mana hanya ada satu input dan tidak memiliki output.

Secara umum, pembuatan Decision Tree (DT) melibatkan berbagai algoritma yang sering digunakan, salah satunya adalah CART (Classification and Regression Trees), yang

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

diimplementasikan dalam pustaka seperti scikit-learn. CART merupakan pengembangan dari algoritma C4.5, yang pada dasarnya adalah penyempurnaan dari algoritma ID3 yang dikembangkan pada tahun 1980. Algoritma ID3 berfungsi dengan memisahkan data menjadi dua kelompok berdasarkan atribut tertentu, menggunakan perhitungan nilai yang disebut entropy. Entropy digunakan untuk menghitung tingkat homogenitas atribut A pada suatu sampel data S. Rumus entropy dapat dituliskan sebagai: [8]

$$Entropy\ (S) = -\sum_{i=1}^{n} \quad p_i\ .log_2(p_i)$$

Di mana pi adalah probabilitas kelas ke-i.
Selain entropy, ID3 juga menggunakan information gain (Gain) untuk menentukan atribut terbaik. Rumus Gain berdasarkan entropy adalah:

$$Gain\ (S,A)\ = Entropy\ (S) = -\sum_{i=1}^{n} \quad \frac{|S_i|}{|S|}\ .Entropy\ (S_i)$$

Dalam metode CART, pemilihan atribut untuk split dapat dilakukan dengan Gini Index atau Classification Error. Rumus Gini Index untuk suatu simpul ttt adalah:

$$Gini\ (t) = -\sum_{i=1}^{n} \quad [p(i|t)]^2$$

Sedangkan Gain menggunakan Gini Index dihitung sebagai:

$$Gain\ (S,A)\ = Gini\ (S) = -\sum_{i=1}^{n} \quad \frac{|S_i|}{|S|}\ .Gini\ (S_i)$$

Classification Error dapat digunakan sebagai alternatif untuk menghitung ketidakhomogenan, dengan rumus:

$$C_{Error}(S) = 1 - max[p(i|S)]$$

2.3 Random Forest

Random Forest adalah metode klasifikasi yang terdiri dari kumpulan pohon keputusan terstruktur, di mana vektor acak independen didistribusikan secara seragam. Setiap pohon keputusan memberikan suara untuk menentukan kelas yang paling umum berdasarkan input yang diberikan [9]. Proses membangun Random Forest melibatkan:

1. Membuat sampel bootstrap dengan memilih sampel data secara acak dengan penggantian untuk setiap pohon keputusan, sehingga sepertiga dari contoh data tidak digunakan.
2. Contoh data yang tidak digunakan ini disebut data Out of Bag (OOB).
3. Data OOB digunakan untuk memperkirakan kesalahan dari setiap pohon keputusan, yang dikenal sebagai estimasi kesalahan OOB.
4. Random Forest juga dapat menghitung tingkat kepentingan variabel dan memberikan perkiraan yang digunakan untuk menangani nilai yang hilang dan outlier.

Berikut adalah tahapan algoritma dalam membangun pohon keputusan menggunakan metode Random Forest: [10]

1. Pengecekan Label Data

P-ISSN: 2986-4178
E-ISSN: 2988-4004
Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

Jika semua label pada data memiliki nilai yang sama, maka pohon akan membentuk daun dengan nilai sesuai label tersebut.

2. Menghitung Nilai Informasi (Entropy)
   Entropi dihitung menggunakan formula:

$$Info\ (D) = -\sum_{i=1}^{m} \quad p_i\ .log_2(p_i)$$

   Di mana pi adalah probabilitas tuple dalam D yang termasuk dalam kelas tertentu. Entropi D menunjukkan rata-rata informasi yang diperlukan untuk mengidentifikasi tuple dalam dataset D.
   Jika atribut A bernilai diskrit, data D akan dipartisi berdasarkan nilai-nilai unik pada atribut A. Setiap cabang hasil partisi diusahakan memiliki data yang homogen. Informasi dari hasil partisi dihitung dengan persamaan:

$$Info_A(D) = -\sum_{j=1}^{v} \quad \frac{|D_j|}{|D|}.Info(D_j)$$

   Di mana $\frac{|D_j|}{|D|}$ adalah bobot partisi ke-j, dan $Info\ (D_j)$ adalah entropi dari partisi tersebut.
3. Memilih Split Point untuk Atribut Kontinu
   Jika atribut A bersifat kontinu, nilai-nilai atribut diurutkan dari yang terkecil hingga terbesar. Split point dihitung dengan mengambil rata-rata dua nilai berurutan. Informasi dihitung untuk setiap split point, dan split point dengan nilai entropi terkecil akan dipilih sebagai titik pembagi.
4. Menghitung Gain Atribut
   Gain untuk setiap atribut dihitung menggunakan formula:

$$Gain(A) = Info(D) - Info_A(D)$$

   Atribut dengan nilai gain tertinggi akan dipilih sebagai cabang untuk pohon keputusan.
5. Pembangunan Cabang dan Daun
   Setelah cabang pohon keputusan terbentuk, langkah dari tahap 1 hingga 4 diulangi untuk data di setiap cabang. Jika jumlah cabang mencapai batas maksimum, atau data di cabang tidak dapat dipartisi lebih lanjut, daun akan dibuat dengan nilai mayoritas dari data pada cabang tersebut.

2.4 GridSearchCV

Grid Search adalah metode eksplorasi matematis yang digunakan untuk mencari kombinasi hyperparameter terbaik guna meningkatkan kinerja model klasifikasi. Metode ini menguji semua kemungkinan kombinasi hyperparameter yang telah ditentukan sebelumnya untuk memilih kombinasi yang paling optimal, yaitu yang memberikan performa terbaik pada model. Proses ini sering digabungkan dengan k-fold cross-validation untuk memastikan bahwa hyperparameter yang dipilih benar-benar tepat. Kombinasi dari kedua pendekatan ini dikenal sebagai Grid Search Cross-Validation atau GridSearchCV [11]. GridSearchCV, yang merupakan fitur dari modul scikit-learn, secara otomatis dan sistematis melakukan validasi untuk berbagai model serta menentukan hyperparameter terbaiknya [12]. Setelah proses selesai dijalankan, GridSearchCV akan menghasilkan model dengan skor kinerja untuk data latih (train score) dan data uji (test score) [13].

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

2.5 Hyperparameter tuning

Pada tahap pembangunan model, penentuan parameter akan dioptimalkan menggunakan metode hyperparameter tuning, yang merupakan teknik untuk mengkombinasikan parameter-parameter yang relevan. Metode hyperparameter tuning yang digunakan adalah grid search, yang merupakan metode dalam pembelajaran mesin untuk menemukan kombinasi terbaik dari parameter-parameter model dengan mencoba semua kemungkinan kombinasi yang sudah ditentukan dalam sebuah grid. Grid search membantu dalam menentukan parameter optimal guna meningkatkan kinerja model [14].

Kinerja model machine learning dalam penelitian ini dievaluasi menggunakan metode confusion matrix. Confusion matrix merangkum jumlah prediksi yang benar dan salah, yang kemudian dikelompokkan berdasarkan masing-masing kelas(Pengenalan Pola Dasar Angka berdasarkan Gerakan Tangan menggunakan Machine Learning). Hasil prediksi tersebut dinilai melalui beberapa metrik, yaitu akurasi (accuracy), presisi (precision), recall, dan F1-Score [15].

1. Akurasi menggambarkan sejauh mana model mampu mengklasifikasikan data secara keseluruhan dengan benar.

$$Akurasi = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

2. Presisi menunjukkan proporsi prediksi positif yang benar-benar akurat.

$$Presisi = \frac{TP}{(TP + FP)}$$

3. Recall mengukur kemampuan model dalam mendeteksi kelas positif dengan benar.

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

4. F1-Score adalah kombinasi harmonis dari presisi dan recall untuk memberikan gambaran yang seimbang tentang kinerja model.

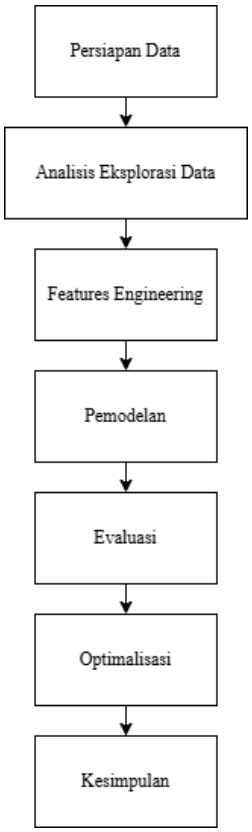
$$F1 - Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Reacall}$$

Keterangan:
$TP = True\ Positive$
$TN = True\ Negatif$
$FP = False\ Positive$
$FN = False\ Negativ$

3. Metodologi Penelitian

Metode yang digunakan dalam penelitian yaitu klasifikasi dengan algoritma Random Forest dan Decision Tree. Berikut diagram proses algoritma pada random forest dan decision tree:

P-ISSN: 2986-4178
E-ISSN: 2988-4004
Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025



Gambar 1. Diagram Alir Penelitian

Berikut langkah-langkah proses analisis:

1) Persiapan Data
Data yang digunakan merupakan dataset yang bersumber dari Kaggle melalui situs www.kaggle.com yaitu data penjualan Supermarket yang berada di UK yang berisi data penjualan per hari dari 9 Januari 2024 hingga update terakhir dengan variabel penelitian antara lain, price, price_unit, unit, names, date, category, dan own_brand.

2) Eksplorasi Analisis Data
Preprocessing merupakan salah satu cara untuk mengatasi masalah-masalah yang ada pada data yang dapat mengganggu hasil dari analisis [16]. Cara yang dapat dilakukan untuk mengatasi masalah tersebut, yaitu mendeteksi dan mengatasi missing values, data duplikat, dan data outlier. Pada masalah missing value, langkah yang dapat dilakukan setelah mengetahui adanya missing value pada dataset yaitu dengan menghapus atau mengisi yang hilang dengan nilai lain. Cara ini dapat ditentukan berdasarkan melihat jumlah data yang hilang, tipe data yang hilang, atau nilai dari kolom tersebut. Data duplikat dapat diatasi dengan menghapus data yang sama. Sedangkan, pada masalah data outlier salah satu cara yang dapat dilakukan untuk mendeteksi outlier yaitu melalui visualisasi boxplot. Data outlier merupakan data yang menyimpang secara signifikan dari data yang lain. Salah satu cara yang dapat dilakukan untuk mengatasi data outlier yaitu dengan menggunakan interkuartil atau IQR. Interkuartil yaitu menghitung rentang antara kuartil bawah (Q1) dan kuartil atas (Q3). Untuk mengatasi data outlier, dapat dilihat terlebih dahulu jumlah data yang menyimpang. Data outlier dapat dihilangkan apabila jumlah outlier tidak terlalu banyak.

3) Feature Engineering
   a. Feature Selection
      Memilih dan menambah fitur-fitur yang akan digunakan dalam analisis. Fitur yang digunakan dalam analisis yaitu category (kategori produk), item_count (jumlah produk), dan keterangan (produk yang laris dan tidak laris).
   b. Mengatasi Data Kategorikal
      Melakukan transformasi atau mengubah data kategorik menjadi numerik menggunakan teknik encoding. Dan pada penelitian ini, teknik yang digunakan yaitu label encoding yaitu menetapkan nilai integer untuk setiap data kategori.
   c. Feature Scaling
      Melakukan scaling yang bertujuan untuk agar data antar fitur mempunyai rentang nilai yang tidak jauh. Pada penelitian ini menggunakan standar scaler yaitu dengan melakukan standarisasi. Standarisasi dilakukan dengan mentransformasikan nilai rata-rata menjadi 0 dan nilai standar deviasi menjadi 1. Kemudian, memusatkan data di sekitar rata-rata dan mengubah skalanya berdasarkan nilai standar deviasi.

4) Pemodelan
   Melakukan splitting data, yaitu membagi data menjadi 80% data training dan 20% data testing. Kemudian melakukan pemodelan pada model decision tree dan random forest.

5) Evaluasi
   Pada tahap evaluasi dilakukan untuk memilih model terbaik dari kedua model yang dianalisis dengan membandingkan matrik evaluasi yaitu precision, recall, F1-score, dan akurasi yang lebih tinggi.

6) Optimalisasi
   Melakukan hyperparameter tuning pada algoritma yang bertujuan untuk mengoptimalkan model guna memperbaiki performa model dengan menentukan kombinasi yang tepat bagi setiap hyperparameter yang ada pada setiap model.

7) Kesimpulan

4. Hasil dan Pembahasan

Pada bagian ini akan membahas hasil dari pengujian yang dilakukan pada dataset penjualan supermarket ASDA. Penelitian ini dilakukan melalui beberapa proses sebagai berikut:

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

4.1. Preprocessing Data

Pada tahap preprocessing data akan dilakukan pembersihan missing value, data duplikat, dan data outlier. berikut ini hasil dari preprocessing data yang dilakukan.

1) Mengatasi Missing Value
   - Menampilkan jumlah missing value

     Pada tahap pembersihan missing value diawali dengan melihat jumlah missing value pada dataset yang digunakan. Berikut ini jumlah missing value pada masing-masing variabel data:

```
supermarket          0
prices_(£)           7
prices_unit_(£)    199
unit               199
names               26
date                 0
category             0
own_brand           26
dtype: int64
```

   - Menghapus missing value pada dataset

     Dataset yang mengandung missing value akan dihilangkan dengan metode dropna yang merupakan sebuah metode pada pustaka **Pandas** yang digunakan untuk menghapus data yang mengandung nilai kosong (*missing values* atau NaN). Berikut ini jumlah data setelah dilakukan pembersihan missing value:

```
asda.shape
```

```
(2456196, 8)
```

     Pada hasil diatas dapat dilihat data memiliki 2.456.196 baris setelah dilakukan pembersihan data, sedangkan data awal berjumlah 2.456.414 baris data.

2) Mengatasi Data Duplikat
   - Menampilkan jumlah data duplikat

```
asda.duplicated().sum()
```

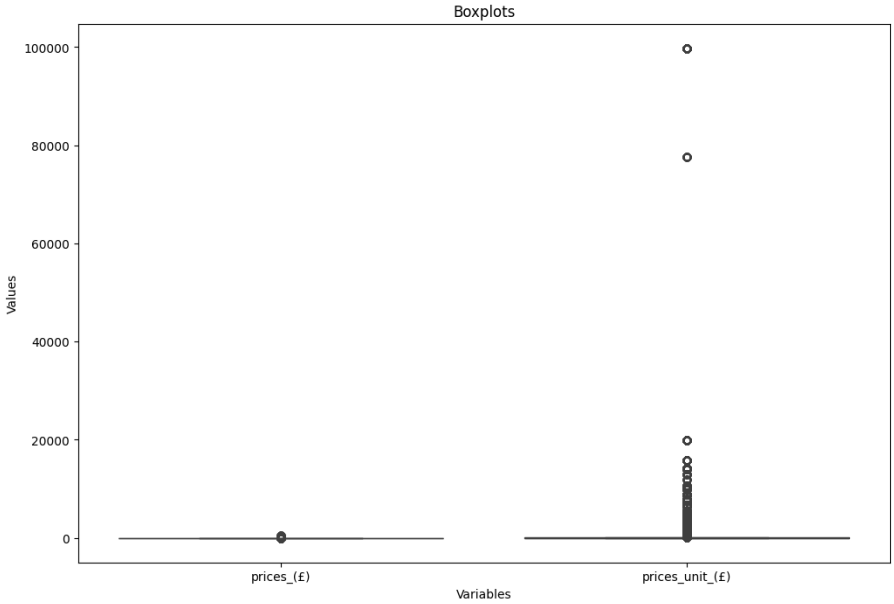
```
0
```

     Berdasarkan hasil di atas dapat dilihat bahwa dataset tidak memiliki data duplikat atau data yang muncul lebih dari sekali dalam dataset.

3) Mengatasi Data Outlier
   - Menampilkan jumlah data outlier

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025



Gambar 2. Boxplot Data Outlier

Berdasarkan gambar boxplot diatas dapat dilihat bahwa terdapat sejumlah nilai ekstrem atau outlier yang jauh dari distribusi utama. Pada variabel prices_(£) sebagian besar data terkonsentrasi di dekat nilai 0 serta menunjukkan nilai median dan rentang antar kuartil yang sangat kecil, pada variabel prices_unit_(£) pada terdapat jumlah outlier yang lebih banyak serta terdapat outlier

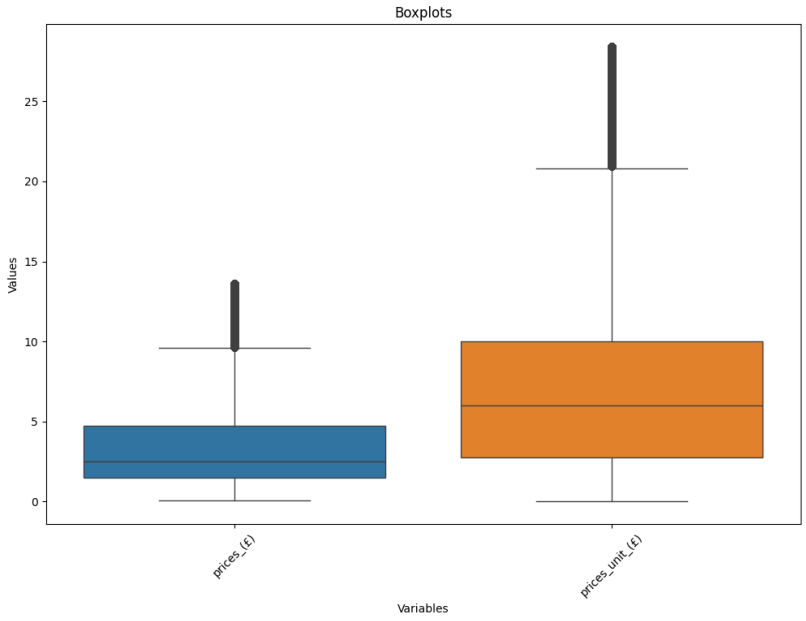
P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

yang memiliki nilai mendekati 100.000. Hal ini menunjukkan bahwa data variabel memiliki outlier yang cukup ekstrim.

- Mengatasi data outlier dengan IQR

  Berikut ini boxplot setelah dilakukan penghapusan outlier yang berada diluar batas atas dan bawah yang ditentukan.



Gambar 2. Boxplot Dataset Tanpa Outlier

Berdasarkan gambar boxplot di atas dapat dilihat bahwa kedua variabel tidak ada lagi outlier pada dataset, yang menunjukkan bahwa outlier sudah berhasil diatasi dengan metode IQR.

4.2. Feature Engineering

Pada tahap Feature Engineering akan dilakukan 3 proses yaitu Feature Selection, mengatasi data kategorikal, dan Feature Scaling. Berikut hasil dari proses tersebut.

1) Feature Selection

   Berikut ini hasil dari fitur selection yaitu berupa penambahan category (kategori produk), item_count (jumlah produk), dan keterangan (produk yang laris dan tidak laris).

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

	date	category	item_count	mean_penjualan
0	2024-01-09	baby_products	661	632.300000
1	2024-01-09	bakery	690	718.637363
2	2024-01-09	drinks	2017	2110.695652
3	2024-01-09	food_cupboard	4442	5046.717391
4	2024-01-09	free-from	627	640.788889
...	...	...	...	...
1002	2024-04-13	frozen	961	629.315217
1003	2024-04-13	health_products	2798	2463.228261
1004	2024-04-13	home	2629	2803.250000
1005	2024-04-13	household	2006	1764.836957
1006	2024-04-13	pets	739	650.108696

1007 rows × 4 columns

Gambar 3. Hasil Feature Selection

Berdasarkan **Gambar 3**, keterangan produk yaitu berupa laris atau tidak laris dilihat dari item_count dan mean_penjualan. Apabila nilai item_count lebih kecil dari mean_penjualan maka produl tidak laris begitupun sebaliknya. **Gambar 4** setelah ditambahkan keterangan produk.

	date	category	item_count	mean_penjualan	keterangan
0	2024-01-09	baby_products	661	632.300000	L
1	2024-01-09	bakery	690	718.637363	TL
2	2024-01-09	drinks	2017	2110.695652	TL
3	2024-01-09	food_cupboard	4442	5046.717391	TL
4	2024-01-09	free-from	627	640.788889	TL
...	...	...	...	...	...
1002	2024-04-13	frozen	961	629.315217	L
1003	2024-04-13	health_products	2798	2463.228261	L
1004	2024-04-13	home	2629	2803.250000	TL
1005	2024-04-13	household	2006	1764.836957	L
1006	2024-04-13	pets	739	650.108696	L

1007 rows × 5 columns

Gambar 4. Dataset dengan Keterangan Produk

2) Mengatasi data ketegorikal

Data kategorikal yang akan dilakukan label encoding yaitu pada variabel category dan keterangan dengan menetapkan nilai integer untuk masing-

P-ISSN: 2986-4178
E-ISSN: 2988-4004
Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

masing variabel. berikut ini tabel hasil label encoding pada kedua variabel tersebut.

	category	item_count	keterangan
0	0	661	0
1	1	690	1
2	2	2017	1
3	3	4442	1
4	4	627	1
...	...	...	...
1002	6	961	0
1003	7	2798	0
1004	8	2629	1
1005	9	2006	0
1006	10	739	0

Gambar 5. Hasil Label Encoding

3) Feature Scaling

Berikut ini hasil dari standarisasi data pada proses scaling yang dilakukan pada penelitian ini.

	category	item_count	keterangan
0	-1.587603	-0.873202	-0.838983
1	-1.271088	-0.853455	1.191920
2	-0.954573	0.050114	1.191920
3	-0.638059	1.701323	1.191920
4	-0.321544	-0.896353	1.191920
...	...	...	...
1002	0.311486	-0.668928	-0.838983
1003	0.628001	0.581905	-0.838983
1004	0.944515	0.466831	1.191920
1005	1.261030	0.042624	-0.838983
1006	1.577545	-0.820091	-0.838983

Gambar 6. Tabel Hasil Standarisasi

4.3. Pemodelan Dasar

Setelah dilakukan proses eksplorasi data langkah selanjutnya yang dilakukan adalah pemodelan dengan menggunakan algoritma Decision Tree dan Random Forest.

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Emerging Statistics and Data Science Journal
Vol 3, No.1, Tahun 2025

Sebelum dilakukan pemodelan dataset akan dibagi menjadi 80% data training yaitu berjumlah 805 data dan 20% data testing yaitu berjumlah 202 data.

Berdasarkan hasil dari pemodelan dengan kedua algoritma tersebut sebelum dilakukan hyperparameter tuning berikut ini hasil yang didapatkan untuk kelas 0 (penjualan laris) dan kelas 1 (Penjualan tidak laris).

Tabel 5. Perbandingan Hasil Pemodelan Dasar

		Decision Tree	Random Forest
Precision			
	0	0,98	0,99
	1	0,97	0,99
Recall			
	0	0,97	0,99
	1	0,98	0,99
F1-Score			
	0	0,98	0,99
	1	0,97	0,99
Akurasi		0,975	0,99

Didapatkan bahwa akurasi dari algoritma Decision Tree sebesar 97,5% dan algoritma Random Forest sebesar 99%, ini artinya dari hasil pemodelan dasar algoritma Random Forest memiliki performa yang lebih baik karena memiliki nilai akurasi yang lebih tinggi. Berdasarkan matrik evaluasi yang digunakan yaitu precision, recall, dan F1-score, algoritma Random Forest juga memiliki nilai yang lebih tinggi.

4.4. Hyperparameter Tuning

Tahap pemodelan kedua yaitu dilakukan hyperparameter tuning pada kedua algoritma yang bertujuan untuk meningkatkan performa model. Hyperparameter tuning pada model menggunakan GridSearchCV untuk mendapatkan kombinasi parameter terbaik. Didapatkan kombinasi parameter terbaik untuk model algoritma Decision Tree adalah max depth: None, min samples leaf: 1, dan min samples split: 2. Kombinasi parameter terbaik yang didapatkan untuk algoritma Random Forest adalah max depth: None, min samples leaf: 1, min samples split: 2, dan n estimators: 50. Berikut ini hasil dari hyperparameter tuning dengan kombinasi parameter terbaik untuk masing-masing model algoritma.

Tabel 6. Perbandingan Hasil Hyperparameter Tuning

		Decision Tree	Random Forest
Precision			
	0	0,98	0,99
	1	0,97	0,99
Recall			
	0	0,97	0,99
	1	0,98	0,99
F1-Score			
	0	0,98	0,99
	1	0,97	0,99
Akurasi		0,975	0,99

Berdasarkan hasil hyperparameter tuning yang dilakukan dapat dilihat bahwa hasil yang didapatkan tidak memiliki perbedaan yang signifikan dengan hasil dari pemodelan dasar walaupun sudah menggunakan kombinasi parameter terbaik. Hasil akurasi dari model algoritma Random Forest masih lebih besar dibandingkan algoritma Decision Tree, dengan akurasi random forest yaitu sebesar 99%, lebih besar dibandingkan dengan akurasi decision tree yaitu sebesar 97.5%. Model algoritma Random Forest memiliki stabilitas dan kemampuan generalisasi yang lebih baik serta mengurangi risiko model untuk terjadi overfitting. Hasil penelitian ini selaras dengan penelitian terdahulu oleh Kusumarini, A. I., Hogantara, P. A., & Chamidah, N. (2021) yang juga membandingkan algoritma Random Forest dengan algoritma lainnya dan mendapatkan hasil bahwa algoritma Random Forest memiliki kinerja yang terbaik.

5. Kesimpulan

Berdasarkan hasil evaluasi, permasalahan pada penjualan produk untuk menentukan produk laris dan tidak laris di Supermarket ASDA dapat diselesaikan dengan teknik data mining dengan model algoritma Random Forest sebagai model terbaik yang didapatkan. Algoritma Random Forest memiliki performa yang lebih tinggi pada matrik evaluasi yang digunakan yaitu precision, recall, dan F1-score serta didapatkan akurasi yang lebih tinggi dengan persentase sebesar 99% sehingga Random Forest dapat memberikan hasil yang lebih akurat dibandingkan Decision Tree. Pada penelitian selanjutnya evaluasi dan validasi model yang lebih mendalam dengan menggunakan cross validation atau k-fold cross-validation dapat dilakukan untuk memastikan konsistensi hasil prediksi dan menghindari efek data yang tidak relevan.

6. Daftar Pustaka

[1] A. W. Wardhana, E. Patimah, A. I. Shafarindu, Y. M. Siahaan, B. V. Haekal, and D. S. Prasvita, “Klasifikasi Data Penjualan pada Supermarket dengan Metode Decision Tree,” in *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya*, 2021, pp. 660–667.

[2] A. K. Lalo, P. Batarius, and Y. C. H. Siki, “Implementasi Algoritma C4. 5 Untuk Klasifikasi Penjualan Barang di Swalayan Dutalia,” *Jurnal Teknik Informatika UNIKA Santo Thomas*, vol. 6, no. 1, pp. 1–12, 2021.

[3] I. Indriyani and A. Bahtiar, “Implementasi Data Mining Untuk Mengklasifikasikan Data Penjualan Pada Supermarket Menggunakan Algoritma Naïve Bayes,” *Jurnal Manajemen Dan Bisnis Ekonomi*, vol. 1, no. 1, pp. 207–220, 2023.

[4] S. A. Wicaksono, I. Kadek, and D. Nuryana, “Perbandingan Klasifikasi Hasil Indeks Kepuasan Masyarakat Terhadap UPT Balai Latihan Kerja Surabaya Menggunakan Algoritma Decision Tree, Random Forest, K-Nearest Neighbor,” *Journal of Informatics and Computer Science*, vol. 05, 2024.

[5] W. Azis Rahmat, S. Madinah Ladjamuddin, and D. Teruna Awaludin, “Perbandingan Algoritma Decision Tree, Random Forest Dan Naive Bayes Pada Prediksi Penilaian Kepuasan Penumpang Maskapai Pesawat Menggunakan Dataset Kaggle,” *Jurnal Rekayasa Informasi*, vol. 12, no. 2, 2023, [Online]. Available: www.kaggle.com,

[6] A. I. Kusumarini, P. A. Hogantara, and N. Chamidah, “Perbandingan Algoritma Random Forest, Naïve Bayes, Dan Decision Tree Dengan Oversampling Untuk Klasifikasi Bakteri E. Coli”. 2021.

[7] T. T. B. Dan, S. W. Sihwi, and R. Anggrainingsih, "Implementasi Iterative Dichotomiser 3 Pada Data Kelulusan Mahasiswa S1 Di Universitas Sebelas Maret," *ITSMART: Jurnal Teknologi dan Informasi*, vol. 4, no. 2, pp. 84–91, 2016.

[8] D. P. Sinambela, H. Naparin, M. Zulfadhilah, and N. Hidayah, "Implementasi Algoritma Decision Tree dan Random Forest dalam Prediksi Perdarahan Pascasalin," *Jurnal Informasi dan Teknologi*, vol. 5, no. 3, pp. 58–64, Sep. 2023, doi: 10.60083/jidt.v5i3.393.

[9] L. Ratnawati and D. R. Sulistyaningrum, "Penerapan random forest untuk mengukur tingkat keparahan penyakit pada daun apel," *Jurnal Sains dan Seni ITS*, vol. 8, no. 2, pp. A71–A77, 2020.

[10] A. U. Zailani and N. L. Hanun, "Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera," *Infotech: Journal of Technology Information*, vol. 6, no. 1, pp. 7–14, Jun. 2020, doi: 10.37365/jti.v6i1.61.

[11] N. Faoziatun Khusna *et al.*, "Implementasi Random Forest dalam Klasifikasi Kasus Stunting pada Balita dengan Hyperparameter Tuning Grid Search," *Seminar Nasional Sains Data*, vol. 2024.

[12] I. Optimasi *et al.*, "Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM," *Online) Teknologi: Jurnal Ilmiah Sistem Informasi*, vol. 13, no. 1, pp. 8–15, 2023, doi: 10.26594/teknologi.v13i1.3098.

[13] M. Fauzi, R. Mahendra, N. L. Azizah, and D. Sumarno, "Implementasi Machine Learning Untuk Memprediksi Cuaca Menggunakan Support Vector Machine", doi: 10.32409/jikstik.23.1.3449.

[14] J. Rusman, B. Z. Haryati, and A. Michael, "Optimisasi Hiperparameter Tuning pada Metode Support Vector Machine untuk Klasifikasi Tingkat Kematangan Buah Kopi," *J-Icon: Jurnal Komputer dan Informatika*, vol. 11, no. 2, pp. 195–202, 2023.

[15] M. Fadli and R. A. Saputra, "Klasifikasi Dan Evaluasi Performa Model Random Forest Untuk Prediksi Stroke Classification And Evaluation Of Performance Models Random Forest For Stroke Prediction," vol. 12, [Online]. Available: http://jurnal.umt.ac.id/index.php/jt/index

[16] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, pp. 393–399, Apr. 2021, doi: 10.29207/resti.v5i2.3008.