



SKRIPSI

PREDIKSI STRATEGI PEMASARAN BERBASIS DATA TRANSAKSI PENJUALAN DENGAN ALGORITMA XGBOOST DAN RANDOM FOREST

DAYINI NUR SABRINA
NPM 21081010322

DOSEN PEMBIMBING
Dr. Basuki Rahmat, S.Si., M.T.
Eva Yulia Puspaningrum, S.Kom., M.Kom.

**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI INFORMATIKA
SURABAYA
2024**



SKRIPSI

PREDIKSI STRATEGI PEMASARAN BERBASIS DATA TRANSAKSI PENJUALAN DENGAN ALGORITMA XGBOOST DAN RANDOM FOREST

DAYINI NUR SABRINA
NPM 21081010322

DOSEN PEMBIMBING
Dr. Basuki Rahmat, S.Si., M.T.
Eva Yulia Puspaningrum, S.Kom., M.Kom.

**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI INFORMATIKA
SURABAYA
2024**

Halaman ini sengaja dikosongkan

LEMBAR PENGESAHAN

PREDIKSI STRATEGI PEMASARAN BERBASIS DATA TRANSAKSI PENJUALAN DENGAN ALGORITMA XGBOOST DAN RANDOM FOREST

Oleh :
DAYINI NUR SABRINA
NPM. 21081010322

Telah dipertahankan dihadapan dan diterima oleh Tim Penguji Skripsi Prodi Informatika
Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jawa Timur Pada tanggal
.....

Dr. Basuki Rahmat, S.Si., M.T.
NIP. 19690723 2021211 002

..... (Pembimbing I)

Eva Yulia Puspaningrum, S.Kom., M.Kom.
NIP. 19890705 2021212 002

..... (Pembimbing II)

Nama Dosen
NIP/NPT

..... (Penguji I)

Nama Dosen
NIP/NPT

..... (Penguji II)

Mengetahui,
Dekan Fakultas Ilmu Komputer

Prof. Dr. Ir. Novirina Hendrasarie, MT
NIP. 19681126 199403 2 001

Halaman ini sengaja dikosongkan

SURAT PERNYATAAN ORISINALITAS

Yang bertandatangan di bawah ini:

Nama Mahasiswa : Dayini Nur Sabrina
Program Studi : Informatika
Dosen Pembimbing : 1. Dr. Basuki Rahmat, S.Si., M.T.
2. Eva Yulia Puspaningrum, S.Kom., M.Kom.

dengan ini menyatakan bahwa isi sebagian maupun keseluruhan disertasi dengan judul:

PREDIKSI STRATEGI PEMASARAN BERBASIS DATA TRANSAKSI PENJUALAN DENGAN ALGORITMA XGBOOST DAN RANDOM FOREST

adalah benar-benar hasil karya intelektual mandiri, diselesaikan tanpa menggunakan bahan-bahan yang tidak diizinkan dan bukan merupakan karya pihak lain yang saya akui sebagai karya sendiri. Semua referensi yang dikutip maupun dirujuk telah ditulis secara lengkap pada daftar pustaka. Apabila ternyata pernyataan ini tidak benar, saya bersedia menerima sanksi sesuai peraturan yang berlaku.

Surabaya,
Yang Membuat Pernyataan,

Dayini Nur Sabrina
NPM. 21081010322

Halaman ini sengaja dikosongkan

ABSTRAK

Nama Mahasiswa / NPM : Dayini Nur Sabrina / 21081010322
Judul Skripsi : Prediksi Strategi Pemasaran Berbasis Data
Transaksi Penjualan dengan Algoritma XGBoost
dan Random Forest
Dosen Pembimbing : 1. Dr. Basuki Rahmat, S.Si., M.T.
2. Eva Yulia Puspaningrum, S.Kom., M.Kom.

ABSTRAK

.

Kata kunci :

Halaman ini sengaja dikosongkan

ABSTRACT

Student Name / NPM : Dayini Nur Sabrina / 21081010322
Thesis Title : Prediction of Marketing Strategies Based on Sales
Transaction Data with XGBoost and Random Forest
Algorithms
Advisor : 1. Dr. Basuki Rahmat, S.Si., M.T.
2. Eva Yulia Puspaningrum, S.Kom., M.Kom.

ABSTRACT

.

Keywords:

Halaman ini sengaja dikosongkan

KATA PENGANTAR

Puji syukur kehadiran Tuhan yang Maha Esa atas segala rahmat, hidayah dan karunia-Nya kepada penulis sehingga skripsi dengan judul **“Prediksi Strategi Pemasaran Berbasis Data Transaksi Penjualan dengan Algoritma XGBoost dan Random Forest”** dapat terselesaikan dengan baik.

Penulis mengucapkan terima kasih kepada Bapak Dr. Basuki Rahmat, S.Si., M.T. dan Ibu Eva Yulia Puspaningrum, S.Kom., M.Kom. selaku Dosen Pembimbing yang telah meluangkan waktunya untuk memberikan bimbingan, nasehat serta motivasi kepada penulis. Dan penulis juga banyak menerima bantuan dari berbagai pihak, baik itu berupa moril, spiritual maupun materiil. Untuk itu penulis mengucapkan terima kasih kepada:

1. Ibu/Bapak..... selaku Dekan Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur.
2. Ibu/Bapak selaku Ketua Program Studi xxxx Fakultas Ilmu Sosial Dan Ilmu Komputer Universitas Pembangunan Nasional “Veteran“ Jawa Timur.
3. Dosen-dosen Program Studi ... dst..

Penulis menyadari bahwa di dalam penyusunan skripsi ini banyak terdapat kekurangan. Untuk itu kritik dan saran yang membangun dari semua pihak sangat diharapkan demi kesempurnaan penulisan skripsi ini. Akhirnya, dengan segala keterbatasan yang penulis miliki semoga laporan ini dapat bermanfaat bagi semua pihak umumnya dan penulis pada khususnya.

Surabaya,

Penulis

Halaman ini sengaja dikosongkan

DAFTAR ISI

LEMBAR PENGESAHAN	3
SURAT PERNYATAAN ORISINALITAS	5
ABSTRAK	1
ABSTRACT	3
KATA PENGANTAR.....	5
DAFTAR ISI.....	7
DAFTAR GAMBAR.....	9
DAFTAR TABEL	11
DAFTAR NOTASI.....	13
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	3
1.3. Tujuan Penelitian	4
1.4. Manfaat Penelitian	4
1.5. Batasan Masalah	5
BAB II TINJAUAN PUSTAKA.....	7
2.1. Penelitian Terdahulu	7
2.2. Data Mining	11
2.3 Machine Learning	14
2.4 Extreme Gradient Boosting.....	16
2.5 Random forest.....	17
2.6 Model Hybrid XGBoost dan Random Forest.....	19
2.7 Strategi Pemasaran	20
BAB III DESAIN DAN IMPLEMENTASI SISTEM	21
3.1 Tahapan Penelitian	21
3.2 Studi Literatur	22
3.3 Pengumpulan Data	22
3.4 Praproses Data.....	23
3.5 Implementasi Model	24
3.5.1 Implementasi Algoritma Random Forest.....	28
3.5.2 Implementasi Algoritma XGBoost	31

3.6	Evaluasi Model.....	34
3.7	Interpretasi Hasil	35
DAFTAR PUSTAKA		37

DAFTAR GAMBAR

Gambar 2.1 Tahap Data Mining.....	12
Gambar 2.2 Skema Algoritma XGBoost	16
Gambar 3.1 Tahapan Penelitian	21
Gambar 3.2 Praproses Data.....	23
Gambar 3.3 Implementasi Model Hybrid	25
Gambar 3.4 Implementasi Random Forest.....	28
Gambar 3.5 Implementasi Model XGBoost.....	31
Gambar 3.6 Mock Up Tampilan Penjualan Perbulan	35

Halaman ini sengaja dikosongkan

DAFTAR TABEL

Tabel 3.1 Sampel Data Transaksi Penjualan	23
---	----

Halaman ini sengaja dikosongkan

DAFTAR NOTASI

Halaman ini sengaja dikosongkan

BAB I

PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi informasi yang terus meningkat seiring dengan kemajuan peradaban manusia juga berdampak pada dunia bisnis. Dalam menghadapi persaingan, pihak usaha dituntut untuk memaksimalkan kemampuannya agar tetap bertahan dan bersaing dengan kompetitor lainnya. Dengan persaingan yang semakin ketat, pihak usaha tidak hanya harus berfokus pada produk yang ditawarkan, tetapi juga perlu mengarahkan perhatian mereka pada konsumen. Pemahaman terhadap kebutuhan dan produk yang diminati menjadi kunci penting dalam menciptakan strategi pemasaran yang efektif untuk tetap bersaing di pasar yang kompetitif (Awalina & Rahayu, 2023).

Seiring dengan berkembangnya pertumbuhan e-commerce, banyak pihak usaha mulai mengadopsi teknologi digital untuk meningkatkan efektivitas penjualan dan pemasaran mereka. Teknologi ini memungkinkan bisnis untuk mengumpulkan, menyimpan, serta menganalisis data transaksi guna memahami produk yang diminati konsumen dengan lebih akurat. Berdasarkan studi terbaru, prediksi berbasis data telah terbukti membantu pihak usaha dalam mengenali preferensi konsumen dan menyusun strategi pemasaran yang lebih tepat sasaran (Alghazzawi et al., 2023).

Infinity Jar sebagai salah satu merek yang berkembang di Surabaya, merupakan toko ritel yang menjual produk dengan fokus pada produk-produk inovatif yang dirancang untuk menarik minat konsumen. Dalam upaya memperluas pangsa pasar dan meningkatkan angka penjualan, pihak usaha perlu menganalisis produk-produk yang diminati konsumen serta mengidentifikasi pembelian mereka. Dengan memanfaatkan data transaksi yang ada, Infinity Jar Surabaya dapat memahami produk-produk dengan permintaan tertinggi dan menganalisis waktu puncak penjualan, sehingga dapat mengembangkan strategi pemasaran yang lebih efektif dan tepat sasaran (Massaro et al., 2021).

Dalam konteks pemasaran, analisis data transaksi sangat penting untuk merancang strategi yang efektif. Dengan menggunakan teknik analisis modern, pihak usaha dapat mengidentifikasi produk-produk dengan permintaan tertinggi yang sering tidak terlihat dengan metode tradisional. Misalnya, analisis ini dapat mengungkap

produk dengan peluang pertumbuhan tertinggi, sehingga memungkinkan alokasi sumber daya yang lebih efisien dan meningkatkan ROI (*Return on Investment*) dari kampanye pemasaran (Wijaya et al., 2024).

Di bidang analisis data, algoritma machine learning seperti XGBoost dan Random Forest telah terbukti efektif dalam memprediksi hasil berdasarkan data historis. XGBoost (Extreme Gradient Boosting) menggunakan teknik boosting untuk meningkatkan akurasi prediksi, sementara Random Forest adalah metode ensemble yang menggabungkan banyak pohon keputusan untuk meningkatkan akurasi dan mengurangi risiko overfitting. Karena kemampuannya menangani beragam variabel yang memengaruhi hasil akhir, kedua algoritma ini sangat ideal untuk analisis data transaksi penjualan yang kompleks (Chen, 2023).

XGBoost (Extreme Gradient Boosting) adalah algoritma machine learning berbasis boosting yang populer untuk analisis prediktif, terutama karena kemampuannya yang unggul dalam menangani data besar, tidak seimbang, dan dengan fitur berkolinear (Siringoringo et al., 2021). Dalam konteks Infinity Jar Surabaya, XGBoost sangat relevan karena mampu mengolah data transaksi dengan kategori produk dan pola pembelian yang beragam, mengidentifikasi pola tersembunyi yang tidak terdeteksi oleh metode tradisional. Selain akurasi yang tinggi, XGBoost menawarkan kecepatan dan efisiensi dalam proses pelatihan model, menjadikannya alat yang tepat untuk analisis penjualan dan optimalisasi strategi pemasaran di Infinity Jar Surabaya.

Random Forest memiliki keunggulan besar dalam interpretabilitas karena algoritma ini membangun banyak pohon keputusan secara bersamaan, lalu menggabungkan hasilnya untuk menghasilkan prediksi yang lebih stabil dan akurat. Setiap pohon dalam Random Forest menganalisis peran variabel-variabel yang berbeda, sehingga memungkinkan penilaian terhadap pentingnya setiap variabel dalam mempengaruhi hasil akhir (Zhou et al., 2023). Kemampuan ini sangat bermanfaat bagi Infinity Jar Surabaya karena memungkinkan mereka untuk mengetahui hal-hal yang memengaruhi pola pembelian konsumen, seperti kategori produk atau frekuensi transaksi. Dengan mengetahui hal-hal ini, mereka dapat menyesuaikan strategi pemasaran untuk meningkatkan efisiensi kampanye.

Dengan menerapkan kedua algoritma ini, Infinity Jar Surabaya dapat memperoleh wawasan yang lebih mendalam tentang pola pembelian konsumen.

Misalnya, analisis data dapat mengungkapkan bahwa pelanggan lebih cenderung membeli produk tertentu pada waktu-waktu tertentu. Informasi ini memungkinkan pihak usaha untuk merancang kampanye pemasaran yang lebih efektif dan meningkatkan loyalitas pelanggan (Ashraf, 2022).

Meskipun ada banyak keuntungan, implementasi algoritma machine learning juga menghadapi tantangan. Salah satu tantangan utama adalah kebutuhan akan data yang berkualitas tinggi. Data yang tidak bersih atau tidak representatif dapat menghasilkan prediksi yang tidak akurat (Song & Liu, 2020). Oleh karena itu, Infinity Jar Surabaya harus memastikan bahwa data transaksi yang digunakan dalam analisis adalah data yang bersih dan relevan untuk menghasilkan wawasan yang akurat.

Dengan memanfaatkan algoritma XGBoost dan Random Forest dalam analisis data transaksi penjualan, Infinity Jar Surabaya dapat mengoptimalkan strategi pemasaran mereka. Harapannya, penggunaan data dan teknologi ini tidak hanya meningkatkan penjualan, tetapi juga memberikan pengalaman yang lebih baik bagi konsumen. Dengan pendekatan yang berbasis data, Infinity Jar Surabaya dapat menjadi pemimpin di pasar dan terus berinovasi dalam produk dan layanan yang ditawarkan.

1.2. Rumusan Masalah

Berdasarkan uraian latar belakang di atas, permasalahan yang dapat dirumuskan adalah sebagai berikut:

1. Bagaimana algoritma XGBoost dalam model hybrid dapat memperbaiki hasil prediksi Random Forest untuk meningkatkan akurasi dalam merumuskan strategi pemasaran?
2. Bagaimana algoritma XGBoost dapat digunakan dalam model hybrid untuk mengidentifikasi produk laku terjual dan kurang laku terjual, sehingga mendukung perancangan strategi pemasaran yang lebih efektif di Toko Infinity Jar Surabaya?
3. Bagaimana hasil prediksi dari model hybrid (XGBoost dan Random Forest) dapat diinterpretasikan untuk merancang strategi pemasaran yang efektif di Toko Infinity Jar Surabaya?

1.3. Tujuan Penelitian

Berdasarkan rumusan masalah yang telah ditetapkan, maka tujuan penelitian ini adalah sebagai berikut:

1. Mengkaji efektivitas algoritma XGBoost dalam model hybrid untuk memperbaiki hasil prediksi dari algoritma Random Forest.
2. Menganalisis penerapan algoritma XGBoost dalam model hybrid untuk mengidentifikasi produk laku terjual dan kurang laku terjual.
3. Menginterpretasikan hasil prediksi dari model hybrid (XGBoost dan Random Forest) untuk merancang strategi pemasaran yang efektif.

1.4. Manfaat Penelitian

Penelitian ini memberikan manfaat bagi sejumlah pihak. Adapun manfaat penelitian ini bagi berbagai pihak terkait adalah sebagai berikut:

1. Bagi Akademisi dan Peneliti
 - a. Menambah pengetahuan tentang penerapan algoritma XGBoost dan Random Forest dalam analisis produk yang diminati untuk perumusan strategi pemasaran berbasis data.
 - b. Menyediakan referensi untuk penelitian lebih lanjut di bidang prediksi penjualan, strategi pemasaran, dan penerapan machine learning dalam bisnis, khususnya dengan algoritma XGBoost dan Random Forest.
2. Bagi Usaha
 - a. Meningkatkan efektivitas strategi pemasaran dengan dukungan prediksi yang lebih akurat.
 - b. Meningkatkan profitabilitas dengan menerapkan strategi pemasaran yang lebih terarah, berdasarkan data pembelian historis dan prediksi berbasis machine learning.
3. Bagi Konsumen
 - a. Meningkatkan kualitas pengalaman berbelanja dengan menyediakan produk yang lebih sesuai dengan preferensi dan kebutuhan konsumen.
 - b. Mempermudah konsumen dalam menemukan produk yang diinginkan melalui strategi pemasaran yang lebih tepat sasaran dan relevan bagi mereka.

1.5. Batasan Masalah

Adapun batasan masalah ini ditetapkan untuk memperjelas fokus dan arah penelitian sebagai berikut:

1. Penelitian ini hanya akan menggunakan data transaksi penjualan yang diperoleh dari Toko Infinity Jar Surabaya selama periode tertentu, yaitu dalam rentang waktu November 2023 hingga Oktober 2024.
2. Penelitian ini terbatas pada penggunaan algoritma XGBoost dan Random Forest untuk memprediksi produk paling laku terjual dan kurang laku terjual, sehingga tidak mencakup perbandingan dengan algoritma lain yang mungkin juga relevan.
3. Analisis akan dibatasi pada variabel-variabel tertentu yang relevan dengan pola pembelian, seperti jumlah pembelian, nomer transaksi, jenis produk, dan waktu pembelian.
4. Penelitian ini akan berfokus pada perumusan strategi pemasaran berbasis data dan hasil prediksi dari model hybrid, tanpa mencakup implementasi atau evaluasi strategi tersebut dalam praktik.
5. Penelitian ini akan mempertimbangkan hasil analisis dan perancangan strategi dalam rangka waktu yang telah ditentukan, dan tidak akan membahas dampak jangka panjang dari strategi yang diusulkan.
6. Hasil penelitian ini hanya digunakan untuk merancang strategi pemasaran pada jenis produk yang paling sering dibeli dan paling jarang dibeli oleh konsumen.

Halaman ini sengaja dikosongkan

BAB II

TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian-penelitian yang telah dilakukan oleh peneliti sebelumnya yang berkaitan dengan penelitian yang akan dilakukan adalah sebagai berikut:

1. Penelitian oleh Kaiwen Zhou, Xiang Li, Tong Sun, dan Bingchun Luo (2023).

Penelitian yang dilakukan oleh Kaiwen Zhou, Xiang Li, Tong Sun, dan Bingchun Luo yang berjudul “Research on Vegetable Sales Prediction and Strategy of Supermarkets Based on XGBoost and Random Forest Modeling”. Penelitian ini mengembangkan strategi pengisian stok dan penentuan harga sayuran segar di supermarket untuk meningkatkan keuntungan. Menggunakan data penjualan dan harga dari 24 hingga 30 Juni 2023, dengan prediksi yang difokuskan untuk periode minggu berikutnya, yaitu 1 Juli 2023. Dengan menganalisis data penjualan dan harga satuan, penelitian menggunakan uji korelasi Spearman untuk melihat hubungan antar produk. Model XGBoost dan Random Forest diterapkan untuk memprediksi volume penjualan dan harga grosir, sementara algoritma PSO mengoptimalkan strategi pengisian ulang dan penentuan harga, memperhitungkan harga grosir, permintaan, dan tingkat pembusukan. Hasilnya menunjukkan bahwa model prediksi memiliki akurasi tinggi dengan rata-rata kesalahan persentase absolut (MAPE) di bawah 20% dan nilai R^2 lebih dari 0,9 dan strategi PSO memberikan rekomendasi optimal untuk memberikan panduan berbasis data yang bermanfaat bagi supermarket.

2. Penelitian oleh Daniyal M. Alghazzawi, Anser Ghazal Ali Alquraishee, Sahar K. Badri, dan Syed Hamid Hasan (2023).

Penelitian yang dilakukan oleh Daniyal M. Alghazzawi, Anser Ghazal Ali Alquraishee, Sahar K. Badri, dan Syed Hamid Hasan yang berjudul “ERF-XGB: Ensemble Random Forest-Based XG Boost for Accurate Prediction and Classification of E-Commerce Product Review”. Penelitian ini mengembangkan model sentiment analysis ERF-XGB (Ensemble Random Forest-Based XGBoost) untuk meningkatkan akurasi klasifikasi sentimen ulasan produk e-commerce. Model ini diterapkan pada dataset IMDB dan

ChnSentiCorp untuk mengklasifikasikan sentimen sebagai positif atau negatif, dengan fokus pada mengatasi ambiguitas kata dan polisemik. Metrik evaluasi seperti akurasi, presisi, recall, dan F1-score menunjukkan bahwa ERF-XGB outperform metode lain seperti LSIBA-ENN dan CNBL. Proses penelitian meliputi pra-pemrosesan data, seleksi fitur menggunakan algoritma Harris Hawk Optimization, dan klasifikasi menggunakan kombinasi Random Forest dan XGBoost dengan hyperparameter tuning. Hasilnya menunjukkan akurasi 98.7% untuk ChnSentiCorp dan 98.2% untuk IMDB, menandakan efektivitas model dalam analisis sentimen. Selain itu, model ini berpotensi digunakan untuk analisis produk ramah lingkungan dan mendukung keputusan konsumen yang berkelanjutan.

3. Penelitian oleh Raja Sekhar Avula, Ergun Gide, dan Indra Seher (2023).

Penelitian yang dilakukan oleh Raja Sekhar Avula, Ergun Gide, dan Indra Seher yang berjudul “Integrated Machine Learning Concept with XG Booster and Random Forest Framework for Predicting Purchase Behaviour by Online Customers in E-Commerce Social Networks”. Penelitian ini mengkaji integrasi algoritma machine learning XGBoost dan Random Forest untuk memprediksi perilaku pembelian pelanggan online pada platform e-commerce dalam media sosial. Tujuannya adalah meningkatkan akurasi prediksi agar platform e-commerce dapat lebih efektif memahami dan memprediksi keputusan pembelian pelanggan. Matriks penelitian melibatkan variabel seperti demografi, aktivitas online, interaksi sosial, dan faktor eksternal lainnya yang dianalisis untuk menentukan pengaruhnya terhadap perilaku pembelian. Dengan metode kuantitatif, algoritma XGBoost dan Random Forest diaplikasikan pada data pengguna e-commerce untuk dianalisis dan dibandingkan efektivitasnya. Temuan menunjukkan bahwa kombinasi kedua algoritma ini memberikan akurasi prediksi yang lebih tinggi, memungkinkan identifikasi pola yang lebih baik dan model prediksi yang lebih andal untuk memahami preferensi pelanggan.

4. Penelitian oleh Huijun Chen (2023).

Penelitian yang dilakukan oleh Huijun Chen yang berjudul “Enterprise Marketing Strategy Using Big Data Mining Technology Combined with XGBoost Model in the New Economic Era”. Penelitian ini membahas strategi

pemasaran perusahaan menggunakan teknologi big data mining dan model XGBoost untuk meningkatkan efektivitas pemasaran di era digital. XGBoost dipilih karena kemampuannya dalam memproses data besar dan akurasi tinggi dalam klasifikasi. Empat strategi pemasaran presisi dirancang, berfokus pada prediksi perilaku pelanggan dan analisis faktor yang memengaruhi aktivasi pelanggan. Data dibagi menjadi set pelatihan (70%) dan pengujian (30%), dioptimalkan melalui cross-validation dan tuning parameter. Hasil penelitian menunjukkan bahwa model XGBoost memiliki akurasi dan recall yang lebih tinggi dibandingkan algoritma lain seperti regresi logistik, SVM, dan random forest. Variabel penting yang diidentifikasi termasuk status pekerjaan dan jenis unit kerja, yang memengaruhi keterlibatan pelanggan. Strategi berbasis big data mining ini terbukti meningkatkan loyalitas pelanggan dan memberikan kontribusi signifikan pada keberhasilan pemasaran yang dipersonalisasi.

5. Penelitian oleh Purvika Bajaj, Renesa Ray, Shivani Shedge, Shravani Vidhate, dan Prof. Dr. Nikhilkumar Shardoor (2020).

Penelitian yang dilakukan oleh Purvika Bajaj, Renesa Ray, Shivani Shedge, Shravani Vidhate, dan Prof. Dr. Nikhilkumar Shardoor yang berjudul “Sales Prediction Using Machine Learning Algorithms”. Penelitian ini menerapkan algoritma machine learning untuk memprediksi penjualan Big Mart menggunakan data historis dan fitur seperti berat barang, konten lemak, visibilitas, tipe barang, harga, tahun pendirian, ukuran, dan lokasi outlet. Metode yang digunakan mencakup Linear Regression, K-Neighbors Regressor, XGBoost Regressor, dan Random Forest Regressor, dengan proses pembersihan data, pengisian data hilang, encoding kategorikal, dan scaling fitur. Visualisasi data seperti heat map menunjukkan bahwa harga barang memiliki dampak terbesar terhadap penjualan. Hasil menunjukkan bahwa Random Forest Regressor memiliki akurasi terbaik (93,53%), membuktikan efektivitasnya dalam menangkap pola kompleks. Pendekatan ini membantu perusahaan merancang strategi yang lebih tepat dan kompetitif.

6. Penelitian oleh Yunxin Liang, Jiyu Wu, Wei Wang, Yujun Cao, Biliang Zhong, dan Zhenkun Chen, Zhenzhang Li (2019).

Penelitian yang dilakukan oleh Yunxin Liang, Jiyu Wu, Wei Wang, Yujun Cao, Biliang Zhong, dan Zhenkun Chen, Zhenzhang Li yang berjudul

“Product Marketing Prediction based on XGboost and LightGBM Algorithm”. Penelitian ini memprediksi volume penjualan produk menggunakan algoritma XGBoost dan LightGBM pada data historis dari Kaggle (Januari 2013–Oktober 2015). Peneliti menganalisis pengaruh parameter *n_estimators* dan *gamma* pada XGBoost, serta *learning_rate* pada LightGBM terhadap akurasi prediksi. Data dibersihkan, dinormalisasi, dan dievaluasi menggunakan RMSE. Hasil menunjukkan *gamma* sangat memengaruhi prediksi pada XGBoost, sedangkan *learning_rate* berpengaruh besar pada LightGBM. Meskipun kedua algoritma menunjukkan hasil RMSE yang mirip dalam prediksi, XGBoost sedikit lebih akurat. Temuan ini menyoroti pentingnya optimasi parameter untuk meningkatkan performa prediksi pada kedua algoritma.

7. Penelitian oleh Farida Titik Kristanti, Mochamad Yudha Febrianta, Dwi Fitrizal Salim, Hosam Alden Riyadh, Yoga Sagama, Baligh Ali Hasan Beshr (2024).

Penelitian yang dilakukan oleh Farida Titik Kristanti, Mochamad Yudha Febrianta, Dwi Fitrizal Salim, Hosam Alden Riyadh, Yoga Sagama, Baligh Ali Hasan Beshr yang berjudul “Advancing Financial Analytics: Integrating XGBoost, LSTM, and Random Forest Algorithms for Precision Forecasting of Corporate Financial Distress”. Penelitian ini meningkatkan akurasi prediksi financial distress perusahaan di Indonesia dengan algoritma XGBoost, LSTM, dan Random Forest, menggunakan data rasio keuangan dari Bursa Efek Indonesia (2013-2022). Untuk mengatasi ketidakseimbangan kelas, digunakan teknik SMOTE dan RUS, yang membantu meningkatkan performa model. Tahapan penelitian mencakup pengumpulan dan pembersihan data, pengembangan model baseline, tuning hyperparameter, dan evaluasi model berdasarkan metrik seperti akurasi, AUC-ROC, precision, recall, dan F1-score. Hasil menunjukkan bahwa XGBoost dengan SMOTE mencapai akurasi tertinggi, yaitu sekitar 98,86%, dengan AUC-ROC sebesar 98,06%, menjadikannya model yang paling efektif untuk prediksi financial distress. Random Forest juga menunjukkan performa kompetitif, sementara LSTM memiliki waktu pelatihan lebih lama dan performa lebih rendah. Rasio "operating profit per share" dan "operating margin" menjadi variabel kunci

dalam prediksi, sehingga temuan ini dapat membantu perusahaan dan regulator dalam mengelola risiko keuangan di Indonesia.

8. Penelitian oleh Xueling Li dan Zhen Li (2019).

Penelitian yang dilakukan oleh Xueling Li dan Zhen Li yang berjudul “A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm”. Penelitian ini mengembangkan model prediksi churn pelanggan e-commerce menggunakan pendekatan hibrida regresi logistik dan XGBoost. Memanfaatkan data dari platform e-commerce, model ini mengandalkan lebih dari 20 indikator, seperti informasi pesanan, profil pelanggan, preferensi, dan adhesivitas, untuk memprediksi churn. Model dasar dibangun dengan regresi logistik, lalu XGBoost meningkatkan akurasi dengan memilih variabel kunci seperti jumlah hari pemesanan, waktu pesanan pertama, sensitivitas promosi, dan daya beli, dan jumlah hari login menjadi variabel utama dalam menentukan apakah pelanggan akan mengalami churn. Hasil dari model hibrida ini menunjukkan akurasi prediksi lebih dari 85%, yang lebih tinggi dibandingkan dengan model regresi logistik saja. Temuan ini menawarkan panduan berharga bagi perusahaan e-commerce dalam meningkatkan loyalitas pelanggan melalui pemahaman yang lebih dalam mengenai pola dan risiko churn pelanggan.

2.2. Data Mining

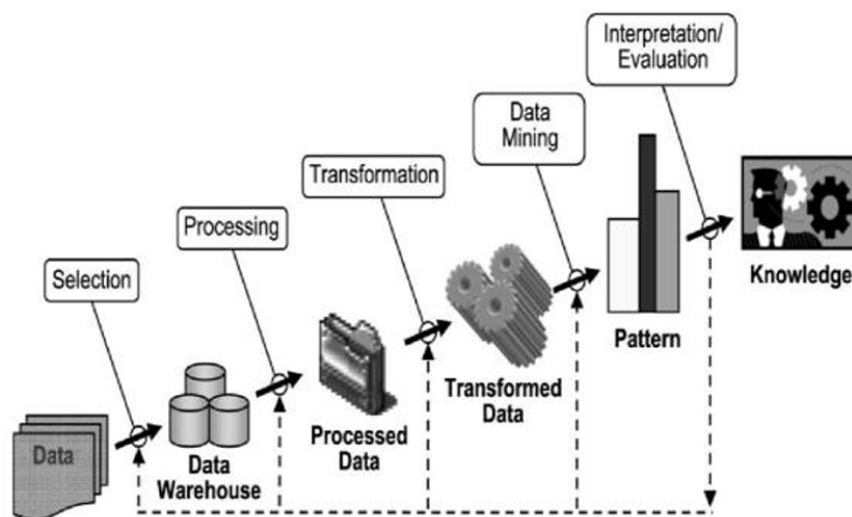
Data mining adalah proses mengidentifikasi pola atau informasi menarik dalam data yang dipilih dengan memanfaatkan berbagai teknik atau metode tertentu. Dalam praktiknya, terdapat beragam metode, teknik, atau algoritma yang dapat digunakan dalam data mining, dan pemilihan metode yang tepat sangat bergantung pada tujuan serta tahap-tahap dalam proses *Knowledge Discovery in Database* (KDD) (Yuli Mardi, 2019). Data mining mulai dikenal sejak tahun 1990-an sebagai metode yang efektif untuk mengekstraksi pola dan informasi yang dapat mengungkap hubungan antar data. Teknik ini digunakan untuk mengelompokkan data ke dalam satu atau lebih cluster, sehingga objek-objek dalam satu cluster memiliki tingkat kesamaan yang tinggi satu sama lain (Zahra et al., 2024).

Menurut (Zahra et al., 2024), data mining memiliki beberapa keunggulan sebagai alat analisis, di antaranya:

- a. Data mining dapat mengelola data dalam jumlah besar dan kompleks.
- b. Data mining juga mampu menangani data dengan berbagai jenis atribut.
- c. Data mining memiliki kemampuan untuk mencari dan memproses data secara otomatis yang disebut semi otomatis karena beberapa teknik dalam data mining membutuhkan input parameter dari pengguna secara manual.
- d. Data mining dapat memanfaatkan pengalaman dan kesalahan sebelumnya untuk meningkatkan kualitas analisis, menghasilkan output terbaik.

Namun, data mining juga memiliki kekurangan. Dalam pencariannya, data mining tidak bekerja pada data secara individu tetapi sebagai satu set individu atau kumpulan yang memenuhi kriteria tertentu.

Terdapat beberapa tahapan dalam melakukan proses data mining. Rincian setiap tahap dalam proses *Knowledge Discovery in Database* (KDD) dapat dilihat pada Gambar 2.1.



Gambar 2.1 Tahap Data Mining
(Sumber: Suliman, 2021)

Menurut (Suliman, 2021), rincian setiap tahap dalam proses Knowledge Discovery dapat dijelaskan sebagai berikut:

1. Data Selection

Data dari sekumpulan data operasional perlu diseleksi terlebih dahulu sebelum memulai tahap eksplorasi informasi dalam *Knowledge Discovery in Database* (KDD). Data hasil seleksi ini nantinya akan digunakan dalam proses data mining dan disimpan dalam berkas terpisah dari basis data operasional.

2. Pre-processing/Cleaning

Sebelum proses data mining dilaksanakan, perlu dilakukan pembersihan pada data yang menjadi fokus *Knowledge Discovery in Database* (KDD). Proses ini meliputi penghapusan data duplikat, pengecekan konsistensi data, dan perbaikan kesalahan, seperti kesalahan penulisan. Selain itu, dilakukan juga proses enrichment, yaitu memperkaya data yang sudah ada dengan informasi tambahan yang relevan, seperti data eksternal yang dibutuhkan untuk *Knowledge Discovery in Database* (KDD).

3. Transformation

Transformasi data atau coding, dilakukan untuk mengubah data yang telah dipilih agar sesuai dengan kebutuhan proses data mining. Proses ini bersifat kreatif dalam *Knowledge Discovery in Database* (KDD) dan sangat bergantung pada jenis atau pola informasi yang ingin ditemukan dalam basis data.

4. Data Mining

Data mining adalah tahap menemukan pola atau informasi yang bermanfaat dalam data yang telah dipilih, menggunakan teknik atau metode tertentu. Teknik dan metode dalam data mining beragam, dan pemilihan metode yang tepat tergantung pada tujuan serta keseluruhan proses *Knowledge Discovery in Database* (KDD).

5. Interpretation/Evaluation

Pola informasi yang dihasilkan dari proses data mining perlu disajikan dalam format yang mudah dipahami oleh pihak terkait. Tahap ini merupakan bagian dari proses *Knowledge Discovery in Database* (KDD) yang disebut interpretasi. Proses ini mencakup evaluasi untuk memastikan apakah pola atau informasi yang ditemukan sesuai atau bertentangan dengan fakta atau hipotesis yang ada sebelumnya. Langkah akhir dalam *Knowledge Discovery in Database* (KDD) adalah menyampaikan pengetahuan dalam bentuk yang mudah dimengerti oleh pengguna.

Data mining dapat dikategorikan ke dalam beberapa kelompok berdasarkan tugas yang dapat dilaksanakan (Yuli Mardi, 2019), yaitu:

1. Description (Deskripsi)

Para peneliti dan analis sering berusaha mencari cara untuk menggambarkan pola dan trend yang tidak terlihat dalam data.

2. Estimation (Estimasi)

Estimasi serupa dengan klasifikasi, namun lebih memusatkan perhatian pada variabel target yang berbentuk numerik bukan kategori. Model dikembangkan dengan memanfaatkan record lengkap yang menyediakan nilai variabel target sebagai nilai prediksi. Selanjutnya, dalam evaluasi berikutnya estimasi nilai dari variabel target ditentukan berdasarkan nilai variabel yang diprediksi.

3. Prediction (Prediksi)

Prediksi memiliki kesamaan dengan klasifikasi dan estimasi, tetapi yang membedakannya adalah nilai yang akan muncul di masa mendatang. Beberapa algoritma dan teknik yang digunakan dalam klasifikasi dan estimasi juga bisa digunakan untuk prediksi dalam kondisi yang tepat.

4. Classification (Klasifikasi)

Dalam klasifikasi variabel, tujuan bersifat kategorik. Sebagai contoh pengklasifikasian persediaan dalam tiga kelas, yaitu persediaan tinggi, persediaan sedang dan persediaan rendah.

5. Clustering (Pengkusteran)

Clustering merupakan teknik pengelompokan record data, pengamatan atau kasus dalam kelas yang memiliki kemiripan. Cluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record lain dalam cluster.

6. Association (Asosiasi)

Mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada satu waktu. Dalam dunia bisnis lebih umum disebut sebagai analisis keranjang belanja.

2.3 Machine Learning

Machine learning adalah bagian dari kecerdasan buatan yang sering digunakan untuk menyelesaikan berbagai jenis masalah. Machine learning dapat diartikan sebagai penerapan algoritma komputer dan matematika yang memungkinkan sistem untuk belajar dari data dan membuat prediksi untuk masa depan. Proses pembelajaran ini melibatkan usaha untuk mencapai kecerdasan melalui dua tahap, yaitu pelatihan (training) dan pengujian (testing) (Roihan et al., 2020).

Menurut (Wijoyo A et al., 2024), algoritma machine learning dibagi menjadi tiga jenis, yaitu:

1. Supervised Learning

Supervised Learning adalah metode machine learning yang menggunakan data berlabel untuk melatih algoritma dalam membuat estimasi terbaik terhadap output (Y) dari input (X). Proses ini efektif untuk tugas klasifikasi dan regresi, dengan contoh algoritma seperti K-NN, Naive Bayes, Decision Tree, dan Support Vector Machine. Meski sederhana dan mudah dipahami, supervised learning membutuhkan data pelatihan yang akurat dan komputasi yang lebih lama dibanding unsupervised learning karena memerlukan pelabelan pada setiap input.

2. Unsupervised Learning

Unsupervised learning adalah algoritma yang bekerja tanpa memerlukan data berlabel. Dalam pendekatan ini, algoritma tidak membutuhkan data pelatihan, melainkan digunakan untuk menemukan pola atau membuat model deskriptif tanpa perlu kategori atau output yang sudah ditentukan. Algoritma unsupervised learning banyak diterapkan dalam clustering dan asosiasi aturan. Keunggulannya adalah fleksibilitasnya dalam mendeteksi pola yang sebelumnya mungkin tidak dikenali. Namun, kekurangannya adalah sulitnya mengidentifikasi informasi spesifik di dalam data karena tidak ada label, sehingga sulit juga membandingkan output dengan input.

3. Semi Supervised dan Reinforcement Learning

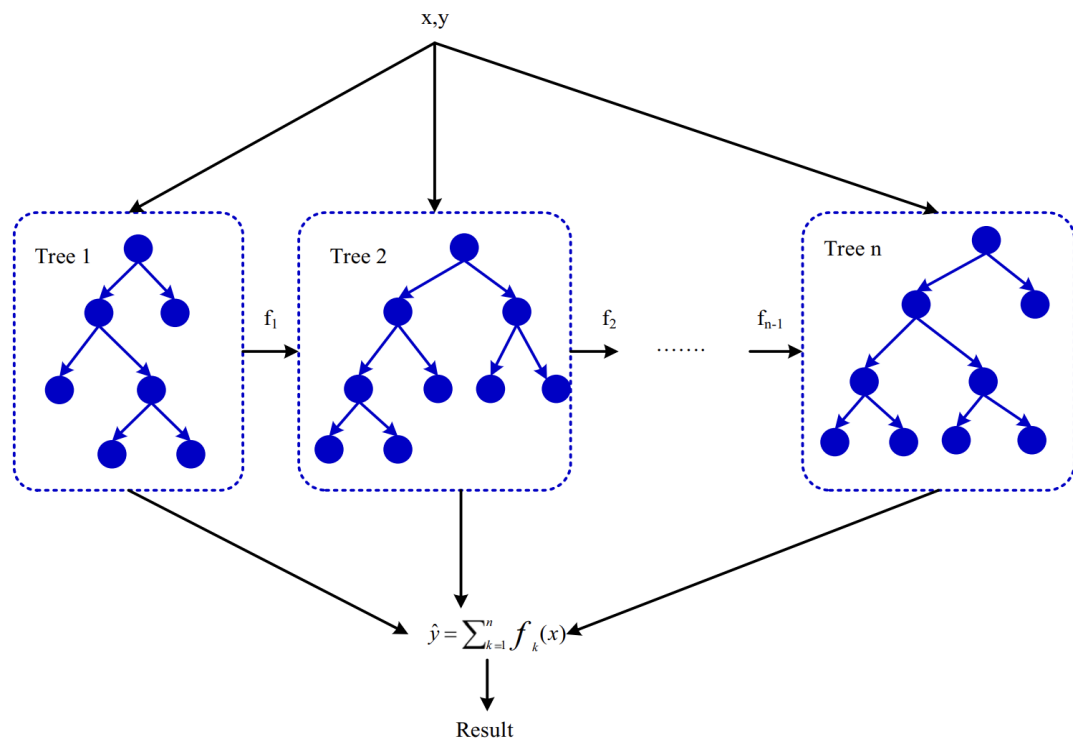
Semi-supervised learning adalah algoritma yang menggabungkan pendekatan supervised dan unsupervised, dengan bekerja pada data besar yang sebagian berlabel dan sebagian tidak. Keunggulannya adalah lebih hemat biaya karena hanya sebagian data yang memerlukan pelabelan dan tidak membutuhkan tenaga ahli untuk pemrosesan. Sementara itu, reinforcement learning bertujuan memaksimalkan hasil dan mengurangi risiko dengan mengamati interaksi agen dengan lingkungannya. Algoritma ini belajar secara berulang, di mana agen mengamati data input, mengambil tindakan, dan menerima "reward" atau umpan balik dari lingkungan. Dengan mengamati

input ulang dan mendapatkan umpan balik tambahan, agen memperbaiki keputusannya secara bertahap untuk hasil yang lebih akurat.

2.4 Extreme Gradient Boosting

XGBoost (Extreme Gradient Boosting) adalah algoritma pengembangan dari Gradient Tree Boosting berbasis ensemble yang dirancang untuk menangani masalah machine learning pada skala besar dengan kinerja dan efisiensi yang tinggi (Saputra et al., 2024). Diperkenalkan oleh Dr. Tianqi Chen pada tahun 2014 (Herni Yulianti et al., 2022), XGBoost menawarkan pengoptimalan yang 10 kali lebih cepat dibandingkan metode boosting lainnya, sehingga banyak diterapkan di berbagai bidang, termasuk kesehatan, risiko kredit, dan metagenomic (Saputra et al., 2024).

Algoritma ini bekerja dengan menyempurnakan parameter secara iteratif untuk meminimalkan *loss function*, yang merupakan ukuran dari kesalahan model. Konsep dasar XGBoost adalah membangun model pohon regresi yang teratur untuk meningkatkan kinerja dan mengurangi kompleksitas model, sehingga dapat menghindari overfitting (Saputra et al., 2024). Proses komputasi algoritma XGBoost dapat dilihat pada Gambar 2.2. Hasil prediksi akhir dari XGBoost dihasilkan dari penjumlahan prediksi dari setiap pohon regresi (Parveen et al., 2024).



Gambar 2.2 Skema Algoritma XGBoost
(Sumber: Parveen et al., 2024)

Fungsi objektif dalam XGBoost dinyatakan dengan rumus:

$$obj(\theta) = L(\theta) + \Omega(\theta)$$

Di mana, $L(\theta)$ adalah fungsi *loss* yang mengukur kesalahan antara nilai aktual dan nilai prediksi. Sedangkan, $\Omega(\theta)$ adalah fungsi regularisasi yang menjaga agar model tetap sederhana. Fungsi *loss* $L(\theta)$ dituliskan sebagai berikut:

$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i)$$

Dengan penjelasan y_i adalah nilai aktual, sedangkan \hat{y}_i adalah hasil prediksi dari model, dan n adalah jumlah data pelatihan. Fungsi ini membantu model untuk belajar dari kesalahan prediksi dan memperbaikinya di setiap iterasi. Sementara itu, fungsi regularisasi bertujuan untuk mencegah model menjadi terlalu kompleks, yang dapat menyebabkan overfitting. XGBoost memiliki keunggulan dalam menangani data dengan fitur kategorikal dan mengatasi data yang tidak seimbang (Saputra et al., 2024).

2.5 Random forest

Random Forest merupakan algoritma machine learning yang menggabungkan beberapa pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan menentukan metode yang lebih efektif dalam memproses. Salah satu keunggulan dari Random Forest adalah kemampuannya untuk menangani dataset besar dengan beragam fitur, serta mengatasi masalah overfitting yang sering muncul pada pohon keputusan tunggal. Selain itu, algoritma ini dapat mempertahankan kinerja yang tinggi dan stabil (Ary Prandika Siregar et al., 2023).

Sebagai pengembangan dari metode Decision Tree, Random Forest terdiri dari beberapa pohon keputusan yang dilatih menggunakan sampel yang berbeda. Setiap pohon memecah atribut berdasarkan subset acak. Random Forest memiliki beberapa manfaat, termasuk meningkatkan akurasi ketika ada data yang hilang, mampu mengatasi outlier, dan efisien dalam penyimpanan data. Selain itu, algoritma ini juga melakukan seleksi fitur yang efektif, sehingga dapat mengidentifikasi atribut terbaik dan meningkatkan performa model klasifikasi. Dengan adanya seleksi fitur, Random Forest dapat beroperasi dengan baik pada data besar dengan parameter yang kompleks (Marlina Haiza et al., 2023).

Menurut (Suci Amaliah et al., 2022), Metode Random Forest pertama kali diperkenalkan oleh Leo Breiman pada tahun 2001 dan termasuk dalam kategori

ensemble learning, yang berarti memanfaatkan kombinasi beberapa model untuk meningkatkan kinerja. Random Forest adalah teknik yang dapat meningkatkan akurasi dengan menghasilkan atribut acak untuk setiap node. Pohon keputusan dibangun dengan menentukan node akar dan diakhiri dengan beberapa node daun untuk mencapai hasil akhir. Proses pembentukan pohon keputusan dalam Random Forest serupa dengan proses pada Classification and Regression Tree (CART), tetapi Random Forest tidak melakukan pemangkasan (*pruning*). Indeks Gini digunakan untuk memilih fitur di setiap simpul internal dari pohon keputusan, dengan nilai Indeks Gini dapat dihitung melalui rumus berikut:

$$\text{Gini}(S_i) = 1 - \sum_{i=0}^{c-1} p_i^2$$

Di mana:

S_i adalah subset data pada node i .

p_i adalah frekuensi relatif dari kelas i di dalam subset S_i .

c adalah jumlah kelas yang ada.

Untuk menghitung kualitas pemisahan (split) pada fitur k ke dalam subset S_i , rumusnya adalah:

$$\text{Gini}_{\text{split}} = \sum_{i=0}^{k-1} \left(\frac{n_i}{n} \right) \text{Gini}(S_i)$$

Di mana:

n_i adalah jumlah sampel dalam subset S_i setelah pemisahan.

n adalah jumlah total sampel di node yang diberikan.

Fungsi margin untuk Random Forest, yang mengukur seberapa baik model dapat mengklasifikasikan kelas tertentu, dinyatakan sebagai:

$$mr(X, Y) = P_{\Theta}(h(X, \Theta) = Y) - \max_{j \neq Y} P_{\Theta}(h(X, \Theta) = j)$$

Di mana:

X adalah fitur input.

Y adalah kelas yang benar.

P_{Θ} adalah probabilitas yang diprediksi oleh model.

Kekuatan himpunan pengklasifikasi dapat diukur dengan:

$$s = E_{X,Y} mr(X, Y)$$

Batas atas kesalahan generalisasi dapat diturunkan sebagai berikut:

$$P_E \leq \frac{\bar{p}(1 - s^2)}{s^2}$$

Di mana:

\bar{p} adalah rata-rata korelasi antara pengklasifikasi.

Rumus-rumus ini memberikan dasar matematis untuk cara Random Forest membangun model dan mengevaluasi kualitas prediksi yang dihasilkan.

2.6 Model Hybrid XGBoost dan Random Forest

Optimalisasi model dalam machine learning dapat dilakukan melalui teknik tertentu atau dengan mengkombinasikan dua model yang dikenal sebagai ensemble. Ensemble Learning adalah metode yang mengintegrasikan beberapa algoritma machine learning, di mana model-model tersebut digabungkan untuk mencapai prediksi yang lebih akurat dibandingkan dengan menggunakan satu model tunggal. Dengan mengkombinasikan berbagai pendekatan machine learning, teknik ini bertujuan untuk mengatasi kelemahan masing-masing model dan menghasilkan prediksi yang lebih konsisten dan akurat. Penggunaan beberapa model telah terbukti efektif dalam mengatasi masalah yang kompleks, terutama dalam situasi yang melibatkan dimensi tinggi serta tantangan dalam regresi dan klasifikasi (Dayera, Musa Bundaris Palungan, 2024).

Menurut (Rayadin et al., 2024), dalam kasus regresi terdapat beberapa metode ensemble yang sering digunakan, yaitu:

1. Bagging (Bootstrap Aggregating)

Metode ini menggunakan beberapa model secara paralel untuk meningkatkan stabilitas dan akurasi. Setiap model dilatih dengan subset acak dari data pelatihan yang diperoleh melalui proses bootstrap, dan hasil prediksi dari setiap model digabungkan untuk menghasilkan prediksi akhir.

2. Stacking

Teknik ini mengkombinasikan beberapa model untuk meningkatkan akurasi prediksi. Dalam stacking, model dasar (base learner) menghasilkan metadata yang kemudian diproses oleh model tingkat atas (meta-learner). Base learner dikenal sebagai level 0 learner, sementara meta-learner yang dihasilkan dari penggabungan base learner disebut sebagai stacking.

3. Boosting

Metode ini membangun model secara bertahap dengan melatih model dasar secara berurutan dan menggabungkan hasilnya untuk membuat prediksi. Setiap model dilatih menggunakan set pelatihan yang diberi bobot, di mana bobot ditentukan berdasarkan kesalahan model sebelumnya. Tujuan dari boosting adalah untuk mengubah pembelajar lemah menjadi pembelajar yang kuat.

Dalam konteks model hybrid, penggabungan antara Random Forest dan XGBoost dapat memanfaatkan kelebihan masing-masing metode, sehingga menghasilkan model yang lebih akurat dan stabil dalam prediksi (Rayadin et al., 2024).

2.7 Strategi Pemasaran

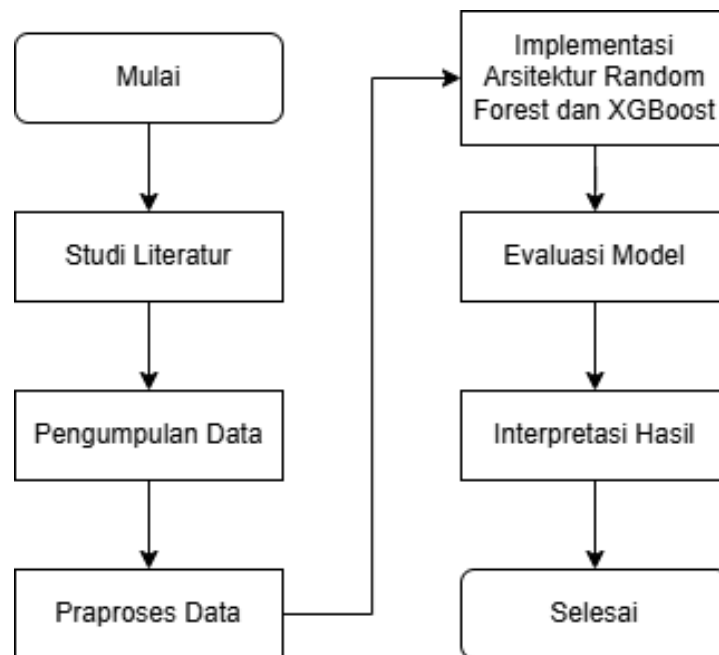
Strategi pemasaran adalah upaya pihak usaha dalam merancang tindakan yang inovatif dan adaptif guna mencapai tujuannya dengan mempertimbangkan perkembangan di lingkungan internal maupun eksternal (Seran et al., 2023). Secara umum, strategi ini berfungsi sebagai arahan dalam memilih pasar yang dituju serta menyusun bauran pemasaran yang tepat (Yolanda & Dwiridhotjahjono, 2021). Sebagai alat untuk mencapai sasaran, strategi pemasaran berperan penting dalam mengarahkan operasional pihak usaha untuk mewujudkan tujuan jangka pendek, menengah, dan panjang (Setyaningsih, 2021). Di tengah persaingan bisnis yang ketat, strategi pemasaran yang baik dapat membantu pihak usaha meraih keunggulan kompetitif secara berkesinambungan, baik dalam sektor produk maupun jasa. Oleh karena itu, penting bagi pihak usaha, termasuk usaha kecil, untuk memprioritaskan perencanaan strategi pemasaran yang matang agar mampu bertahan dan bersaing di pasar (Z et al., 2022).

BAB III

DESAIN DAN IMPLEMENTASI SISTEM

3.1 Tahapan Penelitian

Pada sebuah penelitian, diperlukan tahapan-tahapan yang sistematis untuk mencapai tujuan penelitian secara efektif agar mampu menjawab rumusan masalah serta mencapai tujuan penelitian yang sebagaimana telah dijelaskan pada bagian sebelumnya.



Gambar 3.1 Tahapan Penelitian

Pada Gambar 3.1 menjelaskan alur proses prediksi strategi pemasaran berbasis data transaksi penjualan menggunakan metode ensemble learning. Alur dimulai dengan studi literatur untuk memahami konsep dasar dan metode yang relevan, kemudian dilanjutkan dengan pengumpulan data transaksi penjualan di Infinity Jar Surabaya. Setelah data terkumpul, dilakukan praproses data untuk membersihkan dan menyiapkan data agar siap diolah. Implementasi arsitektur menggunakan Random Forest di layer pertama dan XGBoost di layer kedua diterapkan dalam tahap pemodelan, di mana model dilatih menggunakan dataset yang telah diproses. Setelah pelatihan model selesai, dilakukan evaluasi untuk mengukur kinerja model dalam memprediksi strategi pemasaran yang efektif. Proses diakhiri dengan evaluasi model dan interpretasi hasil yang telah selesai.

3.2 Studi Literatur

Dalam tahap awal penelitian ini, peneliti mengumpulkan sumber literatur untuk memahami, mencari, dan mempelajari permasalahan yang diangkat. Tahap ini sangat penting karena memberikan pemahaman mendalam tentang konsep dan definisi metode yang digunakan, termasuk istilah teknis yang relevan. Dengan memanfaatkan berbagai sumber, seperti jurnal, artikel, buku, dan penelitian terdahulu, peneliti memperoleh landasan teoretis yang kuat mengenai metode prediksi strategi pemasaran berbasis data transaksi penjualan, khususnya algoritma Random Forest, XGBoost, dan pendekatan ensemble.

Prediksi strategi pemasaran menggunakan algoritma machine learning telah banyak dikaji untuk meningkatkan efektivitas dalam memprediksi perilaku pola pembelian. Random Forest dikenal handal dalam menangani data kompleks melalui pendekatan ensemble, sementara XGBoost menawarkan optimasi efisien untuk menghasilkan performa tinggi. Penggunaan metode ensemble yang menggabungkan kedua algoritma ini terbukti meningkatkan akurasi model, memungkinkan pemahaman lebih dalam terhadap pola data transaksi penjualan, dan merumuskan strategi pemasaran yang lebih tepat sasaran.

Studi sebelumnya menunjukkan bahwa pendekatan hybrid dengan Random Forest dan XGBoost memberikan keunggulan dalam interpretabilitas dan performa, serta menawarkan solusi efektif untuk menangani data transaksi yang bervariasi, mendukung pihak usaha dalam merancang strategi pemasaran yang efektif.

3.3 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini berasal dari studi kasus Infinity Jar Surabaya, yang menyediakan informasi tentang transaksi penjualan selama satu tahun terakhir, yaitu dari November 2023 hingga Oktober 2024. Data tersebut akan dikumpulkan dan dirangkum dalam file (.csv). Dataset ini berisi informasi tentang transaksi penjualan produk yang telah dilakukan selama periode tersebut.

Dataset mencakup berbagai atribut, seperti kode produk, jenis produk, nomer transaksi, tanggal transaksi, dan jumlah pesanan. Dengan menggunakan dataset ini, penelitian akan melakukan analisis peramalan untuk memprediksi penjualan produk menggunakan data historis berdasarkan penjualan produk dari data yang terkumpul. Tabel 3.1 menampilkan sampel dari dataset yang telah dikumpulkan terkait transaksi

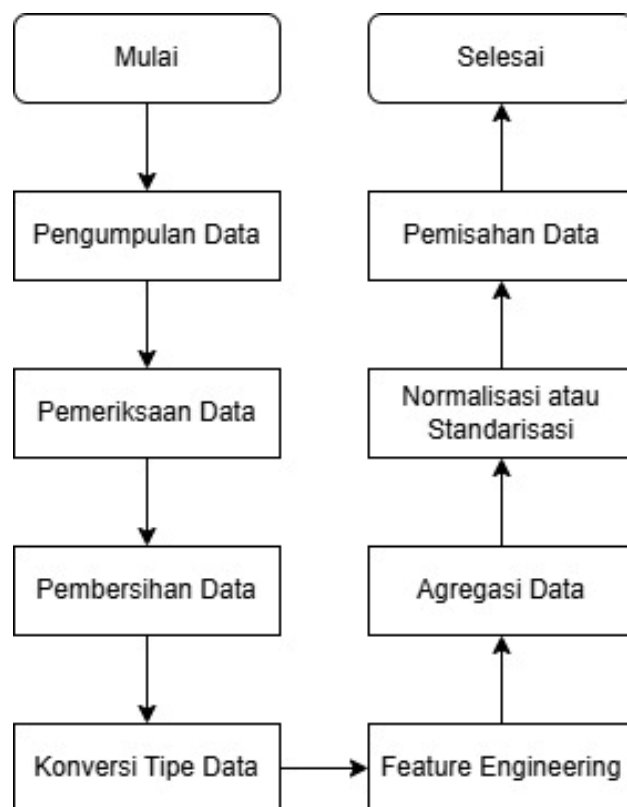
penjualan di Infinity Jar Surabaya selama satu tahun terakhir, mulai dari awal November 2023 hingga akhir Oktober 2024.

Tabel 3.1 Sampel Data Transaksi Penjualan

Kode Produk	Jenis Produk	Nomer Transaksi	Tanggal Transaksi	Jumlah Pesanan
CSP01	Candle Jar Strawberry Pie	JL00007218	01/11/23	1
TRC01	Tracker Book	JL00007219	01/11/23	1
TRC01	Tracker Book	JL00007220	01/11/23	1
NST01	Nutrijar Strawberry	JL00007221	01/11/23	1
NDC01	Nutrijar Dark Chocolate	JL00007221	01/11/23	1
...
CVW01	Candle Jar Vanilla Waffle	JL00009941	31/10/24	1
JRT01	Jartistic	JL00009942	31/10/24	1

3.4 Praproses Data

Praproses data yang baik akan meningkatkan kualitas data yang digunakan dalam model machine learning, sehingga dapat menghasilkan prediksi yang lebih akurat. Dengan mengikuti langkah-langkah pada Gambar 3.2, penelitian ini bertujuan untuk memberikan pemahaman tentang bagaimana data transaksi penjualan dapat diproses.



Gambar 3.2 Praproses Data

Praproses data merupakan langkah penting dalam mempersiapkan dataset untuk analisis dan pemodelan. Proses dimulai dengan pengumpulan data dengan sumber file CSV, yang dapat mencakup informasi tentang transaksi penjualan. Selanjutnya, dilakukan pemeriksaan data untuk memastikan kualitas dan konsistensi data yang dikumpulkan, dengan mengidentifikasi nilai yang hilang, duplikat, dan outlier.

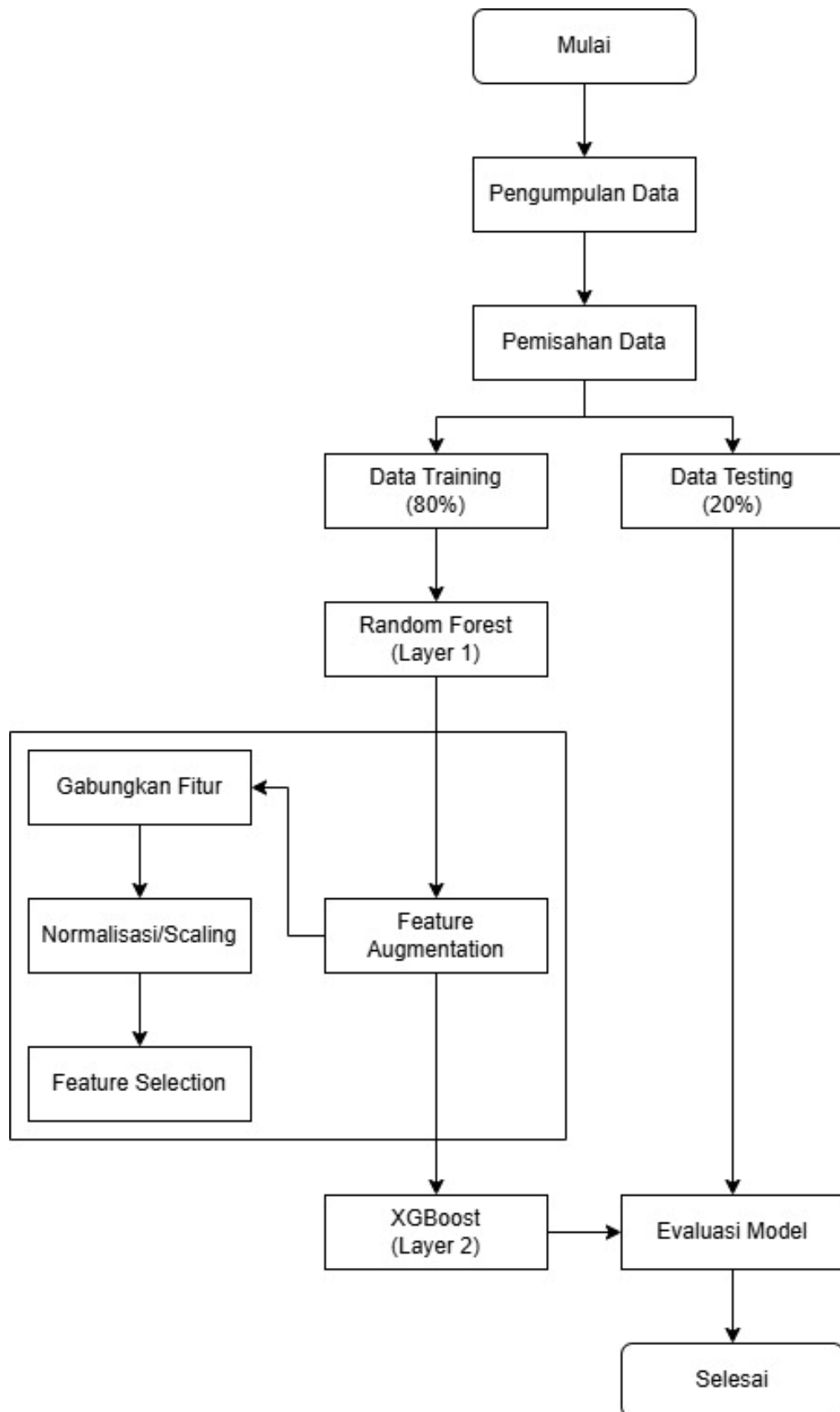
Setelah pemeriksaan, tahap berikutnya adalah pembersihan data, di mana masalah yang teridentifikasi diatasi, seperti menghapus duplikat dan mengisi nilai yang hilang. Kemudian, dilakukan konversi tipe data untuk memastikan kolom atau fitur berada dalam format yang sesuai, seperti mengubah kolom tanggal menjadi format tanggal dan mengonversi kolom kategorikal menjadi format numerik.

Selanjutnya, dalam tahap feature engineering, fitur baru dibuat dari data yang ada untuk meningkatkan model, seperti menghitung total penjualan per bulan. Proses ini diikuti dengan agregasi data, di mana data dari beberapa sumber digabungkan atau dikelompokkan berdasarkan kategori tertentu untuk analisis lebih lanjut.

Tahap terakhir dalam praproses data adalah normalisasi atau standarisasi, di mana skala fitur diubah agar semua fitur memiliki rentang yang sama. Normalisasi digunakan untuk mengubah data ke dalam rentang 0-1, sedangkan standarisasi menghitung nilai z-score. Akhirnya, dilakukan pemisahan data, di mana dataset dibagi menjadi dua bagian, yaitu data latih (training data) dan data uji (testing data) dengan proporsi 80% untuk pelatihan dan 20% untuk pengujian. Proses ini penting untuk memastikan kinerja model dapat diuji setelah dilatih.

3.5 Implementasi Model

Setelah tahap praproses data selesai, tahap selanjutnya adalah implementasi model. Pada tahap ini, model akan dibangun dan dilatih menggunakan data yang telah dipersiapkan. Dalam penelitian ini, digunakan pendekatan hybrid model yang menggabungkan dua algoritma machine learning, yaitu Random Forest dan XGBoost. Penggunaan hybrid model bertujuan untuk memanfaatkan kekuatan masing-masing algoritma, sehingga diharapkan dapat meningkatkan akurasi dan performa prediksi dalam analisis data transaksi penjualan di Infinity Jar Surabaya. Pada Gambar 3.3 merupakan penjelasan untuk setiap tahap dalam alur yang telah direncanakan, sebagai berikut:



Gambar 3.3 Implementasi Model Hybrid

Proses implementasi model hybrid ini meliputi beberapa langkah penting, sebagai berikut:

1. Pemisahan Dataset

Dataset yang tersedia akan dipisahkan menjadi dua bagian, yaitu data training dan data testing. Data training akan digunakan untuk melatih model, sedangkan data testing akan digunakan untuk mengevaluasi kinerja model. Proporsi yang digunakan adalah 80% untuk data training dan 20% untuk data testing. Pemisahan ini untuk memastikan bahwa model dapat menggeneralisasi dengan baik terhadap data yang belum pernah dilihat sebelumnya.

2. Pelatihan Model dengan Random Forest

Model Random Forest akan dilatih menggunakan data training. Algoritma ini efektif dalam menangani data dengan banyak fitur dan mampu memberikan estimasi yang baik melalui pengelolaan beberapa pohon keputusan. Selama pelatihan, model Random Forest juga akan menghasilkan prediksi untuk data training yang akan digunakan dalam tahap berikutnya.

3. Augmentasi Fitur

Hasil prediksi dari Random Forest akan digabungkan dengan fitur asli untuk membentuk dataset baru yang lebih informatif. Proses augmentasi fitur ini bertujuan untuk meningkatkan kualitas dataset sebelum digunakan untuk pelatihan model selanjutnya. Berikut adalah langkah-langkah yang diambil dalam tahapan augmentasi fitur, sebagai berikut:

- a. Gabungkan Fitur Secara Terstruktur

Pada tahap ini, hasil prediksi dari model Random Forest akan digabungkan dengan fitur-fitur asli dalam dataset. Proses penggabungan harus dilakukan secara terstruktur, di mana prediksi yang dihasilkan menjadi fitur baru dalam dataset. Penggabungan ini bertujuan untuk menciptakan dataset yang lebih informatif, di mana fitur-fitur baru yang dihasilkan dari model Random Forest dapat memberikan wawasan tambahan bagi model XGBoost. Dengan cara ini, informasi yang dihasilkan dari analisis model sebelumnya dapat dimanfaatkan untuk meningkatkan performa model yang sedang dibangun.

- b. Normalisasi untuk Menyamakan Skala

Setelah proses penggabungan, penting untuk melakukan normalisasi pada semua fitur yang ada, termasuk fitur asli dan fitur baru hasil prediksi. Normalisasi membantu untuk menyamakan skala dari fitur-fitur yang berbeda, sehingga model dapat mempelajari pola dengan lebih efektif. Proses ini juga mencegah fitur tertentu yang memiliki skala yang lebih besar mendominasi proses pelatihan model. Dengan menyamakan skala, dapat memastikan bahwa semua fitur memiliki pengaruh yang seimbang terhadap hasil prediksi model.

c. Pilih Fitur yang Relevan

Dari kumpulan fitur yang telah diaugmentasi, langkah selanjutnya adalah melakukan pemilihan fitur untuk memastikan hanya fitur yang relevan dan berdampak positif pada model yang akan digunakan. Pemilihan fitur ini sangat penting untuk mengurangi dimensi data, meningkatkan efisiensi model, dan menghindari masalah overfitting. Dalam proses ini, teknik seperti analisis feature importance yang diperoleh dari model Random Forest dapat digunakan untuk mengidentifikasi fitur mana yang paling berkontribusi terhadap hasil prediksi. Dengan memilih fitur yang tepat, model XGBoost akan memiliki fondasi yang kuat untuk belajar dan memprediksi.

4. Pelatihan Model XGBoost

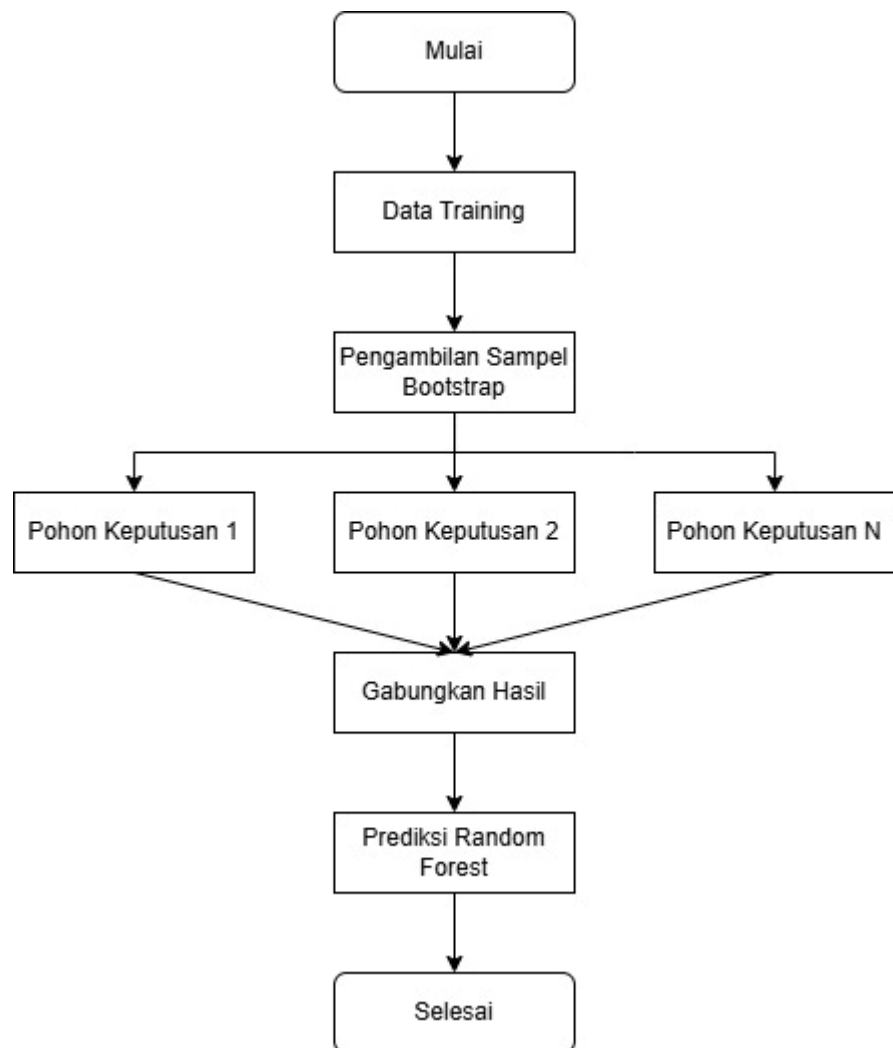
Setelah proses augmentasi, model XGBoost akan dilatih menggunakan dataset yang telah diaugmentasi. XGBoost merupakan algoritma boosting yang digunakan dalam meningkatkan akurasi model dengan mengoptimalkan performa berdasarkan kesalahan yang dibuat oleh model sebelumnya.

5. Evaluasi Model

Setelah pelatihan model selesai, dilakukan evaluasi performa menggunakan data testing. Berbagai metrik evaluasi seperti akurasi, presisi, recall, dan F1-score akan digunakan untuk menilai seberapa baik model dalam memprediksi data yang belum pernah dilihat sebelumnya.

Pada Gambar 3.3, ditampilkan diagram alur dari setiap tahap dalam implementasi model yang telah direncanakan agar memberikan gambaran yang jelas tentang langkah-langkah yang diambil dalam penelitian ini.

3.5.1 Implementasi Algoritma Random Forest



Gambar 3.4 Implementasi Random Forest

Setelah data asli yang terdiri dari atribut Kode Produk, Jenis Produk, Nomer Transaksi, Tanggal Transaksi, dan Jumlah Pesanan dibersihkan, langkah selanjutnya adalah mempersiapkan data untuk pemodelan. Proses ini telah mencakup beberapa tahapan fitur engineering agar data siap digunakan oleh model.

Pertama, variabel kategorikal, seperti Kode Produk dan Jenis Produk, diubah menjadi format numerik menggunakan teknik One-Hot Encoding. Teknik ini mengonversi setiap kategori pada kolom tersebut menjadi fitur biner (0 atau 1) sehingga model dapat mengolah variabel kategorikal secara lebih efektif. Misalnya, jika terdapat beberapa jenis

produk yang berbeda, setiap jenis akan direpresentasikan sebagai kolom fitur biner di dataset.

Selanjutnya, jumlah transaksi per bulan ditentukan sebagai target variabel atau y , yang akan diprediksi oleh model. Ini dihitung dengan mengelompokkan data berdasarkan tanggal transaksi untuk menghitung total transaksi di setiap bulan. Sementara itu, fitur atau X untuk model terdiri dari sisa atribut yang berfungsi sebagai input.

Setelah menentukan fitur dan target, model Random Forest dilatih menggunakan data training. Random Forest adalah metode ensemble learning yang membangun banyak pohon keputusan untuk memperbaiki akurasi prediksi. Pada tahap ini, dua parameter utama dari Random Forest diatur:

- a. Jumlah pohon ($n_estimators$): Menentukan jumlah pohon keputusan yang akan dibuat oleh model. Semakin banyak pohon, semakin baik akurasi model, tetapi dengan risiko waktu komputasi yang lebih lama.
- b. Kedalaman maksimum (max_depth): Mengatur kedalaman maksimal pohon keputusan. Parameter ini penting untuk mencegah overfitting, yaitu ketika model terlalu terfokus pada data training sehingga tidak dapat memprediksi data baru dengan baik.

Dalam setiap pohon pada Random Forest, model akan secara acak memilih subset dari fitur untuk membuat pohon keputusan. Setelah model dilatih, hasil prediksi dari Random Forest pada data training digunakan sebagai fitur tambahan untuk augmentasi fitur.

Augmentasi ini menambahkan prediksi bulanan dari Random Forest ke dalam dataset sebagai fitur baru, yang bertujuan memperkaya informasi yang tersedia bagi model lanjutan, sehingga pola bulanan atau hubungan non-linear dalam data lebih mudah dikenali. Dengan fitur tambahan ini, model selanjutnya dapat memanfaatkan prediksi awal dari Random Forest untuk menghasilkan estimasi yang lebih akurat pada data yang belum terlihat yang bertujuan untuk meningkatkan informasi yang tersedia bagi model selanjutnya. Berikut

adalah penjelasan untuk proses pembuatan model prediksi dengan algoritma Random Forest:

1. Persiapan Data:

- a. Kode Produk: Identitas unik untuk setiap produk.
- b. Jenis Produk: Produk-produk yang dibeli saat transaksi.
- c. Nomor Transaksi: ID unik yang diberikan untuk setiap transaksi.
- d. Tanggal Transaksi: Tanggal di mana transaksi terjadi.
- e. Jumlah Pesanan: Jumlah produk yang dipesan dalam setiap transaksi.

2. Pembentukan Subset Data:

Dari data pelatihan, Random Forest membentuk beberapa subset data secara acak, yang disebut dengan *bootstrap samples*. Setiap subset digunakan untuk membangun pohon keputusan secara independen.

3. Pemilihan Fitur Secara Acak:

Setiap pohon di dalam Random Forest hanya menggunakan subset acak dari fitur. Misalnya, jika ada 5 fitur, setiap pohon mungkin hanya menggunakan 3 di antaranya. Strategi ini membantu membuat pohon-pohon yang bervariasi dan mencegah model dari overfitting.

4. Pembentukan Pohon Keputusan:

Dalam setiap subset data, algoritma membangun pohon keputusan dengan membagi data berdasarkan aturan tertentu (misalnya, jenis produk) hingga mencapai kedalaman maksimal atau data pada node menjadi homogen. Setiap pohon belajar dari subset yang berbeda, yang membantu meningkatkan variasi dan akurasi model secara keseluruhan.

5. Evaluasi Prediksi Setiap Pohon Keputusan

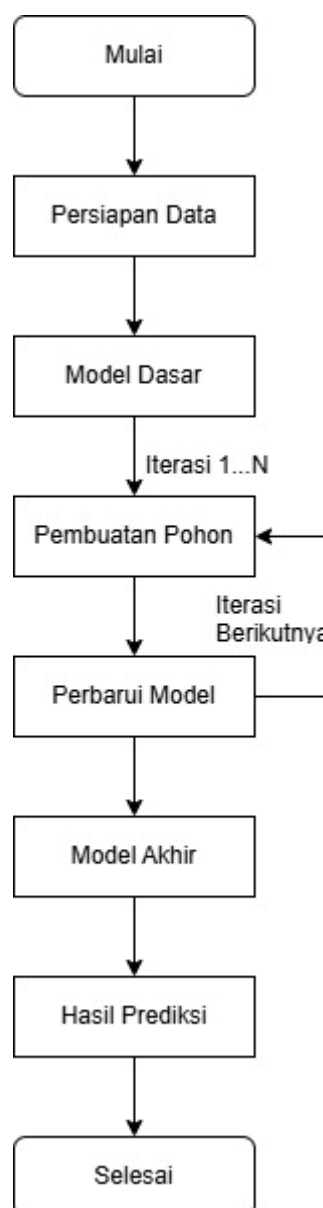
Setelah setiap pohon dalam Random Forest memberikan prediksi terhadap data uji (jumlah transaksi bulanan), hasil prediksi dari semua pohon digabungkan. Untuk regresi, rata-

rata hasil dari setiap pohon digunakan, sedangkan untuk klasifikasi, voting mayoritas dipakai.

6. Output Akhir Random Forest:

Prediksi akhir dari Random Forest adalah jumlah transaksi bulanan setiap produk. Prediksi dari Random Forest ini disimpan sebagai fitur tambahan yang akan digunakan oleh model XGBoost pada tahap berikutnya untuk menghasilkan prediksi yang lebih akurat.

3.5.2 Implementasi Algoritma XGBoost



Gambar 3.5 Implementasi Model XGBoost

Setelah model Random Forest dilatih, hasil prediksi bulanan dari Random Forest pada data training digunakan sebagai fitur tambahan di dataset. Ini dikenal sebagai augmentasi fitur, di mana output model pertama (Random Forest) ditambahkan kembali ke dalam data asli sebagai variabel baru. Penambahan ini dilakukan agar prediksi yang dihasilkan dari Random Forest, yang mungkin menangkap pola-pola penting dalam data, dapat menjadi informasi tambahan yang kaya bagi model selanjutnya. Dengan cara ini, XGBoost dapat memanfaatkan hasil prediksi dari Random Forest, memperkuat kemampuannya dalam mengenali pola kompleks atau hubungan non-linear yang mungkin tidak tertangkap pada model pertama.

Setelah proses augmentasi selesai, model XGBoost dilatih menggunakan dataset yang diperluas, termasuk hasil prediksi Random Forest sebagai salah satu fitur input baru. XGBoost merupakan algoritma boosting yang sangat efektif dalam menangani data besar serta menangkap pola-pola non-linear dalam dataset karena XGBoost beroperasi dengan membangun model secara bertahap dan mengoreksi kesalahan dari model sebelumnya, ia sangat mampu memanfaatkan informasi tambahan dari hasil prediksi Random Forest.

Pada akhirnya, prediksi akhir dari model XGBoost berupa klasifikasi produk yang terbagi menjadi dua kategori yaitu produk yang laku dan produk yang kurang laku. Dengan keunggulannya dalam menangani data besar dan pola-pola kompleks, XGBoost menghasilkan klasifikasi yang lebih akurat berdasarkan pola penjualan historis dan informasi agregat bulanan. Integrasi antara prediksi Random Forest dan proses boosting XGBoost ini meningkatkan kinerja model, membantu menghasilkan prediksi produk laku atau kurang laku dengan lebih akurat.

Pada tahap kedua dalam model hybrid, XGBoost digunakan untuk meningkatkan akurasi prediksi berdasarkan output dari model Random Forest dan fitur-fitur awal lainnya. Berikut adalah alur kerja XGBoost yang bertujuan menghasilkan prediksi produk laku dan kurang laku:

1. Input Data dari Random Forest dan Fitur Asli:

XGBoost menerima dua sumber data yaitu prediksi jumlah transaksi bulanan dari Random Forest (sebagai fitur tambahan untuk digabungkan) dan fitur asli seperti Kode Produk, Jenis Produk, Tanggal Transaksi, dan Jumlah Pesanan. Kombinasi ini memungkinkan XGBoost menangkap pola yang tidak terakomodasi sepenuhnya oleh Random Forest.

2. Pembentukan Model XGBoost:

Model XGBoost menggunakan teknik gradient boosting, membentuk pohon-pohon keputusan secara bertahap. Setiap pohon baru memperbaiki kesalahan pohon sebelumnya, sehingga XGBoost mampu menangkap pola-pola yang kompleks.

3. Training Model XGBoost:

- a. Langkah Pertama: Dimulai dengan membuat pohon keputusan awal berdasarkan data.
- b. Perbaikan Bertahap: Setelah pohon pertama, XGBoost menghitung kesalahan prediksi dan melatih pohon berikutnya untuk memperbaiki kesalahan tersebut.
- c. Iterasi Berulang: Proses berulang hingga tercapai jumlah pohon yang ditentukan atau kesalahan tidak berkurang signifikan.
- d. Regularisasi untuk Mencegah Overfitting: XGBoost menerapkan regularisasi (L1 dan L2) untuk menghindari overfitting, menjaga model tetap sederhana dan lebih mampu menggeneralisasi data baru.

4. Prediksi Akhir:

Setelah seluruh pohon terbentuk, XGBoost menghasilkan prediksi akhir dengan menjumlahkan kontribusi dari semua pohon, memberikan prediksi yang telah diperbaiki secara bertahap untuk hasil yang lebih akurat.

5. Evaluasi Kinerja:

Kinerja model XGBoost dievaluasi menggunakan metrik seperti akurasi, precision, recall, atau F1-score, serta divalidasi dengan teknik seperti cross-validation untuk memastikan konsistensi hasil.

3.6 Evaluasi Model

Evaluasi model merupakan langkah penting dalam pengembangan sistem machine learning yang efektif, terutama dalam konteks analisis data penjualan di Infinity Jar Surabaya. Hasil evaluasi memberikan gambaran yang jelas tentang performa model dalam memprediksi produk paling laris dan tidak laris berdasarkan data transaksi. Model kami dievaluasi menggunakan berbagai metrik, termasuk akurasi, presisi, recall, dan F1-score, yang masing-masing memiliki peran penting dalam menilai kualitas prediksi.

Akurasi memberikan informasi umum tentang seberapa banyak prediksi yang benar dibandingkan dengan total prediksi yang dibuat. Dalam konteks ini, akurasi tinggi menunjukkan bahwa model mampu mengenali pola penjualan dengan baik. Sementara itu, presisi dan recall berfokus pada kinerja model dalam mengidentifikasi produk yang benar-benar laris. Hal ini sangat penting, mengingat kesalahan dalam memprediksi produk yang laris atau tidak laris dapat berdampak signifikan pada strategi pemasaran dan pengelolaan inventaris.

F1-score, sebagai gabungan dari presisi dan recall, memberikan pandangan yang lebih seimbang tentang kinerja model, terutama ketika terdapat ketidakseimbangan antara produk yang laris dan tidak laris. Penggunaan confusion matrix juga diimplementasikan untuk menganalisis prediksi model pada setiap kelas, memberikan wawasan mendalam tentang jenis kesalahan yang terjadi, seperti false positives (produk yang diprediksi laris tetapi sebenarnya tidak) dan false negatives (produk yang diprediksi tidak laris tetapi sebenarnya laris). Informasi ini sangat berharga untuk perbaikan model di masa mendatang.

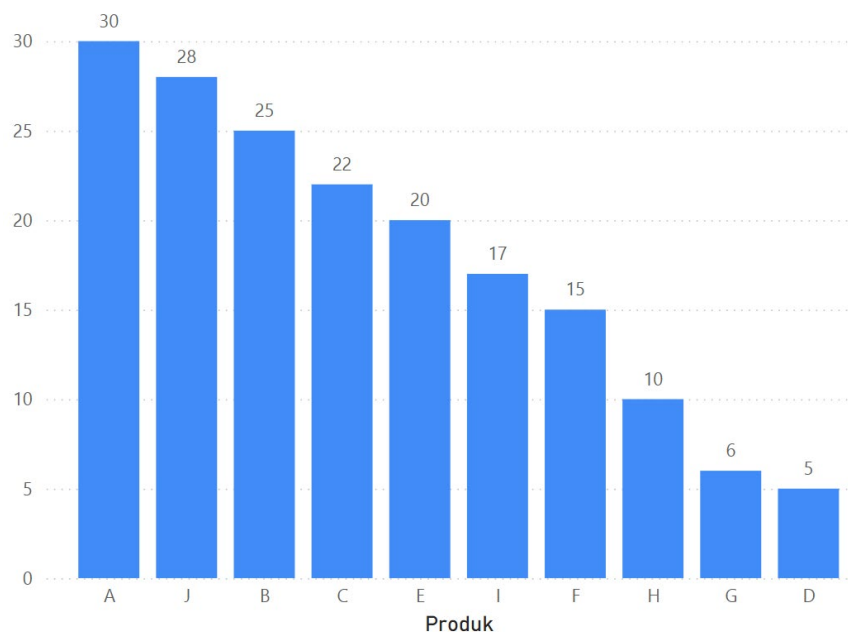
Proses evaluasi dilakukan secara menyeluruh pada validation set dan test set untuk memastikan bahwa model tidak hanya berfungsi dengan baik pada data yang telah dilatih, tetapi juga mampu menggeneralisasi dengan baik pada data yang tidak terlihat sebelumnya. Hal ini penting untuk menghindari masalah overfitting, di mana model terlalu menyesuaikan diri dengan data pelatihan sehingga kehilangan

kemampuannya untuk melakukan prediksi yang akurat pada data baru. Dengan menggunakan validation set, peneliti dapat melakukan penyesuaian parameter dan memilih model yang paling sesuai sebelum menguji kinerjanya pada test set.

Pengujian pada test set memberikan gambaran akhir tentang kemampuan model dalam situasi dunia nyata, khususnya dalam konteks penjualan produk di Infinity Jar Surabaya. Dengan demikian, evaluasi yang komprehensif tidak hanya meningkatkan keandalan model, tetapi juga memberikan dasar yang kuat untuk pengambilan keputusan yang lebih baik dalam penerapan model di berbagai bidang.

3.7 Interpretasi Hasil

Dalam penelitian ini, visualisasi hasil memberikan gambaran yang jelas mengenai pola pembelian produk di Infinity Jar Surabaya. Dengan menggunakan diagram untuk dapat mengidentifikasi produk-produk yang paling sering dibeli oleh konsumen, serta produk yang jarang dibeli. Pada Gambar 3.6 merupakan mock up untuk tampilan penjualan produk per bulannya.



Gambar 3.6 Mock Up Tampilan Penjualan Perbulan

Hasil penelitian ini akan diinterpretasikan sebagai acuan untuk memprediksi strategi pemasaran yang tepat dalam meningkatkan penjualan. Visualisasi ini mencakup produk yang sering dibeli dan produk yang jarang dibeli pada setiap bulannya, serta interpretasi yang mendalam untuk memprediksi strategi pemasaran yang sesuai. Dengan demikian, informasi ini diharapkan dapat digunakan untuk

merancang strategi pemasaran yang lebih efektif, serta mengoptimalkan penawaran yang lebih menarik bagi konsumen.

DAFTAR PUSTAKA

- Alghazzawi, D. M., Alquraishee, A. G. A., Badri, S. K., & Hasan, S. H. (2023). ERF-XGB: Ensemble Random Forest-Based XG Boost for Accurate Prediction and Classification of E-Commerce Product Review. *Sustainability (Switzerland)*, 15(9). <https://doi.org/10.3390/su15097076>
- Ary Prandika Siregar, Dwi Priyadi Purba, Jojo Putri Pasaribu, & Khairul Reza Bakara. (2023). Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke. *Jurnal Penelitian Rumpun Ilmu Teknik*, 2(4), 155–164. <https://doi.org/10.55606/juprit.v2i4.3039>
- Ashraf, D. M. U. (2022). A Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques. *Lahore Garrison University Research Journal of Computer Science and Information Technology*, 6(04), 23–33. <https://doi.org/10.54692/lgurjcsit.2022.0604399>
- Awalina, E. F. L., & Rahayu, W. I. (2023). Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail. *Jurnal Teknologi Dan Informasi*, 13(2), 122–137. <https://doi.org/10.34010/jati.v13i2.10090>
- Chen, H. (2023). Enterprise marketing strategy using big data mining technology combined with XGBoost model in the new economic era. *PLoS ONE*, 18(6 June), 1–22. <https://doi.org/10.1371/journal.pone.0285506>
- Dayera, Musa Bundaris Palungan, F. O. (2024). G-Tech : Jurnal Teknologi Terapan. *G-Tech : Jurnal Teknologi Terapan*, 8(1), 186–195. <https://ejournal.uniramalang.ac.id/index.php/g-tech/article/view/1823/1229>
- Herni Yulianti, S. E., Oni Soesanto, & Yuana Sukmawaty. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *Journal of Mathematics: Theory and Applications*, 4(1), 21–26. <https://doi.org/10.31605/jomta.v4i1.1792>
- Marlina Haiza, Elmayati, Zulus Antoni, & Wijaya Harma Oktafia Lingga. (2023). Penerapan Algoritma Random Forest Dalam Klasifikasi Penjurusan Di SMA Negeri Tugumulyo. *Penerapan Kecerdasan Buatan*, 4(2), 138–143.
- Massaro, A., Panarese, A., Giannone, D., & Galiano, A. (2021). Augmented data and Xgboost improvement for sales forecasting in the large-scale retail sector. *Applied Sciences (Switzerland)*, 11(17). <https://doi.org/10.3390/app11177793>
- Parveen, N., Gupta, M., Kasireddy, S., Ansari, M. S. H., & Ahmed, M. N. (2024). ECG based one-dimensional residual deep convolutional auto-encoder model for heart disease classification. *Multimedia Tools and Applications*, 83(25), 66107–66133. <https://doi.org/10.1007/s11042-023-18009-7>
- Rayadin, M. A., Musaruddin, M., Saputra, R. A., & Isnawaty, I. (2024). Implementasi Ensemble Learning Metode XGBoost dan Random Forest untuk Prediksi Waktu Penggantian Baterai Aki. *BIOS: Jurnal Teknologi Informasi Dan Rekayasa Komputer*, 5(2), 111–119.
- Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. *IJCIT (Indonesian Journal on Computer and Information Technology)*, 5(1), 75–82. <https://doi.org/10.31294/ijcit.v5i1.7951>
- Saputra, A. A., Sari, B. N., Rozikin, C., Singaperbangsa, U., & Abstrak, K. (2024). Penerapan Algoritma Extreme Gradient Boosting (Xgboost) Untuk Analisis Risiko Kredit. *Jurnal Ilmiah Wahana Pendidikan*, 10(7), 27–36.

<https://doi.org/10.5281/zenodo.10960080>

- Seran, R. B., Sundari, E., & Fadhila, M. (2023). Strategi Pemasaran yang Unik: Mengoptimalkan Kreativitas dalam Menarik Perhatian Konsumen. *Jurnal Mirai Management*, 8(1), 206–211. <https://journal.stieamkop.ac.id/index.php/mirai/article/download/4054/2644>
- Setyaningsih, F. (2021). Analisis Strategi Bauran Pemasaran Terhadap Peningkatan Pelanggan Pada Lembaga Rafi Bimbel Tangerang. *JMB : Jurnal Manajemen Dan Bisnis*, 10(1), 221–230. <https://doi.org/10.31000/jmb.v10i1.4233>
- Siringoringo, R., Perangin-angin, R., & Purba, M. J. (2021). Segmentasi Dan Peramalan Pasar Retail Menggunakan Xgboost Dan Principal Component Analysis. *METHOMIKA Jurnal Manajemen Informatika Dan Komputerisasi Akuntansi*, 5(1), 42–47. <https://doi.org/10.46880/jmika.vol5no1.pp42-47>
- Song, P., & Liu, Y. (2020). An xgboost algorithm for predicting purchasing behaviour on e-commerce platforms. *Tehnicki Vjesnik*, 27(5), 1467–1471. <https://doi.org/10.17559/TV-20200808113807>
- Suci Amaliah, Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, 4(3), 121–127. <https://doi.org/10.35580/variensiunm31>
- Suliman, S. (2021). Implementasi Data Mining Terhadap Prestasi Belajar Mahasiswa Berdasarkan Pergaulan dan Sosial Ekonomi Dengan Algoritma K-Means Clustering. *Simkom*, 6(1), 1–11. <https://doi.org/10.51717/simkom.v6i1.48>
- Wijaya, H., Hostiadi, D. P., & Triandini, E. (2024). Meningkatkan Prediksi Penjualan Retail Xyz Dengan Teknik Optimasi Random Search Pada Model Xgboost. *SPINTER (Prosiding Seminar Hasil Penelitian Informatika Dan Komputer)*, 1(2), 829–833.
- Wijoyo A, Saputra A, Ristanti S, Sya'ban S, Amalia M, & Febriansyah R. (2024). Pembelajaran Machine Learning. *OKTAL (Jurnal Ilmu Komputer Dan Science)*, 3(2), 375–380. <https://journal.mediapublikasi.id/index.php/oktal/article/view/2305>
- Yolanda, F., & Dwiridhotjahjono, J. (2021). Strategi Pemasaran dalam Meningkatkan Penjualan Antena HDF Surabaya. *Value : Jurnal Manajemen Dan Akuntansi*, 16(1), 231–239. <https://doi.org/10.32534/jv.v16i1.1841>
- Yuli Mardi. (2019). Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika. *Jurnal Edik Informatika*, 2(2), 213–219.
- Z, M., Zainuddin, M. Z., & Saputra, M. K. (2022). Strategi Pemasaran Untuk Meningkatkan Penjualan Pada Usaha Umkm Sederhana. *Sultra Journal of Economic and Business*, 3(2), 132–149. <https://doi.org/10.54297/sjeb.vol3.iss2.338>
- Zahra, F., Ridla, M. A., & Azise, N. (2024). Implementasi Data Mining Menggunakan Algoritma Apriori Dalam Menentukan Persediaan Barang (Studi Kasus : Toko Sinar Harahap). *JUSTIFY: Jurnal Sistem Informasi Ibrahimi*, 3(1), 55–65. <https://doi.org/10.35316/justify.v3i1.5335>
- Zhou, K., Li, X., Sun, T., & Luo, B. (2023). Research on Vegetable Sales Prediction and Strategy of Supermarkets Based on XGBoost and Random Forest Modeling. *Proceedings - 2023 International Conference on Applied Physics and Computing, ICAPC 2023*, 373–378. <https://doi.org/10.1109/ICAPC61546.2023.00076>