BAB II
TINJAUAN PUSTAKA

2.1 Landasan teori

2.1.1 Prediksi

Prediksi atau peramalan merupakan proses untuk memperkirakan kebutuhan dimasa yang akan datang yang meliputi kebutuhan dalam ukuran kuantitas, kualitas, waktu dan lokasi yang dibutuhkan dalam rangka memenuhi permintaan barang atau jasa (Kushartini & Almahdy, 2016). Menurut (Alfani W.P.R. et al., 2021) *forecasting* merupakan suatu perhitungan untuk meramalkan keadaan di masa mendatang dengan melakukan pengujian terhadap keadaan dimasa lalu.

Peramalan menjadi dasar bagi perencanaan jangka panjang bagi bisnis, dalam area keuangan peramalan memberikan dasar dalam menentukan anggaran dan pengendalian biaya. Bagi bagian pemasaran, peramalan penjualan dibutuhkan untuk merencanakan produk baru, kompensasi tenaga penjual, dan beberapa keputusan penting lainnya. Selanjutnya, bagian produksi dan operasi menggunakan data-data peramalan untuk perencanaan kapasitas, fasilitas, produksi, penjadwalan, dan pengendalian persedian (Rusdiana & Ramdhani, 2014).

Menurut (Rusdiana & Ramdhani, 2014) berdasarkan waktunya, peramalan dapat dibagi menjadi tiga yaitu:

1. Peramalan jangka pendek yang memberikan hasil peramalan satu tahun mendatang atau kurang.

2. Peramalan jangka menengah untuk meramalkan keadaan satu hingga lima tahun ke depan.
3. Peramalan jangka panjang yang digunakan untuk pengambilan keputusan mengenai perencanaan produk dan perencanaan pasar, pengeluaran biaya perusahaan, studi kelayakan pabrik, anggaran, purchase order, perencanaan tenaga kerja dan perencanaan kapasitas kerja, serta pengambilan keputusan yang berhubungan dengan kejadian lebih dari lima tahun yang akan datang.

2.1.2 *Sales Forecasting*

Menurut (Warnaningtyas & Rohmatiah, 2022) dalam bukunya Penganggaran Perusahan menjelaskan *Sales Forecasting* atau peramalan penjualan merupakan suatu kegiatan untuk memperkirakan produk yang akan terjual pada waktu yang akan datang yang dilakukan berdasarkan data penjualan produk yang pernah terjadi. Fungsi dari peramalan penjualan adalah sebagai berikut.

a. Mengestimasi persediaan dari produk.
b. Membuat strategi penjualan.
c. Mengestimasi permintaan produk pada sales perusahaan.

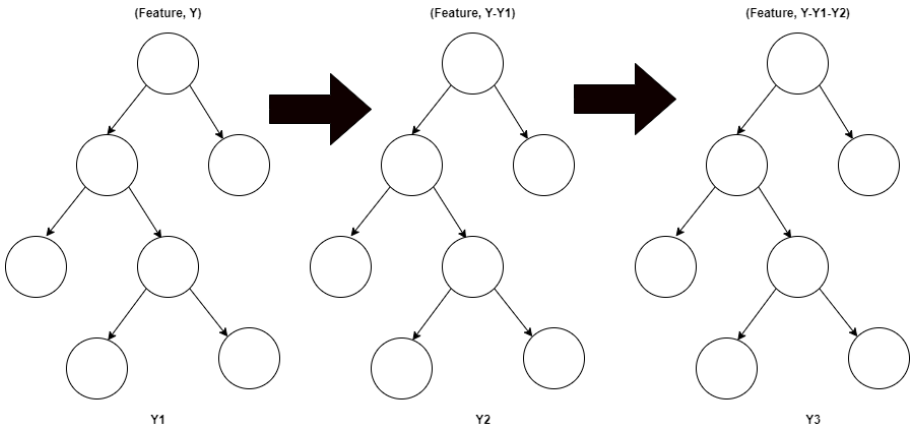
2.1.3 Pembagian Data

Pembagian data merupakan proses untuk membagi dataset yang digunakan menjadi dua kelompok data yaitu data pelatihan dan data pengujian (Mushthofa et al., 2022). Pembagian data atau data *splitting* ini menjadi penting dalam pengolahan data *mining* tertutama dalam pembuatan model (Adinugroho, 2022). Tujuan pembagian data ini untuk memberikan evaluasi terhadap model saat melakukan pengolahan dataset yang digunakan (Eke et al., 2021). Rasio persentasi pembagian

model untuk data pelatihan dan data pengujian sangat beragam (Dankorpho, 2024). Pada penelitian (Tiastama & Budi, 2023) membagi data pelatihan dengan rasio sebesar 70% dan data pengujian sebesar 30%. Penelitian lainnya (Riyyasy, Azfa, Rasikh et al., 2023) membagi data pelatihan dengan rasio sebesar 80% dan data pengujian sebesar 20%.

2.1.4 Algoritma *XGBoost*

Extreme Gradient Boosting atau XGBoost merupakan sebuah model yang dapat menemukan solusi yang optimal untuk berbagai masalah khususnya pada regresi, klasifikasi dan ranking. Konsep dasar dari algoritma ini adalah menyesuaikan parameter pembelajaran secara berulang untuk menurunkan loss function (mekanisme evaluasi atas model). *XGBoost* menggunakan model yang lebih teratur untuk membangun struktur pohon regresi, sehingga dapat memberikan kinerja yang lebih baik dan mampu mengurangi kompleksitas model untuk menghindari *overfitting*. Hasil prediksi akhir dari *XGBoost* adalah penjumlahan hasil prediksi dari setiap pohon regresi (Herni Yulianti et al., 2022). Model *XGBoost* berbentuk pohon keputusan atau pohon regresi (Siringoringo et al., 2021) Proses pembelajaran pohon pertama dapat dilihat pada gambar 2.1 dari data latih (feature, Y) memperoleh hasil estimasi pertama (Y1). Pohon ke dua melakukan proses pembelajaran dari data latih (*feature*, |Y-Y1|), dimana nilai |Y-Y1| merupakan selisih antara label nyata dengan label prediksi pada tahap sebelumnya. Pohon ketiga melakukan proses pembelajaran dari data (*feature*, |Y-Y1-Y2|) dan menghasilkan estimasi Y3. Dengan cara tersebut, nilai error dapat direduksi dengan efektif.



Gambar 2. 1 Pohon *Regresi XGBoost (Siringoringo et al., 2021)*

XGBoost terdiri dari dua pendekatan yaitu pendekatan *gradient descent* fungsinya untuk melakukan optimasi pada *loss function* dan regulasi parameter untuk mencegah *overfriting.* Konsep utama *XGBoost* yaitu untuk meminimalisir fungsi objektif antara *loss function* dan regulasi parameter, berikut rumusnya (Tiastama & Budi, 2023):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \mathrm{l}\left(yi, \hat{y}i^{(t-1)} + (ft(xi)\right) + \Omega(ft)$$

Dimana:

l adalah *loss function.*

Yi adalah nilai aktual.

$\widehat{Y\imath}$ adalah nilai prediksi.

Ft adalah model pohon ke-t

t adalah indeks iterasi selama proses optimasi

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2$$

Dimana:

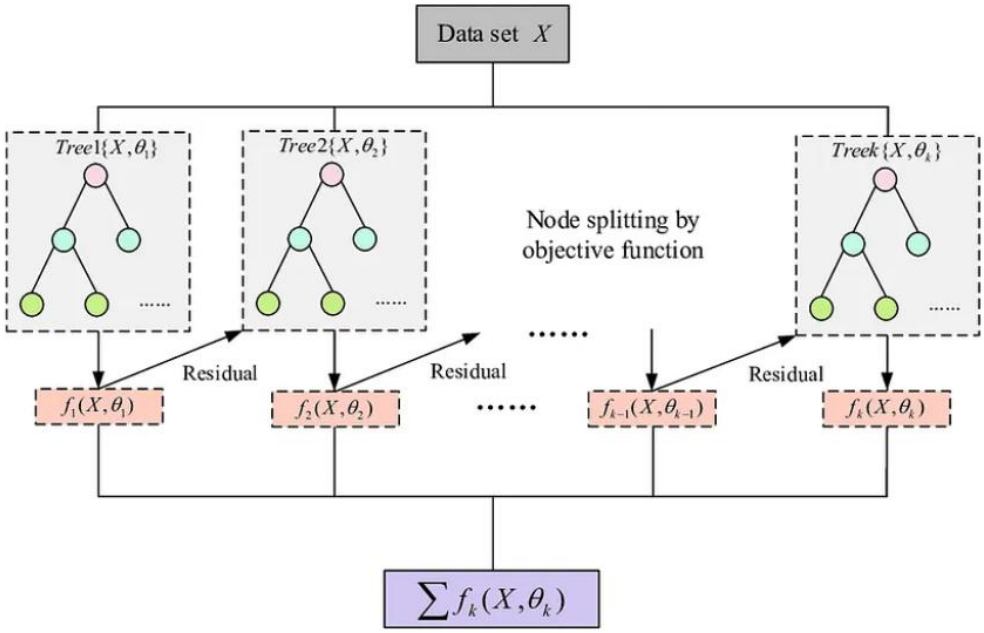
T adalah jumlah *leaf.*

ω adalah bobot *leaf.*

y adalah koefisien nilai default 0.

λ adalah koefisien nilai default 1.

Cara kerja algoritma *XGBoost* sebagai berikut:



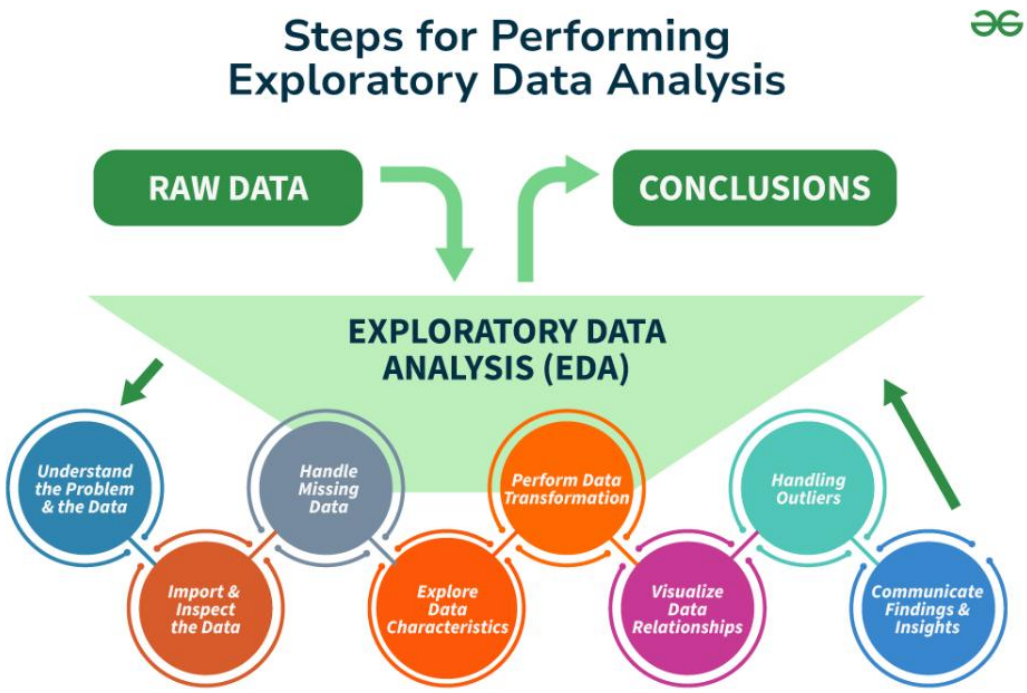
Gambar 2. 2 Cara kerja *XGBoost*
(sumber:https://medium.com/@fraidoonomarzai99/xgboost-regression-in-depth/)

a. Prediksi awal: *XGBoost* dimulai dengan membuat sebuah prediksi sederhana pada saat melakukan data pelatihan, rata-rata menggunakan variabel target.
b. Perhitungan kesalahan atau *residu* yaitu perbedaan antara nilai prediksi dan nilai aktual dalam data pelatihan.
c. Pohon keputusan pertama yaitu pohon pertama pada *ensemble* (gabungan model) yang berfokus pada residu untuk meminimalisir kesalahan.
d. Pohon keputusan berikutnya yaitu masih melanjutkan dari pohon pertama , tetapi disini lebih mencari sisa-sisa kesalahan yang terdapat pada pohon pertama dengan tujuan untuk meningkatkan hasil akurasi.
e. *Loss function* yaitu *XGBoost* memperbaiki fungsi kerugian. Fungsi ini secara matematis menilai seberapa baik model memprediksi sesuai dengan nilai yang sebenarnya. Dengan mengurangi fungsi kerugian, *XGBoost* menjamin bahwa kumpulan model berada di jalur yang tepat untuk menghasilkan prediksi yang akurat.
f. *XGBoost* menambahkan pohon hingga kriteria penghentian tertentu terpenuhi. Kriteria ini bisa berupa jumlah pohon maksimum, peningkatan minimum dalam fungsi kerugian, atau mencapai tingkat akurasi tertentu.

2.1.5 *Exploring Data Analysis (EDA)*

Exploring Data Analysis merupakan tahapan awal untuk melakukan analisis data. *EDA* membantu untuk menggambarkan secara deskriptif mengenai data yang dianalisis yaitu stuktur data, hubungan antara variabel, mendeteksi *outlier*, pengujian asumsi, dan merangkum karakteristik data dengan visualisasi (Wibowo,

2022). Tahapan *EDA* terdiri dari visualisasi grafik *Box Plot*, *Count Plot* dan *Dist Plot*. Pada *Box Plot* dapat menggambarkan beberapa ukuran statistik seperti nilai minimum, kuartil pertama, median kuartil ketiga, nilai maksimum dan ada tidaknya nilai outlier. Pada *Count Plot* digunakan untuk merepresentasikan jumlah observasi variabel kategori untuk setiap grup dengan menggunakan grafik *scatter plot*, diagram batang, diagram ven, *correlation matrix, heatmap*. Pada *Dist Plot* dapat memvisualisasikan distribusi dari data numerik atau menggambarkan variasi sebaran data (Riyyasy, Azfa, Rasikh et al., 2023).



Gambar 2. 3 *Exploring Data Analysis* (sumber: https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/)

2.1.6 *Leave One Out Encoder*

Leave One Out Encoding digunakan untuk mengubah data kategorikal menjadi data numerik dengan cara yang unik dimana metode ini mirip dengan

metode Target *Encoder* perbedaan utamanya yaitu setiap nilai unik tidak langsung digantikan oleh nilai rata-rata dari kolom target, saat menghitung nilai rata-rata, nilai pada perpotongan baris yang sedang diisi dan kolom target tidak dimasukkan ke dalam total jumlah (Bolikulov et al., 2024).

Menurut (Denis Vorotyntsev,2019) cara kerja *Leave One Out Encoding* sebagai berikut.

	category	category_representation	target
0	A	0.666667	0
1	A	0.333333	1
2	A	0.666667	0
3	A	0.333333	1
4	B	0.500000	0
5	B	0.000000	1
6	B	0.500000	0
7	C	0.000000	1
8	C	1.000000	0
9	D	0.500000	1

Gambar 2. 4 Contoh cara kerja *Leave One Out Encoding* (sumber https://towardsdatascience.com/)

Berdasarkan gambar 2.4, contoh cara kerja *Leave One Out Encoding* sebagai berikut.

$$Rata - Rata = \frac{Total\ Target\ A - Target\ ID\ 1}{Jumlah\ A - 1} = \frac{1-0}{4-1} = \frac{1}{3} = 0{,}333333$$

2.1.7 *Mean Squad Eror*

Mean Squad Eror (MSE) merupakan salah satu metode evaluasi metrix dalam regresi untuk mengevaluasi akurasi prediksi. *Mean Squared Error* mengukur

rata-rata perbedaan kuadrat antara nilai prediksi yang diperkirakaan (Nabila et al., 2023). *MSE* digunakan sebagai parameter untuk mengukur akurasi dari model dengan menghitung rata-rata perbedaan kuadrat antara nilai yang diprediksi dan nilai aktual dalam dataset(Arisandi et al., 2023). Semakin rendah nilai *MSE*, semakin baik performa model karena hasil prediksinya lebih mendekati nilai yang benar (Tayyab & Nasim, 2024). Rumus mencari nilai *MSE* (sumber: https://www.geeksforgeeks.org/mean-squared-error/) sebegai berikut.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(Yi - \widehat{Yi})^2$$

Dimana:

N adalah jumlah sampel.

Yi adalah nilai aktual.

$\widehat{Yi}$ adalah nilai prediksi.

2.1.8 *R-squared*

Koefesien determinasi atau R^2 merupakan metode evaluasi untuk mengukur suatu nilai yang menggambarkan seberapa besar berpengaruhnya perubahan dari variabel terikat yang bisa dijelaskan oleh perubahan variasi dari variabel bebas. (Rhamadhani & Saputri, 2023). Menurut (Ghozali 2018, 179) R^2 digunakan untuk mengetahui besarnya nilai variasi yang dihasilkan oleh perubahan dari variabel terikat, hasil uji koefesien determinasi ditentukan oleh nilai R^2 dengan nilai 0 sampai 1. Jika nilai R^2 mendekati 1, artinya model yang digunakan dapat memberikan hasil nilai variabel variasi semua informasi dengan sangat baik.

Sebaliknya jika nilai R^2 mendekati 0, artinya model yang digunakan untuk memberikan hasil nilai variabel variasi informasi terbatas.

Rumus mencari nilai R^2 :(sumber:https://www.geeksforgeeks.org/r-squared/) sebagai berikut.

$$R^2 = 1 - (\frac{SS_{res}}{SS_{tot}})$$

Dimana:

R^2 adalah nilai *R* kuadrat.

SS_{res} adalah jumlah kuadrat residual (selisih antara nilai aktual dan prediksi).

SS_{tot} adalah jumlah kuadrat total (selisih antara nilai aktual dan rata-rata).

2.2 Penelitian terkait

2.2.1 State of The Art

Berdasarkan rumusan masalah dan tujuan penelitian yang telah dibuat, maka dilakukan penyusunan *literature review* dari penelitian sebelumnya yang berkaitan dengan penerapan algortima *machine learning* untuk melakukan peramalan atau prediksi. Beberapa penelitian sebelumnya yang dilakukan adalah, sebagai berikut:

Tabel 2. 1 *State of The Art*

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
1.	(Riando & Afiyati, 2024)	*Implementation of XGBoost Algorithm to Predict The Selling Price of Cayenye Peppers in DKI Jakarta*	*XGboost*	Fluktuasi harga cabai rawit merah di DKI Jakarta dari tahun 2021 hingga 2024 dipengaruhi oleh faktor musiman, ketersediaan pasokan, dan permintaan. Analisis statistik menunjukkan bahwa model XGBoost memberikan kinerja yang sangat baik, dengan nilai R-squared mencapai 99% pada data pelatihan dan 92% pada data uji, yang menunjukkan kemampuan model dalam menjelaskan variabilitas harga secara efektif. Temuan ini	Data yang dibandingkan berasal dari https://www.kaggle.com/

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
				memiliki implikasi penting bagi kebijakanekonomi dan perdagangan di Indonesia, karena stabilitas harga cabai rawit merah berperan penting dalam mengendalikan inflasi dan daya beli masyarakat. Strategi yang disarankan mencakup peningkatan produksi, pengelolaan operasi pasar, dan promosi diversifikasi untuk melindungi kepentingan ekonomi nasional.	
2.	(Tiastama & Budi, 2024)	Perbandingan Random Search dan Algoritma Genetika dalam Penyetelan Hyperparameter XGBoost pada Retail Sales Forecasting	*XGBoost*	Pembagian data pada data latih yang sudah ada menjadi data latih baru dan data validasi dengan rasio 70% dan 30% yaitu berjumlah 675,470 dan 168,868 data. metode random search memiliki performa yang lebih baik dari algoritma genetika dengan nilai RMSE pada proses latih dan proses uji sebesar 0.123 dan 0.122. Sementara itu, nilai RMSE algoritma genetika pada proses latih dan proses uji yaitu 0.333 dan 0.332. Perbedaan nilai	Dataset yang digunakan berasal dari Kaggle yaitu data penjualan pada Rossmann Store, sebuah perusahaan ritel pada bidang farmasi yang mengoperasikan lebih dari 3,000 toko pada 7 negara di Eropa. Dataset Rossmann Store terdiri dari 844,338

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
				yang dihasilkan tidak begitu jauh, sehingga kedua hyperparameter tersebut sama baiknya untuk melakukan peramalan.	data pelatihan dan 168,868 data pengujian.
3.	(Dankorpho , 2024)	*Sales Forecasting for Retail Business using XGBoost Algorithm*	*XGBoost*	Paremeter yang digunakan berdasarkan kategori produk : kecantikan, elektronik, impor fashion, local fashion, dan jam tangan yaitu n_estimator, max depth, eta, dan sub sampel. Selanjutnya pengukuran model menggunakan MAE dan RMSE untuk dibandingkan antara model XGBoost dan metode original masin g-masing kategori produk. Pada kategori **Beauty**, model XGBoost mencapai MAE sebesar 0,49 dan RMSE sebesar 0,61, yang lebih rendah dibandingkan Metode Asli dengan MAE sebesar 1,15 dan RMSE sebesar 1,91. Pada kategori **Electronic**, model XGBoost memiliki MAE sebesar 1,52 dan RMSE sebesar 2,44,	Dataset berasal data penjualan gabungan dari 805 toko dengan tahun transaksi 2013 – 2023.

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
				sedangkan Metode Asli menunjukkan MAE sebesar 1,71 dan RMSE sebesar 2,77. Pada kategori **Imported Fashion**, model XGBoost mencatat MAE sebesar 0,73 dan RMSE sebesar 0,90, dibandingkan dengan Metode Asli yang memiliki MAE sebesar 0,79 dan RMSE sebesar 1,06. Pada kategori **Owned Brand Fashion**, XGBoost memperoleh MAE sebesar 0,46 dan RMSE sebesar 0,63, sementara Metode Asli mencatat MAE sebesar 0,86 dan RMSE sebesar 1,43. Pada kategori **Watches**, MAE pada model XGBoost adalah 1,39 dan RMSE 1,79, sedangkan Metode Asli memiliki MAE sebesar 2,01 dan RMSE sebesar 2,54. Hasilnya rata-rata MAE pada Metode Asli adalah 1,30, sedangkan model XGBoost mencapai nilai yang jauh lebih rendah yaitu 0,92, menunjukkan peningkatan akurasi sebesar 29.23%. Demikian pula	

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
				untuk RMSE, Metode Asli menghasilkan skor rata-rata 1,94, sementara model XGBoost menunjukkan nilai yang lebih rendah yaitu 1,27, yang mencerminkan peningkatan akurasi sebesar 34,54%	
4.	(Tayyab & Nasim, 2024)	*Walmart Sales Prediction by Using Machine Learning Algorithms*	*XGBoost, K-nearest neighbor, Extra Tree Regression, Random Forest*	Dari keempat model yang digunakan, hasil akurasi yang diperoleh yaitu pada model K-NN sebesar 93.81%, model Extra Tree Regressor sebesar 97.51%, model Random Forest sebesar 97.27%, dan model XGBoost sebesar 98.25%. Peramalan penjualan Walmart berdasarkan algoritma yang digunakan semuanya sudah memiliki kinerja yang baik dengan hasil akurasi rata-rata di atas 90% dan model yang terbaik adalah algoritma XGBoost.	Dataset yang digunakan berasal dari Kaggle terdiri dari 6414 data dan 8 fitur.

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
5.	(Gunawan et al., 2023)	*Prediction of Cross-Platform and Native Apps Technology Opportunities for Beginner Developers Using C 4.5 and Naïve Bayes Algorithms*	*C 4.5* *Naïve Bayes*	Data bahasa pemrograman : Dart ,JavaScript , Kotlin , C , C++ , C #, Java, Swift,Objective-c di evaluasi untuk melihat nilai *Precision, recall, and accuracy* dengan menggunakan algoritma *Naïve Bayes* dan *C 4.5.* Hasil uji dengan 60% data *training* dan 40% data *testing* performa algoritma *C 4.5* lebih cepat 1% dari algoritma *Naïve Bayes* dengan akuras rata-ratai 97% dan 96%. untuk kotlin, C, C++, Java, Swift, dan Objective-c kedua algoritma memiliki nilai akurasi yang sama yaitu 100%. C, C++, Java, Swift, dan Objective-c kedua algoritma memiliki nilai akurasi yang sama yaitu 100%. C, C++, Java, Swift, dan Objective-c kedua algoritma memiliki nilai akurasi yang sama yaitu 100%.	Data yang dikumpulkan terdiri dari data bahasa pemrograman yaitu 3375 data dari tahun 2011 sd 2021 yang diambil dari situs https://www.kaggle.com/
6.	(Fieri et al., 2023)	*Introversion-Extraversion Prediction Using*	*Decision Tree, K-Nearest Neighbor, Random Forest,*	Hasil evaluasi pada dataset asli didapatkan hasil untuk algoritma *Decision Tree* hasil *accuracy*:	Data yang digunakan berasal dari dataset Open -

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
		Machine Learning	*SVM*	0.624, *precision*: 0.630, *Recall*: 0.623, dan F1 *score*: 0.626. Algoritma *K-Nearest Neighbor* hasil *accuracy*: 0.698, *precision*: 0.655, *Recall*: 0.698, dan F1 *score*: 0.672. Algoritma *Random Forest* hasil *accuracy*: 0.726, *precision*: 0.704, *Recall*: 0.725, dan F1 *score*: 0.710. Algoritma *SVM linear* hasil *accuracy*: 0.735, *precision*: 0.711, *Recall*: 0.734, dan F1 *score*: 0.716. Dan hasil terbaik adalah algoritma *SVM Liniear*, Hasil evaluasi pada dataset SMOTE didapatkan hasil akurasi terbaik oleh algoritma *Random Forest* dengan nilai *accuracy*: 0.955, *precision*: 0.955, *Recall*: 0.955, dan F1 *score*: 0.955. Hasil evaluasi pada dataset SMOTE-ENN didapatkan hasil	Source Psychometrics.

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
				akurasi terbaik oleh algoritma *Random Forest* dengan nilai *accuracy*: 0.710, *precision*: 0.801, *Recall*: 0.710, dan F1 *score*: 0.721. Hasil evaluasi pada dataset ADASYN didapatkan hasil akurasi terbaik oleh algoritma *Random Forest* dengan nilai *accuracy*: 0.953, *precision*: 0.953, *Recall*: 0.953, dan F1 *score*: 0.953.	
7.	(Li, 2023)	*A Sales Prediction Method Based on XGBoost Algorithm Model*	*XGboost*	Pembagian model untuk data pelatihan sebesar 0,8 / 80% dan data pengujian sebesar 0,2 / 20%. Selanjutnya dilakukan pengukuran *R-squared* pada model dengan tujuan menghasilkan model yang mampu menjelaskan lebih banyak variasi nilai. Nilai yang diperoleh model ini sebesar 0,92 artinya model yang diigunakan sudah cukup baik. Hasilnya dengan melatih model XGBoost, tren dan volume penjualan di masa depan dapat diprediksi dengan lebih akurat.	Dataset penjualan dari tahun 2013 – 2017 berasal dari Kaggle.

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
				Penggunaan model XGBoost dapat memaksimalkan keunggulan model prediksi, yang membantu para manajer bisnis dalam pengambilan keputusan dan membantu perusahaan mengembangkan strategi pemasaran jangka panjang, yang memiliki nilai bisnis penting bagi perusahaan toko.	
8.	(Sim & Wei, 2023)	*XGBoost Regression Algorithms for Efficient Predictions on Inventory Sales and Management*	*XGBoost*	Hasil pengukuran model dengan *R*-squared sebesar 0,5199 artinya model sudah cukup lumayan baik dalam melakukan peramalan pada dataset yang digunakan. Secara teori dalam penelitian ini model prediksi xgboost regression cukup akurat dalam pengujian, tetapi sampel data uji yang ditetapkan dalam proyek ini tidak memiliki kompleksitas penerapan untuk menunjukkan beberapa hasil yang sangat meyakinkan	Dataset yang digunakan berasal dari Kaggle.
9.	(Riyyasy, Azfa, asikh et al., 2023)	Penerapan Algoritma *Machine*	*Random Forest, Logistic Regression, SVM,*	Hasil akurasi Model Logistic Regression sebesar 72.3%, model SVM sebesar 79.7%. Model terbaik	Dataset berasal dari Hugging Face dan mempunyai

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
		Learning Untuk Memprediksi *Term* Deposit Nasabah Perbankan	*XGBoost*	yaitu random forest dan xgboost, dengan akurasi sebesar 91.7%. Hal tersebut tidak menunjukkan adanya overfitting dikarenakan tidak banyaknya perbedaan nilai pada classification report dan hasil nested k-fold. Dengan memahami pola perilaku nasabah terkait keputusan untuk berlangganan term deposit, bank dapat mengoptimalkan strategi pemasaran mereka. Dengan menggunakan model ini, bank dapat secara efektif menargetkan segmen nasabah yang lebih cenderung untuk melakukan investasi dalam term deposit, meningkatkan efisiensi dan mengurangi biaya pemasaran yang tidak perlu.	variabel independent sebanyak 16 kolom yang terdiri age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, dan poutcome.
10.	(Herni Yulianti et al., 2022)	Penerapan Metode *Extreme Gradient Boosting (XGBOOST)* pada Klasifikasi	*XGboost*	Metode XGBoost dengan parameter yang default pada dataset nasabah pengguna kartu kredit menghasilkan model yang dikatakan cukup baik yaitu akurasi model sebesar 80,02% , untuk presisi sebesar 85,32%, recall sebesar 94,86% dan dapat dikategorikan sebagai good	Dataset berasal dari situs *UCI Machine learning Repository*.

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
		Nasabah Kartu Kredit		classification. Untuk percobaan kedua menggunakan teknik optimasi yaitu proses hyperparameter tuning menggunakan 7 hyperparameter dengan memvalidasi data, maka didapatkan hasil hyperparameter tuning yang diperoleh akurasi model sebesar 83,42%, presisi sebesar 85,36%, recall sebesar 95,28%.	
11.	(Dewi et al., 2022)	Penerapan Data Mining untuk Prediksi Penjualan Produk Elektronik Terlaris Menggunakan Metode *K-Nearest Neighbor*	*K-Nearest Neighbor*	Hasilnya adalah data produk yang digunakan sebanyak 25 jenis dengan mengambil dari data penjualan selama satu tahun terakhir. Kemudian dilakukan proses transformasi data dari data produk, selanjutnya dilakukan proses perhitungan jarak tetangga menggunakan *Euclidean Distance.* Lalu melakukan uji coba dengan nilai k=1, k=3, k=5, k=7, untuk menentukan prediksi mengenai produk mana yang paling laris. Dan didapatkan hasilnya yaitu produk yang paling laris dengan nilai akurasi 0,0859 adalah produk seres.	Data berasal dari penjualan produk selama satu tahun.

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
12.	(Wahyudi & Silfia, 2022)	*Implementation of Data Mining Using K-Means Clustering Method to Determine Sales Strategy In S&R Baby Store*	*K-means*	Data yang dipilih untuk cluster awal adalah data 292, 600, 955, 489. Jarak centroid data ke-1 pada *cluster* 1 (C1): 181.492,5, untuk C2: 295.013, untuk C3: 85.250, dan C4: 25.000. Selanjutnya dilakukan tahap evaluasi untuk mengecek apakah dari tiap-tiap operator yang ada pada Rapid Miner dengan tujuan agar operator saling terhubung dan membuat duplikat menjadi empat operator sehingga membentuk empat *cluster*. Nilai *Indeks Davies-Bouldin Setiap* Model *Cluster*: *Cluster 1:* -0,700 *Cluster 2:* -0,692 *Cluster 3:* -0,560 *Cluster 4:* -0,586 Penerapan metode *K-Means* pada S&R Baby Store yaitu dengan mengelompokkan data transaksi penjualan. Kemudian pilih 4 cluster secara acak sebagai centroid awal. Setelah data pada masing-masing cluster tidak mengalami perubahan	Data transaksi penjualan bulan Januari sampai Maret 2022 yang berbentuk dokumen excel sebanyak 1351 data dan 10 atribut.

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
				maka dapat diketahui hasil akhir terdapat 362 produk yang laris manis, 944 produk yang laris manis, yang laris manis, terdiri dari 2 produk dan 43 produk yang tidak laku.	
13.	(Riza, 2022)	Analisis dan Prediksi Data Penjualan Menggunakan Machine Learning dengan Pendekatan Ilmu Data	*Random Forest (RF), LightGBM (LG) dan XGBoost*	Penelitian ini menyajikan analis data penjualan dan dasar-dasar supervised machine learning untuk tugas prediksi penjualan di pusat perbelanjaan Big Mart di lokasi yang berbeda. Berdasarkan hasil evalusi tingkat kesalahan deteksi berdasarkan MAE dan RMSE, algoritma XGBoost dan LightGBM manghasilkan tingkat kesalahan paling rendah dibandingkan algoritma lainnya. Sedangkan hasil evaluasi skor R^2 kedua algoritma ini mencapai nilai tertinggi 0.61 (61%) XGBoost dan LightGBM 0.60 (60%). Selain itu juga, hasil prediksi menunjukkan kegembiraan corr di antara atribut yang berbeda dipertimbangkan dan bagaimana lokasi tertentu dari ukuran	dataset Big Mart Sales Data tahun 2013.

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
				menengah mencatat penjualan tertinggi, menunjukkan bahwa lokasi belanja lainnya harus mengikuti pola yang sama untuk peningkatan penjualan.	
14.	(Alfani W.P.R. et al., 2021)	Prediksi Penjualan Produk Unilever Menggunakan Metode *K-Nearest Neighbor*	*K-Nearest Neighbor*	Dalam penelitian ini dilakukan proses perhitungan jarak tetangga menggunakan *Euclidean Distance,* dilakukan pengujian data ke 1 dengan perhitungan data latih 50 dan 10 data uji dengan nilai k=10 didapatkan nilai akurasi 40%, pengujian ke 2 dengan perhitungan data latih 60 dan 20 data uji dengan nilai k=30 didapatkan nilai akurasi 86,66%. Penentuan dalam nilai k yang tepat dapat berpengaruh terhadap hasil tingkat akurasi.	Data produk berasal dari data pada tahun 2017, 2018, 2019 dengan penjualan perminggu.
15.	(Azhar et al., 2021)	Prediksi pembatalan pemesanan hotel menggunakan optimalisasi hiperparameter pada algoritma	*Random Forest*	Hasil peneliti ini yaitu optimalisasi hiperparameter dapat meningkatkan performa model RF tradisional yang digunakan sehingga meraih nilai akurasi tertinggi sebesar 0,87. Pengukuran performa model lain menggunakan matriks AUC dan sensitivitas juga telah menunjukkan.	data berasal dari dua dataset pemesanan hotel yang berasal dari City Hotel (H1) dan Resort Hotel (H2). Masingmasing

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
		Random Forest		nilai evaluasi tertinggi sebesar 0,87 dan 0,88. Selain itu, prapemrosesan yang digunakan dapat memberikan data latih masukan yang jauh lebih baik dan lebih spesifik lagi untuk kasus prediksi pembatalan pemesanan hotel oleh pelanggan.Model OHRF yang diusulkan juga telah terbukti memiliki kinerja yang lebih baik daripada model lainnya. Hal tersebut ditunjukkan oleh kecenderungan evaluasi model OHRF yang lebih tinggi daripada model lain yang didefinisikan.	dataset mewakili pemesanan hotel yang tiba antara 1 Juli 2015 hingga 31 Agustus 2017 termasuk data pemesanan berhasil dan dibatalkan.
16.	(Yolanda & Fahmi, 2021)	Penerapan Data Mining Untuk Prediksi Penjualan Produk Roti Terlaris Pada PT.Nippon Indosari Corp Indo Tbk	*K-Nearest Neighbor*	Dalam penelitian ini dilakukan proses data *training* dan data *testing* dari data penjualan tiga bulan terakhir. Kemudian peneliti menggunakan beberapa atribut yang dibutuhkan sebagai dasar untuk melakukan prediksi dengan algoritma *K-Nearest Neighbor* yaitu kuantitas produk dan kuantitas terjual. Setelah itu dilakukan proses	Data berasal dari penjualan pada tiga bulan terakhir yaitu periode bulan januari sampai maret.

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
		Menggunakan Metode *K-Nearest Neighbor*		Perhitungan jarak tetangga dengan menggunakan *Euclidean Distance.* Dan nilai k yang digunakan adalah k=5, dengan melakukan perhitungan pada 20 data jenis roti yang terjual. Dalam penentuan tingkat akurasi pada penelitian ini menggunakan kategori naik dan turun, diperoleh suatu hasil dimana roti yang paling diminati adalah roti dengan kriteria naik dimana dengan nilai bobot lebih dari >1,5 yaitu jenis roti tawar, roti sandwich dan roti manis.	
17.	(Siringorin go et al., 2021)	Segmentasi dan Peramalan Pasar Retail Menggunakan *Xgboost* dan *Principal Component Analysis*	*XGBoost*	Segmentasi dan peramalan terhadap penjualan sektor ritel online menggunakan extreme gradient boost (XGBoost). Metode principal component analysis (PCA) diterapkan untuk mereduksi dimensi dataset. Menggunakan kriteria silhoulette, KMC dapat menentukan target data dengan pendekatan klaster yang lebih terpisah dengan baik. Berdasarkan hasil penelitian di	Dataset berasal dari repositori UCI yaitu online retail dataset dengan 525.461 item data dan 8 atribut.

No	Penulis dan Tahun	Judul	Algoritma	Hasil	Dataset
				atas diperoleh kesimpulan bahwa XGBoost dapat melakukan klasifikasi data retail dengan baik	
18.	(Dairu & Shilong, 2021)	*Machine Learning Model for Sales Forecasting by Using XGBoost*	*XGBoost*	Hasil experiment pengukuran dengan *RMSSE* terhadap model *XGBoost* menghasilkan nilai terendah yaitu 0.655, untuk model Regresi Linier Klasik yaitu 0.783, 19,5% lebih tinggi dari model *XGBoost*, dan model Regresi Ridge yaitu 0.744.	Data penjualan selama 1913 hari di ritel Walmart yang diambil dari https://www.kaggle.com/.

2.2.2 Matriks Penelitian

Tabel 2.2 merupakan matriks penelitian yang berisi penelitian yang berfokus untuk menyelesaikan masalah dalam melakukan peramalan atau prediksi. Selain itu, matriks ini dapat memberikan informasi tentang perbedaan penelitian yang akan dilakukan dan penelitian terdahulu.

Tabel 2. 2 Matriks Penelitian

No.	Judul	Penulis dan Tahun	Ruang Lingkup													
			Penerapan Algoritma										Parameter Uji			
			Naïve Bayes	***Random Forest***	***Decision Tree***	***SVM***	***XGBoost***	***XGBoost Regreesion Linear***	***LightGBM***	***Extra Tree Regression***	***K-NN***	***K-Means***	***C 4.5***	***Accuracy***	***Recall***	***Precision***	**F-Measure**
1	*Implementation of XGBoost Algorithm to Predict The Selling Price of Cayenye Peppers in DKI Jakarta*	(Riando & Afiyati, 2024)	-	-	-	-	✓	-	-	-	-	-	-	✓	✓	✓	✓
2	Perbandingan Random Search dan Algoritma	(Tiastama & Budi, 2024)	-	-	-	-	✓	-	-	-	-	-	-	✓	-	-	-

No.	Judul	Penulis dan Tahun	Ruang Lingkup													
			Penerapan Algoritma										Parameter Uji			
			Naïve Bayes	***Random Forest***	***Decision Tree***	***SVM***	***XGBoost***	***XGBoost Regreesion Linear***	***LightGBM***	***Extra Tree Regression***	***K-NN***	***K-Means***	***C 4.5***	***Accuracy***	***Recall***	***Precision***	**F-Measure**
	Genetika dalam Penyetelan Hyperparameter XGBoost pada Retail Sales Forecasting															
3	*Sales Forecasting for Retail Business using XGBoost Algorithm*	(Dankorpho, 2024)	-	-	-	-	✓	-	-	-	-	-	-	✓	-	-	-
4	*Walmart Sales Prediction by Using Machine Learning Algorithms*	(Tayyab & Nasim, 2024)	-	✓	-	-	✓	-	-	✓	✓	-	-	✓	-	-	-
5	*Prediction of Cross-Platform and Native Apps Technology Opportunities for Beginner Developers*	(Gunawan et al., 2023)	✓	-	-	-	-	-	-	-	-	-	✓	✓	✓	✓	✓

No.	Judul	Penulis dan Tahun	Ruang Lingkup														
			Penerapan Algoritma											Parameter Uji			
			Naïve Bayes	*Random Forest*	*Decision Tree*	*SVM*	*XGBoost*	*XGBoost Regreesion Linear*	*LightGBM*	*Extra Tree Regression*	*K-NN*	*K-Means*	*C 4.5*	*Accuracy*	*Recall*	*Precision*	F-Measure
	Using C 4.5 and Naïve Bayes Algorithms																
6	*Introversion-Extraversion Prediction Using Machine Learning*	(Fieri et al., 2023)	-	✓	✓	✓	-	-	-	-	✓	✓	-	✓	✓	✓	✓
7	*A Sales Prediction Method Based on XGBoost Algorithm Model*	(Li, 2023)	-	-	-	-	✓	-	-	-	-	-	-	✓	-	-	-
8	*XGBoost Regression Algorithms for Efficient Predictions on Inventory Sales and Management*	(Sim & Wei, 2023)	-	-	-	-	✓	-	-	-	-	-	-	✓	-	-	-

No.	Judul	Penulis dan Tahun	Ruang Lingkup													
			Penerapan Algoritma											Parameter Uji			
			Naïve Bayes	***Random Forest***	***Decision Tree***	***SVM***	***XGBoost***	***XGBoost Regreesion Linear***	***LightGBM***	***Extra Tree Regression***	***K-NN***	***K-Means***	***C 4.5***	***Accuracy***	***Recall***	***Precision***	**F-Measure**
9	Penerapan Algoritma Machine Learning Untuk Memprediksi Term Deposit Nasabah Perbankan	(Riyyasy, Azfa, Rasikh et al., 2023)	-	✓	-	✓	✓	-	-	-	-	-	-	✓	✓	✓	✓
10	Penerapan Metode *Extreme Gradient Boosting (XGBOOST)* pada Klasifikasi Nasabah Kartu Kredit	(Herni Yulianti et al., 2022)	-	-	-	-	✓	-	-	-	-	-	-	✓	✓	✓	✓
11	Penerapan Data Mining untuk Prediksi Penjualan Produk Elektronik Terlaris Menggunakan Metode *K-Nearest Neighbor*	(Dewi et al., 2022)	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-

No.	Judul	Penulis dan Tahun	Ruang Lingkup													
			Penerapan Algoritma											Parameter Uji			
			Naïve Bayes	*Random Forest*	*Decision Tree*	*SVM*	*XGBoost*	*XGBoost Regreesion Linear*	*LightGBM*	*Extra Tree Regression*	*K-NN*	*K-Means*	*C 4.5*	*Accuracy*	*Recall*	*Precision*	F-Measure
12	*Implementation of Data Mining Using K-Means Clustering Method to Determine Sales Strategy In S&R Baby Store*	(Wahyudi & Silfia, 2022)	-	-	-	-	-	-	-	-	-	✓	-	✓	-	-	-
13	Analisis dan Prediksi Data Penjualan Menggunakan Machine Learning dengan Pendekatan Ilmu Data	(Riza, 2022)	-	✓	-	-	✓	-	✓	-	-	-	-	✓	-	-	-
14.	Prediksi Penjualan Produk Unilever Menggunakan Metode *K-Nearest Neighbor*	(Alfani W.P.R. et al., 2021)	-	-	-	-	-	-	✓	-	✓	-	-	✓	-	-	-

No.	Judul	Penulis dan Tahun	Ruang Lingkup													
			Penerapan Algoritma										Parameter Uji			
			Naïve Bayes	*Random Forest*	*Decision Tree*	*SVM*	*XGBoost*	*XGBoost Regreesion Linear*	*LightGBM*	*Extra Tree Regression*	*K-NN*	*K-Means*	*C 4.5*	*Accuracy*	*Recall*	*Precision*	F-Measure
15.	Prediksi pembatalan pemesanan hotel menggunakan optimalisasi hiperparameter pada algoritma *Random Forest*	(Azhar et al., 2021)	-	-	-	-	-	-	-	-	✓	-	-	✓	✓	✓	✓
16.	Penerapan Data Mining Untuk Prediksi Penjualan Produk Roti Terlaris Pada PT.Nippon Indosari Corpindo Tbk Menggunakan Metode *K-Nearest Neighbor*	(Yolanda & Fahmi, 2021)	-	-	-	-	-	-	-	-	✓	-	-	✓	-	-	-

No.	Judul	Penulis dan Tahun	Ruang Lingkup													
			Penerapan Algoritma											Parameter Uji			
			Naïve Bayes	***Random Forest***	***Decision Tree***	***SVM***	***XGBoost***	***XGBoost Regreesion Linear***	***LightGBM***	***Extra Tree Regression***	***K-NN***	***K-Means***	***C 4.5***	***Accuracy***	***Recall***	***Precision***	**F-Measure**
17.	Segmentasi dan Peramalan Pasar Retail Menggunakan *Xgboost* dan *Principal Component Analysis*	(Siringoringo et al., 2021)	-	✓	-	-	✓	-	-	-	-	-	-	✓	-	-	-
18.	*Machine Learning Model for Sales Forecasting by Using XGBoost*	(Dairu & Shilong, 2021)	-	-	-	-	✓	-	-	-	-	-	-	✓	-	-	-
19.	Prediksi Penjualan Produk Pada Toko Src Menggunakan Algoritma *Xgboost Regression*	(Gustiyandi, Zhehan, 2024)	-	-	-	-	✓	✓	-	-	-	-	-	✓	-	-	-