BAB II

TINJAUAN LITERATUR

2.1 Penelitian Terkait

Sebelum melakukan penelitian, penulis terlebih dahulu melakukan tinjauan pustaka terhadap beberapa penelitian terdahulu terkait prediksi yang diolah menggunakan metode *random forest regresor*:

1. Apriliana Sari (2025), Prediksi Kebutuhan Stok Barang Menggunakan Algoritma *Random Forest* Untuk Meningkatkan Efisiensi Penjualan. Penelitian ini bertujuan untuk memprediksi kebutuhan stok barang menggunakan algoritma *Random Forest* berdasarkan data penjualan sebelumnya guna meningkatkan efisiensi pengelolaan stok dan mendukung peningkatan penjualan. Proses penelitian dilakukan pada platform *Google Colaboratory* melalui tahapan pengumpulan data, preprocessing, pelatihan model, dan evaluasi performa menggunakan metrik *Root Mean Squared Error* (RMSE), *Mean Absolute Percentage Error* (MAPE), dan koefisien determinasi (R^2). Hasil evaluasi menunjukkan bahwa model memiliki performa sangat baik, dengan nilai RMSE sebesar 8.36 pada data latih dan 10.53 pada data uji, MAPE masing-masing sebesar 2.68% dan 7.50%, serta R^2 mencapai 99.00% (latih) dan 98.15% (uji). Model ini terbukti mampu memberikan prediksi yang akurat dalam mengelompokkan kebutuhan stok barang, sehingga dapat digunakan sebagai dasar pengambilan keputusan pemesanan ulang

yang lebih tepat waktu dan sesuai permintaan aktual serta diharapkan dapat membantu Toko Sumini mengurangi risiko kekurangan atau kelebihan stok, meningkatkan kepuasan pelanggan, dan menunjang pertumbuhan penjualan yang lebih konsisten.

2. Muhammad Syahrul Efendi (2024), Penerapan Algoritma *Random Forest* Untuk Prediksi Penjualan Dan Sistem Persediaan Produk. Penelitian ini bertujuan untuk mengoptimalkan prediksi penjualan dan pengelolaan persediaan produk Bolen Crispy di Desa Pekalongan dengan menggunakan algoritma *Random Forest.* Pengusaha Bolen Crispy menghadapi tantangan berupa fluktuasi penjualan yang besar dan kesulitan dalam mengelola persediaan secara efisien. Kesalahan dalam memperkirakan jumlah penjualan dapat menyebabkan kekurangan atau kelebihan stok, yang berdampak pada meningkatnya biaya operasional serta menurunkan kepuasan pelanggan. Masalah *overstocking* dan *understocking* ini berpotensi menimbulkan kerugian finansial. Algoritma *Random Forest* dipilih karena kemampuannya dalam menangani data yang kompleks dan menghasilkan prediksi yang lebih akurat. Dengan memanfaatkan data penjualan historis, algoritma ini diterapkan untuk memprediksi permintaan produk. Pengujian dilakukan menggunakan data penjualan selama satu tahun, dengan pembagian 80% untuk pelatihan dan 20% untuk pengujian. Hasil awal menunjukkan bahwa penggunaan algoritma *Random Forest* mampu meningkatkan akurasi prediksi penjualan hingga 85%, dibandingkan metode konvensional.

3. Farah Nur Farida (2025), Penerapan Model Prediksi Penjualan Pada Usaha Rumah Makan Menggunakan Algoritma *Random Forest*. Penelitian ini bertujuan untuk membangun model prediksi penjualan di usaha rumah makan menggunakan algoritma *Random Forest*, yang dikenal karena kemampuannya dalam menangani data kompleks dan menghasilkan prediksi yang akurat. Penelitian ini menggunakan pendekatan kuantitatif dengan teknik *machine learning,* di mana data penjualan dikumpulkan selama sepuluh bulan (Januari hingga Oktober 2024). Setelah data dikumpulkan, dilakukan tahap preprocessing untuk membersihkan dan mempersiapkan data agar dapat digunakan oleh model. Algoritma *Random Forest* kemudian diterapkan untuk membangun model prediksi penjualan. Hasil evaluasi menunjukkan bahwa model ini mencapai akurasi 97% pada data uji, yang menandakan bahwa model ini sangat efektif dalam mengklasifikasikan kategori penjualan dengan tingkat kesalahan yang minimal.

2.2 Prediksi

Prediksi adalah suatu proses memperkirakan secara sistematis tentang sesuatu yang paling mungkin terjadi di masa akan datang berdasarkan informasi di masa lalu dan sekarang yang dimiliki, agar kesalahannya (selisih antara sesuatu yang terjadi dengan hasil perkiraan) dapat diperkecil (Aletheia, 2019). Peramalan juga merupakan pemikiran terhadap suatu besaran, misalnya ada sebuah permintaan terhadap suatu produk atau beberapa produk pada periode yang akan datang (Hudaningsih, 2020).

Prediksi merupakan suatu jenis kegiatan untuk memperkirakan sesuatu yang akan terjadi dimasa depan. Perasalahan dalam proses menentukan keputusan adalah kendala yang dihadapi, oleh sebab itu prediksi termasuk dari permasalah yang harus dihadapi, dikarenakan prediksi memiliki keterkaitan dengan pengambilan keputusan (Asohi, 2020)

Dari definisi diatas dapat disimpulkan bahwa prediksi yaitu suatu kegiatan yang menggambarkan secara rinci mengenai hasil yang dibuat pada masa mendatang menggunakan data – data dimasa lalu.

2.3 Konsep Permintaan dan Perilaku Konsumen

Teori permintaan adalah teori ekonomi yang menyatakan bahwa harga dipengaruhi oleh permintaan. Oleh karena itu, teori tersebut berasumsi bahwa ketika permintaan di pasar naik, maka harga barang pun akan ikut naik. Tetapi, jika permintaan turun, maka harga pun akan ikut turun. Turunnya permintaan sendiri awalnya disebabkan oleh naiknya, atau terlalu tingginya harga di pasar, sehingga masyarakat berfikir ulang untuk spending money. Maka, ketika masyarakat tidak berminat untuk membeli barang mereka (produsen), maka produsen akan menurunkan harganya, agar masyarakat kembali dapat mengkonsumsi barang yang mereka produksi.

Berdasarkan ciri hubungan antara permintaan dan harga dapat dibuat grafik kurva permintaan. Permintaan adalah kebutuhan masyarakat / individu terhadap suatu jenis barang tergantung kepada faktor-faktor sebagai berikut: harga barang itu sendiri, harga barang lain, pendapatan konsumen, cita

masyarakat/selera, jumlah penduduk, musim / iklim, prediksi masa yang akan datang.

Perilaku konsumen merupakan salah satu aspek penting dalam pemasaran, karena melalui pemahaman tentang peilaku konsumen, pemasar dapat memahami harapan pelanggan tentang produknya, sehingga peilaku pelanggan sebagai fokus bisnis saat ini erat hubungannya dengan permasalahan manusia. Perilaku konsumen adalah tindakan yang langsung terlibat dalam mendapatkan mengkonsumsi, dan menghabiskan produk jasa termasuk proses keputusan yang mendahului dan menyusuli tindakan ini (Setiadi, 2019)

Menurut Schiffman dan Kanuk dalam Tjiptono (2018), perilaku konsumen adalah perilaku yang ditunjukkan oleh konsumen dalam mencari, membeli, menggunakan, mengevaluasi dan menghentikan konsumsi produk, jasa dan gagasan. Sehingga dapat dikatakan dan menghabiskan produk atau jasa termasuk proses pengambilan keputusan. Adapun faktor-faktor yang mempengaruhi perilaku konsumen tersebut adalah faktor sosial, budaya, pribadi dan kekuatan psikologis, dimana faktor budaya dikatakan mempunyai pengaruh yang paling luas dan dalam.

2.4 Prediksi Permintaan Konsumen (*Demand Forecasting*)

Prediksi permintaan adalah proses memperkirakan jumlah produk yang akan dibutuhkan konsumen pada periode tertentu berdasarkan data historis dan variabel terkait. Dalam konteks ritel buah, permintaan bersifat *volatile* karena dipengaruhi oleh faktor musiman, cuaca, tren konsumsi, hari libur,

promosi, dan persediaan. Menurut Chopra & Meindl (2021), *demand forecasting* sangat penting untuk mengurangi ketidakpastian, meningkatkan efisiensi rantai pasok, serta mengoptimalkan keputusan pembelian dan manajemen stok.

Permintaan pada produk buah termasuk kategori *perishable goods*, sehingga kesalahan prediksi dapat menyebabkan overstock (buah rusak, kerugian) atau understock (kehilangan peluang penjualan). Karena itu, pendekatan prediksi dengan model *machine learning* semakin banyak digunakan di sektor ritel buah dan sayur karena mampu menangkap pola kompleks dan non-linear

2.5 Data Mining

Data mining sering *disebut Knowledge Discovery in Database* (KDD), yaitu kegiatan yang meliputi pengumpulan, pemakaian data histori untuk menemukan keteraturan, pola dan hubungan dalam set data berukuran besar. Data mining merupakan teknologi baru yang sangat berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari gudang data mereka. Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data *warehouse*, atau penyimpanan informasi lainnya (Lumbantoruan, 2019)

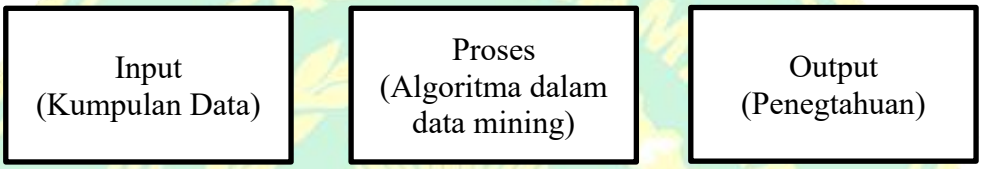
Data mining adalah alat yang digunakan untuk mengakses data dalam jumlah besar. Secara spesifik, data mining adalah alat atau aplikasi yang bergunakan untuk analisis statistika pada data. Data mining dapat digunakan

untuk menggali informasi yang tidak diketahui di dalam database oleh penggunanya (Marisa et al., 2021).

Data mining terdiri dari 2 kata, yaitu:

1 Data merupakan sekumpulan fakta yang tidak memiliki arti karena tidak dapat memberikan informasi

2 Mining yang berarti proses penambangan. Sehingga data mining merupakan proses penambangan data menghasilkan output yang berguna sebagai pengetahuan.



Gambar 2.1 Siklus penyelesaian proses dalam Data Mining(Marisa et al., 2021)

Proses suatu hal terdiri atas 3 fase, yaitu input, proses dan output. Data mining juga mengalami 3 fase tersebut, dengan data sebagai input kemudian data akan diproses menggunakan salah satu algoritma dari data mining dan akan mengeluarkan output berupa pengetahuan.

Terdapat empat pelompokan dalam data mining yaitu klasifikasi, asosiasi, *clustering* dan prediksi. Defiyanti, (2017):

1. Klasifikasi

   Proses klasifikasi didasarkan: Kelas-variabel dependen dari model-yang merupakan variabel kategori mewakili yang 'label' memakai objek setelah klasifikasi.

2. Asosiasi

   Setiap asosiasi antara fitur-fitur yang dicari, bukan hanya satu yang memprediksi nilai kelas tertentu. Pada prinsipnya, penemuan aturan asosiasi/asosiasi mempelajari aturan bagaimana kita memahami proses mengidentifikasi aturan antara ketergantungan yang berbeda dari fenomena kelompok. Dengan demikian, mari kita perkirakan kumpulan set yang kita punya masing-masing berisi sejumlah objek/benda-benda.

3. Clustering

   Cluster adalah menemukan kelompok (kelompok) objek, berdasarkan kemiripan sehingga dalam setiap kelompok ada kemiripan yang besar, sementara kelompok cukup berbeda dari satu sama lain.

4. Prediksi

   Prediksi/perkiraan model yang berkaitan dengan kemampuan untuk memprediksi tanggapan terbaik (output), yang paling dekat ke kenyataan, berdasarkan input data

Adapun data time series dalam penjualan buah pada penelitian ini adalah Data penjualan harian pada toko seperti Toko Nayla Buah merupakan jenis data deret waktu (*time series*) (Hyndman & Athanasopoulos, 2021). Analisis time series membutuhkan teknik yang mampu mengenali pola berulang dan hubungan antar nilai dari waktu ke waktu (*autokorelasi*), yang bisa ditangani oleh model *machine learning* modern.

2.6 Machine Learning untuk Prediksi Permintaan

Machine Learning adalah salah satu dari metodologi ilmiah modern yang dapat melakukan prosedur otomatis untuk menghasilkan prediksi pada suatu fenomena dengan melakukan pengamatan dari kejadian yang terjadi sebelumnya yaitu mencari pola pada suatu kumpulan data yang diberikan. Salah satu *algoritma machine learning* adalah *algoritma random forest* t (Shalev-Shwartz, 2019).

Machine learning (ML) adalah cabang kecerdasan buatan (AI) yang memungkinkan sistem komputer belajar dari data, mengenali pola, dan membuat keputusan secara mandiri tanpa diprogram secara eksplisit. *Machine learning* digunakan untuk melakukan prediksi menggunakan pola dari data historis dengan meminimalkan error. Beberapa model yang umum digunakan pada prediksi permintaan antara lain: *Linear Regression, Decision Tree, Random forest*, SVM, dan *Deep Learning* seperti CNN dan LSTM.). *Machine learning* dapat bekerja lebih baik daripada metode statistik klasik ketika pola data bersifat non-linear, berfluktuasi tajam, atau memiliki banyak fitur prediktor (multivariat). (Hahn, 2020).

2.7 Python

Python adalah bahasa pemrograman komputer, sama seperti bahasa pemrogramanlain, misalnya C, C++, pascal, java PHP, Perl, dan lain-lain. Sebagai bahasa pemrograman, *python* tentu memiliki dialek, kosakata atau kata kunci (*keyword*), dan aturan tersendiri yang jelas berbeda dengan bahasa pemrograman lainnya. *Python* adalah salah satu pemrograman interpretatif

multiguna dengan filosofi perancangan yang berfokus pada tingkat keterbacaan kode. Sehingga *python* diklaim sebagai bahasa yang menggabungkan kapabilitas, kemampuan, dengan sintaksis kode yang sangat jelas, dan dilengkapi dengan fungsionalitas pustaka standar yang besar serta komprehensif (Syahrudin, 2018).

Berdasarkan definisi diatas dapat dijelaskan bahwa *python* adalah sebuah bahasa pemrograman yang multiguna dan dapat digunakan untuk berbagai keperluan perangkat lunak.Dengan perkembangan bahasa pemrograman *python* yang sangat pesat ini dapat menarik minat google untuk membuatkan *Integrated Environment Development* (IDE) secara online yang biasa disebut *google colab*.

2.8 Algoritma Random Forest

Random Forest adalah algoritma supervised learning yang dikeluarkan oleh Breiman pada tahun 2001. *Random forest* adalah algoritma ensemble yang terdiri dari banyak *decision tree* yang digabungkan menggunakan teknik *bagging*. *Random Forest* biasa digunakan untuk menyelesaikan masalah yang berhubungan dengan klasifikasi, regresi, dan sebagainya. Algoritma ini berupa kombinasi dari beberapa tree predictors atau bisa disebut *decision trees* dimana setiap *tree* bergantung pada nilai *random vector* yang dijadikan sampel secara bebas dan merata pada semua *tree* dalam *forest* tersebut. Hasil prediksi dari *Random Forest* didapatkan melalui hasil terbanyak dari setiap individual *decision tree* (voting untuk klasifikasi dan rata-rata untuk regresi) (Louppe, 2016). Untuk RF yang terdiri dari N trees dirumuskan sebagai:

$$I(y) = \text{argmax}_c \left(\sum_{n=1}^{N} I_{hn\ (y0=c)}\right)$$

Dimana I adalah fungsi indikator dan hn adalah tree ke-n dari RF (Liparas, 2018).

Random Forest memiliki mekanisme internal yang menyediakan estimasi dari *generalization error*-nya sendiri yang disebut *out-of-bag (OOB) error estimate*. Dalam pembentukan *tree* hanya 2/3 dari data asli yang digunakan dalam pengambilan sampel *bootstrap* sedangkan 1/3 sisanya diklasifikasikan oleh *tree* yang terbentuk dan digunakan untuk menguji performanya. *OOB error estimation* adalah rata-rata dari kesalahan prediksi untuk setiap kasus *training y* menggunakan *tree* yang tidak mengikutsertakan y dalam sampel *bootstrap*-nya. Kemudian, saat RF dibuat, semua *training cases* menyusuri setiap pohon dan matriks kedekatan setiap kasus dihitung berdasarkan pasangan kasus yang sampai di terminal *node* yang sama (Liparas, 2018).

Banyak penelitian yang membuktikan bahwa *Random Forest* memiliki performa prediksi yang baik dalam regresi serta klasifikasi di berbagai bidang seperti prediksi finansial, remote sensing, serta analisa genetik dan biomedis. RF juga menunjukan performa yang lebih baik saat disandingkan dengan metode lain seperti, *partial least squares regression*, *support vector machine* dan *neural network* (Xu, 2018)

Menurut Breiman (2021), *Random forest* bekerja dengan:

1. Membuat banyak *decision tree* secara acak (*random subset of data* + *random subset of features*).
2. Mengambil rata-rata prediksi seluruh *tree* (untuk regresi).

Kelebihan *Random forest* adalah mampu menangani data non-linear, tahan terhadap *overfitting*, dapat memberikan *feature importance*, cocok untuk data tabular seperti penjualan, harga, cuaca, hari, dan promo. Adapun kelemahannya adalah tidak terlalu efektif menangkap pola berurutan (*sequential pattern*) pada time series, kurang optimal jika data sangat dipengaruhi pola temporal jangka panjang.

Adapun tahapan cara kerja *random forest* antara lain sebagai berikut:

1. Pengambilan Sampel Data (*Bootstrap Sampling*)

   Random forest dimulai dengan mengambil beberapa sampel acak (dengan pengembalian) dari dataset asli untuk membentuk beberapa subset data. Proses ini disebut *bootstrap sampling* dan setiap subset digunakan untuk melautih satu pohon keputusan.
2. Pembentukan Pohon Keputusan (*Decision Tree*)

   Untuk setiap subset data, sebuah *decision tree* dibentuk. Dalam proses ini, hanya sebagian adcak dan fitur yang dipertimbangkan disetiap *node* pembelahan. Hal ini memastikan bahwa pohon-pohin yang dihasilkan memiliki variasi dan tidak saling tergantung. Proses pembentukan pohon melibatkan langkah-langkah berikut:

a) Pemilihan Fitur secara Acak

Di setiap *node* (simpul) dari pohon, sejumlah fitur dipilih secara acak dari semua fitur yang tersedia. Hal ini berbeda dari pohon keputusan biasa yang menggunakan semua fitur untuk memecahkan *node*.

b) Pemisahan *Node*

Dari fitur-fitur yang dipilih secara acak, algoritma mencari fitur dan titik pemisahan terbaik yang memaksimalkan pemisahan data berdasarkan kriteria tertentu menggunakan Gini Index untuk klasifikasi.

Gini Indeks = $1 - \sum_{i=1}^{n}(Pi)^2$

$= 1 - [(P+)^2 + (P-)^2]$

Keterangan:

n = Jumlah dari masing-masing atribut

Pi = jumlah atribut dari masing-masing kelas atau labelnya

P+ = Probabilitas Positif Class

P- = Probabilitas Ngatif Class

Untuk mengukur seberapa sering elemen yang dipilih secara acak dari kumpulan data, Gini Index melakukan pemisahan optimal antara simpul akar dan simpul berikutnya.

c) Pembangunan Cabang atau Pembelahan *Node*

Node tersebut dipecah menjadi dua cabang, dan proses ini berulang sampai pohon mencapai kedalaman tertentu atau *node* tidak bisa dipecah lagi (misalnya, semua data dalam *node* adalah homogen).

3. *Majority Voting*

Setelah semua pohon dalam hutan terbentuk, *Random forest* membuat prediksi akhir dengan cara klasifikasi. Klasifikasi setiap pohon memberikan suara (prediksi) untuk kelas tertentu, dan kelas dengan suara terbanyak di antara semua pohon dipilih sebagai prediksi akhir.

4. Evaluasi Model

Random forest dievaluasi berdasarkan kinerja prediksinya pada data uji yang tidak digunakan dalam pelatihan. Teknik ini melibatkan pengukuran akurasi,precision, recall, dan F1-score. Tahapan proses perhitungan algoritma *random forest* dalam penelitian ini adalah sebagai berikut:

a) Model akan mengambil 100 subset data secara acak dan mungkin terdapat sampel yang bisa terpilih lebih dari sekali dan bisa saja ada yang tidak terpilih.

b) Pada setiap subset, model akan membangun 1 decision tree dengan mencapai kedalaman maksimal 10.

c) Kemudian internal *node* akan ditemukan dengan mencari nilai weighted gini index terkecil. Rumus untuk mendapatkan weighted gini index terkecil, yaitu $\frac{N_1}{N}\ gini1 + \frac{N_2}{N}\ gini2$ Untuk mendapatkan gini dari rumus tersebut yaitu dengan menggunakan rumus gini = 1-$\sum_{i-1}^{k} p2_1$

d) Untuk mendapatkan variabel y prediksi, maka akan dilakukan voting dari setiap pohon. Jika mayoritas pohon memprediksi 1, maka *random forest* akan menghasilkan angka 1

e)

2.9 K-*Nearest Neighbor* (KNN)

Metode K-*Nearest Neighbor* merupakan metode non parametrik yang dapat digunakan untuk pengklasifikasian berdasarkan tetangga k terdekat dan regresi. Algoritma K-*Nearest Neighbor* merupakan metode untuk mengklasifikasikan objek yang paling dekat dengan objek berdasarkan data pembelajaran Algortima K-*Nearest Neighbor* bekerja dengan mencari jarak antara data yang akan dievaluasi (data training) dan himpunan terdekat dari k tetangga (neighbor) terdekat dalam data baru (data testing) (Harun dkk, 2020).

Kelebihan KNN adalah mudah diimplementasikan dan tidak membutuhkan proses training berat, Cocok untuk data dengan pola lokal, seperti pola permintaan harian. tidak memerlukan asumsi distribusi data (non-parametric), dpat bekerja baik jika data cukup banyak dan variabel jelas. Adapun **k**ekurangan KNN adalah sensitif terhadap skala data sehingga butuh normalisasi, waktu prediksi cukup lambat jika dataset besar (karena menghitung jarak ke semua data), sensitif terhadap data noise dan outliers, pemilihan nilai *k* sangat berpengaruh terhadap akurasi (Jet al., 2021).

KNN bekerja berdasarkan prinsip bahwa objek serupa berada dalam jarak dekat satu sama lain. Dalam klasifikasi, KNN memilih kelas dari K tetangga terdekat dan memilih kelas yang paling umum. Dalam regresi, KNN memprediksi berdasarkan rata-rata atau fungsi dari K tetangga terdekat.

Mekanisme Algoritma KNN adalah sebagai berikut:

1. Penentuan Nilai K: Menentukan jumlah tetangga terdekat K.
2. Perhitungan Jarak: Menghitung jarak antara titik data baru dengan titik data dalam set pelatihan menggunakan metrik seperti Euclidean.
3. Pemilihan Tetangga Terdekat: Mengidentifikasi k titik data dalam set pelatihan yang paling dekat dengan titik data baru.
4. Penentuan Kelas (untuk Klasifikasi): Memilih kelas yang paling umum di antara tetangga terdekat.
5. Prediksi Nilai (untuk Regresi): Menghitung rata-rata atau fungsi lain dari nilai tetangga terdekat. (Amma, et. al, 2023)