BAB II
LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian ini akan merujuk pada berbagai studi terdahulu yang membahas penerapan *Business Intelligence*, analitik data besar (*big data)*, dan *machine learning* dalam mendukung proses bisnis, khususnya di industri ritel. Tinjauan literatur ini berfokus pada bagaimana analitik data digunakan untuk mengubah data mentah menjadi Keputusan strategis.

Berikut adalah tabel yang merangkum beberapa penelitian terdahulu yang relevan dengan topik ini. Tinjauan ini berfokus pada penerapan big data seperti data lake, model prediktif seperti forecasting dan customer churn, serta platform visualisasi interaktif dalam mendukung Keputusan strategis di berbagai industri, termasuk ritel.

Tabel 2. 1 Perbandingan Penelitian Terdahulu

No	Judul Jurnal	Edisi Volume	Masalah	Metode	Hasil Penelitian
1	An innovative and adaptive deep Markov random field-based retailing prediction framework in big data analytics with improved mud ring optimizer [4]	Expert Systems with Applications, 2025	Keterbatasan model prediktif dalam BDA ritel; arsitektur lama tidak efisien	Hadoop MapReduce, PCA, ADMRV, IMRA	Meningkatkan akurasi dan efisiensi prediksi penjualan ritel dibandingkan metode tradisional

No	Judul Jurnal	Edisi Volume	Masalah	Metode	Hasil Penelitian
2	Data Analytics in Sales and Marketing: A Comprehensive Methodology for Business Analysts [5]		mendukung keputusan strategis pemasaran dan penjualan	visualisasi dan data science	berbasis data dan menghindari kesalahan umum big data
3	A Brief Analysis of Palantir Gotham: A Collaborative and Interactive Big Data Visualization Analysis Software Based on Dynamic Ontology [6]	Int. Conf. on Big Data and Information Analytics, 2024	Kurangnya pemahaman mendalam tentang bagaimana visualisasi dan kolaborasi data besar dilakukan dalam tools seperti Palantir Gotham	Analisis fitur Palantir: dynamic ontology, arsitektur kolaboratif, teknik visualisasi interaktif	Mengungkapkan bagaimana Palantir mendukung analis bisnis dalam analisis kolaboratif dan pemetaan data besar
4	Conceptual modeling of big data SPJ operations with Twitter social medium [7]	Social Network Analysis and Mining, 2023 (5)	Masalah dalam proses ETL (Extract–Transform–Load) pada data media sosial yang berdampak pada pengambilan keputusan	Model transformasi berbasis MapReduce; implementasi dengan Talend for Big Data	Model memungkinkan proses seleksi, proyeksi, dan join data sosial media dalam arsitektur NoSQL dan lingkungan terdistribusi
5	A Proposed Big Data Architecture Using Data Lakes for Education Systems [8]	IFIP AICT Series, 2022	Arsitektur data lama tidak mampu menangani 3V data pendidikan	Arsitektur data lake berbasis lapisan; integrasi data heterogen	Mengoptimalkan proses analitik pendidikan melalui konsolidasi data dengan data lake

No	Judul Jurnal	Edisi Volume	Masalah	Metode	Hasil Penelitian
6	Predicting Customer Churn in Insurance Industry Using Big Data and Machine Learning [9]	ICAEECI, 2023	Tingkat churn tinggi di industri asuransi; sulit diprediksi secara akurat	LGBM + XGBoost hybrid, Logistic Regression, Random Forest	Model hybrid meningkatkan akurasi; LR terbaik menurut AUC; Memberi kan wawasan perilaku pelanggan
7	Quarry: A User-centered Big Data Integration Platform [10]	Information Systems Frontiers, 2021	Kompleksitas integrasi data dari berbagai sumber oleh pengguna non-teknis	Hypergraph metadata, platform visualisasi, integrasi otomatis	Quarry memudahkan Business Analyst melakukan eksplorasi, integrasi, dan deployment data
8	Retailing and retailing research in the age of big data analytics and risk of management modern [11]	International Journal of Research in Marketing, 2020	Bagaimana big data mempengaruhi riset dan manajemen ritel modern	Analisis perspektif teoritis dan praktis lima pihak (manajer, peneliti, investor, pengembang sistem, dan pengguna akhir).	Menyoroti peluang dan tantangan baru dari revolusi big data dalam konteks ritel

9	Interactive Dashboard of Flood Patterns Using Clustering Algorithms [12]	ICIC Express Letters, Part B: Applications, Vol. 10, No. 5, 2019	Kurangnya sistem analisis banjir yang mudah dipahami oleh pengguna non-teknis di wilayah Tangerang	K-Medoids, DBSCAN, X-Means, Power BI	Menghasilkan dashboard interaktif yang menampilkan pola kenaikan level air sungai berdasarkan waktu dan lokasi, memudahkan analisis banjir oleh pengguna non-teknis

Berdasarkan Tabel 2.1.1 Perbandingan Penelitian Terdahulu, penelitian-penelitian terdahulu menunjukkan bahwa pemanfaatan big data analytics dalam bisnis umumnya mencakup tiga fokus utama yang saling berkaitan, yaitu pengolahan data berskala besar, pemodelan dan analisis data, serta penyajian hasil analisis melalui visualisasi dan sistem pendukung keputusan.

Fokus pertama berkaitan dengan pengolahan dan pengelolaan data. Beberapa penelitian menekankan pentingnya infrastruktur dan arsitektur data dalam menangani volume, kecepatan, dan keragaman data. Studi An Innovative and Adaptive Deep Markov Random Field-Based Retailing Prediction Framework in Big Data Analytics serta Conceptual Modeling of Big Data SPJ Operations with Twitter Social Medium menunjukkan bahwa teknologi seperti Hadoop, MapReduce, NoSQL, dan data lake berperan penting dalam memastikan data berskala besar dapat diproses secara efisien. Hal ini diperkuat oleh penelitian A Proposed Big Data Architecture Using Data Lakes for Education Systems, yang menyimpulkan bahwa pendekatan data lake mampu mengatasi keterbatasan sistem tradisional dalam mendukung analisis data yang kompleks dan beragam.

Fokus kedua berkaitan dengan pemodelan dan analisis data untuk menghasilkan insight bernilai bisnis. Penelitian mengenai prediksi ritel berbasis ADMRV serta studi Predicting Customer Churn in Insurance Industry Using Big Data and

Machine Learning menunjukkan bahwa penerapan machine learning dan analisis statistik mampu meningkatkan akurasi prediksi serta pemahaman terhadap perilaku pelanggan. Sementara itu, jurnal Data Analytics in Sales and Marketing menegaskan bahwa hasil analitik akan memberikan manfaat optimal apabila diintegrasikan dengan konteks dan kebutuhan bisnis, bukan hanya berorientasi pada performa model semata.

Fokus ketiga berhubungan dengan penyajian dan pemanfaatan hasil analisis. Penelitian A Brief Analysis of Palantir Gotham dan Quarry: A User-centered Big Data Integration Platform menekankan pentingnya visualisasi interaktif, kolaborasi manusia–komputer, serta antarmuka yang ramah pengguna dalam membantu analis dan pengambil keputusan memahami data yang kompleks. Hal ini sejalan dengan penelitian Interactive Dashboard of Flood Patterns Using Clustering Algorithms, yang menunjukkan bahwa dashboard interaktif berbasis Power BI mampu menyederhanakan interpretasi hasil analitik bagi pengguna non-teknis.

Perbedaan Penelitian dengan Penelitian Terdahulu

Berdasarkan tinjauan tersebut, dapat disimpulkan bahwa sebagian besar penelitian terdahulu masih berfokus pada satu aspek tertentu, seperti pengembangan infrastruktur big data, pemodelan prediktif, atau visualisasi data secara terpisah. Penelitian ini memiliki perbedaan utama dengan mengintegrasikan ketiga fokus tersebut ke dalam satu alur kerja Business Intelligence yang utuh. Penelitian ini mengombinasikan analisis data menggunakan Python untuk pengolahan statistik, klasterisasi, dan peramalan, dengan implementasi Power BI sebagai sarana visualisasi sekaligus penyajian insight berbasis Natural Language Generation (NLG). Dengan pendekatan ini, penelitian tidak hanya menghasilkan hasil analisis, tetapi juga menyajikan rekomendasi strategis yang siap digunakan oleh manajemen ritel dalam pengambilan keputusan.

2.2 Teori tentang Topik Skripsi

2.2.1 Business Intelligence (BI)

Business Intelligence (BI) adalah sebuah payung terminologi yang komprehensif, yang mencakup arsitektur, *tools*, basis data, aplikasi, dan metodologi. BI adalah proses berbasis teknologi untuk menganalisis data dan menyajikan informasi yang dapat ditindaklanjuti (*actionable information*) guna membantu para eksekutif, manajer, dan pengguna akhir dalam membuat keputusan bisnis yang lebih baik. BI berevolusi dari sistem pelaporan statis (menjawab "Apa yang terjadi?") menjadi sistem analitik interaktif (menjawab "Mengapa terjadi?") dan kini, seperti dalam penelitian ini, bergerak menuju analitik preskriptif (menjawab "Apa yang harus dilakukan?").

Secara fundamental, BI adalah *engine* yang dirancang untuk mengubah data mentah menjadi kebijaksanaan. Proses ini sering digambarkan sebagai piramida DIKW [13] (Data -> Information -> Knowledge -> Wisdom). Evolusi BI dapat dikategorikan menjadi empat tingkatan:

1. Analitik Deskriptif (*Descriptive Analytics*): Menjawab pertanyaan "Apa yang telah terjadi?" Fase ini melibatkan pelaporan, *dashboarding*, dan visualisasi data historis, seperti yang dilakukan oleh *dashboard* Power BI untuk menampilkan Total Penjualan dan Tren.
2. Analitik Diagnostik (*Diagnostic Analytics*): Menjawab pertanyaan "Mengapa hal itu terjadi?" Fase ini melibatkan teknik *drill-down*, *data mining*, dan analisis akar masalah (misalnya, identifikasi klaster toko berkinerja rendah dan validasi penyebabnya melalui uji Chi-Square).
3. Analitik Prediktif (*Predictive Analytics*): Menjawab pertanyaan "Apa yang kemungkinan akan terjadi di masa depan?" Fase ini melibatkan *forecasting* dan *machine learning* (misalnya, peramalan penjualan menggunakan ARIMA).
4. Analitik Preskriptif (*Prescriptive Analytics*): Menjawab pertanyaan "Apa yang harus kita lakukan?" Fase ini adalah puncak dari BI, yang menyediakan rekomendasi *actionable* secara otomatis, seperti yang dicapai

melalui *engine* Natural Language Generation (NLG) di Power BI.

2.2.2 Data Transaksi Harian Ritel

Data transaksi harian ritel adalah rekaman digital terperinci dari setiap aktivitas penjualan yang terjadi di titik penjualan (*Point of Sales*/POS). Data ini mencakup atribut penting seperti Transaction Date, Store ID, Net Price, Open (Qty), dan Product Category. Data ini merupakan aset strategis primer karena berfungsi sebagai proksi langsung untuk permintaan konsumen dan efisiensi operasional.

Dalam penelitian ini, data transaksi harian ritel digunakan sebagai *input* fundamental untuk dua tujuan analitis utama

1. Segmentasi Toko: Data agregat (misalnya, total Net Price dan rata-rata Open Qty per toko) digunakan sebagai variabel dalam algoritma K-Means untuk membagi populasi toko menjadi kelompok-kelompok homogen (klaster).
2. Analisis Runut Waktu (*Time Series*): Data penjualan diagregasi per periode waktu (misalnya, bulanan) untuk menjadi *input* dalam model ARIMA guna memprediksi tren penjualan di masa depan

2.2.3 Metrik Kinerja Kunci

Dalam analisis ritel, pemahaman yang mendalam tentang metrik kinerja adalah prasyarat untuk pengambilan keputusan yang efektif. Penelitian ini menggunakan dua variabel utama, yang merupakan proksi langsung dari kinerja toko, sebagai dasar untuk algoritma *clustering* K-Means:

A. Total Nilai Penjualan Bersih (*Total Net Sales*)

Total Net Sales (Net Price) adalah metrik vital yang mewakili total pendapatan yang dihasilkan oleh toko dari penjualan, setelah dikurangi diskon, retur, atau PPN (jika ada). Metrik ini menjadi indikator langsung dari kontribusi pendapatan toko. Toko dengan *Net Sales* tinggi menunjukkan volume penjualan yang besar atau harga jual rata-rata (*Average Selling Price*/ASP) yang tinggi. Dalam konteks *clustering*, *Net Sales* digunakan sebagai dimensi vertikal untuk memisahkan klaster toko berkinerja tinggi dari toko berkinerja rendah.

B. Volume Penjualan (*Open Quantity - Qty*)

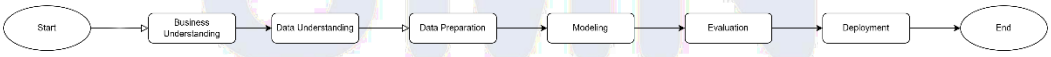
Volume Penjualan (Open Qty) adalah metrik yang mewakili jumlah unit produk yang berhasil dijual oleh toko dalam periode waktu tertentu. Metrik ini memberikan *insight* mengenai seberapa sukses suatu toko dalam mendorong *traffic* dan mengonversi kunjungan menjadi transaksi, terlepas dari nilai rupiahnya. Kombinasi *Net Sales* (nilai moneter) dan *Open Qty* (volume unit) sangat penting karena memungkinkan identifikasi strategis:

1. **Klaster High Performers:** Kelompok toko dengan nilai rata-rata Net Price dan Open Qty yang paling tinggi. Toko-toko dalam klaster ini adalah penyumbang pendapatan terbesar dan merupakan prioritas utama untuk strategi retensi.
2. **Klaster Mid-High Performers:** Kelompok dengan kinerja yang kuat tetapi berada satu tingkat di bawah *High Performers*. Klaster ini menunjukkan potensi untuk diangkat (*upselling*) menjadi klaster tertinggi melalui intervensi yang tepat.
3. **Klaster Mid-Low Performers:** Kelompok toko yang memiliki kinerja di bawah rata-rata total, namun relatif stabil. Klaster ini menunjukkan risiko penurunan dan memerlukan strategi intervensi untuk menjaga *traffic* dan frekuensi kunjungan agar tidak jatuh ke klaster terendah.
4. **Klaster Low Performers:** Kelompok toko dengan nilai rata-rata Net Price dan Open Qty yang paling rendah. Toko-toko dalam klaster ini memerlukan perhatian dan intervensi strategis yang paling agresif, dengan fokus utama pada peningkatan volume transaksi.

2.3 Teori tentang Framework/Algoritma yang digunakan

2.3.1 CRISP-DM (Cross Industry Standard Process for Data Mining)

CRISP-DM adalah metodologi standar industri yang paling banyak digunakan untuk proyek *data mining* dan analitik. *Framework* ini menyediakan alur kerja yang terstruktur dan iteratif, yang terdiri dari enam fase:



Gambar 2. 1 Flowchart Diagram CRISP-DM (Sumber Draw.io)

1. *Business Understanding* (Pemahaman Bisnis)

   Fase *Business Understanding* adalah langkah awal dan paling krusial, di mana fokus utama diletakkan pada pemahaman tujuan dan kebutuhan proyek dari perspektif bisnis. Aktivitas pada fase ini mencakup pendefinisian tujuan bisnis (misalnya, meningkatkan efisiensi operasional dan akurasi pengambilan keputusan manajerial), penentuan kriteria

keberhasilan bisnis (misalnya, tersedianya rekomendasi yang *actionable* dan terintegrasi), dan penilaian situasi saat ini (misalnya, identifikasi masalah *Dashboard Dilemma*—dashboard yang hanya menyajikan data deskriptif tanpa *insight* mendalam).

2. *Data Understanding* (Pemahaman Data)

   Setelah tujuan bisnis ditetapkan, fase *Data Understanding* dimulai dengan pengumpulan data dan eksplorasi mendalam terhadap set data yang tersedia. Aktivitas utama mencakup pengumpulan data transaksi harian ritel, pemuatan data ke lingkungan analitik (seperti Python atau Power BI), serta melakukan pemeriksaan kualitas data awal. Tujuan fase ini adalah memperoleh pemahaman yang komprehensif mengenai struktur, format, kualitas, dan potensi informasi yang terkandung dalam data, termasuk mengidentifikasi variabel-variabel kunci seperti Net Price, Open Qty, Transaction Date, dan kategori produk yang akan digunakan sebagai masukan untuk pemodelan K-Means dan ARIMA.

3. *Data Preparation* (Persiapan Data)

   Data Preparation merupakan fase yang menghabiskan waktu paling banyak, di mana data mentah diubah menjadi set data akhir yang bersih dan siap untuk dimodelkan. Proses ini melibatkan ETL *pipeline* yang komprehensif, mencakup pembersihan data (penanganan *missing values* atau *outliers*), integrasi data (menggabungkan data fakta transaksi dengan data dimensi toko), serta transformasi data. Transformasi data krusial dalam penelitian ini adalah agregasi data harian ke level toko (untuk *clustering*) dan level bulanan (untuk *forecasting*), serta normalisasi data (menggunakan *StandardScaler* di Python) untuk memastikan semua variabel memiliki skala yang sama sebelum dimasukkan ke dalam algoritma K-Means.

4. *Modeling* (Pemodelan)

Pada fase *Modeling*, berbagai teknik *data mining* dan *machine learning* diterapkan untuk menemukan pola tersembunyi dan menghasilkan *insight* prediktif yang relevan dengan tujuan bisnis. Berdasarkan kebutuhan proyek, dua model utama diimplementasikan di lingkungan Python *backend*: pertama, K-Means *Clustering*, yang digunakan untuk mengelompokkan toko berdasarkan matriks kinerja penjualan, dan kedua, ARIMA (*AutoRegressive Integrated Moving Average*), yang digunakan untuk peramalan runut waktu penjualan ritel. Fase ini mencakup penentuan parameter optimal untuk kedua model tersebut, seperti penentuan jumlah klaster (*K*) yang paling efektif melalui *Elbow Method* dan *Silhouette Score*.

5. *Evaluation* (Evaluasi)

   Fase *Evaluation* bertujuan untuk menilai kualitas model dari sudut pandang teknis (akurasi) dan signifikansi temuan dari sudut pandang bisnis. Penilaian signifikansi temuan difokuskan pada validitas klaster yang terbentuk. Dalam penelitian ini, validasi klaster dilakukan secara statistik menggunakan Uji ANOVA untuk membuktikan perbedaan signifikan rata-rata kinerja antar klaster, dan Uji Chi-Square untuk mengidentifikasi hubungan antara klaster dengan variabel kategorikal (seperti kategori produk atau metode pembayaran). Jika hasil evaluasi tidak memuaskan, proses akan kembali ke fase *Modeling* atau *Data Preparation* (bersifat iteratif).

6. *Deployment* (Penyajian/Implementasi)

   Fase terakhir, *Deployment*, adalah tahap di mana hasil akhir dan model yang telah divalidasi dan disetujui diintegrasikan ke dalam lingkungan operasional untuk digunakan oleh pengguna akhir, yaitu manajemen ritel. Dalam proyek ini, Power BI *frontend* digunakan sebagai platform penyajian, di mana hasil dari Python (*klaster toko* dan *hasil forecast* ARIMA) diimpor. Puncak dari fase *Deployment* adalah implementasi Natural Language Generation (NLG) berbasis DAX *measure* yang secara otomatis menerjemahkan data dan hasil klasterisasi menjadi narasi

preskriptif, sehingga memastikan bahwa *insight* dan rekomendasi strategis yang dihasilkan dapat langsung ditindaklanjuti (*actionable*) oleh pengambil keputusan.

Dalam penelitian ini, CRISP-DM digunakan sebagai kerangka kerja utama yang memandu alur kerja *hybrid*. Fase *Modeling* dieksekusi di *backend* (Python) untuk K-Means dan ARIMA, sedangkan fase *Deployment* dieksekusi di *frontend* (Power BI) untuk menyajikan *dashboard* dan *engine* NLG.

2.3.2 K-Means Clustering

K-Means *Clustering* adalah algoritma *machine learning unsupervised* (tanpa pengawasan) yang paling populer, yang bertujuan untuk mempartisi sekumpulan n data (toko) ke dalam sejumlah K klaster, di mana data dalam satu klaster memiliki kemiripan yang tinggi satu sama lain.

A. Proses Algoritma:

K-Means bekerja secara iteratif untuk meminimalkan inertia, yaitu jumlah kuadrat jarak antara setiap titik data dengan centroid klasternya. Jarak ini biasanya diukur menggunakan Jarak Euclidean ($D(x, y)$).

$$\min_{c} = \sum_{i=1}^{n} \min_{\mu j \in C} \|xi - \mu j\|^2$$

Rumus 1 K- Means Clustering

Di mana x_i adalah titik data, μ_j adalah *centroid* klaster j, dan C adalah himpunan semua *centroid*.

B. Metode Penentuan Jumlah Klaster (K):

Karena K-Means memerlukan penentuan nilai K di awal, penelitian ini menggunakan dua metode validasi secara bersamaan:

1. Metode Siku (*Elbow Method*): Metode ini mengamati nilai *Within-Cluster Sum of Squares* (WCSS) atau *inertia* sebagai fungsi dari jumlah klaster (K). Nilai K yang optimal adalah titik di mana kurva WCSS mulai mendatar (*siku*), menunjukkan bahwa penambahan klaster tidak lagi memberikan manfaat pengurangan varians yang signifikan.
2. Skor Siluet (*Silhouette Score*): Metode ini mengukur seberapa baik sebuah objek data (toko) "cocok" dengan klasternya sendiri dibandingkan dengan klaster tetangga. Skor yang mendekati +1 menunjukkan klaster yang padat dan terpisah dengan baik, dan skor mendekati 0 menunjukkan *overlap* antara klaster. Nilai K yang menghasilkan skor siluet tertinggi adalah nilai K yang paling optimal.

2.3.3 Uji Validasi Statistik

Setelah *clustering* dilakukan, validasi statistik diperlukan untuk membuktikan signifikansi klaster yang terbentuk.

A. Analisis Variansi (ANOVA Test)

ANOVA (ANalysis Of VAriance) digunakan untuk menguji apakah terdapat perbedaan yang signifikan secara statistik antara nilai rata-rata dari dua atau lebih kelompok yang berbeda. Dalam konteks ini, ANOVA digunakan untuk memvalidasi klaster toko.

- Hipotesis Nol (H_0): Rata-rata metrik kinerja (misalnya, Net Price) dari semua klaster adalah sama ($\mu_1 = \mu_2 = \mu_3 = \mu_4$).
- Aplikasi: Jika nilai p (*p-value*) dari uji F-statistik sangat kecil (misalnya,

$p < 0.05$), maka H_0 ditolak, dan disimpulkan bahwa terdapat perbedaan **signifikan** pada rata-rata kinerja antar klaster. Hal ini membuktikan bahwa klasterisasi K-Means yang dilakukan berhasil membagi populasi toko berdasarkan perbedaan kinerja numerik yang nyata.

B. Uji Chi-Square (X^2)

Uji Chi-Square digunakan untuk menentukan apakah terdapat hubungan (asosisasi) atau independensi antara dua variabel kategorikal. Dalam penelitian ini, Uji Chi-Square sangat penting untuk mendefinisikan karakter setiap klaster.

- Hipotesis Nol (H_0): Variabel Klaster Toko independen terhadap variabel kategorikal lain (misalnya, Product Category atau Paid Method).
- Aplikasi: Jika nilai p besar (misalnya, $p > 0.05$), H_0 diterima, dan disimpulkan bahwa tidak ada hubungan signifikan antara Klaster Toko dengan variabel tersebut. Hasil ini memberikan *insight* strategis krusial: misalnya, jika klaster toko High Performers dan Low Performers memiliki proporsi pembelian Product Category yang sama, maka strategi manajemen yang efektif seharusnya tidak berfokus pada perubahan produk, melainkan pada metrik lain (misalnya, peningkatan *traffic* atau jumlah transaksi).

2.3.4 ARIMA

ARIMA (*AutoRegressive Integrated Moving Average*) adalah model statistik yang digunakan untuk menganalisis data runut waktu (*time-series*) dan memprediksi nilai di masa depan tanpa mempertimbangkan komponen musiman secara eksplisit. Model ini sesuai digunakan pada data dengan periode pengamatan yang relatif pendek, di mana pola musiman belum terbentuk secara jelas.

Model ARIMA dinyakan dalam bentuk ARIMA (p, d, q).

- AR (p / P): *AutoRegressive*: Menggunakan nilai observasi sebelumnya (lag) untuk memprediksi nilai saat ini.
- I (d / D): *Integrated*: Mengacu pada proses *differencing* (perbedaan) yang

diperlukan untuk membuat data menjadi stasioner (mean dan variansnya tidak berubah seiring waktu). Uji Augmented Dickey-Fuller (ADF Test) biasanya digunakan untuk memvalidasi stasionaritas data.

- MA (q / Q): *Moving Average*: Menggunakan galat (*error*) dari model peramalan sebelumnya sebagai variabel prediktor.

Dalam penelitian ini, model ARIMA (diimplementasikan di Python) digunakan untuk memprediksi Total Net Sales ritel untuk 2 bulan ke depan, memberikan informasi Analitik Prediktif kepada manajemen.

2.4 Teori tentang Tools

2.4.1 Natural Language Generation (NLG)

Natural Language Generation (NLG) adalah teknologi kecerdasan buatan yang bertindak sebagai "penerjemah" otomatis, mengubah data terstruktur (angka, tabel, klaster) menjadi narasi atau teks yang dapat dibaca dan dipahami manusia.

Peran NLG dalam *Business Intelligence* sangat strategis karena:

1. Mengatasi Kelelahan Data (*Data Fatigue*): Manajer tidak perlu menghabiskan waktu menafsirkan grafik atau tabel yang kompleks. NLG secara otomatis menyimpulkan *insight* penting.
2. Menyediakan Analitik Preskriptif: Dengan menggabungkan hasil statistik (*clustering*, ANOVA, Chi-Square) dan logika bisnis, NLG dapat menghasilkan paragraf yang bersifat "Apa yang harus dilakukan" (*prescriptive*), bukan hanya "Apa yang terjadi" (*descriptive*).

2.4.2 Power BI

A. Power BI: Power BI berfungsi sebagai lapisan *deployment* (penyajian) dari *Business Intelligence*. Perannya adalah mengintegrasikan hasil analisis *backend* Python (*clustering*, ARIMA) dan menyajikannya dalam *dashboard* interaktif. Fitur utama Power BI adalah kemampuannya untuk melakukan *Data Modeling* (membuat hubungan antara tabel, Gambar 4.1) dan *Visualisasi Interaktif* (Gambar 4.9-4.13).

B. DAX (*Data Analysis Expressions*): DAX adalah bahasa formula yang

digunakan di Power BI, Power Pivot, dan SSAS Tabular. DAX sangat berbeda dari SQL atau Python karena DAX bersifat kontekstual. DAX digunakan untuk:

1. Perhitungan Metrik Kompleks: Menghitung KPI dinamis seperti Total Net Sales, *Gross Margin %*, dan perbandingan antar periode waktu (misalnya, YOY, MOM).
2. *Engine* Logika NLG: DAX *measure* digunakan untuk menulis logika bersyarat (IF, SWITCH, CALCULATE) yang menginterpretasikan hasil klaster (*centroid* klaster, skor ANOVA/Chi-Square) dan menampilkannya sebagai teks naratif, membentuk inti *engine* NLG

Jupyter Notebook digunakan sebagai *platform backend* yang mengandalkan beberapa pustaka Python utama untuk pemrosesan data dan pemodelan:

1. Pandas: Pustaka fundamental untuk struktur data dan alat analisis data. Digunakan untuk manipulasi, pembersihan, dan agregasi data transaksi (misalnya, pengelompokan data per toko atau per bulan).
2. NumPy: Pustaka yang mendukung array dan matriks berdimensi besar, yang penting untuk komputasi numerik berkecepatan tinggi, terutama yang terkait dengan statistik dan matematika dalam *machine learning*.
3. Scikit-learn: Pustaka terkemuka untuk *machine learning* di Python. Digunakan khusus untuk implementasi algoritma K-Means Clustering dan alat bantu seperti StandardScaler (untuk menormalisasi variabel sebelum *clustering*) dan Silhouette Score.
4. Statsmodels / SciPy: Pustaka yang menyediakan kelas dan fungsi untuk model statistik, estimasi, dan uji hipotesis, yang digunakan untuk menjalankan Uji ANOVA dan Uji Chi-Square serta pemodelan ARIMA.