**1. Methodology Chapter**

**This chapter describes the steps, tools, and techniques used to address the problem, organized into statistical, machine learning, and deep learning approaches.**

**1.1. Introduction**

* **The primary goal of the project is to predict customer value and understand customer behavior through survey data and transaction data integration.**
* **The data sources include transaction data from online retail sales and survey data capturing customer preferences and demographics.**

**1.2. Statistical Methods**

**This section covers the data preparation and exploratory analysis steps.**

**1.2.1. Data Loading and Exploration**

* **The transaction and survey datasets were loaded into the system for analysis.**
* **Initial exploration involved:**
  + **Examining the structure of datasets, including column names, data types, and overall size.**
  + **Identifying and addressing missing values, duplicates, and invalid records to ensure data quality.**
  + **Generating summary statistics to understand the data’s distribution and relationships.**

**1.2.2. Data Cleaning and Feature Engineering**

* **Cleaning steps included removing duplicates, addressing outliers using statistical thresholds, and dropping irrelevant records.**
* **Feature engineering added meaningful variables to enhance the dataset:**
  + **Total\_Amount: The product of Price and Quantity, representing transaction value.**
  + **Time-based Features: Extracted variables like Hour, Day\_of\_Week, Month, and Is\_Weekend to capture temporal patterns.**
* **Recency, Frequency, and Monetary (RFM) Analysis was conducted to analyze customer purchasing behavior:**
  + **Recency: Time since the last purchase.**
  + **Frequency: Number of purchases made by a customer.**
  + **Monetary Value: Total amount spent by a customer.**
* **Additional metrics, such as averages, minimums, and maximums, were computed to provide further insights into customer patterns.**

**1.2.3. Correlation Analysis**

* **Correlation analysis was performed to identify relationships between key features, focusing on factors influencing customer value and behavior.**

**1.3. Machine Learning Methods**

**This section discusses the use of the XGBoost algorithm for customer value prediction.**

**1.3.1. XGBoost Model Development**

* **XGBoost, a gradient boosting algorithm, was chosen for its efficiency with structured data and ability to handle feature interactions.**
* **Hyperparameter Tuning: The Optuna framework optimized the model’s parameters, such as:**
  + **max\_depth: Maximum depth of the decision trees.**
  + **learning\_rate: Step size for gradient updates.**

**1.3.2. Feature Importance Analysis**

* **The XGBoost model was analyzed to determine the most impactful features, providing valuable insights into factors driving customer value.**

**1.3.3. Model Evaluation**

* **The XGBoost model’s performance was assessed using metrics such as:**
  + **Root Mean Squared Error (RMSE): Measures average prediction error.**
  + **Mean Absolute Error (MAE): Reflects average absolute differences between predictions and actual values.**
  + **R2 Score: Indicates how well the model explains variance in the target variable.**

**1.4. Deep Learning Methods**

**This section focuses on the development and evaluation of a Neural Network model.**

**1.4.1. Neural Network Development**

* **A three-layer deep neural network was designed to process the data and capture complex patterns:**
  + **Architecture: The network consisted of three dense layers with non-linear activation functions.**
  + **Overfitting Prevention: Early stopping was implemented to monitor validation performance and stop training when improvements ceased.**

**1.4.2. Learning Curve Analysis**

* **Learning curves were plotted to assess the Neural Network’s training and validation performance, identifying any overfitting or underfitting issues.**

**1.4.3. Model Evaluation**

* **The Neural Network’s performance was evaluated using the same metrics as XGBoost for consistency:**
  + **RMSE**
  + **MAE**
  + **R2 Score**

**1.5. Survey Data Processing and Integration**

The survey data was processed and integrated with the transaction data to provide a unified dataset for analysis:

* Numerical features were standardized, and categorical variables were encoded for compatibility with machine learning models.
* The survey data was merged with transaction data, enabling deeper insights into customer preferences and purchasing behavior.

**1.6. Advanced Analysis**

This section details the analysis steps used to enhance model performance and insights:

* **Feature Importance (XGBoost)**: Identified key features driving predictions, providing actionable business insights.
* **Learning Curves (Neural Network)**: Evaluated the model’s learning progress to refine its training.
* **Clustering and Customer Segmentation**: KMeans clustering on survey data revealed distinct customer segments, enabling targeted behavioral analysis.