ABSTRACT

This study aims to improve the predictive accuracy of sales and satisfaction analysis by addressing missing data using mean imputation and machine learning models. Numerical missing values were handled with mean imputation, while categorical missing values were excluded. The dataset was split into 80% for training and 20% for testing. Predictive models were built using Decision Tree, k-Nearest Neighbors (KNN), and Random Forest algorithms. Model performance was evaluated using metrics such as AUC, accuracy, F1-score, precision, recall, and MCC. Results demonstrate that the imputation process significantly enhanced model performance, with Random Forest achieving the highest AUC (0.999) and classification accuracy (0.982). This highlights the critical role of imputation in improving data quality and predictive reliability. Moreover, the study establishes Random Forest as a robust method for handling missing values and achieving superior predictive outcomes in similar datasets.

Keywords: Desicion Tree, Imputation, Missing Data, K-Nearest Neighbors, Random Forest