ABSTRACT

A COMPARATIVE ANALYSIS OF IMPUTATION TECHNIQUE PERFORMANCE FOR HANDLING MISSING VALUES (A CASE STUDY ON DIABETES MELLITUS DATASET)

By

Vincentius Sagi Alban Anindyajati

21110037

Diabetes mellitus is a chronic disease with a continuously increasing prevalence in the 21st century. Early detection of this disease is crucial but is often hindered by the issue of missing values in medical datasets. This research aims to analyze and compare the performance of various imputation techniques in handling missing values in the diabetes mellitus dataset. The study utilizes the 'Pima Indians Diabetes Database' dataset from Kaggle, focusing on a comparison of the MICE and MissForest imputation techniques, with KNN Imputer as the benchmark. The research methodology includes data processing, developing a model using Random Forest, where hyperparameter optimization was specifically performed on KNN-imputed data using Random Search, and these optimal parameters were then applied to models trained with MICE and MissForest imputed data then evaluate the model's performance using accuracy, precision, recall, and F1-score metrics. The results show that MICE and MissForest significantly outperformed KNN Imputer. MissForest was identified as the best imputation technique, providing an increase in the F1-Score (Class 1) of +0.005829 compared to KNN. More importantly, MissForest achieved the highest F1-Score (0.807273), highest Recall (0.828358) (equivalent to KNN Imputer) , and highest Precision (0.787234) for the 'Diabetes' class, indicating the best balance between precision and recall, which is crucial for medical applications. These findings provide an evidence-based recommendation for an imputation technique to improve the prediction accuracy of diabetes mellitus, thereby contributing to the development of more reliable medical data analysis systems.

Keywords: *Diabetes Mellitus, Early Detection, Machine Learning, Missing Value, Imputation*