ABSTRACT

Chronic diseases such as diabetes mellitus, cancer, stroke, chronic kidney failure, hypertension, and heart disease are major health problems that continue to increase in Indonesia, especially among the elderly. Preventive efforts are an important step to reduce the number of cases in the future, one of which is by using data from general checkups that are analyzed to detect potential disease risks. In current practice, doctors rely on medical interviews, physical examinations, and supporting tests, which are processed through clinical reasoning to produce a diagnosis. However, as medical data becomes more complex and grows in volume, this process requires support from machine learning models to improve the accuracy and efficiency of clinical decision-making.

However, most previous studies have used machine learning to predict only a single type of disease, without directly integrating clinical knowledge into the preprocessing stage or the predictive modelling process. In addition, other major challenges include issues related to data quality, such as missing values, outliers, and imbalanced data. Therefore, this dissertation proposes a new approach by integrating domain knowledge into the preprocessing process and developing an ensemble learning model based on modified tree-based classifiers.

This research uses the GCU dataset from Yakes Telkom covering the period from 2019 to 2021, consisting of medical checkup data for six chronic diseases. The research stages include: 1) data exploration and handling of imbalanced data using Weighted SMOTE designed based on domain knowledge, 2) developing an ensemble learning model by incorporating feature weights obtained from domain knowledge into Random Forest, XGBoost, and AdaBoost algorithms, and 3) evaluating the model's performance using F1-Score, Balanced Accuracy Score, and ROC-AUC metrics.

The results show that the proposed approach significantly improves prediction performance compared to standard models. The model integrating domain knowledge within Weighted SMOTE and ensemble learning achieves an average F1-Score above 85%, which is at least 15% higher than the baseline models without *domain knowledge*. Moreover, the model is capable of predicting multiple types of chronic diseases and provides insights to assist doctors in developing more targeted preventive actions.

The main contribution of this dissertation is the development of a preprocessing method and a machine learning prediction model integrated with domain knowledge to improve the accuracy and reliability of chronic disease predictions. This approach is expected to provide real benefits for patients, healthcare professionals, and insurance providers, supporting better and more efficient clinical decision-making.

Keywords: *Chronic Diseases, Domain Knowledge, Machine Learning, Imbalanced Data, Ensemble Learning*