Abstract

Missing value is a value loss case of the data that may occur in a dataset. When it reaches a high rate of a dataset, it would require further treatment to the case of the missing value. Handling can be done is by imputation, in which missing values from a data filled with a certain value. One of imputation methods that can be used is the Hot Deck method. In this method, the value of a record on the same dataset will be taken to fill in the value of the record having missing value. In this study, the selection of the nearest record can be done by Chebyshev distance, Squared Euclidean, Canberra, and Random Hot Deck for numerical datasets. Meanwhile, for categorical dataset different distance Hot Deck was used. Performance measurement is done by measuring the RMSE on numerical dataset and measuring precision, recall, and f-measure on categorical dataset. From the results performed on a numerical dataset, Hot Deck based Squared Euclidean seem to give the most stable accuracy, while Random Hot Deck featuring the worst accuracy. In the categorical dataset, accuracy produced by Random Hot Deck show no significant difference from accuracy produced by distance Hot Deck.

Keywords : missing value, imputation, hot deck.