IMPUTATION MISSING VALUES FOR DIABETES DATA USING AN ALGORITHM

ANAS MUDHAFAR AHMED AHMED
Ministry of Oil –Republic of Iraq
Scientific supervisor: Ali Ibrahim Lawah, Ph.D.

Relevance. To calculate missing values in a database of people with diabetes. The missing values are compensated by using an algorithm inspired by nature, which is the gray wolf algorithm. Accuracy is achieved through the use of three distinct classifiers, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayesian Classifier (NBC).

Goal of the work – The objectives of this article to design an imputation algorithm based on Grey Wolf Algorithm with three different classification models, KNN, SVM, and NBC. and to impute the missing values of PIDD dataset using the proposed algorithm. and last step to verification and testing of the validity of the proposed algorithm using various evaluation scales.

Result analysis – The effectiveness of the suggested imputation algorithm was determined by the results of two primary trials. First, by using cross validation with 5 folds, and then in the second experiment, the algorithm was tested by using the holdout validation method with the dataset divided into a training set of 65% and a testing set of 35%. In each of the separate experiments, a total of sixteen different tests were performed and analyzed. the effectiveness of the algorithm was evaluated based on a variety of factors, including the number of iterations, as well as the size of the swarm, each search agent in the swarm represents of stands in for a potential solution to the problem at hand. Here, we employ the imputation technique based on Grey Wolf Algorithm (GWO). The resulting missing values are then evaluated using one of three classifiers: The K Nearest Neighbors (KNN), the Support Vector Machine (SVM), or the Naive Bayesian Classifier (NBC). the IGWO-KNN algorithm had the best overall results; but, the IGWO-SVM algorithm produced better results on average. The standard deviation demonstrated that both IGWO-SVM and IGWO-NBC are more stable than IGWO-KNN. The distinction was made using this metric.

Conclusion As a solution to the issue of missing data, a number of imputation methods have been offered in the literature. Some common methods include ignoring or eliminating samples with missing values, substitution with zeros, average, or random values generation. In recent research, optimization-based imputation techniques were proposed to be used in place of traditional imputation methods. These algorithms seek out optimal values to substitute for missing data, rather than relying on linear mathematics or complete randomization.