CSE 523 ML Project

Faculty mentor:
Prof. Mehul Raval

Project Number: 9

Project Title: Data-driven imputation scheme for human-subject-based dataset

Group name: Syntellect

Group member details:

ENROLLMENT NUMBER	NAME
Kashish Jethmalani	AU2140029
Srushti Thakar	AU2140117
Priyal Patel	AU2140204
Riya Patel	AU2140214

Progress report : (Week 3)

Date: 15-02-2024 to 22-02-2024

Tasks completed

- Understanding Problem statement: Firstly we analyzed the problem statement. It is about developing an imputation scheme to fill the missing values.

- Literature review: We read some recent papers which used different imputation techniques like KNN, Random forest and imputation using mean/median.

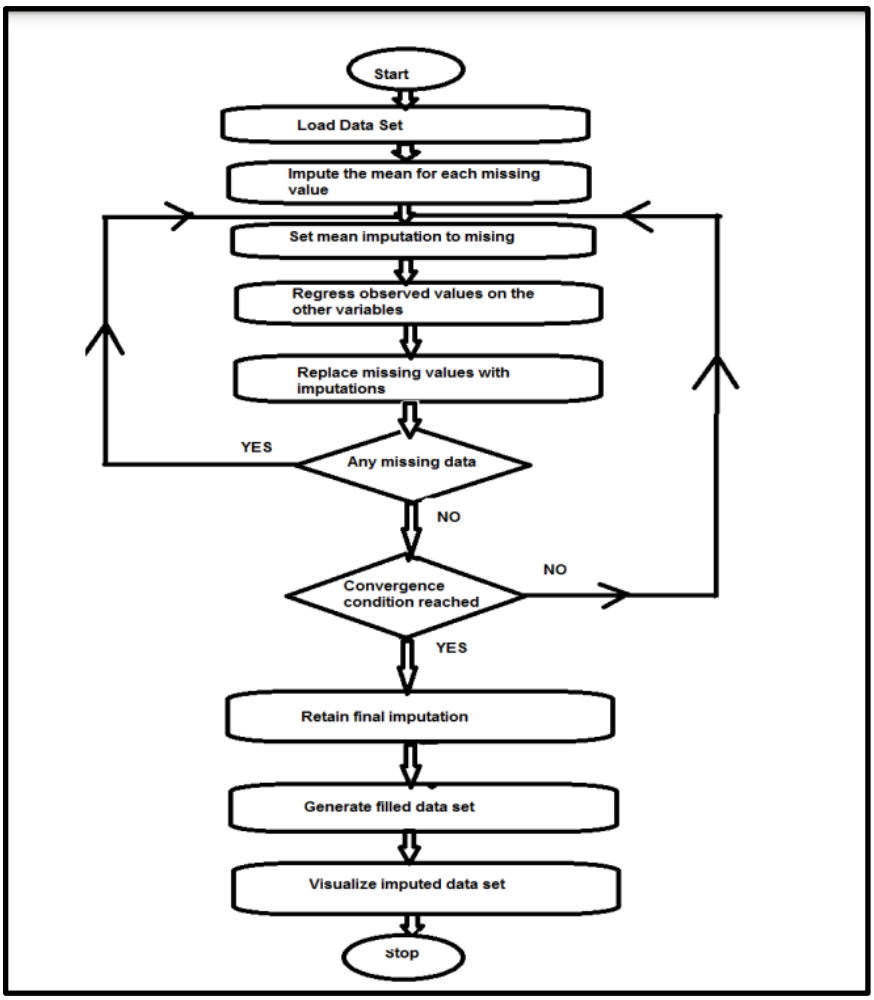
- Exploring dataset: Further we explored the structure, format and characteristics of the dataset.This includes their sleep pattern, training details, cardiac rhythm pattern, emotional-mental state information, game score, weekly readiness scores and jump-data.

Progress Summary

This week we explored many imputation schemes starting from simple mean and median based techniques to Random forest, MICE and SICE.

Mean imputation (MI): It simply means filling missing values using the mean of remaining data. This is not a good imputation technique for our dataset as it can introduce bias and can reduce the variance of data. It is also very sensitive to outliers.

MICE: Multivariate Imputation by Chained Equations: MICE operates under the assumption that given the variables used in the imputation procedure, the missing data are Missing At Random (MAR). First step is to identify variables in the dataset with missing values, it could be sleep pattern, training details,game score, etc. Then we can build a model for each missing variable and can iteratively impute the missing values.
Below chart shows how the MICE algorithm works



SICE: Single Center Imputation from Multiple Chained Equations.
It is a modified version of MICE. Here we have two variants of SICE, namely SICE-Categorical and SICE-Numeric. Following Algorithm 1: SICE-Categorical imputes missing values of categorical attributes such as binary or ordinal attributes. For better understanding, a flowchart of the SICE, which is applicable for both categorical and numeric versions are there. It executes the MICE algorithm for user-defines m times and adds the results in an array. Then a missing value is replaced with the most frequent item of the array.Algorithm 2: SICE-Numeric imputes missing values for numeric attributes. It executes MICE algorithm for a user defined m times and adds the results of each iteration in an array. Then each missing value is replaced by the mean of its corresponding imputed value from the array.

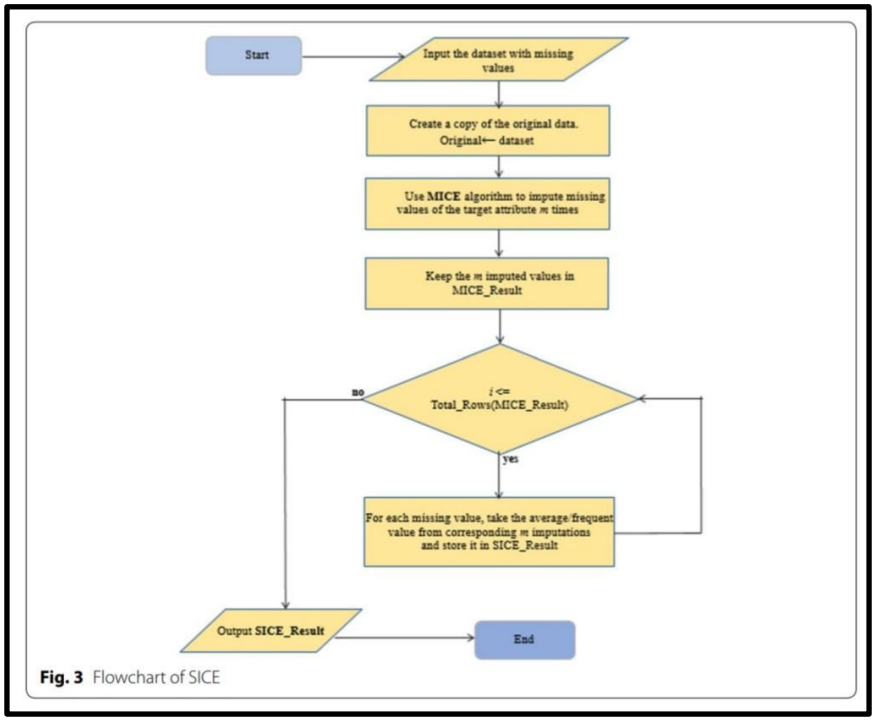


Fig. 3 Flowchart of SICE

Random forest: It is a supervised machine learning algorithm which is popular for classification and regression problems. This technique include splitting the dataset into two parts: one with complete data and another with missing values. Then train a random forest model using complete data and predict the missing values in incomplete data. The target variable for the model will be the variable with missing values, and the other variables in the dataset will be used as features. For each instance with missing values, the random forest model will predict the missing value based on the observed values of other variables.

References:

http://www.ijarse.com/images/fullpdf/1494989187_L1043ijarse.pdf

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00313-w