

Weekly Project Report:

Project 9: Data-driven imputation scheme for human-subject based dataset

Name	Roll No.
Shubham Apat	AU2140220
Aaditya Yadav	AU2140094
Helly Ullas Thakkar	AU2020130
Pal patel	AU2020113

The purpose of this weekly report is to provide an update on the progress made in our project and discuss the recent developments and findings. Our project aims to [briefly summarize the project objectives and goals].

1. KNN Imputation vs. MICE Imputation:

- In the previous work, we conducted an in-depth comparison between KNN (K-Nearest Neighbors) imputation and MICE (Multiple Imputation by Chained Equations) imputation techniques.
- We implemented both methods to handle missing values in our dataset and evaluated their performance based on imputation accuracy, computational efficiency, and overall imputation quality.
- The results of our analysis were shown in the mid sem presentation.

2. Feature Selection using Correlation Matrix:

- We utilized the correlation matrix to assess the pairwise correlations between features and selected a subset of features based on predefined correlation thresholds.
- The selected features were then used as input for subsequent imputation and analysis steps.

3. K-Means Clustering for Imputation:

- Building upon our previous work, we employed K-means clustering as an alternative imputation strategy for handling missing values.

- K-means clustering is an unsupervised machine learning algorithm that partitions data into clusters based on similarities between data points.

- We applied K-means clustering to identify clusters within our dataset and imputed missing values by replacing them with the cluster centroids.

Results and Findings:

K-Means Clustering for Imputation:

- The use of K-means clustering provided an alternative approach to imputation and allowed us to explore the accuracy of imputation techniques. The result of accuracy were more accurate and nearer to MICE imputation than compare to CORR feature selection + KNN strategy.

Next Steps:

1. Further Analysis and Evaluation:

We are going to further experiment more like this with the combination of feature selection + Imputation strategy. We are going to explore more imputation techniques like Regression Imputation, Stochastic regression, Interpolation. And for feature selection we can apply selection techniques like XGBOOST, Random Forest, etc.

In conclusion, our recent efforts have focused on comparing imputation techniques, exploring feature selection methods, and experimenting with clustering for imputation. Through these endeavors, we aim to enhance our understanding of the dataset and develop robust methodologies for handling missing data and selecting relevant features. Moving forward, we remain committed to advancing our project objectives and achieving meaningful outcomes.