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 2)

Date: 8-02-2024 to 15-02-2024

We tried to understand the concept of imputing missing data values and reviewed the related literature. We even explored the existing and commonly used methods for imputing data values in a multimodal data set.

Previous research in imputation techniques for multi-modal datasets, such as the Division I women's basketball dataset, has emphasized robust methods tailored to the dataset's characteristics and research objectives. Researchers have explored various approaches to address missing data effectively.

Imputation Methods:

1. Commonly Used Methods:

- **Mean/Median Imputation**: Replace missing values with the mean or median of observed values. We can use weighted mean for better efficiency, where deciding the weights would be crucial.
- **Forward/Backward Fill:** Propagate the last observed value forward or backward to fill missing values.
- **Linear Interpolation:** Interpolate missing values linearly based on surrounding observed values.
- **Multiple Imputation by Chained Equations (MICE):** Generate multiple imputations by modeling each variable with missing data conditional on others.
- **K-Nearest Neighbors (KNN) Imputation:** Fill in missing values by averaging the values of the K-nearest neighbors.
- **Matrix Factorization Techniques:** Decompose the dataset into latent factors to approximate missing values.

2. Enhancing the quality of imputation using methods as stated below:

- **Feature Engineering:** Preprocessing steps like scaling or transformation to enhance imputation effectiveness.
- **Domain-Specific Considerations:** Incorporating athlete-specific characteristics or performance trends to improve accuracy.
- **Handling Imputation Uncertainty:** Techniques like bootstrapping or sensitivity analysis to manage uncertainty.

3. Assessing Imputation Appropriateness:

Model Evaluation:

Once candidate imputation models are identified, researchers evaluate their performance using appropriate metrics. Standard evaluation metrics include:

Mean Squared Error (MSE): This metric measures the average squared difference between the imputed and actual values. A lower MSE indicates better imputation accuracy.

Correlation Coefficients: Researchers may calculate Pearson's correlation coefficient or similar measures to assess the linear relationship between imputed and actual values. A higher correlation coefficient indicates a more robust agreement between imputed and observed data.

Cross-Validation and Validation:

Researchers often employ cross-validation techniques to assess the generalizability of imputation models. Cross-validation involves partitioning the dataset into multiple subsets, training the imputation model on a subset, and evaluating its performance on the remaining data. This process is repeated numerous times to obtain reliable estimates of model performance.

Additionally, researchers may validate the imputation models on independent datasets to ensure their applicability beyond the original dataset used for model development. Validation on independent datasets helps assess whether the imputation models generalize well to unseen data and different contexts.