Assignment 1

Tanushi Dubey- 22111063

12th July 2024

Question:

Describe 5 data imputation techniques with case study and codes (python).

1) Mean/Median Imputation

Mean or median imputation is a simple method for handling missing numerical data. It involves replacing missing values with the mean (for symmetrically distributed data) or median (for skewed data) of the available data for that feature.
Case Study: Housing Price Prediction Imagine we have a dataset of housing prices in a city, and some houses are missing the 'number of bedrooms' information. We want to predict house prices, so we need to handle these missing values.

2) Multiple Imputation by Chained Equations (MICE)

MICE is a more sophisticated imputation method that creates multiple imputations for each missing value. It works by imputing missing values in a round-robin fashion for each feature, using the other features as predictors.
Case Study: Medical Research on Diabetes Consider a medical dataset with information about patients, including age, BMI, blood pressure, and glucose levels. Some patients have missing values for different features.

3) K-Nearest Neighbors (KNN) Imputation

KNN imputation fills in missing values by finding the K most similar data points (neighbors) and using their values to impute the missing data.
Case Study: Customer Segmentation Imagine a retail company wants to segment its customers based on various attributes, but some customers have missing data for certain features.

4) Regression Imputation

Regression imputation uses the relationship between variables to predict missing values. It involves fitting a regression model on the observed data and using it to predict the missing values.
Case Study: Predicting Employee Salaries A company wants to analyze its employee salary data, but some employees have missing information for years of experience.

5) Random Forest Imputation

Random Forest imputation is an ensemble method that uses multiple decision trees to impute missing values. It can capture complex relationships and interactions between variables.
Case Study: Environmental Data Analysis Scientists are analyzing environmental data, including temperature, humidity, and pollutant levels. Some sensors occasionally fail, leading to missing data.