# MISSING DATA HANDLING

- Disadvantages on ML model
- Techniques
- Resources

Missing data can pose significant challenges in data analysis, especially in the context of machine learning. Some of the key disadvantages are,

- **Reduced accuracy :** Reduce the accuracy of the model's predictions

- **Bias :** It can introduce bias into the model's predictions, leading to incorrect results.

- **Incomplete insights :** Which can make it difficult to draw valid conclusions from the model.

- **Reduced model performance :** Lead to reduced model performance and increased error rates, making it less useful for real-world applications.

- **Difficulty in data interpretation :** Which can make it challenging to make informed decisions based on the model's predictions.

However, with the right strategies and techniques, we can effectively handle missing data and create accurate, high-performance models. Some of them are,

## Deletion

In this method, we simply remove the rows or columns that contain missing data.

**Use case :** When the amount of missing data is small.

## Imputation

In this method, we estimate missing values using statistical techniques. Some popular techniques and their use cases are,

- **Mean / Median imputation :** When the missing values are relatively small in number.

- **Mode imputation :** When the data is categorical in nature and the majority of the values are missing.

- **K-Nearest neighbor imputation :** When there is a clear pattern to the missing data and the dataset is not too large.

## Regression imputation :

This method involves using a regression model to estimate missing values based on the values of other variables.

**Use case :** When there is a linear relationship between the variable with missing values and other variables in the dataset.

## Multiple imputation

This method involves creating multiple imputed datasets and averaging the results to obtain a final dataset.

**Use case :** When there is a moderate amount of missing data and the missing values can be imputed using a regression model based on the other available variables.

When dealing with missing data, there are several key considerations that should be kept in mind. Here are some of the most important considerations:

- **Understand the nature and pattern of missing data :** Different types of missing data require different handling methods.

- **Assess the percentage of missing data :** High percentages of missing data can negatively impact the accuracy of the model.

- **Determine the reason for missing data :** The reason for missing data can help in choosing the appropriate handling technique.

- **Avoid deleting too many records with missing data :** Removing too many records can result in loss of valuable information.

- **Choose appropriate imputation method :** Different imputation techniques are suitable for different scenarios.
- **Assess the impact of imputation on the model :** Imputing missing data can introduce bias in the model, and it is important to assess the impact.

- **Consider using a combination of techniques :** Multiple imputation and other hybrid techniques can be used for handling missing data.

- **Validate the imputed data :** The imputed data should be validated to ensure it is accurate and reliable.

- **Document the missing data handling process :** Keeping a record of the missing data handling process can help in reproducibility and transparency of the model.

Resources for handling missing data by using Python packages such as,

- **Pandas :** It provides various functions such as dropna(), fillna(), and interpolate() for handling missing data.

- **NumPy :** It provides the np.nan value to represent missing data and various functions to handle them such as isnan(), sum(), and mean().

- **SciPy :** It provides the interpolate() function for handling missing data.

- **Scikit-learn :** It provides the SimpleImputer() class and IterativeImputer() class for handling missing data.

- **Datawig :** It is a Python library that uses deep learning models for imputing missing data.

- **KNNImputer :** It is a scikit-learn imputer that imputes missing values using k-nearest neighbors.

- **XGBoost :** It is an implementation of the gradient boosting machine learning algorithm that can handle missing data.

In conclusion, missing data can be a challenge in machine learning, but by using the right strategies and techniques, we can effectively handle missing data and create accurate, high-performance models.

# THANK YOU

@ Gangadhar Neelam