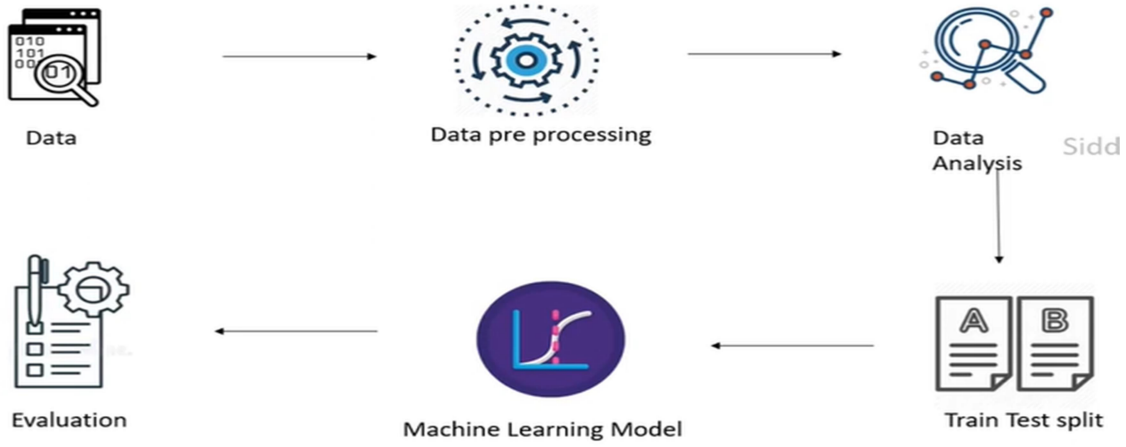
Data Collection and Pre-Processing in Machine Learning

Overview

This study guide focuses on the fourth module of the Machine Learning course, which covers **Data Collection** and **Data Pre-Processing**. Understanding these concepts is crucial for building effective machine learning models.



Key Concepts

1. Machine Learning Workflow

- **Data Collection**: The first step in any machine learning project is to gather relevant data based on the problem statement. For example, to predict diabetes, we need medical data from diabetic and non-diabetic individuals.
- **Data Pre-Processing**: After collecting data, it must be processed before feeding it into a machine learning algorithm. This includes handling missing values and normalizing data.

2. Data Analysis

- Analyse the data to gain insights. This involves identifying important features that contribute to predictions.

3. Train-Test Split

- **Definition**: The train-test split is a method to divide the dataset into two parts: training data (used to train the model) and testing data (used to evaluate the model's performance).
- **Typical Split Ratio**: Commonly, 80% of the data is used for training and 20% for testing.
- **Importance**: Testing data is crucial for evaluating how well the model generalizes to unseen data. It ensures that the model is not just memorizing the training data.

4. Implementation Example

(Check colab file uploaded in same folder for clear information and code)

- **Using Python's train_test_split**:
    - Import the function from sklearn.model_selection.
    - Split the dataset into features (X) and target (y).
    - Specify the test size (e.g., test_size=0.2 for 20% test data) and a random state for reproducibility 6.

5. Example Code Snippet

```
from sklearn.model_selection import train_test_split


# Assuming X and y are defined

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=2)
```

This code splits the dataset into training and testing sets, ensuring that the model can be evaluated effectively 7.

Important Takeaways

- Always preprocess your data before training a model.
- Use train-test split to evaluate model performance accurately.
- Analyze data to identify key features that influence predictions.

Questions to Consider

- What steps are involved in data pre-processing?
- Why is it important to have a separate test dataset?
- How do you implement train-test split in Python?