BAZE UNIVERSITY ABUJA IN PARTNERSHIP WITH DOMINEUM

IDEAS Emerging Technology Skills Scholarship Program

Data Processing

Presented by: Khadijah Saad Mohammed

SUPPORTED BY THE WORLD BANK NITDA

Innovation Development
and Effectiveness
in the Acquition of Skills

Introduction to Data Processing

Data processing in machine learning involves preparing raw data into a suitable format that enhances the performance of machine learning models.

Importance of Data Processing.

Effective data processing is crucial as it directly impacts the accuracy and efficiency of predictive models by ensuring clean, relevant, and well-formatted data

Preparing Data and Engineering Features

Data Pre-processing

- Cleaning: Removing errors and inconsistencies from the data.
- Transformation: Scaling or normalizing features to ensure uniformity.
- Encoding: Converting categorical variables into numerical representations.
- Feature Engineering: Creating new features or transforming existing ones to improve model performance.

Data Cleaning

Definition

Data cleaning involves identifying and correcting inaccuracies and inconsistencies in data to improve its quality.

Handling Missing Values

Techniques such as imputation (filling missing values with statistical measures like mean or median) or removal of records ensure completeness..

Dealing with Duplicates

Identifying and removing duplicates to prevent biased machine learning outcomes.

Correcting Errors

Fixing data entry errors and outliers to maintain data integrity.

.

Data Transformation

Transforming data normalizes scales and formats data into a uniformly understandable format for algorithms

Scaling and Normalization: Standardizing features to a uniform scale.

Encoding categorical data: Converting categories to numerical values.

Transforming dates and text: Extracting usable features from complex data types.

Feature Engineering

Feature engineering creates predictive features from data, significantly impacting model accuracy.

Creating Interaction Features

- Combining two or more features to capture interaction effects not observed independently..

Dimensionality Reduction

- Reducing the number of input variables using PCA to simplify the model without losing essential information.

Data Splitting

.

Reason for Splitting

- ○ Data is split into separate sets to train models, validate accuracy, and test before deployment.

Methods of Splitting

- ○ Random, stratified, and time-based splits cater to different types of data and distribution needs.

Cross-Validation

- ○ Using techniques like k-fold cross-validation to validate model performance with limited data

Machine Learning - Data-Driven

1. Collect data of images and their labels/classes
2. Use machine learning to train this to get a classifier
3. Test this classifier on a different set of data (Excluding the ones used for training)

CONCLUSION

- Recap:
    - Proper data processing is foundational to successful machine learning, ensuring clean, relevant, and well-structured data.
    - Embrace data processing practices to enhance your predictive modeling efforts

1.

BAZE UNIVERSITY ABUJA

IN PARTNERSHIP WITH

DOMINEUM

THANK YOU

Q&A