DAILY WORK REPORT

TR-02

INFOWIZ

15 JUNE 2024

Day 9: Data Cleaning and Preprocessing

Summary: Today's focus was on mastering essential techniques in data cleaning and preprocessing using Pandas and NumPy, crucial steps in preparing raw data for machine learning tasks. We explored strategies to handle missing data, encode categorical variables, and scale numerical features, ensuring data quality and compatibility with machine learning algorithms.

Key Learnings:

1. **Handling Missing Data:**
    - **Identifying Missing Values:** Utilized Pandas methods such as isnull(), notnull(), and info() to identify missing data entries (NaN values) within datasets.
    - **Strategies for Handling Missing Data:** Implemented techniques like data imputation using mean, median, or mode values fillna method, or dropping rows/ columns with missing data dropna() method, depending on dataset characteristics and analysis requirements.

2. **Categorical Data Encoding:**
    - **One-Hot Encoding:** Explored one-hot encoding technique using Pandas' get_dummies() function to convert categorical variables into binary indicator variables, enabling machine learning models to interpret categorical data appropriately.
    - **Label Encoding:** Learned about label encoding using scikit-learn's LabelEncoder for converting categorical labels into numeric form, suitable for algorithms that require numerical inputs.

3. **Feature Scaling:**
    - **Standardization and Normalization:** Implemented feature scaling techniques using scikit-learn's StandardScaler and MinMaxScaler to standardize numerical features to a common scale, mitigating the impact of feature magnitude differences on model performance.
    - **Application in Machine Learning:** Discussed the significance of feature scaling in improving model convergence speed, performance, and interpretability across various machine learning algorithms like logistic regression, SVMs, and neural networks.

4. **Practical Applications:**
    - Applied data cleaning and preprocessing techniques to real-world datasets:
        - Addressed missing values through appropriate imputation strategies.
        - Converted categorical variables into a suitable format for machine learning models using one-hot encoding and label encoding.
        - Standardized numerical features to ensure consistent input ranges for model training and evaluation.