Title

ai-gen.pdf

Address Quality Classification

Objective:

The objective of this assignment is to classify customer addresses into three categories: good, medium, and bad. Candidates can use any open-source data or API to enhance the address data and perform the classification.

Assignment Description:

You are provided with a dataset containing customer addresses. Your task is to classify each address into one of three categories based on the quality and completeness of the address information.

Dataset:

You will receive a CSV file containing the following columns:

1. `index`: Unique identifier for each customer
2. `Address`: Address of the customer
3. `City`: The city of the address
4. `State`: The state of the address
5. `Pincode`: The postal code of the address

Tasks:

1. Data Cleaning:
    - Handle missing values in the dataset.
    - Standardize the address format (e.g., ensuring consistent capitalization, removing special characters, etc.).
2. Data Enhancement:
    - Use any open-source data or API (OpenStreetMap API) to validate and enhance the

address data.

3. Feature Engineering:

- Create features that can help in classifying the addresses. Some possible features include:

- Address completeness (whether all fields are filled).

- Address validity (whether the address exists).

4. Classification Model:

- Define the criteria for classifying addresses into good, medium, and bad categories. For example:

- **Good**: Address is complete, valid

- **Medium**: Address is mostly complete, may have minor validation issues, - **Bad**: Address is incomplete, invalid

- Build a classification model to categorize the addresses based on the defined criteria. You can use any classification algorithm (e.g., decision trees, random forests, logistic regression).

5. Evaluation:

- Evaluate the performance of your classification model using appropriate metrics (e.g., accuracy, precision, recall).

Deliverables:

1. Python or Jupyter Notebook containing your code and analysis.

2. The enhanced and classified dataset in CSV format.