Data Processing Requirements (DPR) Document

Phishing Detection System

Table of Contents

---

1. Introduction

The Data Processing Requirements (DPR) document outlines the data processing pipeline and requirements for the project. It describes the steps involved in data ingestion, validation, transformation, and storage to ensure reliable and accurate predictions.

2. Data Ingestion

- The system should support data ingestion from various sources, such as CSV files, databases, or real-time API integration.
- The data ingestion process should include mechanisms for validating the data integrity, consistency, and format.
- Error handling and logging should be implemented to track any issues or anomalies during the data ingestion phase.

3. Data Validation

- System perform data validation checks to ensure the correctness and quality of input data which were pre-defined requirements.
- Common validation checks include checking for missing values, data type consistency, range validation, and outlier detection.

4. Data Transformation

- Data transformation step prepare the data for model training and prediction.

- This involve feature engineering, normalization, encoding categorical variables, and handling missing values.
- Performed data up-sampling on target feature to balance the data and ensure that the data doesn't create over-fitting or bias to ML model.

5. Data Storage

System uses MongoDB to store data for this project.

6. Conclusion

The Data Processing Requirements (DPR) document outlines the key considerations and steps involved in data processing for the project. It covers data ingestion, validation, transformation, storage, privacy, security, and monitoring aspects. This document serves as a reference for the development team to ensure effective and reliable data processing within the system.