# Model Input Data Documentation

Project Name: [Insert Project Name]

Model Name: [Insert Model Name]

Date: [Insert Date]

Version: [Insert Version Number]

**1. Introduction**

This document provides detailed information about the input data used for training and validating the [Insert Model Name] within the [Insert Project Name]. It ensures that the data meets governance, quality, and compliance requirements, and serves as a reference for data provenance, preprocessing, and usage in model development.

**2. Purpose and Scope**

**2.1 Purpose**

The purpose of this documentation is to:

* Describe the data sources, formats, and structures used as input for the [Insert Model Name].
* Outline the preprocessing steps applied to the data before model training and validation.
* Ensure that the data adheres to governance, quality, and compliance standards.
* Provide a reference for the reproducibility of model training and validation processes.

**2.2 Scope**

This documentation covers all datasets used in the training and validation of the [Insert Model Name], including structured, unstructured, and semi-structured data. It includes details on data collection, preprocessing, storage, and usage within the model development pipeline.

**3. Data Sources**

**3.1 Data Source Overview**

This section provides a comprehensive overview of the data sources used in the [Insert Model Name].

* Source Name: [Insert Data Source Name]
* Data Provider: [Internal/External; specify the original provider or system]
* Data Type: [Structured/Unstructured/Semi-structured]
* Description: [Brief description of the data source and its relevance to the model]
* Origin: [e.g., Database, API, Third-Party Vendor, Internal Collection]
* Update Frequency: [e.g., Real-time, Daily, Weekly, Monthly]
* Data Owner: [Insert Data Owner Name]
* Format: [e.g., CSV, JSON, XML, SQL Table]

**3.2 Data Acquisition**

This section describes how data is acquired from the sources mentioned above.

* Acquisition Method: [e.g., API Calls, Data Pipelines, Manual Uploads]
* Ingestion Tools: [e.g., ETL Tools, Data Integration Platforms]
* Acquisition Schedule: [e.g., Scheduled, On-Demand]
* Data Collection Date Range: [Specify the date range of the data collected]
* Data Storage Location: [Insert Location, e.g., Data Warehouse, Data Lake, Cloud Storage]

**4. Data Characteristics**

**4.1 Data Structure**

This section provides detailed information on the structure of the datasets used.

* Dataset Name: [Insert Dataset Name]
* Number of Records: [Insert Number of Records]
* Number of Features/Columns: [Insert Number of Features/Columns]
* Feature Description:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature Name | Data Type | Description | Example Values | Source |
| [Feature 1] | [Type] | [Description] | [Example Values] | [Source] |
| [Feature 2] | [Type] | [Description] | [Example Values] | [Source] |
| [Feature 3] | [Type] | [Description] | [Example Values] | [Source] |
| [Feature 4] | [Type] | [Description] | [Example Values] | [Source] |

**4.2 Data Volume**

This section describes the size and volume of the data used in training and validation.

* Total Records: [Insert Total Number of Records]
* Total Data Size: [Insert Data Size, e.g., GB, TB]
* Training Data Volume: [Insert Number of Records/Size for Training Data]
* Validation Data Volume: [Insert Number of Records/Size for Validation Data]
* Test Data Volume (if applicable): [Insert Number of Records/Size for Test Data]

**5. Data Quality and Integrity**

**5.1 Data Quality Checks**

This section outlines the data quality checks applied to ensure that the data is fit for use in the model.

* Accuracy: [Describe methods used to ensure data accuracy]
* Completeness: [Describe checks for missing values or incomplete records]
* Consistency: [Describe how data consistency is maintained across sources]
* Validity: [Describe how data validity is ensured, including format checks]
* Timeliness: [Describe how data timeliness is maintained]

**5.2 Data Cleaning and Preprocessing**

This section describes the data cleaning and preprocessing steps performed before model training.

* Missing Data Handling: [e.g., Imputation, Removal, Filling with Defaults]
* Outlier Detection and Treatment: [e.g., Z-Score, IQR Method, Removal]
* Data Normalization/Standardization: [e.g., Min-Max Scaling, Z-Score Standardization]
* Categorical Encoding: [e.g., One-Hot Encoding, Label Encoding]
* Feature Engineering: [Describe any feature creation or transformation steps]

**6. Data Governance and Compliance**

**6.1 Data Governance Policies**

This section ensures that the data used meets governance requirements.

* Data Ownership: [Describe data ownership and any relevant licensing agreements]
* Data Access Controls: [Describe who has access to the data and how access is managed]
* Data Lineage: [Briefly describe the lineage of the data from source to usage in the model]
* Audit Trails: [Describe any audit trails maintained for data access and modification]

**6.2 Compliance with Regulations**

This section outlines the compliance of the data with relevant regulations and standards.

* GDPR Compliance (if applicable): [Describe how GDPR requirements are met]
* HIPAA Compliance (if applicable): [Describe how HIPAA requirements are met]
* Data Privacy Measures: [e.g., Anonymization, Pseudonymization, Data Masking]
* Consent Management: [Describe how consent for data usage is managed and documented]

**7. Data Usage in Model Development**

**7.1 Training Data**

This section details the data used for training the model.

* Training Data Description: [Describe the data used for training, including any specific preprocessing applied]
* Training Data Source: [List the specific datasets used for training]
* Data Split Method: [e.g., Random Split, Stratified Split, Time-Based Split]
* Training Data Size: [Insert Size or Number of Records]

**7.2 Validation Data**

This section details the data used for validating the model.

* Validation Data Description: [Describe the data used for validation, including any specific preprocessing applied]
* Validation Data Source: [List the specific datasets used for validation]
* Validation Method: [e.g., Cross-Validation, Hold-Out Validation]
* Validation Data Size: [Insert Size or Number of Records]

**7.3 Test Data (if applicable)**

This section details the data used for testing the model.

* Test Data Description: [Describe the data used for testing, including any specific preprocessing applied]
* Test Data Source: [List the specific datasets used for testing]
* Test Method: [e.g., Final Hold-Out Test, Cross-Validation]
* Test Data Size: [Insert Size or Number of Records]

**8. Data Lineage and Traceability**

This section outlines the data lineage and traceability from source to model input.

* Data Lineage Diagram: [Include or reference a diagram showing the data flow from source to final usage in the model]
* Data Provenance: [Describe the origin of the data and any transformations applied along the way]
* Traceability: [Describe how data changes are tracked and documented]

**9. Data Security**

This section describes the security measures in place to protect the data.

* Data Encryption: [Describe encryption methods used for data at rest and in transit]
* Access Controls: [Describe the roles and permissions associated with data access]
* Data Breach Protocols: [Outline the steps to be taken in the event of a data breach]
* Data Backup: [Describe the backup procedures for the data]

**10. Document Control**

* Document Owner: [Insert Name, Role]
* Approval Date: [Insert Date]
* Next Review Date: [Insert Date]
* Version History:
  + Version [Insert Version Number] - Initial Document - [Insert Date] - Approved by [Insert Name]