

**Horizon Europe**

**Data Management Plan Template**

**Version 1.0**

**15 April 2025**

|  |  |  |
| --- | --- | --- |
| **HISTORY OF CHANGES** | | |
| **Version** | **Publication date** | **Changes** |
| 1.0 | 15.04.2025 | * Initial version |
|  |  |  |

**Action Number:** [insert project reference number]

**Action Acronym:** FD-DPHA

**Action title:** Fair Dataset for Disease Prediction in Healthcare Applications

**Date:** 15-04-2025

**DMP version:** 1.0

The Horizon Europe Model Grant Agreement requires that a data management plan (‘DMP’) is established and regularly updated. The use of this template is recommended for Horizon Europe beneficiaries. In completing the sections of the template the requirements for research data management of Horizon Europe as described in article 17 and analysed in the Annotated Grant Agreement, article 17, must be addressed.

## Data Summary

This project does not reuse any existing external datasets. Instead, a new dataset was uploaded to TU Wien's DBRepo system, created for the specific purpose of developing and evaluating a machine learning model for disease risk prediction based on health metrics.

The dataset is in CSV format, and contains structured tabular data with the following fields: age, blood\_pressure, cholesterol, bmi, glucose, smoker, exercise\_hours, and disease\_risk. It was divided into training, validation, and test subsets, each assigned a unique Persistent Identifier (PID) by DBRepo.

The purpose of generating this data is to train, evaluate, and validate a supervised machine learning classifier (Random Forest) that can predict an individual's disease risk based on common health indicators. This directly supports the project’s goal of demonstrating FAIR-compliant data handling in applied ML experiments.

The total size of the dataset (all splits combined) is approximately under 1MB, making it lightweight and suitable for educational and experimental use.

The origin of the data is synthetic or anonymized health data uploaded by the student, with no personal or sensitive information involved. Metadata and descriptive information were added during the upload process via the DBRepo graphical interface.

This dataset could be useful for:

Students and educators looking for small, interpretable datasets for ML training and evaluation,

Researchers building lightweight health-prediction models,

Anyone interested in learning how to apply the FAIR principles in a machine learning context.

## FAIR data

### **Making data findable, including provisions for metadata**

All datasets used in this project are assigned a Persistent Identifier (PID) by DBRepo, TU Wien’s institutional research data repository. These PIDs allow for permanent, citable, and traceable references to each dataset (training, validation, and test sets).

Rich metadata has been provided during the upload process through the DBRepo graphical interface. This includes the dataset title, author, license (CC BY 4.0), description, keywords, creation date, and format. Metadata is aligned with FAIR4ML, CodeMeta, and Croissant standards, which are compatible with both disciplinary and general metadata practices.

Keywords such as “machine learning,” “disease prediction,” “health data,” and “classification” were added to ensure the datasets are easily discoverable and reusable by others. All metadata is structured in a way that it can be harvested, indexed, and retrieved via DBRepo’s API and internal search infrastructure. This supports both human-readable and machine-actionable discovery mechanisms.

### Making data accessible

All datasets are deposited in TU Wien’s DBRepo, which is a trusted institutional repository designed for long-term research data preservation. The repository automatically assigns a Persistent Identifier (PID) to each uploaded dataset, and these identifiers are resolvable, linking directly to the digital object through the DBRepo portal.

The data generated for this project is openly accessible, as it does not contain any sensitive, personal, or proprietary information. There are no legal, ethical, or contractual restrictions that would limit its availability. No embargo period has been applied, and the data is available immediately after submission to the repository.

Access to the datasets is provided through standardized and free access protocols offered by the DBRepo API and its web-based interface. Authentication via TU credentials is required for private access; however, this project’s data is openly shared, and no additional authentication or data access committee is required.

Metadata is published under a CC0 (public domain dedication) license, as per the Grant Agreement guidelines. Metadata includes all necessary information to locate, identify, and reuse the datasets, including titles, authorship, licenses, data descriptions, keywords, creation date, and file format. Even if the underlying data were to be removed, metadata will remain persistently available to ensure that discovery and citation remain possible.

Documentation and any necessary software references (e.g. Python version, required libraries such as scikit-learn, pandas, etc.) are included in the project's GitHub repository. The code used for data processing and model training is available as open-source and can be reused to replicate or extend the analysis.

### Making data interoperable

The dataset is provided in CSV format, a widely accepted and interoperable format that facilitates reuse across disciplines and software platforms. Metadata follows the FAIR4ML, CodeMeta, and Croissant standards, which are community-endorsed for describing machine learning models, datasets, and code in a structured, machine-readable format.

The use of standard, self-explanatory column names (e.g., age, glucose, disease\_risk) ensures that the data is understandable and processable without the need for domain-specific vocabularies. No uncommon or project-specific ontologies were created. Therefore, no mappings to external vocabularies are necessary. If future refinement requires controlled vocabularies, appropriate mappings (e.g. to HL7, SNOMED CT) could be added in later iterations.

Qualified references to related data are included: each data subset (train, validation, test) is linked by its PID, and these are cited within the model’s metadata. This supports clear traceability and enables reusers to understand the full data lineage.

### Increase data re-use

To ensure reproducibility and reusability, the project provides detailed documentation, including a README file that describes the data structure, methodology, feature descriptions, units of measurement, and code used for model training and evaluation. All preprocessing, imputation, model parameters, and evaluation procedures are transparently documented in the Jupyter Notebook and the GitHub repository.

All datasets and outputs (trained model, evaluation metrics, visualizations, and predictions) are made freely available under the CC BY 4.0 license, permitting broad reuse with attribution, in accordance with the Grant Agreement. Metadata is published under CC0, enabling unrestricted indexing and reuse.

The data and outputs will remain available and usable by third parties after the end of the project, as they are hosted on TU Wien’s trusted repositories (DBRepo and TUWRD) with persistent identifiers and long-term availability guarantees. The provenance of the data is thoroughly documented, referencing the original source PIDs for each subset (training, validation, test), along with the version of the code and libraries used.

To ensure data quality, all inputs were validated for missing values, outliers, and type consistency. Quality assurance steps included imputation, label encoding, and performance validation using classification metrics and visualizations like confusion matrices and feature importance plots.

This DMP also incorporates considerations related to other research outputs (e.g., models, visualizations), allocation of resources, security, and ethics, ensuring full alignment with the FAIR principles and the Grant Agreement obligations.

## Other research outputs

In addition to structured datasets, this project generates several digital research outputs: a trained Random Forest model (.pkl), performance evaluation metrics (.json), visualizations (confusion matrix and feature importance .png files), and prediction results (recommendations.csv). These outputs support transparency, reproducibility, and reuse of the machine learning workflow.

All outputs have been uploaded to TUWRD, TU Wien’s research data repository, and are accompanied by rich metadata that adheres to the FAIR4ML, CodeMeta, and Croissant metadata standards. This ensures that the research outputs are findable, accessible, interoperable, and reusable.

Outputs are licensed under CC BY 4.0, allowing others to reuse them freely with appropriate attribution. Each file is documented with information on how it was generated, including the source datasets (with PIDs), software versions, and model configuration. All relevant code has been published in a public GitHub repository, allowing others to reproduce or extend the experiment.

No physical outputs were produced in this project. All digital outputs are stored in a standardized format and made accessible via open repositories, with long-term findability ensured through persistent identifiers and metadata retention.

## Allocation of resources

The costs associated with making the data and other research outputs FAIR in this project were minimal to none, as the infrastructure and tooling were provided by TU Wien. Services such as DBRepo for dataset management and TUWRD for storing research outputs were used at no cost to the student, and no external archiving or commercial platforms were required.

No direct or indirect costs were incurred for data storage, sharing, or licensing. All work related to data curation, metadata generation, documentation, and FAIR compliance was performed by the student (Sufiyan Yousaf) as part of the course assignment.

Data management responsibilities were carried out entirely by the student, including uploading data to repositories, ensuring metadata completeness, and complying with repository and project standards. Guidance was taken from the course materials and TU Wien documentation.

Long-term preservation is ensured by the institutional repositories—DBRepo and TUWRD—which support persistent identifiers (PIDs), public metadata, and long-term hosting. These repositories are managed by TU Wien and designed to preserve research data and outputs for extended periods, even after project completion. The data and outputs will be retained indefinitely unless otherwise specified by the repository’s retention policies.

## Data security

All data is stored and managed using trusted institutional repositories: DBRepo for dataset storage and TUWRD for derived research outputs. These platforms are maintained by TU Wien and offer secure storage, access control, versioning, and backup mechanisms to prevent data loss or corruption.

Although no sensitive or personal data is used in this project, secure access (including authentication via TU credentials and VPN access) was used during upload and management. All transfers occurred over secure HTTPS connections. In the event of repository maintenance or failures, institutional backup and recovery policies ensure data integrity and long-term availability.

## **Ethics**

There are no ethical or legal concerns related to data sharing in this project. The data used is fully anonymized or synthetically generated and contains no personal identifiers or sensitive information.

Therefore, informed consent was not required, and no ethical clearance was needed. The dataset is used purely for academic, non-commercial purposes as part of a university assignment. Ethical compliance with institutional data handling norms was ensured.

## Other issues

This project follows the data management procedures outlined by TU Wien and adheres to guidelines provided by the course and repositories (DBRepo and TUWRD). No additional national, funder-specific, or sectoral data policies were required.

All repository use, metadata creation, and FAIR compliance were aligned with TU Wien infrastructure standards and supported documentation.