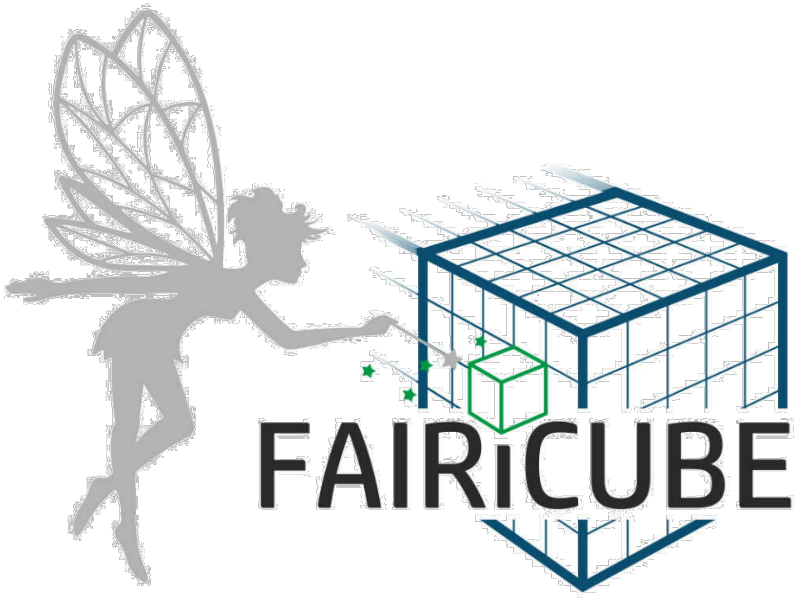

FAIRICUBE – F.A.I.R. INFORMATION CUBES

Work Package 3: Process
Milestone 6: Machine learning strategy specific for each use case released

Deliverable Lead: NIL
Deliverable due date: 30/06/2023

Version: 1.0
2024-11-21

Document Control Page

Document Control Page	
Title	Machine learning strategy specific for each use case released
Creator	Stefan Jetschny
Description	M6 Machine learning strategy specific for each use case released
Publisher	"FAIRICUBE – F.A.I.R. information cubes" Consortium
Contributors	Mohamed-Bachir Belaid
Date of delivery	30/06/2023
Type	Text
Language	EN-GB
Rights	Copyright "FAIRICUBE – F.A.I.R. information cubes"
Audience	☒ Public ☐ Confidential ☐ Classified
Status	☐ In Progress ☒ For Review ☐ For Approval ☐ Approved

Revision History			
Version	Date	Modified by	Comments
0.1	27/06/2023	Stefan Jetschny	Initial draft with headlines
0.2	28/08/2023	Stefan Jetschny	Full draft for internal review
1.0	30/08/2023	Stefan Jetschny, Mohamed-Bachir Belaid, Kathi Schleidt	Internal review and minor updates, draft for external review
1.0	13/10/2023	Jaume Targa	Final review performed

Disclaimer

This document is issued within the frame and for the purpose of the FAIRICUBE project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101059238. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.

Table of Contents

List of Tables

1 Introduction

The use cases executed under the FAIRiCUBE Hub and during the project duration are not formally a key delivery from the project but nevertheless serve a significant role: to test, develop, improve, and co-define the FAIRiCUBE hub which bundles all cloud-based services to be executed while solving typical data science research questions. As the use cases (UCs) are truly diverse, operate on different scale length & time ranges and cover various scientific fields, we thereby cover a wide range of potential data science tasks. Providing all online services such as accessing, ingesting, storing, processing, and sharing data, in a FAIR, efficient, streamlined, and user-friendly way is a key project component and only by executing our UCs, we can determine which components of the FAIRiCUBE Hub are already sufficiently working and where is room for improvement.

As part of the data science work in WP3 and in close collaboration with the WP2 domain experts, we have first studied the nature and characteristics of the data identified by the UC owners as being relevant to solve the UC research question(s). This was achieved as part of M5 "*Use cases exploratory data analysis released"* and is effectively the basis of M6. Driven by this exploratory data analysis, a machine learning strategy was developed for each use case which efficiently balances the UC needs, the appropriate machine learning algorithm and the computational efforts and resources required. After all, the background of applying a machine learning algorithm is to identify and exploit scientific relationships without using strict scientifical equations. Instead, an ML algorithm is using relevant input data, either in a raw state or processed as features, to approximate the desired output. Through a training phase, each ML algorithm is automatically determining its own parametrization to perform this prediction. Finally, a good data science practice is to start with simple, fast and efficient ML models/algorithms, then use advanced models only if needed, e.g., when the accuracy of the prediction is not meeting the expectations. This approach will allow for the creation of an initial ML model that tests the suitability of the input data, the feature engineering, i.e., the calculation of predictors from the input data, and ability to generalize the problem of using predictors/features to approximate the output/target. The initial ML mode is serving as a baseline model and further tuning of the ML model or advancing to more advanced ML method can be compared against this baseline model both in terms of increased efforts to train the model and improved accuracy.

Scaling and costs of executing the ML method will also impact the development of an ML strategy. In addition to solving the UC problems, we also collect and make available documentation on "which" ML method was chosen "why", how well it performs and scales, and how much resources are needed. The documentation will either be captured as processing meta data (WP4, task *Enable FAIRiCUBE Processing/Analysis*) or in the knowledge base (WP3, task *Processing knowledge base*) both accessible as applications of the FAIRiCUBE Hub (WP4, task *Community collaboration platform*). This knowledge transfer is a valuable input to future data science applications and will help to plan and manage the utilization of ML methods in future projects (not limited to the FAIRiCUBE Hub).

Further, milestones M5 and M6 function as indicators of UC progress, and simultaneously demonstrate the data science and machine learning capabilities enabled by the FAIRiCUBE Hub.

2 Deliverables contributing to M6

There are two formal deliverables contributing to Milestone 6, deliverable D3.2 and D3.3 as listed in Table 1. Note that the UC exploratory analysis describes the current state each UC, with updates foreseen as the UC execution progresses.

Description	Lead Beneficiary	Type	Dissemination level	Due date
D3.2 Machine learning strategy specific for each use case	NIL	R	Public	30.06.2023
D3.3 Processing and ML applications	NIL	R	Public	30.06.2023

Table 1: Formal deliverables contributing to M6.

Bundled in this milestone is the documentation of both the algorithmic approach as well as the resource consumption. The algorithmic approach documented in D3.2 briefly reviews the research question, the findings from the exploratory data analysis and describes the most appropriate initial data processing approaches, including the application of ML algorithms. Suggestions are made how to further improve the accuracy of prediction by advancing to more demanding ML methods. The resource documentation in D3.3 both describes how we measure the computational efforts and the actual resources needed to execute the data science tasks as described in D3.2. The listings are not limited to the ML applications – even though this potentially being a dominant task – but additionally covers data processing steps, e.g., during the feature engineering step. As part of the work in D3.3, we aim to standardize the way we measure resources across the different use cases and provide programming solutions for automatization of the collection of resource metrics. Some of the collected metrics, such as monetary cost of execution and CO2 footprint, are forward looking to enable cloud resource budgeting as well as environmental assessment of the data science processing. In return, the measures collected in D3.3 can affect the choice of ML method described in D3.2. In fact, the measures will give insight on whether the ML model should be tunned/improved to have a more optimal runtime, available computational resources/costs, environmental impact, etc.

Together, D3.2 and D3.3 give insights into what processing and ML algorithms have been applied to the input data and which resources were consumed to provide an answer to the UC specific research questions.

3 Summary

Each use case (UC) has defined domain-specific research questions, documented as use case analysis plans, and they are being addressed as data science tasks. As part of the data science work and following the thorough analysis of the suitability and characteristics of the input data (exploratory data analysis) a use case specific machine learning strategy has been developed and partly executed. The result of this data science processing, including the application of machine learning algorithms, is a snapshot of the progress of each UC and will be updated during the project duration as required. Key elements of the milestone will be made available as part of the UC scientific publishing task and on our project website.