D5.1 Verification and Validation Plan

Deliverable 5.1

Dispatcher3

Grant:	**886461**
Call:	**H2020-CS2-CFP10-2019-01**
Topic:	**JTI-CS2-2019-CfP10-SYS-01-16**
Consortium coordinator:	**University of Westminster**
Dissemination level:	**Public**
Edition date:	**31 January 2022**
Edition:	**01.01**

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Julia De Homdedeu / Universitat Politècnica de Catalunya	Project member	11 August 2021
Jovana Kuljanin / Universitat Politècnica de Catalunya	Project member	11 August 2021
Damir Valput / Innaxis	Project member	11 August 2021
Luis Delgado / University of Westminster	Project member	11 August 2021

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
Luis Delgado / University of Westminster	Project member	11 August 2021
Damir Valput / Innaxis	Project member	11 August 2021
Ralph Schultz / PACE	Project member	11 August 2021
Héctor Fornes / Vueling Airlines	Project member	11 August 2021
Saranne Verhellen / skeyes	Project member	11 August 2021

Approved for submission to the CSJU

Beneficiary	Position	Date
University of Westminster	Project coordinator	31 January 2022

Document History

Edition	Date	Status	Author	Justification
01.00	11 August 2021	Release	Dispatcher3 Consortium	New document for review by Topic Manager
01.01	31 January 2022	Release	Dispatcher3 Consortium	Accepted by Topic Manager as is (no changes with respect to 01.00)

The opinions expressed herein reflect the authors' view only. Under no circumstances shall the Commission or Clean Sky Joint Undertaking be responsible for any use that may be made of the information contained herein.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

EUROPEAN UNION

Dispatcher3

INNOVATIVE PROCESSING FOR FLIGHT PRACTICES

This deliverable is part of a project that has received funding from the Clean Sky Joint Undertaking under grant agreement No 886461 under European Union's Horizon 2020 research and innovation programme.

Abstract

In this deliverable, we present a verification and validation plan designed to carry out all necessary activities along Dispatcher3 prototype development. Given the nature of the project, the deliverable points to a data-centric approach to machine learning that treats training and testing models as an important production asset, together with the algorithm and infrastructure used throughout the development. The verification and validation activities will be presented in the document.

The proposed framework will support the incremental development of the prototype based on the principle of iterative development paradigm. The core of the verification and validation approach is structured around three different and inter-related phases including data acquisition and preparation, predictive model development and advisory generator model development which are combined iteratively and in close coordination with the experts from the consortium and the Advisory Board. For each individual phase, a set of verification and validation activities will be performed to maximise the benefits of Dispatcher3. Thus, the methodological framework proposed in this deliverable attempts to address the specificities of the verification and validation approach in the domain of machine learning, as it differs from the canonical approach which are typically based on standardised procedures, and in the domain of the final prospective model. This means that the verification and validation of the machine learning models will also be considered as a part of the model development, since the tailoring and enhancement of the model highly relies on the verification and validation results.

The deliverable provides an approach on the definition of preliminary case studies that ensure the flexibility and tractability in their selection through different machine learning model development.

The deliverable finally details the organisation and schedule of the internal and external meetings, workshops and dedicated activities along with the specification of the questionnaires, flow-type diagrams and other tool and platforms which aim to facilitate the validation assessments with special focus on the predictive and prospective models.

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Table of Contents

Abstract **3**
Executive summary **6**
1 Introduction 9
1.1 Overview of the verification and validation activities **10**
1.2 Technological Readiness Level and context within **11**
1.3 Deliverable structure **12**
2 Verification and validation concept and approach 14
2.1 Verification and validation phases approach **17**
2.2 Management and tracking of the verification and validation activities **20**
3 Data acquisition and preparation verification and validation 22
3.1 Cleaning and preparation **23**
3.2 Labelling and feature engineering **25**
4 Preliminary data requirements 26
4.1 Machine learning algorithm selection **26**
4.2 Training the model and cross-validation **27**
4.3 Model validation **32**
5 Prospective model verification and validation 37
5.1 Verification and Integration **37**
5.2 Validation **38**
6 Research Questions 41
7 Scenarios and case studies 49
7.1 Methodology to define case studies **50**
7.2 Potential values for definition of case studies **51**
8 Schedule of development and verification, integration and validation 53
9 Conclusions 56
10 Next steps and look ahead 58
11 References 60
12 Acronyms 61
Appendix A - Questionnaires for validation 63

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

EUROPEAN UNION

List of figures

Figure 1. TRL levels and transition phases (Source:[4]) 12
Figure 2. Overall concept of the verification and validation process 15
Figure 3. Machine Learning development pipeline (Based on: [1]) 23
Figure 4. Typical errors in quick access recorder data (Source: [11]) 24
Figure 5. Internal and Expert-driven validation loop 26
Figure 6. Machine learning algorithm selection logic-path (Source: [13]) 27
Figure 7. Confusion matrix 33
Figure 8. The relationship between bias and variance in the context of model accuracy 34
Figure 9. Underfitting and overfitting 35
Figure 10. Example of underfitting and overfitting 35
Figure 11. Definition, selection, instantiation and evaluation of case study 50
Figure 12. Model development and verification and validation activities 53
Figure 13. Verification and validation Gantt diagram 54

List of tables

Table 1. Different cross validation techniques 29
Table 2. Research questions 42
Table 3. Examples of targeted KPI per role and time-frame 47
Table 4. Potential support envisioned per Role 47
Table 5. The composition of case-study and sub-case study relevant for the project 52
Table 6. Case studies based on the advisory board feedback and scope of the project: 52
Table 7. Methods to address different RQs defined in VA1 63
Table 8. Methods to address different RQs defined in VA2 64
Table 9. Methods to address different RQs defined in VA3 67

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Executive summary

The verification and validation plan describes the project verification and validation activities, which aim to reflect the data-driven nature of the project and ensure the proper development of the final prototype according to the principle of an iterative development methodology. Unlike traditional software development methodologies which contains a set of standard procedures, the verification and validation activities are highly entwined with the machine learning model development. Thus, an **iterative "design-train-test-validate" process with a lot of back and forth between different experts is crucial for maximising the success of Dispatcher3**. From the data acquisition stage up to design of advice generator, a batch of verification and validation activities will be conducted and structured around **three different and inter-related phases**, which have to be combined iteratively and in close coordination with the experts from the consortium and the Advisory Board, namely:

- **Data acquisition and preparation -** the verification and validation activities within this phase need to ensure that the process and tools used for data preparation are executed properly to obtain a dataset of a satisfiable quality for training specific machine learning algorithms. Therefore, the verification and validation activities performed to the methodologies and tools used, will be present in the following data acquisition and preparation activities:
    - **Cleaning and preparation:** process performed to the historical data (i.e., raw data), ensuring its consistency and readiness before it is processed for model-development purposes.
    - **Labelling and feature engineering:** process performed to the cleaned data to build datasets to validate the model. This process can be complex.
- **Predictive model development -** standard **validation activities** in a machine learning pipeline will be performed to ensure the model performs well (accuracy) and according to the expectations. That means that for each machine learning model developed a reference level of accuracy or a benchmark model needs to be established as a part internal validation. Naturally, a set of standard metrics commonly used in machine learning will be used in Dispatcher3 as well to evaluate the model results. The output of the predictive models will also be **assessed by the experts** to evaluate if the development is moving in the "right direction" as a part of the validation activities performed with experts. In this light, the validation of the models will be two-folded:
    - **Internal validation**: ensuring that the technical particularities of the model are set (e.g., hyper-parametrisation, algorithm selection, etc.). In this phase, the validation dataset is used to measure the performance of the trained the ML model against the performance of some benchmark. The whole process will be performed in a continuous and iterative manner which can be triggered by the need for the model adjustment, but also by the acquisition and preparation of new datasets until predefined threshold is met.
    - **Expert-driven validation**: driving the model configuration following experts' knowledge for the features selection, acceptability thresholds, etc. The industrial experts within the consortium and the experts from the Advisory Board will assess the outcome of the model performance relying on the presented metrics and visualisations

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

- **Advisory generator model development** - the last phase encompasses **verification** and **validation** activities that will be conducted using fully developed prototype that will, relying on the predictive models' output, present a series of advice for targeted users of a particular model/case study. Similar to the machine learning model development, the Advice generator will be developed in an iterative manner and will be based on the outcome of the machine learning model. Therefore, the industrial partner within the consortium, Vueling, skeyes and PACE, will be mainly involved in assessing the benefits of the advice generated, as the latter will be tailored based on the outcome of the machine learning models that use their data.

In order to achieve the goals envisaged by each of the three phases a set of supporting activities are required:

- **Definition and selection of scenarios and case studies**. The use cases that will be implemented will depend on **data availability** and driven by the first analysis of the datasets. The analysis of available data may support the prioritisation of some machine learning techniques. Simplified test cases might be required for some of the verification activities and specific scenarios and case studies will be considered for the validation of the machine learning techniques. Thus, feature engineering will drive the specification of required datasets while the available datasets will define which case study can be candidates as model input.

- **Definition of the assessment framework containing the research questions, hypotheses and success criteria.** The assessment of the results obtained by a specific machine learning model will be performed internally within the consortium, but also in a close interaction with the experts. In addition, the assessment of the results obtained by respective prospective model will be subject of internal validation as well as the expert validation. In order to facilitate and drive the assessment, a set of research questions and their corresponding functioning hypotheses are designed aiming to estimate the operational benefits of the model predictions provided by the advice generator per each role and time-frame.

- **Workshops and dedicated validation activities**. The consortium members will organise **at least two internal technical seminars** which will assess the results obtained by different prototype versions and different set of case studies. The external validation campaign envisions the organisation of **one external workshop** which will gather the **Advisory Board members, Topic Manager and other experts and stakeholders**. In addition to the external workshop, the extensive interaction with the external experts will be conducted by the means of dedicated validation activities (**on-line workshops, site visits and questionnaires**).

When **defining a case study** in the domain of machine learning, one needs to make sure that it captures the problem or a questions that can be effectively answered by **some predictive model** and **the historical datasets available** at hand, i.e., one predictive model essentially learns one mapping between the input (features) and the output (label, targeted indicator) and a case study should be defined in such a way that we can expect a predictive model to be able to learn from the given data (e.g., for routes between a given origin-destination pairs or for flights arriving at a specific airport).

Concretely, a particular case study in Dispatcher3 might **focus on one concrete route**, as the function that a trained model learned most likely would not perform well if the same trained model was tried to be used on a different route (due to the concept known as **data drift**).

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

In order to structure all these different considerations mentioned above, **the definition of case study** will highly depend on **two different elements**, namely:

- **Routes/airports of interest** - all the machine learning models developed within the project will be either route-based or airport-based. However, the experts within the consortium (i.e., Vueling and skeyes) may show a particular interest on the specific route/airport which is worth analysing from the operational point of view.

- **Prediction horizons** - for the given route and specified targeted indicators, different model outputs can be obtained at different time-frames. Therefore, these predictions-horizons should be considered when collecting the datasets that will be used to characterise the available information at the given time-frame. This means that the same dataset might be required at different time-horizons with different resolution. The predictions at different time-frames will tackle different **targeted indicators** and **roles** involved in flight management process.

It is worth mentioning that the specific case studies that will be modelled in Dispatcher3 will be continuously refined alongside the data management and descriptive analysis processes and in the close interactions with the industrial partner from the consortium and the Advisory Board.

In order to effectively manage and track the progress of the verification and validation campaign, the results of the different machine learning models as well as prospective model together with the feedback obtained (including suggestions, recommendations, limitations) during the workshops will be stored in the dedicated page created in collaborative tool (e.g., inGrid).

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

1 Introduction

Dispatcher3 aims at developing a prototype for the acquisition and preparation of historical flight data in order to give support on the optimisation of future flights providing predictive capabilities and advice to relevant stakeholders (e.g., dispatchers and pilots). This will be done considering airline preferences and impact of flight missions on the overall airline objectives.

The dispatching process typically entails the generation and submission of flight plan in order to ensure the safety of the flight. The role of dispatchers may differ across Europe and the North America with the differences that mainly stem from the distinctive operational environment. For instance, dispatchers in North America are assigned with more responsibilities and airlines tend to have a large number of dispatchers working on the day of operations. Contrary, the dispatching process is highly automatised in Europe with few staff in charge of the supervision of these activities. This is partially due to the use of flight plan generators (e.g., Lufthansa LIDO software) which are generally fed with many constraints and pre-defined optimisation parameters (Cost Index estimated by the back-office for specific routes), and a reduced number of dispatchers to generate all the flight plans for the airline.

However, as indicated, the dispatching process understood as the management of the fleet on the day of operation has increased the relevance of longer lookahead decision making process. Identifying potential disruptions early in the day might provide possibilities to plan for solutions beyond adjustment of flight plans (e.g., aircraft swapping). Finally, independently on the automation, dispatchers preparing the flight plans might still manually intervene to adjust and modify solutions when non-nominal situations arise (e.g., avoiding a turbulence region by using a different flight level or an ATFM regulation by re-routing).

As expected, there are variations between planned flights and their execution due to internal and external events (e.g., holdings due to congestion at arrival airport, shorter routes than planned). Experienced dispatchers can consider different flight plan alternatives to select one which best captures the airline's policies. However, the variations of a single flight plan might have a limited impact on the overall experienced airline's performance and fleet management actions can be considered (e.g., aircraft swapping) to minimise the impact of disruptions in the network. These processes, however, rely on individual expertise and automatisation, and lack of the benefit of systematically considering historical performances of flights on same routes under similar conditions. Tactically, pilots might lack an understanding of the changes ahead and therefore are not provided with specific advice on how to operate a given flight considering the impact of the current operational environmental conditions, such as weather, air traffic congestion, time of the day, etc.

Flight operations generate a large set of data from different sources: from planned activities, such as flight plan, forecast weather at the moment of dispatching the flight or expected airspace and airport congestion, to actual realisations, such as flight performance data (QAR), actual weather or holding times. The scope of Dispatcher3 is to consider all these data in order to produce predictions on the outcome of individual flight plans on the different airline's KPIs, which could be used to generate the

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

operational flight plan considering the expected trade-offs involved and, which could provide advice to pilots on how to operate the given flight considering the precursors of the different variances expected.

1.1 Overview of the verification and validation activities

One of the main objectives of Dispatcher3 is therefore to **improve these dispatching and flight operating processes by providing an infrastructure able to leverage on historical data and machine learning techniques** to systematically estimate the variability between planned and executed flight plans, providing expected results of flight plans and advice to the flight planning processes and pilots. This tool will help create more informative flight plans better aligned with the airline policies. Dispatcher3 will also enable airlines to find a suitable solution to fly as efficiently as possible within the known constraints, to ensure the robustness of the airline network against disturbances and environmental conditions (e.g., adverse meteorological conditions, network capacity) and the airline's network-wide impact (changes request in the planning or pilot behaviour). Dispatcher3 has as objective to lead to **a more robust network and a better operational outcome for the European airlines**.

High-level requirements for Dispatcher3 are identified in D1.1 - Technical Resources and Problem definition [7]. Bearing in mind that Dispatcher3 is a data-driven project, the consortium needs to ensure that adequate verification and validation activities will be tailored to address the distinctive nature of the project. In order to properly address these high-level requirements and specificities of the verification and validation approach in the machine learning domain, the following activities will be performed in an iterative manner during the course of the project through the development of the prospective model:

- **Data acquisition and preparation** - the verification and validation activities within this phase need to ensure that the process of data preparation is executed properly in order to obtain a dataset of a satisfiable quality for training specific machine learning algorithms; However, one has to make a clear distinction between data verification and data validation as these processes may have a tremendous impact on the performance of the given machine learning models:
    - The role of **data verification** in the machine learning pipeline is that of a gatekeeper. It **ensures accurate and updated data** over time. Data verification is made primarily at the new data acquisition stage and aims to identify duplicate records and perform deduplication, and to clean mismatch.
    - On the other hand, **data validation** ensures that the incremental data that is added to the learning data is of good quality and similar (from a statistical property perspective) to the existing training data. For example, this includes **finding data anomalies** or detecting **differences between existing training data and new data** to be added to the training data. Otherwise, any data quality issue/statistical differences in incremental data may be missed and training **errors may accumulate over time** and **deteriorate model accuracy**. Thus, data validation detects **significant changes (if any) in incremental training data** at an early stage that helps with root cause analysis.
- **Predictive model development** - the **validation activities** at this phase will be performed for different machine learning (ML) models under development as part of WP4. In other words,

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

for each machine learning model that is going to be developed a reference level of accuracy or a benchmark model will be established in order to ensure that the model performs well (accuracy) and according to the expectations.

- **Advisory generator model development** - the **verification** and **validation** activities that will be conducted using fully developed prototype that will, relying on the predictive models output, present a series of advice for targeted users of a particular model and case study considered. Firstly, the verification of the prototype as a system will be performed against the system requirements defined in D1.1 [7]; and eventually, the validation of Dispatcher3 as a whole system will also be performed in the close interaction with the experts from the Advisory Board and the consortium members in an iterative manner.

The research questions/hypothesis that will be answered by the predictive model are largely underpinned by the case study with the scope and targeted indicators to be predicted. Note that the machine learning model development is highly entwined with the verification and validation activities mentioned above, and thus, some backward feedbacks are planned and expected (e.g., model development affected by result of validation activities conducted together with the experts, case study prioritisation affected by the data availability, etc.,). In addition, such approach will enable the identification of potentially new hypotheses/research questions, or the modification of those initially defined, as results of the validation activities. The similar approach applies to the advisory generator which will be iteratively modified/updated based on the feedback obtained from the experts as a part of the validation activities.

1.2 Technological Readiness Level and context within

Clean Sky 2 (CS2) is a Joint Technology Initiative (JTI) that aims to develop and mature breakthrough 'clean technologies' for Air Transport. The CS2 Programme, will serve society's needs, contributing to Europe's strategic environmental priorities and simultaneously, promoting competitiveness and sustainable economic growth. It will enable cutting edge solutions for further gains in decreasing fuel burn (and CO_2) and reducing NO_x and noise emissions. It will contribute strongly to the renewed ACARE Strategic Research and Innovation Agenda. The CS2 Programme consists of four different elements [5]:

- three Innovative Aircraft Demonstrator Platforms (**IADPs**) for Large Passenger Aircraft (LPA), Regional Aircraft and Fast Rotorcraft, operating demonstrators at vehicle level;
- three Integrated Technology Demonstrators (**ITDs**), looking at Airframe, Engines and Systems, using demonstrators at system level;
- the Technology Evaluator (**TE**), assessing the environmental and societal impact of the technologies developed in the IADPs and ITDs; and
- two Transverse Activities (Eco-Design, Small Air Transport), integrating the knowledge of different ITDs and IADPs for specific applications.

Dispatcher3 fits within the activities of **CS2 Systems ITD WP1.3 "FMS and functions"** of the systems (SYS) ITD, and it addresses some of the high-level objectives and challenges for this ITD defined by the CS2 Joint Technical Programme [5], in particular **the extension of FMS capabilities.** SYS-ITD aims to further mature some of the incipient developments and demonstrators done in CS SGO in the first CS programme, raising them to TRL5 or TRL6, while accommodating the needs of the next generation of aircraft, such as those foreseen in **CS2 IADPs** (Innovative Aircraft Demonstrator Platforms), and

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

considering the specificities of air transportation in different key performance areas (KPA) involving a diversity of stakeholders.

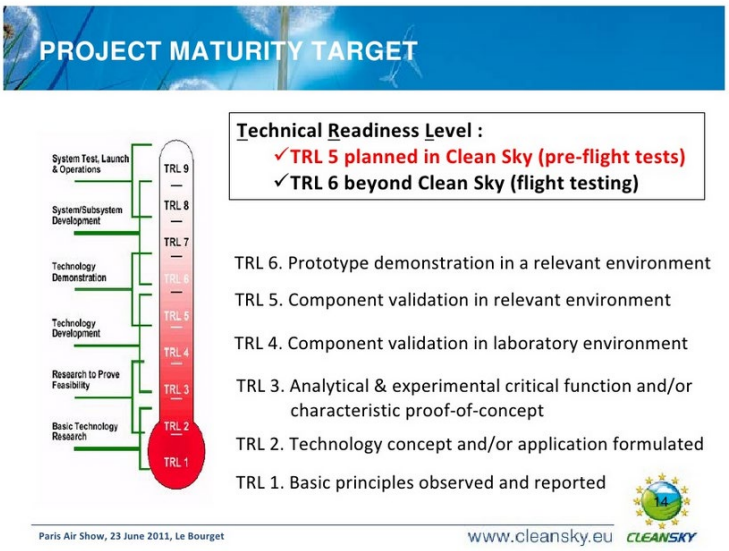


Figure 1. TRL levels and transition phases (Source:[4])

Dispatcher3 highly relies on the application of the machine learning to predict performance indicators (KPIs) or to estimate operational uncertainties which entails that TRL will remain at a lower level as it encompasses more exploitative research activities aiming at presenting a proof-of-concept. Dispatcher3 aims at reaching a TRL level between the 3 and 4 (see **Figure 1**). The reason for the inter-level identified lies in the fact that a full integration (stand-alone system) prior to formal prototyping is not aimed. Nevertheless, Dispatcher3 framework consists of more than a proof of concept, since real data and sources are going to be used and feedback and development input from end users will be applied.

1.3 Deliverable structure

This document is organised in 12 sections and one annex:

- Section 1 introduces the context of Dispatcher3 decision support tool for tactical planners, duty managers, dispatchers and pilots in the time-frame ranging from the day prior the flight to minutes before off block. Then it also presents the definition and approach considered for verification and validation. The targeted TRL is also presented.
- Section 2 lays out the verification and validation concept and approach that will be followed in Dispatcher3. Preliminary considerations and definitions are also described in this section.
- The different activities that will be performed for the verification and validation within data acquisition and preparation are presented in Section 3.
- Section 4 describes the actions that could be conducted to validate the predictive models with the experts involved in the ML model development as well as the other experts from the consortium and the Advisory Board.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

- Section 5 provides the verification and validation activities that will be performed once the prospective model is in place.
- The validation activities aim at answering some research questions. These are described in Section 6.
- Section 7 presents the overview of the potential case studies that will be considered in the project.
- The materialisation of the verification and validation approach into a time schedule is presented in Section 8.
- The document closes with some conclusions (in Section 9) and next steps and look ahead (Section 10).
- References and acronyms are provided in Sections 11 and 12 respectively.
- Finally, an annex (Appendix A - Questionnaires for validation) is provided with questionnaires to be used for the different validation activities.

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

2 Verification and validation concept and approach

This section describes the verification and validation approach planned for Dispatcher3. As explained in Section 1, the verification and validation activities need to be tailored to reflect the iterative development of the prototype and to ensure that:

1. Data wrangling activities containing data preparation and cleaning are performed adequately for each machine learning model developed;
2. Descriptive analysis is correctly performed in order to identify patterns and relations between target variables and data, and extract most useful features for each indicator (target variable);
3. Each machine learning model achieves a satisfying level of accuracy (compared to some reference level or a benchmark model),
4. The prototype meets user requirements and it is assessed against the set of research questions and their corresponding hypotheses.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

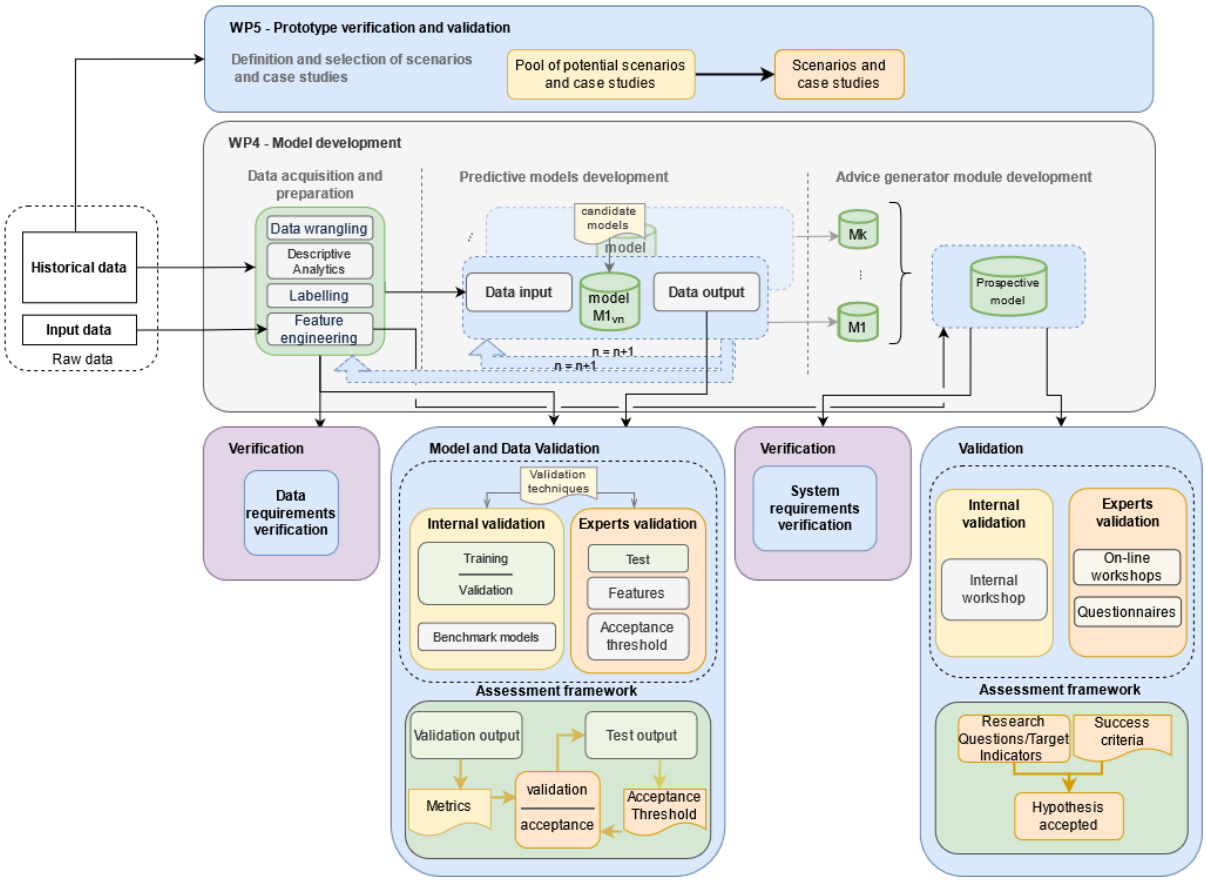


Figure 2. Overall concept of the verification and validation process

Figure 2 illustrates the overall concept of the verification and validation process adopted in this project. As observed from **Figure 2**, the verification and validation of the machine learning models will also be considered as a part of the model development, since the tailoring and enhancement of the model highly relies on the verification and validation results. In this light, the core of the verification and validation approach is structured around three different and inter-related phases, which must be combined iteratively and in close coordination with the experts from the consortium and the Advisory Board. Namely:

- **Data acquisition and preparation -** the verification and validation activities within this phase need to ensure that the process of data preparation is executed properly in order to obtain a dataset of a satisfiable quality for training specific machine learning algorithms. A large amount of raw acquired data will need to be verified and validated prior to further processing or feeding it to the machine learning models. For this to be executed successfully and efficiently, a good data management plan is crucial, and the one adopted and specifically designed for Dispatcher3 needs is described in Deliverable 2.1 [8]. This plan consists of a sequence of actions that includes all the common process in a data science pipeline, covering techniques from data wrangling, descriptive analytics and target variable labelling; and it is executed in an iterative process according to the needs of particular case studies. This covers **verification** activities to ensure that the data requirements defined in D1.1 are properly met, as well as **validation** activities that will be performed to make sure that the labelling and feature

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

engineering activities are properly driven by the experts' input and user requirements. Additional data preparation models that may be required, such as decoders have already been validated in previous projects [7, 12].

- **Predictive model development -** This phase will consist of the **validation** of the different ML models under development as part of WP4. Therefore, for each specific machine learning model, standard validation activities in a machine learning pipeline (see [10]) will be performed to make sure the model performs well (accuracy, precision and recall) and according to the expectations. That means that for each machine learning model developed a reference level of performance or a benchmark model needs to be established, thus in some cases different ML models will be developed for the same KPI and time-frame for cross-validation purposes. This can be as simple as creating a random (dummy) predictive model. The concrete nature of these validation activities depends heavily on the class of machine learning models used, as it is usually the case in machine learning. Naturally, a set of standard metrics commonly used in machine learning will be used in Dispatcher3 as well to evaluate the model results. The output of the predictive models will also be assessed by the experts to see if the development is moving in the "right direction". For example, depending on the particular use case, an expert might define a threshold on a certain metric that needs to be achieved by a model so that it can be considered usable in their operational (business) model. With that objective in mind, a close interaction with experts is considered as crucial during the model development and validation. Additionally, to facilitate this exchange of information between model development team and experts, we will rely on visualisation techniques to present the key information to experts that can then evaluate the performance of a model against their expectations and experience.

- **Advisory generator (AG) model development** - the last phase encompasses **verification** and **validation** activities that will be conducted using fully developed prototype that will, relying on the predictive models' output, present the results on a dashboard for targeted users of a particular model/case study. It is important to stress that the AG will also be used as an integrator of the models, converting the input data into usable features for the models and finally identifying which models (and prediction horizons) would satisfy best fit for the required input time frame. If required, the Advisory generator will combine the output of different machine learning models. Firstly, the verification of the prototype as a **system** will be performed against the system requirements defined in D1.1 [7]; and ultimately, the validation of Dispatcher3 as a whole system will also be performed in the close interaction with the experts from the Advisory Board and the consortium members in an iterative manner.

Note that due to that fact this project is highly data-driven, the different activities involving data collection, data preparation, development and verification and validation are very inter-related, and an **iterative "design-train-test-validate" process with back and forth between different experts is crucial for maximising the success of the project**. Naturally, this sometimes can blur the lines between development and validation, especially when compared to more traditional software development methodologies. Taking all these into account, it must be emphasised that the verification and validation approach within ML domain does not strictly follow the principal of the canonical verification and validation approach which are mostly based on standardised procedures.

1. **Definition and selection of scenarios and case studies**. As indicated in **Figure 2** the use cases that will be implemented will depend on data availability and driven by the first analysis of the datasets. The analysis of available data may support the prioritisation of some machine

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

learning techniques. Simplified test cases might be required for some of the verification activities and specific scenarios and case studies will be considered for the validation of the machine learning techniques. Thus, feature engineering will drive the specification of required datasets while the available datasets will define which scenarios can be candidates as model input. The specific scenarios and case studies that will be modelled in Dispatcher3 will be continuously refined alongside the data management and descriptive analysis processes and in the close interactions with the industrial partner from the consortium, Vueling, skeyes, and the Advisory Board.

2. **Definition of the assessment framework containing the research questions, hypotheses and success criteria**. The assessment of the results obtained by a specific machine learning model will be performed internally within the consortium, but also in a close interaction with the experts. As already mentioned, the experts will define **an acceptance threshold** on a certain metric that needs to be achieved by a model and thus, driving the potential tuning of the model. Moreover, the assessment of the results obtained by respective prospective model will be subject of internal validation as well as the expert validation. In order to facilitate and drive the assessment, a set of research questions and their corresponding functioning hypotheses are designed aiming to estimate the operational benefits of the model predictions provided by the advice generator per each role and time-frame. These hypotheses/research questions might be subject to modifications as the results obtained by the respective machine learning models could generate the need for additional hypotheses/research questions. The experts will assess the results of the scenarios and case studies selected for the given prospective model based on **proposed metrics** (computed from the execution of the prototype) or specific **questionnaires** (e.g., using 6-point Likert scale) to capture the impression of the experts.

3. **Workshops and dedicated validation activities**. The workshops present the main collaborative instruments which will enable to gather the feedback from the great number of experts involved during the validation campaign. The consortium members will organise **at least two internal workshops** (or technical seminars) which will assess the results obtained by different prototype versions and different set of scenarios and case studies. The external validation campaign envisions the organisation of **one external workshop** which will gather the **Advisory Board members, Topic Manager and other experts and stakeholders**. In addition to the external workshop, the extensive interaction with the external experts will be conducted by the means of dedicated validation activities (**on-line workshops, site visits and questionnaires**).

2.1 Verification and validation phases approach

This section provides a brief description of the approach followed for the verification and validation phases. The timely definition of the different activities is detailed in Section 8.

2.1.1 Data verification and validation

This phase will consist of three main activities: **Data wrangling** (preparation and cleaning), **Descriptive analytics** and **Labelling and features engineering**. Therefore, collected data must be cleaned and prepared so that it can be used for the data analytics. Data mining techniques will be used to compute the KPIs that will be used as target variable (variables to predict using ML models). It will also focus on

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

identifying the precursors (e.g., based on variables correlation and aviation experts) that will be used to train the model. These techniques per se do not for part of validation and verification activities; however, during the process of data wrangling and feature engineering for a particular ML model, the analyst will make sure that all the erroneous data entries are fixed or removed, that the data distribution is according to the expectations and that any outliers are removed from the training dataset, etc.

Finally, **target variable labelling and features engineering** activities need to be initiated. KPIs are the target variables to predict, and a list of potential precursors will be used as input features to feed the models. Therefore, the activity will focus on ensuring the correct definition of the labels and making sure the features of the datasets are correctly calculated. The computed KPIs which will represent the labels of the training examples will have to be checked for correctness and consistency. These activities are performed as part of model development and training, as needed for each ML model, but will be reported as part of data validation activities hence following **the philosophy of data-centric model development** (together with similar data validation activities performed in data preparation part). Moreover, when creating training and validation/test datasets, special attention has to be paid to those activities to make sure there is no data leakage. Again, although performed during model development by the analysts working on the ML models, this will form another validation activity and will be reported as such.

2.1.2 Predictive model validation

This phase will consist of the standard validation of the ML models during development. As it can be observed in **Figure 2**, several ML models will be developed in parallel for:

- The **same KPI;**
- Different **time-frames or routes;**
- Different **routes or airports.**

The definition of case-studies in the context of Dispatcher3 project, is further explained in Section 7.

Therefore, multiple ML models will end up existing for the same KPI when predicting on different time-frames or routes. Even for the same indicator and time-frame different ML models which use different input data could be tested and developed. In some cases, the algorithms will focus on the operational environment rather than in a particular route. For example, when predicting the runway in use the machine learning model could be generic enough to be valid for any flight arriving to this destination, i.e., airport dependent; however, when assessing differences on fuel consumption between planned flight plan and realised flight, the specific origin-destination pair, i.e., route, could be relevant.

Predictive model testing and validation: The majority of (or even all) proposed machine learning problems in Dispatcher3 can be solved using **supervised learning** strategies. These activities will aim at ensuring that the designed ML model generalises well to unseen data when giving a KPI forecast. The validation of the ML model will be performed in two phases:

- **Internal validation** - in this phase, the validation dataset is used to measure the performance of the trained the ML model against the performance of some benchmark. We expect to obtain models that outperform the baseline model as much as possible, which is achieved usually through additional data wrangling (**data-centric development**) and/or hyperparameters tuning (**model-centric development**). Note that in the absence of a baseline model, a naive model will be created as it is a common practice in the context of machine learning. The whole

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

process will be performed in a continuous and iterative manner which can be triggered by the need for the model adjustment, but also by the acquisition and preparation of new datasets until predefined threshold is met. Lastly, the model is tested on the test dataset (after tuning on the validation dataset). Alternatively, a cross-validation of some sort might be performed (in case of a smaller training dataset), as explained in the subsequent sections.

- **Validation with experts** - in this phase, the validation will be conducted in a close interaction with the industrial experts within the consortium and the experts from the Advisory Board who will assess the outcome of the model performance relying on the presented metrics and visualisations. In addition, they will provide feedback on the relevance of some important aspects of the ML model, namely:
    - Are the most relevant features, as selected by the trained model, in line with their expectations and expert (domain) knowledge?
    - Are there any additional features that should be included in the ML model?

2.1.3 Advisory generator model verification and validation

This phase aims at verifying and validating advice generator module that relies on the output of the predictive models. This will be conducted throughout two activities:

- The first activity consists of **verifying** that the different ML models can work together as part of the same Advice generator model and that the prospective module has been developed correctly according to the high-level requirements specified in WP1 and low-level requirements and operational work-flow specified in WP4 by applying software engineering principles. Note that the objective of these functional tests is to support the identification of errors in the code, and verify some lower-level functionalities of Dispatcher3, and not the validation of the output of the prototype.

- The second activity has two objectives: to **validate the functionalities of the components** of Dispatcher3 and to **quantify the benefits** of Dispatcher3 in order to understand if the main goals of the Advice generator have been successfully achieved. This will be performed considering different **case studies**. Similar to the ML model development, the Advice generator will be developed in an iterative manner and will be based on the outcome of the ML model. Therefore, the industrial partner within the consortium, Vueling, will be mainly involved in assessing the benefits of the advice generated, as the latter will be tailored based on the outcome of the ML models that use their data.

The different activities that will be considered as part of this internal validation include:

- the **validation of the different predictions** for role/time-frame;
- the **assessment of the benefits** of Dispatcher3 in terms of advice and quantitative information given to the user;
- the **validation of the proposed interface**.

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

2.2 Management and tracking of the verification and validation activities

This section explains the methodology adopted to effectively manage the progress achieved in the verification and validation activities. The adequate execution of the verification and validation campaign requires substantial effort to synchronise and monitor the large number of activities defined above.

2.2.1 Management and tracking of data acquisition and preparation

Due to its complexity and interdependencies with other tasks within WP5, as well as those defined in WP4, WP3 and WP2, the data acquisition and preparation has to be efficiently managed to ensure the seamless progress of all activities performed. The management of all data acquisition and preparation activities has to be based on the concept of transparency which requires the certain level of information sharing among the members within consortium and providing relevant information for each of them. **BeSt** by DataBeacon, a multi-sided, data storage and processing platform, will take the role of the Data Infrastructure layer. In addition, the execution of the activities in the data acquisition and preparation calls for a high level of coordination between the partners in order to fulfil the goals in time efficient and cost effective manner. In addition, this phase has the largest expected workforce required and will highly steer the creation of scenarios and case studies. D2.1 Data definition and processing report deliverable [8] defines the thorough data collection plan, focusing on the data life-cycle since the data collection, then data storage, and finally different processes required to successfully manage acquired datasets.

In order to provide the transparent insight into the progress of the data acquisition and preparation activities, all experts within consortium must have access to specifically designed files in the collaborative decision support tool (i.e., inGrid).

As aforementioned, data mining techniques will be used to extract the KPIs and precursors for the different ML models, which may require **workshops** for the experts to supervise that the results of the activities is aligned with their expectations.

2.2.2 Management and tracking of Predictive model development

As already explained in Section 2.1, different ML models will be iteratively validated and developed. WP5 leader will setup the **living document** by creating a dedicated page in a collaborative tool (e.g., a dedicated inGrid page or shared spreadsheet), which will help to monitor the progress of the validation activities in a transparent manner. A usage of other collaborative and version tracking tools such as **GitHub** and **Databricks** [6] will be considered and adapted as required as well.

In addition to tracking validation activities in the live document, the consortium members will periodically carry out **internal meetings** in order to:

- Ensure that all aimed ML models are being framed within the development process;
- Share the information of common interests among different teams that compose the Dispatcher3 consortium;
- Identify the potential bottleneck which can occur during data acquisition and preparation;

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

- Identify the potential bottleneck which can occur during the development of particular ML models;
- Identify the potential new scenarios and/or prioritise the existing ones;
- Identify the new tasks required for different modules of Dispatcher3 and/or re-prioritise the existing ones.

Since the process of ML models development is expected to require constant iteration, validation and tailoring, at least **monthly** internal meetings are expected for this phase.

2.2.3 Management and tracking of Advisory generator model development

In order to successfully verify and validate the final prospective model, the consortium members will need to carefully perform the following tasks:

- **Integration** and **Verification** of the Advisory generator model. For this campaign, a **living document** will be created to track both: the definition and execution of the verification tests against System Requirements defined in D1.1 [7].
- Design of dedicated activities to **identify and select scenarios and case studies** to be modelled in Dispatcher3. From the pool of scenarios further feedback and information will be gathered by the use of dedicated site visits (or teleconferences) with members of the Advisory Board.
- Design of the **validation workshop** (with the Advisory Board members, Topic Manager and other experts and stakeholders): The main aim of the workshop is twofold - first, to **briefly introduce the capabilities of the tool** to the experts and second, to **validate the first release of the prospective model**. The results from the validation will be used to gather from the experts using **questionnaires**, for example, based **on the six-point Likert scale** and **flow-chart diagrams** which will be distributed during the workshops. These validation activities will also include the validation of the HMI.

As with the internal validation, the management of the Advice generator verification and validation campaign has to be based on the principle of transparency. For this purpose, the collaborative tool (e.g., inGrid) will be used:

- to store all **results** obtained during the Advice generator integration and verification campaign.
- to store all feedback information obtained during the **internal/external** validation campaign.
- the results of the different case studies and feedback obtained (including suggestions, recommendations, limitations) during the Advice generator validation activities will be stored in the dedicated page created in inGrid.

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

3 Data acquisition and preparation verification and validation

The role of **data verification and validation** in machine learning pipeline is to **ensure accurate and relevant data** is used for the development of the model [2]. Data verification is primarily triggered already at the new data acquisition stage i.e., at step 1 to 4 of the ML pipeline, as shown in **Figure 3**, although in general it is a task to which one has to come back whenever needed during the ML model development. The objective of verification and validation in the context of this deliverable is, among others, to ensure that the methodologies used to acquire the data that will be fed into ML models are correct (e.g., plotting data after being processed to identify potential outliers). Further explanation of this process will be presented in the following sections. At a high level, the pipeline ingests historical data, passes it to data processing, then pipes it to a set of features and labels engineering modules and finally reaches the model training and validation. These pipelines typically work in a continuous manner: a new batch of data arrives periodically, which triggers a new run of the pipeline. In other words, each batch of input data will trigger a new run of the data-validation logic and hence potentially a new set of data anomalies. Thus, the verification and validation of the different tools and methodologies (e.g., scripts or models) should ensure data consistency at all levels of the data pipeline.

As depicted in **Figure 3**, with different color-coding, the Data acquisition and preparation verification and validation is two-folded:

- **Cleaning and preparation:** process performed to the historical data (i.e., raw data), ensuring its consistency and readiness before it is processed for model-development purposes.
- **Labelling and feature engineering:** process performed to the clean data in order to build datasets to validate the model. This process can be complex in those cases where obtaining the labelling and features a particular model is required.

Therefore, the following two sections aim at the definition of verification and validation methodologies that are envisioned. It is important to stress that the potential tools, scripts or models are not yet defined and under an **iterative methodology**, the verification and validation campaign will be defined in the most generic approach to fit all potential scenarios.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

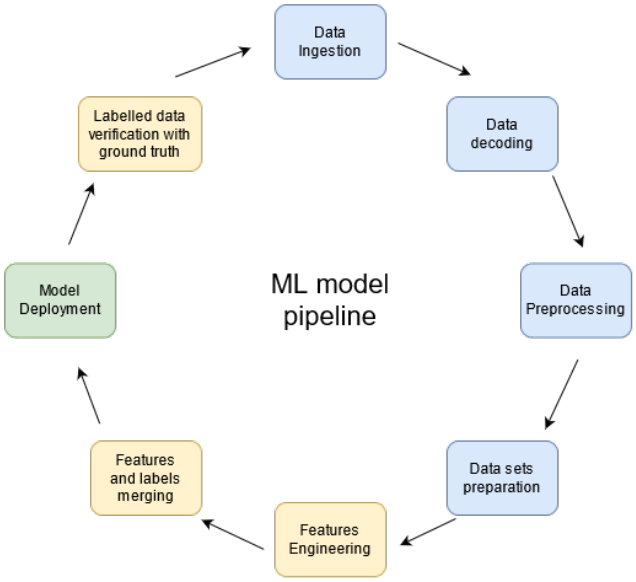


Figure 3. Machine Learning development pipeline (Based on: [1])

Exhaustive definition of the two mentioned activities has been defined in deliverable D2.1 [8] and will be further elaborated in deliverable D3.1. Nevertheless, the aim of these two activities within the present document framework will consist on the verification and validation of the different methodologies used for the execution of the activities. Following the iterative approach, a high-level definition of the methodology will be proposed analysing the different actions that may potentially be required. This is due to the fact that the specific software used and validation approach will be defined ad-hoc and revised iteratively throughout the model development life-cycle.

3.1 Cleaning and preparation

Proper data cleaning is a very critical aspect for the success of machine learning applications as it has a direct impact on the creation of a reliable dataset. The main aim of data cleaning is to identify and remove erroneous and duplicate data as it this enables the model to learn properly. In principle, data cleansing is a highly **iterative** process which completely depends on the type of data at hand. In this sense, we have to ensure that all the activities within data cleaning specified in WP3 are properly verified and validated.

The most common **data errors** (**Figure 4**) that one must look for are:

- **Outliers**: data points that differ significantly from other observations. Normally, as outliers are not indicators of repetitive historical patterns, often we opt for their removal from the data observations that will be used.
- **Constant**: source of error that causes measurements to deviate consistently from their true value
- **Missing value**: data is not available for a particular field. Missing values should be interpolated, labelled specifically or the observations that contain missing value should be removed from the training dataset.
- **Corrupt data**: for example, data values out of expected range due to wrongful sensor readings.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

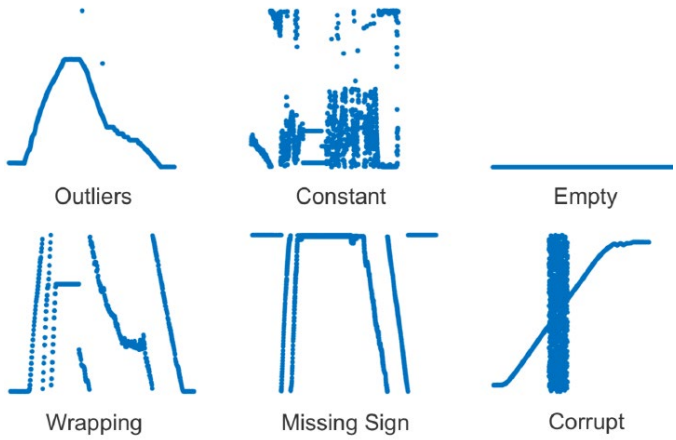


Figure 4. Typical errors in quick access recorder data (Source: [11])

3.1.1 Techniques

There are mainly two distinct approaches for identifying and classifying data errors [3], namely:

- **Qualitative techniques** - this approach contains different set of methods such as rules, constraints, and patterns to identify errors. One common way of specifying those patterns or constraints is by using data quality rules expressed in some integrity constraint languages; and errors are captured by identifying violations of the specified rules. To clean a dataset using rule-based qualitative data cleaning techniques, one first need to design data quality rules that reflect the semantics of the data.
- **Quantitative techniques** - the approach employs statistical techniques to identify errors in the trained data. An expert may analyse the data using the values of **mean, standard deviation, range** and **clustering algorithms** in order to find values that are unexpected and thus erroneous. The correction of such data is difficult to be performed since the true value is not known, but still it can be resolved by setting the values to an average or other statistical value. Statistical methods can also be used to handle missing values which can be replaced by one or more plausible values, which are usually obtained by extensive data augmentation algorithms

Once the errors are identified using the above-mentioned techniques, they can be rectified by making changes through a script or through manual handling (i.e., human intervention), or with a combination of both. Therefore, the verification and validation campaign will consist on the verification and validation of the potential tools used to performed the mentioned techniques.

3.1.2 Methodology and tools

The data will be stored in AWS cloud storage, and the analysis will be run using Databricks notebooks, so all the consortium partners can access and work simultaneously and in a collaborative manner on the data. This platform also will facilitate the data verification and validation through various embedded tools (e.g., for visualisation).

Whenever a software or script is developed for a particular data source (outcome of the techniques mentioned in Section 3.1.1), unit testing will be performed internally by the developer or someone within the same team. It is important to stress that the data cleaning and preparation process tends to require creativity and experience in the given domain in order to drive the best solution.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

3.2 Labelling and feature engineering

As mentioned above, the fine-grained definition of the activities related to “Labelling and feature engineering” are scope of D3.1. Therefore, the preparation of the datasets and the creation of labels and features is thoroughly defined outside of the present document content. Thus, in this section only the verification and validation pertinent to the activities defined in D3.1 will be presented.

The important concept to stress is that obtaining the full set of labels and features to feed a model may require different techniques depending on the data and format available. The obtaining of this data could be deemed from a more complex label or feature to a more trivial, a fact that will determine the verification and validation activities that need to be performed. In the next subsection, we will present the most common techniques.

3.2.1 Techniques

There are multiple techniques to verify and validate the outcome of the labelling and feature engineering activities. Listed below the ones foreseen as candidates for Dispatcher3 project:

- **Cross-validate** results obtained (i.e., Vueling flights from Vueling FDM data vs ADS-B data).
- **Benchmark** refers to testing the software or algorithm against a reference point to quantify whether the difference could deem the tool as valid.
- **Different software** for label and feature dataset calculation. Software can also be used to validate the historical data (i.e., UPC’s trajectory optimiser Dynamo for fuel computation to ensure that estimations from FDM are sound).
- **Feedback from experts** on some labels or features where the importance of validity can be deemed by experience.
- **Visualisation** for data validation, normally used for those labels and features which require visualisation in order to assess whether they are valid or not.

The list of techniques proposed, can be used separately on combined with other techniques were there exist synergies.

3.2.2 Methodology and tools

The data will be stored in AWS cloud storage, and the analysis will be run using Databricks notebooks, so all the consortium partners can access and work simultaneously and in a collaborative manner on the data.

Whenever a software or script is developed for a particular data source (outcome of the techniques mentioned in Section 3.2.1), unit testing will be performed internally by the developer or someone within the same team. It is important to stress that the data cleaning and preparation process tends to require creativity in order to drive the best solution.

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

4 Preliminary data requirements

This section will be primarily focused on the understanding of the machine learning principles with respect to their validation. The validation of the models will be two-folded (see **Figure 5**):

- **Internal validation**: ensuring that the technical particularities of the model are set (e.g., hyper-parametrisation, algorithm selection, etc.)
- **Expert-driven validation**: driving the model configuration following experts' knowledge for the features selection, acceptability thresholds, etc.

The former is based on different cross validation techniques to perform hyperparameter selection process. The latter will employ standard metrics, such as prediction accuracy, precision or recall, and tools such as confusion matrix to display and analyse the results in the close interaction with the members of the consortium, and with the Topic Manager.

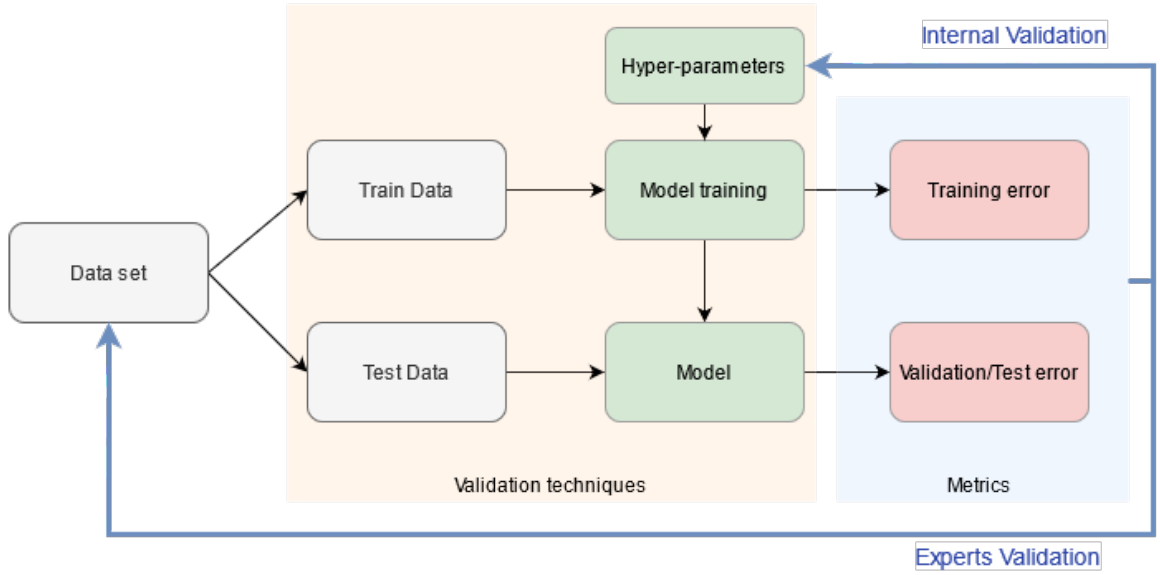


Figure 5. Internal and Expert-driven validation loop

4.1 Machine learning algorithm selection

Selection of the appropriate algorithm that will be used to train machine learning model is not a straightforward task. It will be largely driven by the ability to define the problem in terms of machine learning (e.g., identifying the problem as classification or regression) on one side, and the selection of the predictive model, on the other. Thus, it is of high importance to explore the main characteristics of different machine learnings techniques (i.e., algorithms) and their major limitations. In D1.1 an extensive research on machine learning models was performed and a list of initial candidates was

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

proposed [7]. Nevertheless, for validation purposes, this section is aimed to make a clear distinction between different types of the existing machine learning algorithms (indicated by the green boxes in **Figure 6**) and the machine learning models which will be trained on some data by using the specific algorithm. In this regard, the term "algorithm" will be used in general sense, while the term “model” refers to a concrete machine learning model that is trained on some data (i.e., an instance of an algorithm).

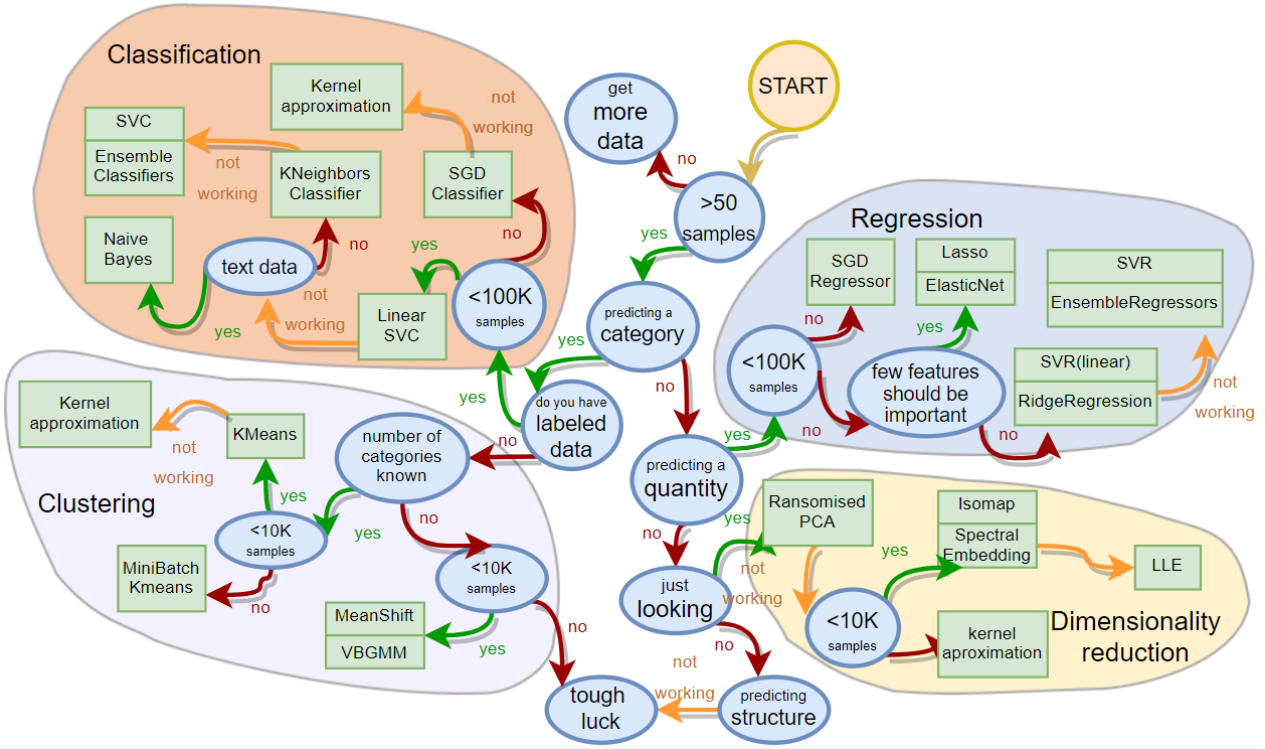


Figure 6. Machine learning algorithm selection logic-path (Source: [13])

As already mentioned, definition of the problem in terms of the machine learning context presents a first step in the development of machine learning models. In this light, some key issues have to be addressed in order to scale down the selection of the model. **Figure 6** provides a general approach which may facilitate the process of selection of the appropriate algorithm. In the context of supervised learning that will be mostly used to tackle the problems in Dispatcher3, the first step is to identify the problem as either classification or regression. As explained in D1.1, the KPIs that we want to predict will largely drive the process of the final selection of the machine learning algorithm [7].

4.2 Training the model and cross-validation

Once the appropriate machine learning algorithm has been selected, we move onto the process of training, testing and validation of the model. As expected, learning the parameters of a model and testing it on the same data will cause the problem of **overfitting**. In other words, a model that would just repeat the labels of the samples that it has just seen would have **a perfect score** but would fail to predict anything useful on yet-unseen data [13]. In order to avoid this issue, it is common practice when performing a (supervised) machine learning validation to use only part of the data as **training**

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

set and hold out part of the available data as a **test set**. Different types of **cross validation** methods are typically conducted during the process of training the dataset in order to avoid the perfect score.

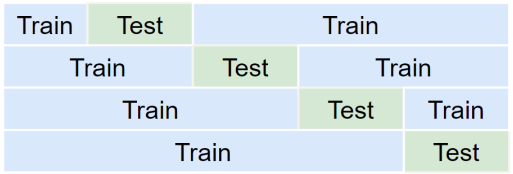
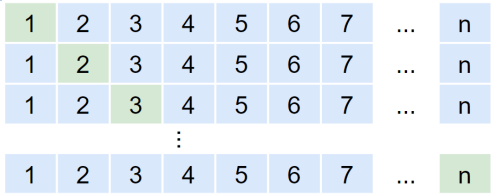
The existing methods have also different characteristics in terms of the time required to prepare and execute the test. The methods mainly involve partitioning a sample of data into complementary subsets by performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the testing set). In order to reduce the variability, in most of the methods multiple rounds of cross-validation are performed using different partitions. The validation results are then combined (e.g., averaged) over the rounds to obtain an estimation of the model's predictive capabilities.

Among all cross-validation techniques mentioned in **Table 1**, k-fold cross validation is the ones widely used to minimize sampling bias.

Figure 5 provides a flowchart of typical cross validation workflow in model training. As observed from **Figure 5**, one has to determine a set of optimal hyper-parameters by using different techniques (e.g., grid search, randomised search, etc.,). Hyper-parameters are parameters that are not directly learnt within estimators and have to be tuned so that the model can optimally solve the machine learning problem. They are often passed as arguments to the constructor of the estimator classes. Note that it is common that a small subset of those parameters can have a large impact on the predictive or computation performance of the model while others can be left to their default values. It is possible and recommended to search the hyper-parameter space for the best cross validation score.

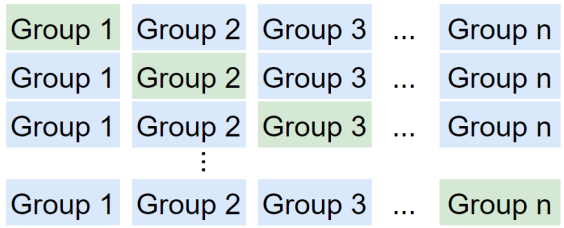
© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Table 1. Different cross validation techniques

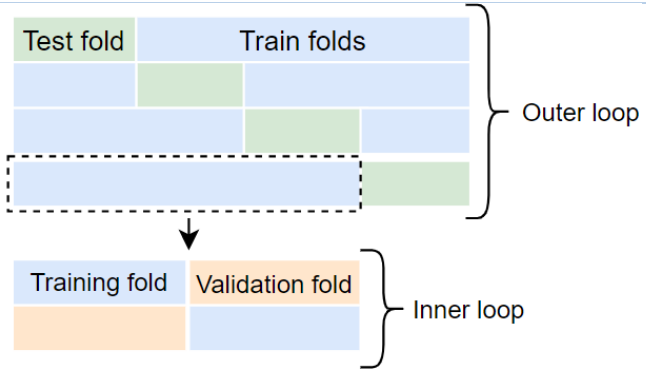
Method	Characteristics	Pros	Cons
k-fold cross-validation	• The original sample is randomly partitioned into *k* equal sized subsamples. Of the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining *k* − 1 subsamples are used as training data. • The cross-validation process is then repeated *k* times, with each of the *k* subsamples used exactly once as the validation data. The *k* results can then be averaged to produce a single estimation.	• All observations are used for both training and validation, and each observation is used once for validation. • The method involves either *i*= 5 or *k*=10 as they find a nice balance between computational complexity and validation accuracy	• One might also look at the **variance** or **standard deviation** of the resulting folds as it will provide information about the stability of the model across different data inputs
			
Leave-one-out (LOO) cross-validation	• Each learning set is created by taking all the samples except one, the test set being the sample left out. • Thus, for samples, we have different training sets and different tests set.	• LOO procedure is data efficient as only one sample is removed from the training set	• In terms of accuracy, LOO often results in high variance as an estimator for the test error.
			

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Method	Characteristics	Pros	Cons
Leave-p-out (LpO) cross-validation	• It involves using *p* observations as the validation set and the remaining observations as the training set. • This is repeated on all ways to cut the original sample on a validation set of *p* observations and a training set	• A variant of LpO cross-validation with p=2 known as leave-pair-out cross-validation has been recommended as a nearly unbiased method for estimating the area under ROC curve of binary classifiers	• LpO cross-validation requires training and validating the model C_p^n times • For $p > 1$ and for even moderately large *n*, LpO CV can become computationally infeasible.



© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Method	Characteristics	Pros	Cons
Nested cross-validation	It allows to separate the hyperparameter tuning step from the error estimation step. To do this, two k-fold cross-validation loops are nested: • The inner loop for hyperparameter tuning and • The outer loop for estimating accuracy.	• **Nested cross-validation** cross-validation is often used to train a model in which hyperparameters also need to be optimized. **Nested** cross-validation estimates the generalization error of the underlying model and its (hyper)parameter search.	N/A
			

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

4.3 Model validation

4.3.1 Internal validation

The internal validation campaign is driven by the experts from the consortium which are directly involved in the development of the machine learning models. The process is highly machine learning model specific and will be performed in continuous manner as illustratively shown in **Figure 2**. As observed, the experts will need to perform the model selection which refers to the process of selecting the right model that fits the data. This is done using test evaluation matrices. The results from the test data are passed back to the hyper-parameter tuner to get the most optimal hyper-parameters. Essentially, this enables us to efficiently control the over-fitting and under-fitting of the model.

In order to obtain the best hyperparameters the following steps are followed:

1. For each proposed hyperparameter setting the model is evaluated
2. The hyperparameters that give the best performance are selected

As already mentioned, there is a variety of techniques which can help to determine the best parameters. Among all, three of them have been extensively used:

- **Grid search** picks out a grid of hyperparameter values and evaluates all of them. Guesswork is necessary to specify the minimum and maximum values for each hyperparameter.
- **Random search** randomly values a random sample of points on the grid. It is more efficient than grid search.
- **Smart hyperparameter** tuning picks a few hyperparameter settings, evaluates the validation matrices, adjusts the hyperparameters, and re-evaluates the validation matrices. Examples of smart hyper-parameter are Spearmint (hyperparameter optimization using Gaussian processes) and Hyperopt (hyperparameter optimization using Tree-based estimators).

During the assessment of the algorithm selection, multiple models may be developed for the same KPI and time-frame in order to cross-validate the results (e.g., in terms of error, precision, etc.) between the different models and deem the best algorithm fit.

4.3.2 Metrics

The purpose of this section is to provide a list of metrics used to quantify the quality of predictions.

- **Regression**

  MSE (Mean Squared Error): $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$

  MAE (Mean Absolute Error): $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(x_i)|$

- **Classification**

  In the case of supervised machine learning algorithms, and in particular classification algorithm, tools such as confusion matrix are going to be used to display the results and analyse how well the models are performing. Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class (or vice versa) (see **Figure 7**).

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

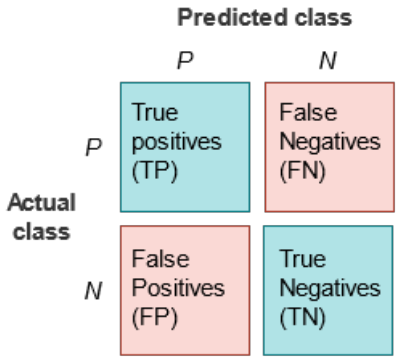


Figure 7. Confusion matrix

The elements of the matrix report the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). This allows more detailed analysis than mere proportion of correct classifications (accuracy). The list of metrics potentially derived from the matrix are given as follows:

- **Accuracy**: The simplest metric, it does not represent the error if the model is bad (i.e., model that does not find the sick people, if the population size is 1000 and there are 3 sick people, the accuracy will be almost 100% despite the model does not find the sick ones), Correct/ Total.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**: Focused on **ONLY** finding the True positives.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**: Opposite of Precision, it is centred on finding ALL the True Positives.

$$Recall = \frac{TP}{TP + FN}$$

- **F1 score**, combination of Recall and Precision

$$F1\ Score = \frac{2TP}{2TP + FN + FP}$$

4.3.2.1 Overfitting and Underfitting

The main purpose of the machine learning models is to achieve a good generalisation level. Generalization refers to the model's ability to give sensible outputs to sets of input that have not been included in the modelling process. Moreover, performance of the model as well as the application as a whole relies heavily on the generalization of the model. Based on this notion, terms like **overfitting and underfitting** refer to deficiencies that the model's performance might suffer from. This means that knowing "how off" the model's predictions are a matter of knowing how close it is to overfitting or underfitting. The issue of overfitting and underfitting is strongly related to bias and variance as depicted in **Figure 8** and **Figure 9**.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

However, every estimator has its advantages and drawbacks. Its generalization error can be decomposed in terms of:

- Bias - is an average error for different training sets,
- Variance - indicates how sensitive the estimator is to varying training sets,
- Noise - is a property of the data.

Both **bias** and **variance** are inherent properties of estimators and thus, it is important to select learning algorithms and hyperparameters so that both bias and variance are as low as possible. **Figure 8** depicts different relationships between desired accuracy, validation and training accuracy.

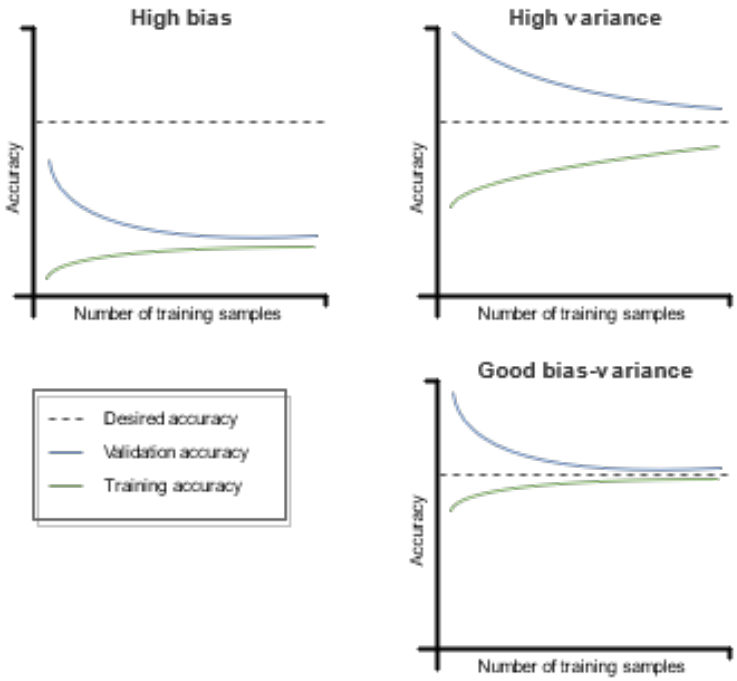


Figure 8. The relationship between bias and variance in the context of model accuracy

High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting), while high variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting). The bias-variance trade-off is one of the important aspects to consider in supervised learning. One would ideally select a model that accurately captures both the regularities in its training data, but also generalises well to unseen data. As these two requirements cannot be achieved simultaneously, the trade-off between them needs to be formulated.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

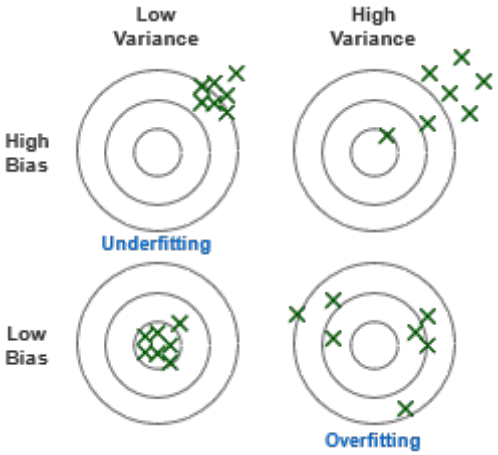


Figure 9. Underfitting and overfitting

Managing overfitting of the model

A good indication of **overfitting** can be the difference between training error and the validation error in which case training error will be usually lower than the validation error. **Figure 10** shows the example of the model that learns even the error patterns leading to overfitting. As observed, despite the training error equals to zero, the model is not a good generalisation of the reality.

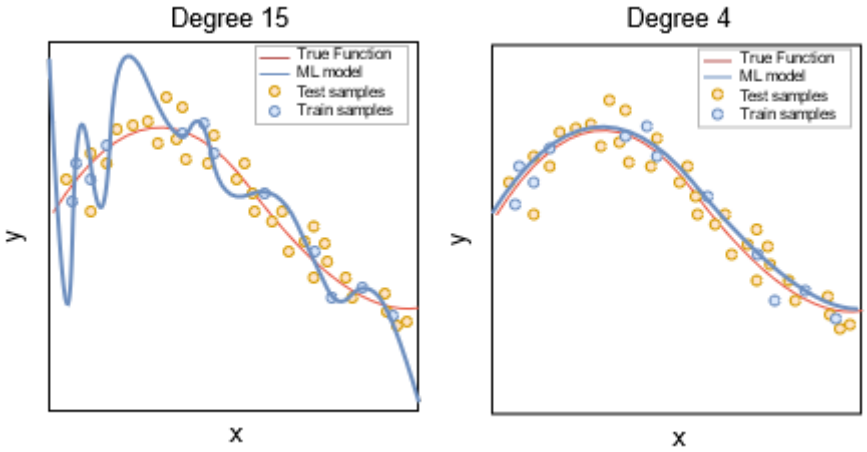


Figure 10. Example of underfitting and overfitting

There are two common approaches to mitigate the effect of overfitting, namely:

1. **Using more training data** in order to reduce the variance of a model. However, collecting more training data is reasonable only when the true function is too complex to be approximated by an estimator with a lower variance.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

2. **Selecting the features with the best performance**. This approach may require some of the additional activities, for instance:
    a) removal of the redundant features or those that are not relevant.
    b) generating new features based on the others.
    c) simplification of the model based on a subset of the features that are deemed truly informative.

Managing underfitting of the model

Underfitting usually implies that the selected model is inappropriate. An indication for underfitting are low scores of both training and validation sets. In addition to considering the application of alternate machine learning algorithms, the analyst may try several other options that could efficiently handle the model underfitting including:

1. Using more training data.
2. Increasing the size or number of parameters in the model.
3. Increasing the complexity of the model.

4.3.3 Experts' validation

The aim of this action is to validate the results of the predictive models development interacting with the external experts, in order to obtain an initial feedback. As the initial action in the external validation campaign, the main goal is to put all the external experts in the context by introducing them with several important aspects of the given predictive model, such as:

- the KPIs obtained by the predictive model (i.e., "*How good the model is performing in terms of KPIs obtained?*"),
- the importance of specific features (i.e., "*What kind of precursors does the predictive model take into account?"*),
- the metrics obtained and their acceptance threshold (i.e., "*Do the metrics obtained provide the satisfactory results?*").

At this stage of the project, the specific model has been already fine-tuned by the machine learning model experts within the consortium. The particular aim will be to receive the valuable feedback on the given aspects from the experts who was not involved into the machine learning model development and who already have some extensive experience in the domain of operations.

The purpose of the first validation action is to show to the expert's committee some results obtained with the specific machine learning model with respect to the KPI selected in order to obtain feedback. For the second action, the experts will assess whether the features identified for the model should be of relevance or not according to their expertise. And finally, the third action will consist on their definition of thresholds in terms of the performance of the models (e.g., definition of minimum error rate).

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

5 Prospective model verification and validation

This section aims at defining how the verification and validation of the Advice generator module will be conducted within Dispatcher3 framework. Differently from the ML model validation, the activities described in this section follow a more conventional software model verification and validation.

Although the distinction between verification and validation are clear enough from their ultimate objective it very often remains quite vague when it comes to their practical applications. For instance, some authors consider that verification should remain as a static process which focus at reviewing the code and the system, while the dynamic execution of the software and the analysis of its output could be considered as part of the validation activities; while others incorporate some dynamic testing as part of the verification to ensure that the code is error free. Therefore, the border between verification and validation becomes fuzzy when the details which specific activities should be performed as part of verification or as part of validation are defined [9]. This is particularly the case in the domain of testing code at a functioning and system level. However, the agreed view is the ultimate objective of these activities:

- Software **verification** aims at answering the question “Are we building the product right?”; that is, does the software conform to its specifications?
- Software **validation** aims at answering the question, “Are we building the right product?”; that is, does the software do what the user really requires?

As defined in prior Section 2, the verification and validation of the prospective model will be performed once some (or most of the) ML predictive models for a particular time-frame have been developed and its validation activities have been satisfactory concluded.

Note that the Advice generator will use the machine learning models selecting the right models to use for producing some predictions for a particular new input. It will therefore follow a more model-driven approach. In some cases, complex predictions, e.g., reactionary delay could be computed as the combination of the outcome of lower-level machine learning driven estimators.

5.1 Verification and Integration

The **dynamic verification of the software** is conducted to verify the working of the software. As described in the introduction, some views are that this should be considered as part of the validation and that the verification should remain static. However, we consider that these low-level dynamic tests are required to verify the code developed and basic functionalities. Note that the full system execution and analysis of functionalities will be performed as part of the internal validation activities. Therefore, we define the dynamic verification of the software as an incrementally process on complexity and integrated verifications:

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

1. **Integration testing**: which aims at verifying that the different modules operate correctly jointly.
2. **System testing**: the objective is to perform a set of activities in order to verify high-level requirements for different prototype release.

5.1.1 Integration

As an upper second layer in the verification of the working software, integrating testing aims to test the combinations of individual software modules. In other words, integration testing checks whether different modules are working fine when combined together as a group. In contrast to unit testing that considers checking a single component of the system, integration testing aims at checking integrated modules in the system.

The aim of this activity is two-fold:

- Verify integrated models and data. In Dispatcher3 project, the focus will be put on the correct integrated processing of features from raw data input targeting specific ML models.
- Ensure and verify that different modules that compose Dispatcher3 are correctly integrated. In Dispatcher3 project, the focus will be put on the correct integration of ML models and the integration with the chosen dashboard interface.

The methodology applied here will follow the similar approach as in the case of unit testing, although integration testing aims at checking different combinations of the individual modules.

5.1.2 System testing

System requirements test. System testing considers **the complete, integrated system as a whole**. Thus, the given test has to be performed on some of the matured versions of the prototype. System Testing is important as it verifies that the application meets the technical, functional, and business requirements that were set at the onset of the project. The main goals at this level is to evaluate if the system has complied with all of the high-level requirements specified in D1.1. (i.e., SYS requirements) and lower-level requirements specified for each module composing Dispatcher3 (as captured in WP4) to see if it meets quality standards [7].

5.2 Validation

The validation will be conducted using fully functional versions of the prospective (prescriptive) model and based on the results of case studies performed during the predictive models development. Dedicated validation activities (e.g., workshop) will be carried out internally in the consortium, but also presenting the prototype to airlines and other experts from the Advisory Board in a one-to-one interaction and as part of a workshop once the first prototype is created. Thus, the validation of the prototype will be performed both internally in the consortium and externally with the Advisory Board experts through two main types of actions:

- **Presentation of advice generator based on the predictive analytics obtained** - the objective of this validation action is to validate the **decision framework** which should aid all the final customers to understand the predictions produced taking into account the differences in their specific requirements.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

- **Visual demonstration of the results obtained and overall capabilities** - the activity aims to validate the presentation of results, how the different end-users can benefit from the presentation and to show the overall capabilities of Dispatcher3 prototype.

The validation is carried out through different validation actions (VA), which can be grouped between:

- **Actions aiming at assessment the benefit of Dispatcher3**
    - **VA1 - Dispatcher3 performance at prediction of individual KPI** - the objective of this step is to assess the benefits of Dispatcher3 model prediction at single KPI level.
    - **VA2 - Dispatcher3 performance at prediction of set of KPIs** - the aim of this action is to assess the benefit of Dispatcher3 model prediction at a set of KPIs for different roles and/or time-frames.
    - **VA3 - Evaluate the dashboard results from prototype** - the aim of this action is to assess the benefit of the visualisation of results obtained by Dispatcher3 and presented in a dashboard. Note that a possible overlap between the advice generator evaluated in VA2 and the dashboard of VA3 might exist.

This section presents the different internal validation actions with detail on the methodology and metrics that will be generated for the assessment of the research questions presented in Section 6.

5.2.1 Purpose

The purpose of this validation action is to show to the experts the capabilities (and results) of the advice generator in order to obtain feedback from stakeholders. The advice generator will transform the outcome of the predictive engine into advice according to airline policies. As the outcome of the advice generator may provide the actionable indications for different roles within airline and at different time-frames in the flight planning process, the involvement of stakeholders with different background (e.g., pilots, dispatchers, etc.) will be essential to capture the operational benefit of the tool.

5.2.2 Tools and methodology

As already explained in Section 2, the goal of the advice generator module is to collect all the information from the predictive analytics obtained, including information about the quality of the prediction (accuracy, precision, recall, etc.) and build a **decision framework**. Note that specific predictive and advisory capabilities will be delivered for each specific user based on the independent user-oriented models and focus on specific case studies which will be defined around specific origin and destination routes. This module will be the entry point to the use of the different machine learning models.

Based on the results presented, the experts will be able to:

- provide the feedback on the understanding of the predictive analytics provided and its probabilistic nature (i.e., machine learning interpretation)
- provide the feedback on the understanding of how the probabilistic outcome of the predictors is translated into qualitative suggestions,
- provide the feedback on the usability of the overall concept with respect to operational perspective (i.e., different time-frames targeted at specific roles),

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

- assess the benefit of the tool with the advice available when making the decision

In order to facilitate the assessment of the results presented, the experts will **be asked to provide their feedback on a set of statements** addressing their personal attitudes towards operational benefits of the advices provided. For this purpose, a six-point Likert scale is suggested to be employed. In addition, a flow-chart acceptance diagram is also designed to assess the experts' opinion on the benefit of Dispatcher3 tool in terms of the solutions obtained.

After obtaining the feedback for different scenarios proposed (encompassing different time-frames, the results will be gathered and the tool in general will be validated based on defined success criteria.

5.2.3 Main metrics for validation

- Average outcome of a six-point Likert scale will the following items: "Strongly disagree", "Disagree", "Slightly Disagree", "Slightly Agree", "Agree" and "Strongly agree";
- Outcome of flow-chart diagram.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

6 Research Questions

Table 2 summarises the research questions for the validation activities. The validation activities will aim at quantifying some of the results of the different planned case studies. A set of research questions (RQs) and their corresponding hypotheses (HPs) are designed to address the benefits of Dispatcher3. The RQs aim at being qualitative rather than quantifiable, as we try estimate the real operational benefits of the tool from the perspective of the final customer. Objective and quantifiable success criteria will be defined for each RQ in order to validate or refute the corresponding hypothesis. As previously indicated, there will exist different models to predict the same KPI for different time-frames, and also there will exist different advice for the same case-scenario for a different role. Therefore, in some validation activities, the same question will be used for as many case-studies and KPIs as there exist at the validation execution, which will highly depend on the data availability and individual requests obtained from the partners within consortium as well as from other stakeholders. Modifications to the hypotheses/research questions or the inclusion of new ones might be required in in the light of the obtained feedback.

This section summarises the different research questions for the validation actions that will be performed for VA1 (Validation Activities 1) once the individual ML model is trained in close cooperation with experts' supervision, and VA2 (Validation Activities 2) once the Advice generator prototype has been defined for a specific case-study and time-frame. The RQs and hypothesis will cover all three aspects of the actions that were explained in Section 5.

VA1: Evaluate predictions at:

- Individual KPI prediction level

VA2: Evaluate the benefit of:

- KPI- group level prediction
- Advice generator

VA3: Evaluate the dashboard results from prototype, which might overlap and could be reconsidered if required based on the outcome of VA2 (Advice generator validation activities).

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Table 2. Research questions

Validation Action	ID	Rationale	Research question (RQ)	Hypotheses (HP)	Success criteria	Methodology
VA1	**D3-RQ-IV-010**[1]	Validate that advanced prediction on a given KPI is relevant for different case studies tailored for specific role/time frames.	Based on the output provided for each given **role** and **time-frame**, do experts find Dispatcher3 as a tool worth incorporating in regards to the predicted **KPI**?	The information on the predicted KPI for each role and time-frame will add value to the decision-making progress for each particular role/time-frame.	• The **majority** of the respondents should "slightly agree" that Dispatcher3 aids the pilot to take a more informed decision • None of the respondents should indicate "strongly disagree" and "disagree" option	6-point Likert scale for the "" questionnaire. (see Appendix A - Questionnaires for validation)
VA2	**D3-RQ-IV-020**	Validate that advanced prediction on a group of KPIs are relevant for the specific role/time-frame	Are the KPI-group prediction provided by Dispatcher3 meaningful enough in the case of the case study presented and with respect to the given role/time-frame?	Dispatcher3 will efficiently deal with a variety of issues imposed by different operational context that define the particular case study by providing the set of meaningful KPIs predictions at a given role/time-frame.	• The **majority** of the respondents should "agree" that Dispatcher3 aids the pilot to take a more informed decision • None of the respondents should indicate "strongly disagree" and "disagree" option	6-point Likert scale for the "Goodness of solutions in the given operational context" questionnaire (see Appendix A - Questionnaires for validation)"

[1] This RQ will be applied to each combination of role/time-frame KPI (see Table 3) available.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Validation Action	ID	Rationale	Research question (RQ)	Hypotheses (HP)	Success criteria	Methodology
	D3-RQ-IV-030	Validate that the airlines' policies are captured in the advice provided by Dispatcher3.	For a given triggering event for a particular role/time-frame, will Dispatcher3 provide appropriate advice/KPI predictions?	Dispatcher3 will provide benefits to the airline industry as it will capture airline policies that will lead to different operational decisions for the same problem.	• The **majority** of the respondents should "agree" that Dispatcher3 aids the pilot to take a more informed decision • None of the respondents should indicate "strongly disagree" and "disagree" option	6-point Likert scale for the "Goodness of solutions in the given operational context (i.e., roles and time-frames)" questionnaire (see Appendix A - Questionnaires for validation)"

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Validation Action	ID	Rationale	Research question (RQ)	Hypotheses (HP)	Success criteria	Methodology
	D3-RQ-IV-040	Validate that the advice provided by the Advice generator is relevant for decision making process within specific role/time frames.	Will the advice provided by the Advice generator aid the final user to make a more informed decision?	The advice provided by the Advice generator will add value to the decision making process by providing the actionable indications based on the probabilistic outcome of the predictors.	• The **majority** of the respondents should "agree" that Dispatcher3 aids the pilot to take a more informed decision • None of the respondents should indicate "strongly disagree" and "disagree" option	6-point Likert scale for the "Specific role's acceptance of the tool with respect to the Advice generator's capabilities" questionnaire (see Appendix A - Questionnaires for validation)"
	D3-RQ-IV-050	Validate overall acceptance of Dispatcher3 decision framework as a complementary aid in the flight planning process.	Will Dispatcher3 (advice provide by Advice generator) be easily incorporated into the daily operation of the specific role? (i.e., could it be easy to use together with the normal workload/stress/ etc.)	The required time and easiness of interaction with Dispatcher3 on an operational-basis will allow it to be integrated within the role job performances.	• The **majority** of the respondents should "agree" that Dispatcher3 aids the pilot to take a more informed decision • None of the respondents should indicate "strongly disagree" and "disagree" option	6-point Likert scale for the "Integration of the system into the operation" questionnaire (see Appendix A - Questionnaires for validation)"

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Validation Action	ID	Rationale	Research question (RQ)	Hypotheses (HP)	Success criteria	Methodology
VA3	**D3-RQ-IV-060**	Validate the overall acceptance of Dispatcher3 by experts of the dashboard.	From a very general point of view and based on the visual representation and information displayed in the dashboard, do experts find Dispatcher3 as a tool which is worth (or useful) having in the airline company?	Given its user-friendly interface as well as a broad amount of well-structured information provided, Dispatcher3 is deemed as a very desirable decision support tool for commercial use by the airlines with different business models.	• The final score provided by the individual experts should range between 8 and 10	Flow-chart diagram for global acceptance. (see Appendix A - Questionnaires for validation)

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Validation Action	ID	Rationale	Research question (RQ)	Hypotheses (HP)	Success criteria	Methodology
	D3-RQ-IV-070	Validate the simplicity but completeness of the information presented to specific role.	Is the information given by the dashboard to the specific role simple (or concise) enough to allow their prompt reaction?	The information presented by the dashboard will be simple and, as much as possible, predictable in its presentation, which means that appropriate balance will be found in terms of the amount of information so that the specific role can easily conceive (process) it.	• The **majority** of the respondents should "agree" that Pilot3 provides clear information to the pilot • None of the respondents should indicate a "strongly disagree" and "disagree" option	6-point Likert scale for the "General acceptability" questionnaire. (see Appendix A - Questionnaires for validation)
	D3-RQ-IV-080	Validate the facility of the dashboard to convey the information computed by Dispatcher3.	Is the information given by the dashboard to the specific role informative enough and helps to take a more informed decision for a given operational context?	The AG visualisation will ensure that the specific role can easily understand the advice which is based in the predictive analytics (i.e., KPIs prediction).	• The **majority** of the respondents should "agree" that Dispatcher3 aids the specific role to take a more informed decision • None of the respondents should indicate "strongly disagree" and "disagree" option	6-point Likert scale for the "Easiness of understanding of the information" questionnaire. (see Appendix A - Questionnaires for validation)

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

EUROPEAN UNION

As aforementioned in prior sections, several preliminary case-studies have been presented as a starting point for development. Nevertheless, and following the agile methodology, the final combination of estimated KPI models for route and time-frame may be revised throughout the project life-cycle, extracted from D2.1 [8]. Therefore, table Table 3 presents some examples of KPI per role and time-frame in order to facilitate the envisioned research questions that will rely on that combination.

Table 3. Examples of targeted KPI per role and time-frame

Prediction horizon (time-frame)	Targeted role	Target indicators (KPI)
Day prior operations (D-1)	Tactical planner	• Congestion in airspace • Flight affected by congestion • Congestion at airports
Hours prior the flight (- 10/9 H)	Duty manager	• Fuel deviations (taxi, flight times, block fuel) • Fuel tankering • Time deviations (taxi, flight times, block times) • Holdings at arrival
Few hours prior flight (- 4/3 H)	Dispatcher	• Fuel deviations (taxi, flight times, block fuel)
	Duty manager	• Fuel tankering
Before push-back (30')	Pilot	• Time deviations (taxi, flight times, block times)

In previous deliverable D2.1, the potential support that was envisioned for Dispatcher3 was described collaborating with the different stakeholders [8]. Therefore, it is of upmost importance, that while the different outputs from Dispatcher3 prototype are being analysed, the very first aim of the tool is being satisfied. Therefore, **Table *4*** extracted from D2.1 captures de initial objective of Dispatcher3 for each role in order to support the validation activities [8].

Table 4. Potential support envisioned per Role

Role	Potential support
Schedule planner[3]	• Out of scope of Dispatcher3 but the project will create the infrastructure needed to store and process planned and actual historical flight and operational environmental data. This will allow strategic decisions to be further developed based on these data, e.g., modifying airline flight policies. • Dispatcher3 could provide advice on which flights, and in which conditions, are more prone to variance between schedules and execution blocks.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Tactical planner[2]	• Identify which flight plans are more likely to be disrupted. • Estimate already block times, fuel usage and impact on reactionary delay. • Support the estimation of benefit of alternatives such as aircraft swapping, crew rotations, etc.
Duty manager[2]	• This position might benefit from enhanced predictive capabilities, not aimed at improving a given flight, but at identifying which flights might suffer from disruptions in the network with a few hours of look-ahead. • The goal is to highlight, identify which flights will be prone to have disruptions and propagate them through the network.
Dispatcher[1]	• Providing enhanced metrics on the result of the flight might be useful, but in most cases, it will not be able to use the information as the actions that can be performed are very limited (e.g., cost index tends to be fixed strategically by the airlines and not modified when generating the flight plans). • There are some particular instances when these enhanced capabilities might be useful: assessment of different flight plans when avoiding areas with turbulence or with ATFM regulations, estimating the fuel required for tankering activities, expected holding times in non-nominal conditions. • Identify the precursors of the different variations between planning and execution in order to highlight the factors influencing these variabilities.
Pilot[1]	• Crews will appreciate having a better understanding on the rationale between some of the decisions performed at dispatching (e.g., fuel on-board for holdings). • It would be beneficial to have an indication of the variances that they can expect during their flight and follow up rotations. • Dispatcher3 could provide information on the expected variance between the flight plan and the execution while indicating the precursors for these changes and advice on some flight operations (such as the possibility to recover some delay in the air).
Back-office analyst[3]	• Dispatcher3 will set up an infrastructure which enables the analysis of past flights to better identify situations and operations which could be optimised (e.g., selecting different baseline cost indexes).

[1] Main scope of Dispatcher3: predictive capabilities based on advanced machine learning and advice generator modules will be created, with a focus on flight analysis.
[2] May be considered in Dispatcher3: predictive capabilities for flights but with greater focus on the network, and identification of disruptions.
[3] Out of scope of Dispatcher3: will benefit from Dispatcher3 infrastructure and capabilities.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

7 Scenarios and case studies

As mentioned previously, Dispatcher3 is a data-driven project. That means that the definition of case studies is highly driven by the data availability; and the available data sources and their analysis will steer to a great extent the definition of case studies. It also means that case studies will be reviewed on-the-go alongside the data wrangling processes as new data insights are discovered. The definition of the case studies will be triggered by the research questions to be answered by the model and this will further define the scope of the case study and targeted indicators (KPIs) to be predicted. The indicators and features need to be estimated in the historical datasets from different data sources defined in the data catalogue. Moreover, as it was already mentioned in D3.1 deliverable, when defining a case study in the domain of machine learning, one needs to make sure that it captures the problem or a questions that can be effectively answered by some predictive model and the historical datasets available at hand, i.e., one predictive model essentially learns one mapping between the input (features) and the output (target, KPI) and a case study should be defined in such a way that we can expect a predictive model to be able to learn from the given data (e.g., for routes between a given origin-destination pairs or for flights arriving at a specific airport).

Concretely, a particular ML case study in Dispatcher3 might have to focus on one concrete route since a lot of the datasets we will be working with is route-dependent and the function that a trained model learned most likely would not perform well if the same trained model was tried to be used on a different route (due to the concept known as "**data drift**"). Additionally, independently of the route considered, specific case studies will focus on developing ML models for specific airport in order to predict the KPIs of interest such as the holding time at arrival, the runway in use and others. In this light, the project will make a clear distinction between **route-based case study** and **airport-based case study** although the outcome of both models may complement each other and used to provide more informative advice to final decision makers.

In order to structure all these different considerations mentioned above, **the definition of case study** will highly depend on **two different elements**, namely:

- **routes/airports of interest** - all the ML models developed within the project will be either route-based or airport-based. However, the experts within the consortium (i.e., Vueling and skeyes) may show a particular interest on the specific route/airport which is worth analysing from the operational point of view (i.e., the flights on this route are systematically prone to variations, the airports featured by highly congested TMA space)
- **prediction horizons** - for the given route and specified targeted indicators, different model outputs can be obtained at different time-frames. Therefore, these predictions-horizons should be considered when collecting the datasets that will be used to characterise the available information at the given time-frame. This means that the same dataset might be required at different time-horizons with different resolution. The predictions at different time-frames will tackle different roles involved in flight management process.

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Once the case study has been defined in terms of the route/airport and the time-frame considered above, a set of different **targeted indicators** will be estimated (i.e., holding at arrival, fuel deviation, etc.) forming a particular **sub-case study**. Some indicators might be more relevant than the others and they will be prioritised during the ML model development. For each indicator considered, different algorithms can be tested together with different data inputs.

Therefore, each single ML model will be characterized by the two components (as shown in **Table *5***):

1. **Case-study:** The route (or airport) and prediction horizon framing the data that ultimately will feed the model.
2. **Sub-case study**: The targeted indicator to be predicted.

7.1 Methodology to define case studies

As explained in Section 2, the definition and selection of the case studies play an important role during the machine learning development and validation; and as described in D3.1 deliverable, an interrelation between the specification of particular case studies and data acquisition and preparation exists, as each case study will trigger the process of data processing and preparation in order to be ingested by the predictive models. For this reason, a consultative approach is suggested. The consultation with the experts from the consortium (i.e., Vueling) and the feedback received from the first Advisory Board meeting will define the high-level situations to be considered with the expected outcome that could be obtained with each of them.

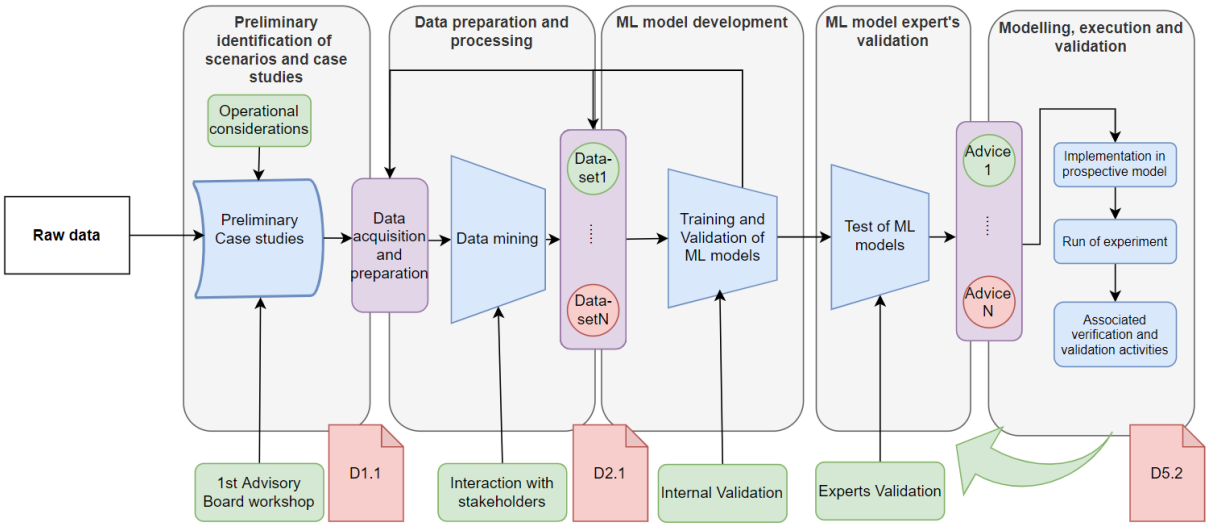


Figure 11. Definition, selection, instantiation and evaluation of case study

Figure 11 presents the approach followed to create the pool of potential case studies to be modelled and evaluated in Dispatcher3. First, the consortium carried out a **preliminary identification of case studies** by identifying a set of potential situations to consider when defining case studies to evaluate in Dispatcher3. These included elements such as route to be considered, the time-frame in which the prediction should be performed and the targeted indicators to be estimated. During the first Advisory Board meeting (held online on the 9th of October 2020), feedback was gained from airlines and experts on which operational aspects are more relevant. These included, for instance, the consideration of the key performance indicators to be estimated such as time deviation and fuel deviation. In addition to

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

these principal indicators, the Advisory Board also indicated the set of other relevant indicators which need to be considered in the machine learning models (e.g., holdings, outcome of fuel tankering, ATFM regulations, etc.,), thus identifying opportune time points when the Dispatcher3 prototype would be queried. Information was also gathered on how to present advice to dispatchers and pilots once the predictive model is ready. With all this information, a list of potential scenarios/case studies was created and reported in D1.1 - Technical Resources and Problem definition [7].

Since the data provided by Vueling will be mainly used for the purpose of machine learning development, the selection of case studies will be highly driven by its availability. Considering the prioritisation of time-frame (4 to 3 hours), roles (dispatcher) and indicators defined in D1.1 [7] for Dispatcher3, and the feedback from the Advisory Board and Vueling, a set of specific case studies are identified in D2.1 [8]. These specific case studies allow us to define the scope required for the data acquisition (as they focus on specific routes) and serve as a kick-off for the upcoming development activities. Following an incremental development approach, the prediction of relevant indicators in specific routes will be of the particular focus. Subsequently, a number of features used for the prediction will be analysed as part of the feature importance analysis and model validation. This means that incremental inclusion (and acquisition) of datasets could be performed as required in order to increase the potential performance of the models. It is worth mentioning that the interaction within the consortium (particularly with Vueling and skeyes) and with the Advisory Board (e.g., ad-hoc site visits) will be used to assess the relevance of the features and the performance of the predictions obtained by a machine learning model using the test dataset for a given route-time frame environment.

Finally, the outcome of the predictive models will be processed by advice generator module for planning activities which will transform it into advice considering airline policies. The assessment of the advice provided for the given KPI will be conducted in close interaction with Vueling who will be able to provide their feedback based on their operational experience/practice. Note that this process might be performed iteratively until the satisfactory level of advice is produced for a particular role and a given operational context. Nevertheless, the number of different case studies executed will increase as the project progresses and the prospective model matures. This will be also driven by the availability of new dataset along the project and the feedback obtained during the expert's validation activities.

7.2 Potential values for definition of case studies

As already mentioned, creating a new case-study requires a significant amount of effort, as data acquisition, preparation and model training of the KPIs. For this reason, we tried to define a reasonable number of case study along with the priority ranking of each. Considering case-studies which are operated (or similar to operated routes) by members of the Advisory Board is also considered of relevance as more in-depth feedback might be acquired from the results during the preparation of the case studies and the validation activities.

Nevertheless, producing an exhaustive and strict definition of all the potential case studies is avoided to enable the continuous refinement of specific case studies alongside the data management and descriptive analysis processes based on the principles of agile development methodology.

Table *6* provides preliminary specification of the case studies that was initially selected as those of a high relevance from the operational perspective. Note that as discussed, when the parameter to

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

predict is an operational factor (e.g., runway in use, holdings at arrival), these might not be route dependent but only airport dependent and therefore the prediction horizon might be also independent of a particular flight schedule reference, e.g., defining time-horizons for which the reference is the time at arrival at the airport rather than time prior departure.

Finally, due to the involvement of Vueling and skeyes in the project, the potential case studies will rely on their expertise and input and besides specific routes from Vueling flights, analysis of operations at EBBR might be considered.

Table 5. The composition of case-study and sub-case study relevant for the project

Prediction horizon (time-frame)	Route	Indicators to estimate (in order of priority per route)	Potential support to
24 h prior operation	Destination airport - Arrival airport	• Holding (non-nominal conditions)	• Tactical planner • Duty manager
10h to 9h prior SOBT		• Identify flights potentially affected by disruptions (delay)	• Dispatcher • Pilot
		• Identify congestion in network impacting flights	
4h to 3h prior SOBT		• Time deviations	
		• Fuel deviations	
		• Taxi times and fuel	
		• Impact on reactionary delay	

Table 6. Case studies based on the advisory board feedback and scope of the project:

Prediction horizon (time-frame)	Route	Indicators to estimate (in order of priority per route)	Potential support to
4h to 3h prior SOBT	LEBL – EGLL	London airports are located in a dense traffic TMA, the estimation of holdings is key to understand traffic in the area and possible impact on arrival delays.	• Holding time • Holding fuel • Time deviation • Fuel deviation
	LEBL - GCXO	Flights between Barcelona and Tenerife North are long and affected by changing weather conditions which might affect the total flight duration and arrival procedure to be used.	• Time deviation • Runway in use • Fuel deviation • Arrival procedure

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

8 Schedule of development and verification, integration and validation

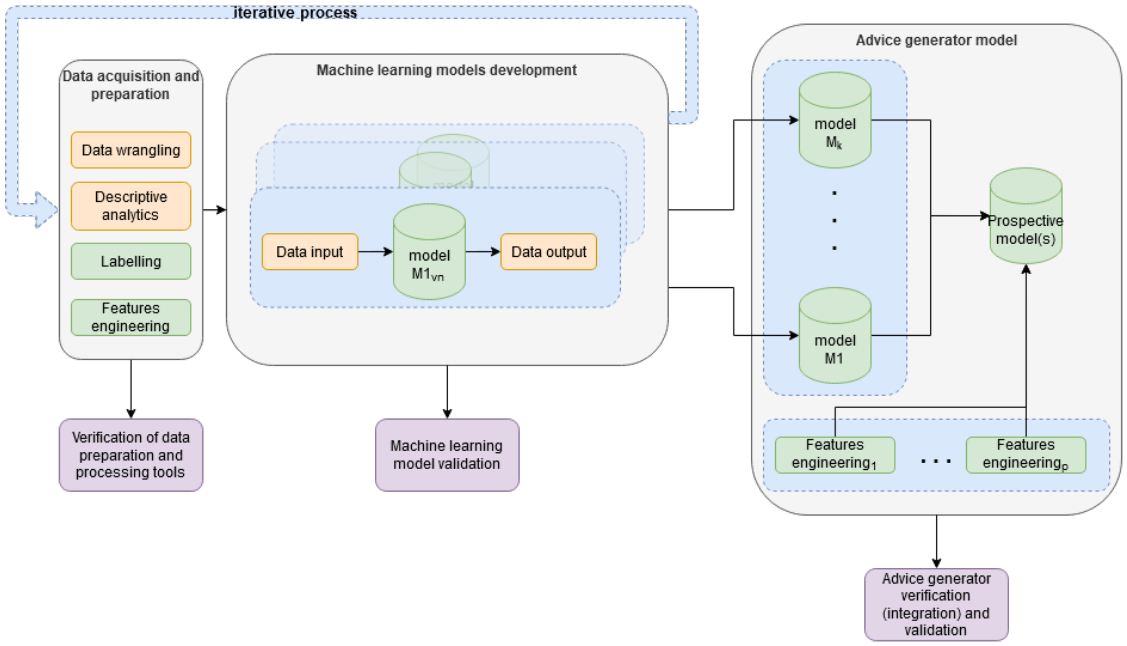


Figure 12. Model development and verification and validation activities

Figure 12 presents a simplified version of **Figure 2**. Two different phases can be observed on the model development of Dispatcher3:

- Machine learning models development (predictive models): this will require the implementation of algorithms to prepare the data required to train the models (i.e., data wrangling, descriptive analytics, and labelling and features engineering). These algorithms will be verified as indicated in Section 3, and these verification activities will be performed in parallel to the development which will be use case driven. Then the machine learning candidate models will be developed and evaluated as explained in Section 4. This iterative process will be conducted for different targeted indicators, time-frames and operational conditions as described in Section 7.
- Advice generator model(s) development (prospective model): the outcome of the previous activities will be, in general, low-level predictors, e.g., an estimator of the probability of having holdings at arrival at destination when estimating 3 hours prior departure. The advice generator will integrate these machine learning models in order to produce more aggregated estimation which could in some cases be model-driven. For example, an estimation of

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

probability of breaching a curfew could be generated by using several lower-level machine learning predictors; or a prediction of probability of having a holding at arrival requested 7 hours prior departure might combine the outcome of two machine learning models, one trained 3 hours prior departure and another model trained 9 hours prior departure. The prospective model will therefore, based on the 'raw' input data, select which machine learning models to use and combine their outcome to generate the desired estimation. This will require the integration and use of different feature engineering algorithms which are able to process the 'raw' data into the input required by the different models. As indicated in Section 6, different verification and integration activities will be carried out to ensure the quality of this process.

The capabilities of the advice generator will be limited by the different machine learning models previously trained but the advice generator being the layer interfacing the end user will be considered when providing the releases of the different versions of Dispatcher3 prototype.

Finally, note that the advice generator will also consider how to present the outcome of the predictions to the user and is the layer that will be considered for the different external validation activities which rely on the research questions defined in Section 6 as explained in Section 5.

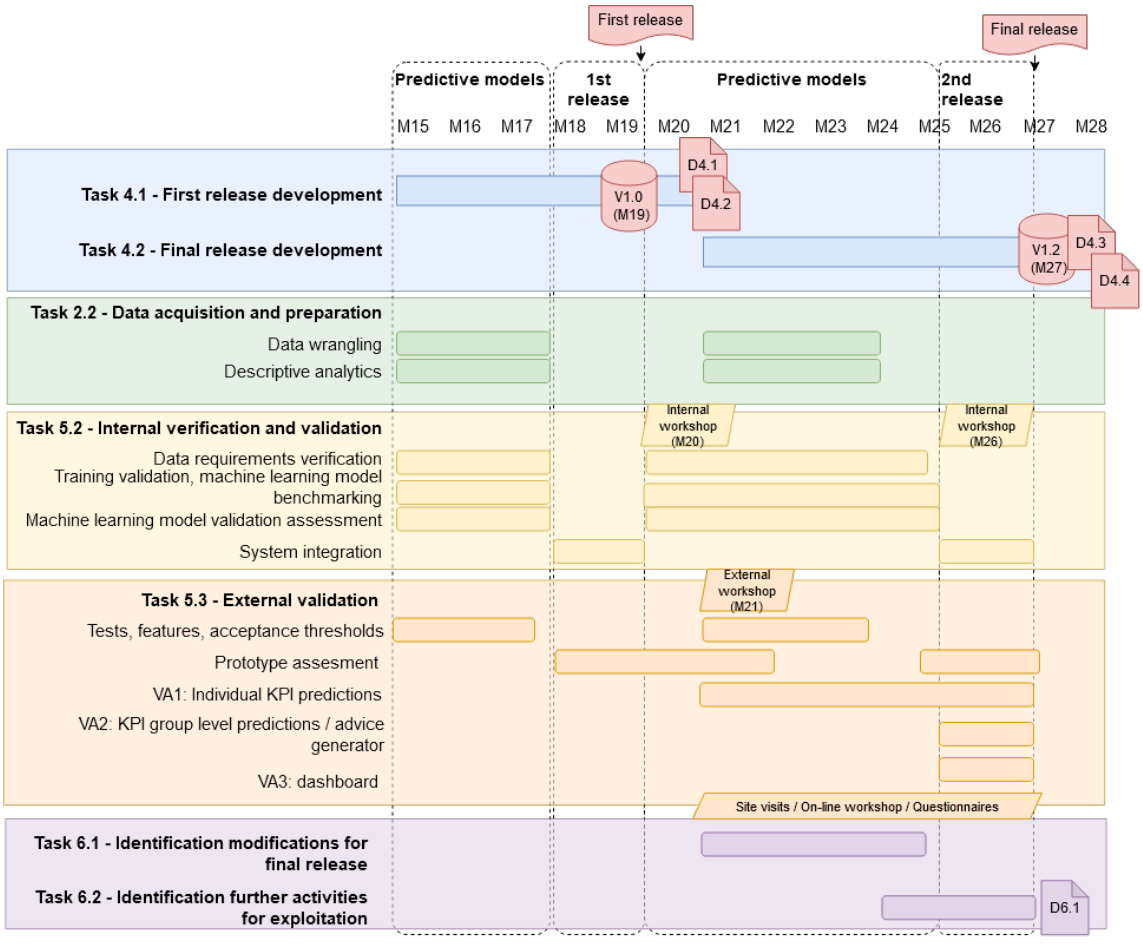


Figure 13. Verification and validation Gantt diagram

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

As previously indicated, in a data-driven project the activities of verification and validation are developed in parallel to the development of the different machine learning models. **Figure 13** presents a Gantt diagram of the different activities which are grouped into predictive model development tasks, which focus on the development of the machine learning models, and releases tasks, which focus on the integration and development of the advice generator.

From a development point of view both activities (the implementation of the machine learning models and of the advice generator) will be performed in parallel, but from a verification and validation perspective the focus will shift during different times in the project development.

The first development activities will focus on implementing the pipelines for preparing and processing the data while documenting and tracing the different models and datasets. This will be done with some small case studies and an iterative process will be conducted to develop further machine learning models for different indicators. Internal verification and validation activities focused on the verification of the data processing algorithms and on the development and selection of the machine learning models will be conducted. The external validation activities might require the interaction with members of the Advisory Board to validate the features and acceptance thresholds defined. This first activities are planned to last until end of M17 (October 2021). Then during M18 and M19 the different models implemented will be integrated into a first version of the advice generator albeit limited due to the rather small number of machine learning models developed. This will lead to the first prototype of Dispatcher3 aimed by M19 (December 2021).

An internal workshop will be performed to assess the first prototype and prepare the external workshop with the Advisory Board aimed at M21 (February 2022). Then, with the feedback obtained, the whole cycle will be repeated with further development of machine learning models (M21-M25) and integration in the prospective model (aimed to be delivered by M27 (August 2022)).

This second cycle will be supported by site visits, on-line interactions and questionnaires with the Advisory Boa to both gather feedback on the models development and to support the validation activities. Particularly the research questions. An internal workshop will be organised to assess the validation of the prototype and if deemed necessary and beneficial a second workshop with the Advisory Board could be organised to support these final validation activities.

Further potential improvements to be performed after the project finalisation to bring Dispatcher3 toward industrialisation obtained from the validation and interaction with stakeholders will be collected in D6.1 -- System evolution and uptake report (M27).

Finally, note that the current planning aims at delivering the final version of the prototype of Dispatcher3 by M27. Three months delayed with respect to the Grant Agreement, due to the delays experienced on data acquisition during the first part of the project due to COVID-19, and but still maintains a three-month buffer with respect to the 30 months of the action. This plan will be reviewed once the first prototype is produced.

For more information on the next steps on the project development see Section 10.

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

9 Conclusions

The verification and validation plan is a complex document which aims to provide guidelines for the validation of the machine learning models and the advice generator. Unlike traditional software development methodologies which typically follow the principal of the canonical verification and validation approach, the lines between development and verification and validation are often more blurred in the domain of the machine learning. The major reason for this resides in the nature of data driven projects in which the different activities involving data collection, data preparation, development and verification and validation are very inter-related. In order to encounter all these aspects, the verification and validation plan is based on the principle of an **iterative methodology** which assume that the "design-train-test-validate" process will be performed in an iterative manner. Such approach will trigger a backlog of back and forth tasks between different experts in the consortium in order to maximise the success of Dispatcher3.

As acknowledged by a great number of experts who deal with the verification and validation activities in the machine learning domain, the enhancement of the machine learning model highly relies on the verification and validation results. In addition, the acquisition and preparation of the historical data are crucial for developing high performance machine learning models. Thus, the sequence of the verification activities will be performed to ensure that the data feeding the machine learning models complies with the requirements specified in D1.1 [7], while dedicated **validation actions** are defined to ensure that the labelling and feature engineering activities are in line with the inputs obtained from the experts and final users. The remaining validation actions focus on quantifying the performance of the specific machine learning (predictive) model by using a set of standard metrics and will be conducted in close interaction with the experts who will monitor whether the development is evolving properly.

Finally, the set of verification and validation activities will be performed once the fully developed prototype has been developed relying on the output of the predictive models. The verification of the prototype as a **system** will be performed against the system requirements, while the validation of a whole system will also be performed in the close interaction with the experts from the Advisory Board and the consortium members in an iterative manner. Specific **research questions** have been defined and will be answered during these validation campaigns. The aim of these validation activities is to **assess the benefits** of the **decision framework** which aims to provide predictive capabilities and advice to the final customers.

The realisation of the verification and validation plan proposed in this document will be primarily underpinned through the **interaction** with the consortium industrial partners, Vueling and skeyes, but also with the Advisory Board and more broadly relevant stakeholders. These interactions are ensured by organising a **workshop** and **dedicated validation activities** (e.g., site visits, bi-lateral meetings) which will support the prioritisation of the case studies, the refining of the machine learning models, defining the thresholds of the certain metrics, while gaining more detailed information on operational context, airline policies and potential datasets. The case studies defined in this deliverable will be

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

highly driven by the initial input obtained from the consortium partners as well as the data availability and might be subject to modification/prioritisation during the course of the project (as a part of the iterative methodology adopted).

Overall, the verification and validation plan ensures that the project team can promptly identify the bottlenecks for both the machine learning model development and the data acquisition and preparation. Note that two internal workshops and interaction between the members of the consortium are also planned to support this promptly detection of potential issues.

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

10 Next steps and look ahead

The deliverable presents the comprehensive framework which defines the actions for the verification and validation of the prototype. As mentioned in this plan, the project will follow an incremental development for different machine learning models and case studies to be modelled. This mainly stems from the distinctive nature of the project which requires a large amount of datasets to be processed during the duration of the project. The consortium has already identified the datasets of interests and reported them in the data definition and processing report [8]. Moreover, the consortium has just internally signed the required Data Protection Agreement and Data Protection Annexes with Vueling so that their data can be used in the project.

Meanwhile, the development team has initiated the process of gathering the minimum dataset required for the creation of initial training datasets to develop machine learning models focusing on different target indicators (e.g., holding, time and fuel deviations, etc.) on a specific route and time-frames. The initial set of routes which will be considered has been already selected based on the feedback obtained from the industrial partners in the consortium, Vueling and skeyes, which datasets will be used in the project. For this purpose, the routes such as Barcelona (LEBL) - London Heathrow (EGLL) and Barcelona (LEBL) - Tenerife Norte (GCXO) have been identified as suitable initial route candidates for the analysis as the consortium experts already owned the substantial knowledge on the operational aspects of these routes. In particular, Vueling point out that these routes persistently experience some issue and are worth of analysing further, while skeyes' expertise as an ANSP will help us to properly understand potential operational constraints and challenges at Brussel's TMA for which dedicate case studies might be implemented. Additionally, we attempt to maximise the synergy with Pilot3, the Innovative Action project, which runs in parallel, and which also applies machine learning techniques for prediction of some indicators on the Barcelona (LEBL) - Frankfurt (EDDF) route.

In WP3, a review on predictive modelling techniques is under development with specific focus on the targeted variables and features to be considered on the different models to be developed in Dispatcher3. WP3 will also include a design of dashboard that could be used to visualise the advice from the predictions for dispatchers and pilots. All these techniques will be summarised in D3.1 - Data engineering and analytic techniques report planned for the end of September 2021 (M16). The validation activities will start closely monitoring the progress in WP4, and validation actions will start with the development of certain machine learning model, defining the thresholds on a specific metric and monitoring the model performance evolution over time.

The consortium aims at having a set of first main functionalities of the prospective model implemented and validated by the end of December 2021 (M19) (a brief description of the first release of the model will be produced as D4.1 - Technical documentation (first release) and D4.2 -Prototype package (first release)), accounting for three month delay period occurred as a result of delay in data acquisition. We will consider that the first prototype is ready when the Advice generator model has the integrated capabilities to transform new data to features (input for the required ML models) and to integrate the required time-frame ML outputs through visualisation of the results . Once the first prototype of the

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

different models is in place the external validation workshop will take place in February 2022 (M21) to obtain the feedback on the operational benefits and limitations of the results obtained from the experts that were not involved in the prototype development . In addition to the non-development part of the consortium team, the workshop will also gather different stakeholders including the Advisory Board members, the Topic Manager and other relevant experts.

Between the external validation workshop and the end of the project, a continuous interaction with the Advisory Board, the Topic Manager and the consortium members is planned in order to obtain support on the prioritisation of case studies and on the validation of the further developed model. Due to the COVID-19 outbreak, all these interactions might be arranged through online meetings. This will enable us to create the final release of the prototype (D4.3 - Architecture and prototype description (final release) and D4.4 - Prototype package (final release)). Further modifications and improvements to the system required in order to facilitate its industrialisation will be compiled in D6.1 - System evolution and uptake report.

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

11 References

1. Analytics Vidhya, 2021. Description available at: https://www.analyticsvidhya.com/ (accessed August 2021)
2. Breck, E., Polyzotis, N., Roy, S., Whang, S. and Zinkevich, M., 2019, April. Data Validation for Machine Learning. In MLSys.
3. Chu, X., 2017. Scalable and Holistic Qualitative Data Cleaning. PhD thesis at the University of Waterloo, Ontario, Canada.
4. Clean Sky 2, 2015 (Mar), Joint technical programme. Technical report. Version 5.
5. Clean Sky 2 Joint Undertaking, 2019, Horizon 2020 Clean Sky 2 Joint Undertaking bi-annual work plan and budget 2020-2021
6. Databricks platform. Available at: https://databricks.com/ (accessed August 2021)
7. Dispatcher3 Consortium, 2020 (Dec), Technical Resources and Problem definition, Tech. Report. Deliverable D1.1. Ed. 01.01
8. Dispatcher3 Consortium, 2021 (May), Data definition and processing report, Tech. Report. Deliverable D2.1. Ed. 01.02
9. Plutora, 2019. Verification vs Validation: Do You know the Difference? Available at: https://www.plutora.com/blog/verification-vs-validation (accessed August 2021).
10. Pullum, L., Steed, C., Jha, S.K. and Ramanathan, A., 2018. Mathematically rigorous verification & validation of scientific machine learning. Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States).
11. SafeClouds Consortium, 2019 (Sep), SafeOps Validation, Tech. Report. Deliverable D6.1. Ed. 01.00.
12. SafeClouds, 2020. European project description available at: https://cordis.europa.eu/project/rcn/206420/en (accessed August 2021)
13. Scikit Learn, 2021. Choosing the right estimator. Available at: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html (accessed August 2021)

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

12 Acronyms

AB: Advisory Board

ACARE: Advisory Council for Aviation Research and innovation in Europe

ADS-B: Automatic Dependent Surveillance – Broadcast

AG: Advice Generator

ANSP: Air Navigation Service Provider

ATFM: Air Traffic Flow Management

AWS: Amazon Web Services

CI: Cost Index

COVID-19: Coronavirus disease – 2019

CS2: Clean Sky 2

CSJU: Clean Sky 2 Joint Undertaking

DX.Y: Deliverable number (X=workpackage, Y=deliverable numbering within workpackage)

EBBR: Brussels Airport

EDDF: Frankfurt Airport

EGLL: London Heathrow Airport

FDM: Flight Data Monitoring

FMS: Flight Management System

FN: False Negatives

FP: False Positives

GCXO: Tenerife Norte Airport

H2020: Horizon 2020 research programme

HMI: Human Machine Interface

HP: Hypotheses

IADP: Innovative Aircraft Demonstrator Platform

ITD: Integrated Technology Demonstrators

JTI: Joint Technology Initiative

KPA: Key Performance Area

EUROPEAN UNION

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

KPI: Key Performance Indicator

LEBL: Barcelona Airport

LOO: Leave-one-out

LPA: Large Passenger Aircraft

LpO: Leave-p-out

ML: Machine Learning

QAR: Quick Access Recorder

RQ: Research Question

SGO: Systems for Green Operations

SOBT: Scheduled Off-Block Time

TBD: To Be Defined

TE: Technology Evaluator

TMA: Terminal Manoeuvring Area

TN: True Negatives

TP: True Positives

TRL: Technology Readiness Level

VA: Validation Actions

VAx: Validation Activity x

WP: Workpackage

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Appendix A - Questionnaires for validation

Table 7. Methods to address different RQs defined in VA1

RQs ID	Questionnaires/Flow chart diagram
P3-RQ-VA-010	"Specific role's acceptance of the tool with respect to individual KPI's prediction"

P3-RQ-IV-010:

1- Specific role's acceptance of the tool with respect to individual KPI's prediction

This set of statement is particularly **designed for the each specific role and with respect to individual KPI's prediction**:

Please indicate, by ticking the bullets, whether you agree or disagree with the statements given below when considering the statements designed to assess your general acceptance of the tool with respect to individual KPI's prediction.

1- Specific role's acceptance of the tool with respect to individual KPI's prediction	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1. With the KPI prediction provided, the **<role>** will have better awareness of his/her actions than in the case he/she needs to take the decision by him/herself	O	O	O	O	O	O
2. In the absence of the KPI prediction provided by Dispatcher3, the given KPI will not be intuitively easy to predict by the experts?	O	O	O	O	O	O

Please indicate any additional comments relevant for the above set of statements

_____

_____

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

Table 8. Methods to address different RQs defined in VA2

RQs ID	Questionnaires/Flow chart diagram
P3-RQ-VA-020	"Specific role's acceptance of the tool with respect to the group KPIs' prediction"
P3-RQ-VA-030	"Goodness of solutions in the given operational context (i.e., roles and time-frames)"
P3-RQ-VA-040	"Specific role's acceptance of the tool with respect to the Advice generator's capabilities"
P3-RQ-VA-050	"Interaction with the system (or Integration of the system into the operation TBD)"

P3-RQ-IV-020

2- Specific role's acceptance of the tool with respect to the group KPIs' prediction.

2- Specific role's overall acceptance of the group KPIs' prediction	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1. The KPI-group prediction provided by Dispatcher3 will facilitate the **<role>**'s action to take the appropriate decisions	O	O	O	O	O	O
2. With the KPI-group prediction provided, the **<role>** will have better awareness of his/her actions than in the case he/she needs to take the decision by him/herself	O	O	O	O	O	O
3. In the absence of the KPI-group prediction provided by Dispatcher3, the given KPI-group will not be intuitively easy to predict by the experts?	O	O	O	O	O	O

Please indicate any additional comments relevant for the above set of statements

_____

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

EUROPEAN UNION

P3-RQ-IV-030

3 - Goodness of solutions in the given operational context (i.e., roles and time-frames)

Please indicate, by ticking the bullets, whether you agree or disagree with the statements given below when considering the entire results obtained for the given scenario:

3- Goodness of solutions in the given operational context	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1. In the light of the obtained KPI-group prediction and advice within the operational context of the given case study, do you believe that Dispatcher3 is worth acquiring by your company?	O	O	O	O	O	O
2. In the light of the obtained KPI-group prediction and advice within the operational context of the given case study, do you believe that Dispatcher3 will contribute to better translate the policy of your company to the operational level?	O	O	O	O	O	O
3. The obtained KPI-group predictions generated by Dispatcher3 will be more accurate than the prediction based on the experience and judgement of the given expert ?	O	O	O	O	O	O

Please indicate any additional comments relevant for the above set of statements

_____

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

4- Specific role's acceptance of the tool with respect to the Advice generator's capabilities

4- Specific role's overall acceptance of the Advice generator's capabilities	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1. The information provided by the Advice generator will facilitate the **<role>**'s action to take the appropriate (and more informed) decisions	O	O	O	O	O	O
2. The information provided by the Advice generator contains a good trade-off between accuracy and interpretability which allows the expert to better understand the probabilistic nature of the information	O	O	O	O	O	O
3. The information provided by the Advice generator could be well aligned with specific airline policies.	O	O	O	O	O	O

Please indicate any additional comments relevant for the above set of statements

_____

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

EUROPEAN UNION

5- Interaction with dashboard

Please indicate, by ticking the bullets, whether you agree or disagree with the statements given below when considering acceptability of the mechanism for interaction between the pilot and the tool

5- Interaction with the system	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1. The Advice generator user workflow is appropriate and easy to use	O	O	O	O	O	O
2. The advice provided by the Advice generator can be easily integrated within the **<role>**'s job performances.	O	O	O	O	O	O
3. The advice provided by the Advice generator can be adopted in high stress levels of operation.	O	O	O	O	O	O
4. The mechanism which allows configuration of the Advice generator is appropriate and easy to use.	O	O	O	O	O	O

Please indicate any additional comments relevant for the above set of statements

_____

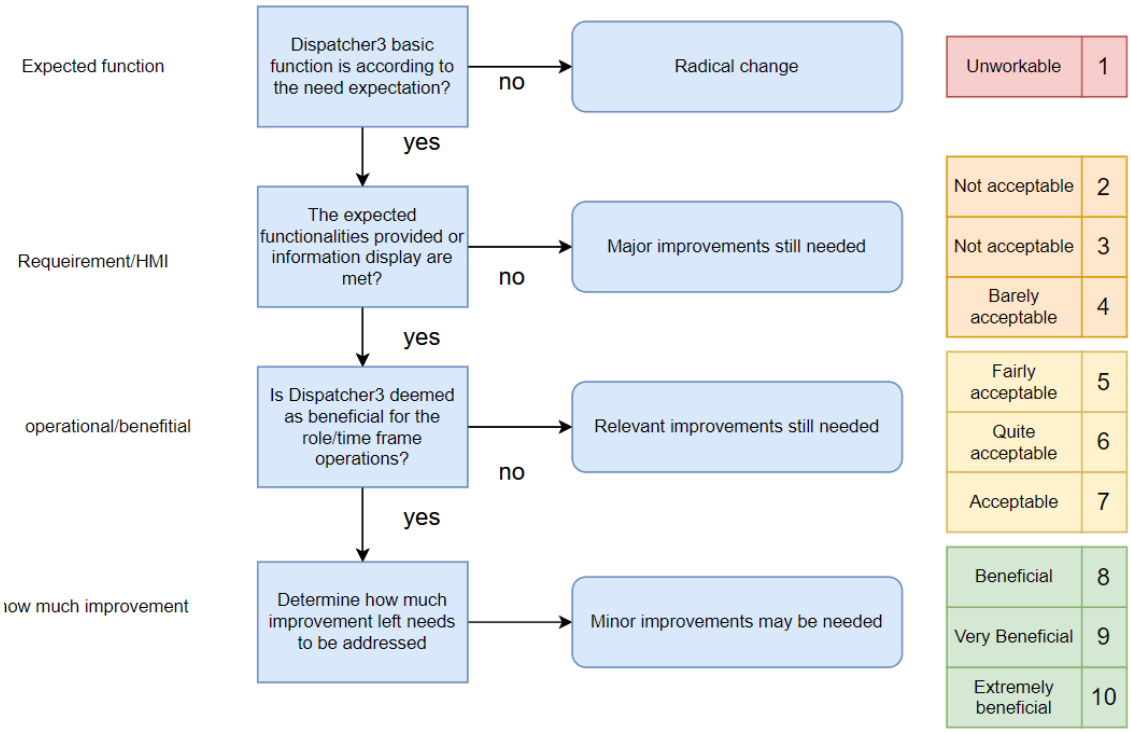
Table 9. Methods to address different RQs defined in VA3

RQs ID	Questionnaires/Flow chart diagram
P3-RQ-VA-060	"Flow-chart diagram for global acceptance"
P3-RQ-VA-070	"General acceptability"
P3-RQ-VA-080	"Easiness of understanding of the information"

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

1- Flow-chart diagram for global acceptance

Considering the solutions presented for the given scenario, express your overall acceptance of the Dispatcher3 HMI prototype by going through the scheme given below, indicating the final score by circling the appropriate numeric value (on the provided 1 – 10 scale).



Please indicate any additional comments relevant for the above set of statements

_____

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

2- Easiness of understanding of the information

Please indicate, by ticking the bullets, whether you agree or disagree with the statements given below when considering the easiness of understanding of the information provided

2- Easiness of understanding of the information	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1. 1. Information on the KPIs prediction and its impact on the advice provided is easy to understand	O	O	O	O	O	O
2. The information on the confidence level provided for each KPI prediction is clear and easy to understand	O	O	O	O	O	O

Please indicate any additional comments relevant for the above set of statements

_____

3- General acceptability

Please indicate, by ticking the bullets, whether you agree or disagree with the statements given below when considering general acceptability of the tool with the respect to quantity of information provided to the pilot.

3- General acceptability	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1. The information provided to the pilot is simple and concise enough	O	O	O	O	O	O
2. The amount of information presented to the pilot is well balanced (or is not overflowed)	O	O	O	O	O	O
3. The information provided to the specific role is predictable in its presentation	O	O	O	O	O	O
4. The visual representation of the advice together with KPIs predictions is clearly presented and well organised	O	O	O	O	O	O

Please indicate any additional comments relevant for the above set of statements

_____

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.

-END OF DOCUMENT-

© – 2022 – University of Westminster, Universitat Politècnica de Catalunya, Innaxis, PACE Aerospace Engineering and Information Technology, Vueling Airlines, skeyes. All rights reserved.