A Predictive Analytics Infrastructure to Support a Trustworthy Early Warning System

David Baneres 1,2,*, Ana Elena Guerrero-Roldán 1,2, M. Elena Rodríguez-González 1,2 and Abdulkadir Karadeniz 1,3

Abstract: Learning analytics is quickly evolving. Old fashioned dashboards with descriptive information and trends about what happened in the past are slightly substituted by new dashboards with forecasting information and predicting relevant outcomes about learning. Artificial intelligence is aiding this revolution. The accessibility to computational resources has increased, and specific tools and packages for integrating artificial intelligence techniques leverage such new analytical tools. However, it is crucial to develop trustworthy systems, especially in education where skepticism about their application is due to the risk of teachers' replacement. However, artificial intelligence systems should be seen as companions to empower teachers during the teaching and learning process. During the past years, the Universitat Oberta de Catalunya has advanced developing a data mart where all data about learners and campus utilization are stored for research purposes. The extensive collection of these educational data has been used to build a trustworthy early warning system whose infrastructure is introduced in this paper. The infrastructure supports such a trustworthy system built with artificial intelligence procedures to detect at-risk learners early on in order to help them to pass the course. To assess the system’s trustworthiness, we carried out an evaluation on the basis of the seven requirements of the European Assessment List for trustworthy artificial intelligence (ALTAI) guidelines that recognize an artificial intelligence system as a trustworthy one. Results show that it is feasible to build a trustworthy system wherein all seven ALTAI requirements are considered at once from the very beginning during the design phase.

Keywords: predictive analytics; artificial intelligence; trustworthy early warning system; standards and guidelines; software engineering in e-learning

1. Introduction

Educational technology has helped in the past years to evolve the way education is delivered in online settings [1] and on-site settings [2]. It can be used in different ways: providing a platform where the teaching and learning process is performed or providing tools for helping learners acquire skills or competencies, amongst others. In the former, the teacher is responsible for conducting the teaching process. However, in the latter, technology can partially or fully automate the learning process by assessing the learners’ requirements to complete the course. Here, it is inevitable that the learners’ behavior is monitored. Although privacy has been a sensitive topic in the past years, monitoring learners could effectively move education to a different dimension with a significant leap [3]. Monitoring learners produces a vast amount of data that when conveniently processed with educational data mining techniques allow for inferring valuable information about learners’ performance and behavior, and, in the end, it provides a personalized learning experience [4].
Moreover, this information can be further analyzed and exploited to extract knowledge about learners’ behavior [5] and therefore can be used to create intelligent adaptive systems for future learners to guide them to complete their learning outcomes successfully. Artificial intelligence (AI) techniques are more accessible during the past years through specific libraries for programming languages. Models to estimate unknown information about learners such as performance, acquired knowledge, or grades [6,7] can be trained and used to identify at-risk learners early. The identification could lead to applying some intervention mechanism to increase retention [8] or dropout reduction [9].

However, the application of AI-powered systems in education has some reluctance from practitioners [10,11]. Many systems are conceived as black boxes without knowing the complete set of features they offer. Moreover, the predictive models trained by the AI libraries do not offer an explanation about the performed predictions. Such reasons derive from an untrusting feeling towards their utilization. Therefore, trustworthy artificial intelligence (TAI) must be developed [12]. Workshops [13], challenges [14], recommendations [15], guidelines [16], self-assessment procedures [17], and standardization efforts [18] have been developed to build these TAI systems in the past years.

One of these intelligent adaptive systems is being developed through a project denoted as learning intelligent system (LIS) [19]. It aims to assist the learners during their learning process. The project focuses on developing an adaptive system (i.e., the LIS system) to be applied independently of the academic field or academic program in the custom learning management system (LMS) at the Universitat Oberta de Catalunya (UOC) to help learners succeed in their learning process. The system intends to provide learning analytics capabilities to analyze the learners’ progression for all stakeholders (i.e., learners and teachers) and predictive analytics capabilities to assess the likelihood to pass or drop out of a course. Additionally, the system can enhance likelihood predictions with automatic feedback and recommendations regarding learning paths, self-regulation, or learning resources.

The system is being developed using a mixed research methodology that combines action research methodology with a design and creation approach. Action research [20] allows for investigation and improvement of own practices, while design and creation [21] are especially suited for developing new information technology (IT) artifacts. The latter can be reduced to a problem-solving approach that uses an iterative process to improve the artifact on each iteration on the basis of conducting evaluations of the developed artifact on real settings, evaluating the results, and proposing improvements for the next iteration. This paper focuses on the second iteration of the project when the infrastructure was redesigned, aiming to create a trustworthy system from the first iteration [22]. During this iteration, the infrastructure was built to gather the data, process them, and perform predictions to evaluate the likelihood of failing a course. Such a predictive model has been used to create an early warning system (EWS), which aims to early detect learners at risk of failing or dropping out of a course, applying intervention mechanisms to prevent a worse situation.

A critical issue is to deliver a trustworthy EWS for a heterogeneous campus where multiple disciplines are taught, and where different learning methodologies and assessment models are used. In the past years, we can find some EWS developments [23–25], but there is no reference about extending as far as we know to a trustworthy system. This paper presents the build system that is considered trustworthy because it follows the requirements stated in the European self-assessment Assessment List for trustworthy artificial intelligence (ALTAI) guidelines [17]. Although several guidelines can be considered, this paper follows the ALTAI ones because our institution is located in Europe. The paper also discusses the main results coming from the self-assessment analysis carried out.

The paper is structured as follows: Section 2 summarizes the related work to TAI systems, and Section 3 describes the used materials and methods for the self-assessment analysis. Section 4 introduces the EWS, the infrastructure that supports the system, and the predictive models to have evidence to justify the system’s trustworthiness. Finally, the
results derived from the analysis are discussed in Section 5, while conclusions and future work are presented in Section 6.

2. Related Work

It is essential to focus on data management aspects when AI systems are designed. Firstly, it is mandatory to fulfill the data privacy regulations of the country or region where the data are managed when a system is handling sensitive personal data. An example can be found in the European Union with the General Data Protection Regulation (GDPR) [26], but most countries have their own legislation about data privacy management. Secondly, it is also essential to consider a data policy. However, data policies have not been standardized, and there is not a unique policy to apply. If we focus on higher education institutions, each organization is applying its own policy on the basis of the challenges they faced during the analysis of the data it possesses [27–29]. Nevertheless, data policies are not enough [30] when we face predictive analytics infrastructure built upon AI requirements. AI needs to consider more requirements because data are not only used for analyzing and visualizing information. Data are also used for impacting people’s decisions and behavior. Thus, ethical issues appear as a new relevant dimension. Governments and institutions are investing efforts to define guidelines to develop TAI systems and standards that should positively impact citizens [13–18]. Although there are different definitions, recommendations, challenges, guidelines, and standardization procedures, all are working on similar concepts to define the requirements to be analyzed.

The TAI system concept has been studied in several works [31–34]. The authors of [35] (p. 5) define a TAI system as “a system that is developed, deployed, and used in ways that not only ensure its compliance with all relevant laws and its robustness but especially its adherence to general ethical requirements.” The general ethical requirements include inclusive growth, sustainable development, and well-being; human-centered values and fairness; transparency and explainability; robustness, security, and safety; and accountability [36]. The authors of [37] (p. 29) proposed that “in order to make it trustworthy, computer scientists need to consider measures . . . , robustness to data set shift, robustness to poisoning, fairness, interpretability, end-to-end service-level provenance and transparency, and application for social good”. In other words, a “TAI system should provide beneficence, non-maleficence, autonomy, justice, and explicability” [38] (p. 696).

In all these definitions, the requirements can be reduced to a single word: trust. Stakeholders should trust the system, and trustworthiness is accomplished by understanding and making explicit system capabilities explicit. Without knowing all the features, stakeholders will have concerns about feeling completely confident with the system. Note that visual interface design is a key issue to improve confidence. For instance, the author in [39] identified emotional and usability criteria factors that directly impacted interface acceptance. Thus, a good user interface design could lead to a greater confidence level. This effect is even more critical in education. Teachers guide learners, and they tend to fully control all the learning process, from the materials and contents to the activities and feedback provided to the learners. Introducing in the learning process an instructional technology with AI involves a new element in the equation. Learners may discover adverse effects on personalization by feeling that the system is not treating all learners equally [40], and teachers may have doubts about the applicability value of AI. However, there are clear positive effects on the utilization of TAI systems in education. The authors of [41] (p. 3) describe the positive benefits: “personalized learning, the support of learners with special needs, . . . reduce dropout, and assessing new skillsets.” Moreover, TAI systems can contribute to the learning design research field [42] by producing technological tools that are easy to share and reuse [43]. Best practices on dashboard visualization or data management can be shared among practitioners and engineers to produce other systems compliant with TAI requirements.
TAI systems should be designed, built, assessed, and tested in real settings. Previous guidelines could help on their development by combining the knowledge and the experts’ experience on education and technology. However, it is critical to know the hidden infrastructure behind the system [44] to trust it better. As far as we know, some infrastructure designs on AI in education can be found [45–48], but no study case supports a TAI system. Some systems focus on security [49], explainability [50], privacy [51], ethics [52], fairness [53], human well-being [54], and trust [55]. Some of the previous works even analyze together different requirements related to a TAI system at once. However, no one extends the system to a trustworthy one, although some recent work [56] suggests the importance of considering such characteristics in the software life cycle development (from the system design to its maintenance) when educating software engineers. This paper focuses on self-assessing the trustworthiness of an EWS by applying the European ALTAI guidelines based on the performed infrastructure design.

3. Materials and Methods

3.1. University Description

The UOC, from its origins (1995), was conceived as a purely online university that used information and communication technology (ICT) intensively for both the teaching/learning process and management. UOC has a custom LMS where massive data are generated from the interactions between learners, teachers, and administrative staff. The educational model and the design of the online classrooms are centered on the learner and the activities across the courses. The assessment model focuses on continuous assessment with different activities during the semester, combined with a summative assessment at the end of the semester with a face-to-face final exam. The learner receives qualitative grades and personalized feedback. The qualitative grades are based on the following scale: A (very high), B (high), C+ (sufficient), C− (low), and D (very low), where a C− or D grade means failing the assessment activity. In addition, another grade (N, non-submitted) is used when a learner does not submit the assessment activity.

The learner profile at UOC is particularly different when comparing with on-site universities. Learners are aged up with family and partial or full-time employment. They usually devote time to learn to get a better position in their job or enhance their knowledge with another bachelor’s degree. Learners at UOC are competence-oriented, but they require tight self-regulation to combine personal and professional duties. New learners have significant difficulties self-regulating, and it is a critical skill to be learned in the first semesters. For such reasons, providing an EWS to guide them personally depending on their background and profile could enhance their learning process to pass the courses successfully.

3.2. Data Source: The Universitat Oberta de Catalunya Data Mart

The massive amount of data produced by the custom LMS are distributed among different operational data sources. The institution performed an effort to merge all these data into a centralized database: the UOC data mart [57]. An ETL (extract, transform, load) process collects and aggregates data from the different data sources solving common problems on distributed systems as fragmentation, duplication, and different identifiers. Additionally, the ETL process anonymizes the sensitive data (i.e., personal data are obfuscated, and the different identifiers of the same person are changed to a unique one [27]).

In the end, the data mart incorporates data about enrollment, accreditation, assessment, and transcripts, among others, but not demographic data. Moreover, data about navigation, interaction, and communication within the learning spaces available at the custom LMS are stored. As a result, the UOC data mart offers: (1) historical data from previous semesters, and (2) data generated during the current semester. Conceptually, the relevant data are stored as a tuple containing five attributes (U[D], T, S, R, X) based on xAPI specification [58]. The tuple represents that “user U (optionally, using device D) at time T employs service S on Resource R with result X.” All these tuples are stored to a single entity that is related
to others that stores data about users (U), resources (R), and results (X), following a star schema [59].

The data mart is implemented in Amazon DynamoDB. Although this database engine was initially designed as a key-value NoSQL [60] database, it evolved into a document-oriented one to store JSON-formatted documents. This structure is helpful to store semi-structured data and, following cloud features supported by Amazon Web Services, facilitates horizontal scaling and distributed processing through the Hadoop ecosystem [61]. Although data are persistently stored in the DynamoDB database, data are daily and semesterly exported on an S3 Bucket to avoid direct access to the database. Tuples are split into different datasets (i.e., files) on the basis of the value of the resource attribute R in CSV format.

3.3. European ALTAI Guidelines

The High-Level Expert Group on Artificial Intelligence designated by the European Commission published the Ethics Guidelines for Trustworthy Artificial Intelligence in 2019 [36]. Within the guidelines, an assessment list was published to assess an AI system as a TAI one. The system must adhere to the seven requirements for being considered a TAI system:

1. Human agency and oversight;
2. Technical robustness and safety;
3. Privacy and data governance;
4. Transparency;
5. Diversity, non-discrimination, and fairness;
6. Societal and environmental well-being;
7. Accountability.

The ALTAI guidelines [17] aims to protect people's fundamental rights. It is intended for organizations to understand what a TAI system is and the risks that may generate on their stakeholders. It is focused on minimizing those risks while maximizing the benefits the TAI system can bring.

This paper focuses on self-assessing the trustworthiness of the presented EWS by applying the European ALTAI guidelines on the basis of the performed infrastructure design. Before the analysis is performed, the infrastructure and the EWS are presented.

4. The Predictive Analytics Infrastructure and the Early Warning System

In this section, the predictive analytics infrastructure is presented. The design decisions are based on which type of infrastructure and architecture are used, which final design has been performed at UOC, and which languages have been used to leverage the challenge to read heterogeneous data from the UOC data mart. Additionally, the EWS is presented with the used predictive model and the information delivered to teachers and learners. All this information will help in the next section to perform the self-assessment of the TAI system.

4.1. The Predictive Analytics Infrastructure

4.1.1. Microservices Infrastructure

The infrastructure design is crucial for an effective implementation but also for the success of a TAI system. Although monolithic infrastructures are still used and have some advantages, infrastructures based on microservices are recognized as a growing good practice [62] because of their major benefits. The system is transformed into packages of small services, each one independently deployable, horizontally scalable, and running its own processes. Moreover, the benefits can be seen during development because there is flexibility to use different technologies and a reduction in development cycles and time-to-market [63]. However, some challenges must be faced because system complexity increases. Microservices need to communicate, and lightweight and secure channels must
be provided. Moreover, testing each microservices in an isolated way is more complex, and fault tolerance must be enforced while interacting with other microservices.

Currently, many technology companies are following this paradigm to develop their products. Some relevant examples can be found in [64,65]. The developed predictive analytics infrastructure follows the microservices infrastructure style supported by DevOps practices:

- **Microservices**: Each service has been developed independently and scoped on a single purpose. Docker technology [66] has been used to containerize each service.
- **Continuous integration and delivery**: The code is managed with GitLab [67], and containers are created automatically within the GitLab server and stored in a private Docker Registry to allow continuous delivery [68].
- **Infrastructure as a code**: Complete infrastructure is managed by Docker Compose [69] with a single configuration file obtaining the latest version of each service from the Docker Registry.
- **Monitoring and Logging**: Although Docker Compose provides a simple logging process to monitor the services, the client service provides several dashboards to supervise batch tasks run by the different microservices.

We used an on-premise instance to deploy the infrastructure due to the research nature of the project. Nevertheless, the system can be easily reproducible on any cloud environment supporting Docker, Docker Compose, and GitLab technologies.

4.1.2. Four-Tier Architecture

Multitier architecture is a well-known model used to develop client/server applications. The most widespread architecture is the three-tier one with the presentation, application, and data management tiers [70]. Our system proposed a four-tier architecture where the application tier is divided into a data-dependent (DD) and a data-independent (DI) one. The architecture is depicted in Figure 1.

![Four-tier architecture](image)

**Figure 1.** Four-tier architecture.
This architecture allows for splitting processing data services from loading and serving data ones. This idea adds simplicity to the infrastructure and prevents dependency on institutional data representation. The DD tier is responsible for the ETL process to load data from the institution, for the messaging service (i.e., send recommendations to learners), and for presenting the data to the different stakeholders through the client tier. DI tier oversees processing the data, training models related to predicting learners’ behaviors, and performing predictions on the basis of those models.

DI tier can hold multiple models and can support data from multiple data sources. This is possible because the DD tier has been designed to create a view from the available sources coming from the university Student Informational System (SIS). This view is stored and secured in the Persistence data tier. Then, the DI tier can create and handle different models from the available data in the view. As aforesaid, this approach allows high flexibility since any change in the university SIS data model has only to be informed in the view’s specification on the DD tier without affecting the DI tier models.

4.1.3. UOC Predictive Analytics Infrastructure

This section describes the development of the infrastructure for UOC on the basis of the previous four-tier architecture. The infrastructure with the internal microservices is depicted in Figure 2. The client tier is the only access point to the system by accessing the public web service. This service is responsible for visualizing the information to the users, and there is a view for each role. Currently, five roles are available:

- The learner can see the warning level of failing a course, and they have access to personalized recommendations.
- The instructor teacher guides and monitors the learner’s learning process throughout a specific online classroom in a course. They can also send recommendations and see the assigned warning levels in their classroom.
- The coordinating teacher designs the course and is ultimately responsible for guaranteeing that the learners receive the highest quality teaching. They can configure the system for the course, manage recommendations, and have a full view of the learners’ progress.
- The administrator manages the whole system. They can configure available models for courses and can access the logs and monitoring tools of the system.
- The developer manages the low-level details of the infrastructure. They can run specific operations on the DI tier, schedule tasks, check models, and access the logging information of the system.

Additionally, Traefik [71] is used as a reverse proxy to create a namespace for the web client. Moreover, Traefik can be used as a load balancer when high traffic demand is expected by running multiple containers for the web service.

The DD tier is responsible for forwarding all operational requests related to gathering the data from the university SIS and serving data on the public web service. Currently, the tier has embedded three services. The loader service gathers the data from the university SIS on the basis of the specification file that describes the view to be created (a complete description of the custom query language can be found in Section 4.1.4). Such a view is created through Pentaho Data Integration jobs [72] by reading the data from the UOC data mart and storing the view as a collection on a MongoDB instance. Note that the service also performs a validation step for cleaning inconsistent and incomplete data because such validation is not currently performed on the UOC data mart and it highly affects the models’ accuracy.
4.1.3. UOC Predictive Analytics Infrastructure

This section describes the development of the infrastructure for UOC on the basis of the specification file that describes the view to be created. Such a view is created through a custom query language used to define the predictive models. The computational service is responsible for reading the data from the UOC data mart and training/predict models using AI classification techniques. The computational service manages these operations and forwards all the operations related to the use of the data (i.e., read from SIS and train/predict models using AI classification techniques). The computational service that manages these operations can be reviewed on [77]. Note that a second specification file is used to maintain the adaptiveness to the institution data view. Section 4.1.5 highlights the custom language used to define the predictive models.

The DD tier also provides a service to run statistical analysis operations with R language [73] and a service to send messages to learners. Although the R service is not crucial for learners and teachers, it is a valuable resource to perform statistical analysis for administrators and developers. Currently, R-based scripts are hardcoded in the system, but it is expected to run specific scripts from the web client in the future. As we will explain in Section 4.2, the EWS provides a warning level to learners on the basis of the likelihood of failing the course and an intervention mechanism to help the learners succeed in the course. This intervention mechanism is crucial to effectively provide learners explanations, information, and recommendations about passing the course. The intervention mechanism is implemented as a recommender service that sends messages to learners using GMAIL API (the institution has G Suite support). The sending process can be performed manually by the teacher or automatically by the system using the teachers’ email account. This decision design aims to increase the message’s effectiveness since a learner will take into more consideration a message received from her teacher than a message received from a generic email account. Note that the teacher must grant access rights to the recommender service to use her email account.

We can observe that the DD tier has only an endpoint to access data, and it is controlled by a Restful API [74]. This API provides access to all operations that can be forwarded to the DI tier, such as serving data to the web client, configuring the system for the institution, and running operations related to the predictive models. Although all tiers, except the client one, are in private networks, the communication is secured by OAuth 2.0 protocol [75]. Related to serving information to the web service, the data are deanonymized using the university’s anonymizer service. Since considerable delays have been detected on courses with a high number of enrollments, we used a Redis cache [76] to leverage the operation.

The DI tier has only an access point from the Restful API. It is responsible for reading sensitive data and forward all the operations related to the use of the data (i.e., read from SIS and train/predict models using AI classification techniques). The computational service that manages these operations can be reviewed on [77]. Note that a second specification file is used to maintain the adaptiveness to the institution data view. Section 4.1.5 highlights the custom language used to define the predictive models.
Finally, the persistence data tier supports all the infrastructure by using different databases. Sensitive data about learners are encrypted and stored in a MongoDB instance. The data from the web client related to the courses’ distribution, courses’ activities, and users’ preferences are stored in a MySQL instance. Finally, the specification files are also stored in the persistence data tier on disk storage.

4.1.4. JSON-Based Query Language for ETL Process

One challenge of the infrastructure is to provide adaptiveness to different institutional data models. To simplify the task, we assumed the data as the UOC data mart currently provides them. Data are stored in a secured URL path where a dataset in CSV format represents data coming from one service of the institution (i.e., enrollment, submission, grading, etc.).

One option is to use a standard query language using an intermediate format as JSON [78–80] or directly querying the CSV format files [81,82]. However, such approaches need a highly experienced user to build the queries. We propose a more straightforward method that even a non-expert practitioner or researcher on query languages may use.

We propose a JSON-like format language to perform queries. The language can perform the most common operations on any database query language as select-from-where sentences combined with joins, aliases, and aggregated fields with max/min/count/average functions. Figure 3 describes the grammar in extended Backus-Naur form (EBNF).

```
views ::= ['<view> (<view>)* ']
view ::= ['<view> '<name_view> ',' '<fields_view> ']
name_view ::= 'Name View:' '<identifier>
fields_view ::= 'Fields View:' ['<field_view> (' '<field_view>)* ']
field_view ::= ['<select> | '<aggregation> ',', '<from_source> ',', '<from_dataset> ']
select ::= '<list_field> (',' '<list_alias> ')? (',' '<conditional> )?
list_field ::= 'Fields: [' '<identifier> (',' '<identifier>)* ']
list_alias ::= 'Alias: [' '<identifier> (',' '<identifier>)* ']
aggregation ::= '<single_field> ',' '<single_alias> (',' '<conditional> )? ',' '<aggregate>
single_field ::= 'Field: [' '<identifier> ']
single_alias ::= 'Alias: [' '<identifier> ']
conditional ::= '<where> (',' '<collectby>)?
where ::= 'Condition: '<Java_like_condition>
collectby ::= 'CollectBy: [' '<identifier> (',' '<identifier>)* ']
aggregate ::= 'Aggregate: (MAX | MIN | COUNT | AVG )
from_source ::= 'Source: '<identifier>
from_dataset ::= 'Name Dataset: '<identifier>
identifier ::= ( [0-9] | [A-Z] | [a-z] | '_' )+
```

Figure 3. Extended Backus-Naur Form (EBNF) for the query language for the ETL process.

Multiple views can be defined in the same file. A single view name_view includes different fields field_view that can gather fields from different sources from_source and datasets from_dataset. As aforesaid, a dataset can be queried in different ways. Fields
can be selected (i.e., list_field), aliased with a different name (i.e., list_alias), filtered by a condition (i.e., where), collected by a set of fields (i.e., collectby), and aggregated by a function (i.e., aggregate). The views are managed by the Pentaho Data Integration jobs system embedded in the Loader service. The result is a collection for each view in MongoDB, where each document in the collection stores the data for one learner identified by their anonymized identifier. Each document has embedded subdocuments with the from_dataset as key-value, and a second level of embedded documents can be found when the collectby operation is used. As a schemaless database, MongoDB offers the flexibility to store only the available data for each learner. Furthermore, it also allows recovering learner data efficiently by avoiding join operations.

4.1.5. JSON-Based Query Language for Model Creation

Similar to creating the views, we need a simple method to query the views from the MongoDB collections and generate the datasets for training and testing the predictive models. Using hardcoded MongoDB aggregation pipelines will break the data independence philosophy. Thus, a second language has been defined to create the datasets. This language is used to create a dynamic aggregation pipeline specifying the view, the datasets (i.e., the key values of the embedded documents in the view collection), and fields (i.e., fields within the embedded documents). Figure 4 describes the grammar in EBNF.

Figure 4. EBNF for the query language for the model creation process.

```plaintext
models ::= ['< model >', '< model >)* ']
model ::= ['< name_model >', '< name_view >', '< outcome >', '< slice >',
            '< fields_model >']
name_model ::= 'Name Model : < identifier >
name_view ::= 'Name View : < identifier >
outcome ::= 'Outcome : < field_model >
slice ::= 'Slice : < field_model >
fields_model ::= 'Fields Model : ['< field_model >', '< field_model >)* ']
field_model ::= ['< name_dataset >', '< list_field >', '< where >)? ('< collectby >)?
             '< default_values >']
name_dataset ::= 'Name Dataset : < identifier >
list_field ::= 'Fields : ['< identifier >', '< identifier >)* ]'
where ::= 'Condition : <Python like condition>
collectby ::= 'CollectBy : ['< identifier > ']
default_values ::= 'Default : ['< default_value >', '< default_value >)* ']
default_value ::= (< integer > | < float > | < string > | < boolean > )

integer ::= ([0-9])+  
float ::= ([0-9])+ '. ([0-9])+  
string ::= '"' 'identifier '"'  
boolean ::= '"' ('true' | "false") '"'  
identifier ::= ([0-9] | [A-Z] | [a-z] | _ )+
```

Figure 4. EBNF for the query language for the model creation process.
The model needs a name (i.e., name_model), a collection to search the data (i.e., name_view), an outcome to predict (i.e., outcome), and a list of features (i.e., fields_model). The dataset is created by gathering the different fields from the different embedded subdocuments in the learners’ document. Note that the schemaless features of the MongoDB database may imply empty fields for some subdocuments. The specification file allows for setting default values for such empty fields (i.e., default_value). The slice argument helps to cut the data among all the embedded subdocuments by a field. For instance, data can be gathered for a course and activity slicing by the course identifier and the activity number.

Like the query language for the ETL process, embedded subdocuments can be queried in different ways: (1) querying by single or multiple fields (i.e., list_field), (2) filtering by the value of a field (i.e., where), or (3) creating new features in the dataset on the basis of the values of a field (i.e., collectby). For example, let us assume a course with three activities that have already been graded. The information is stored in the MongoDB instance, as observed in Figure 5a. Using the query described in Figure 5b, we created the dataset of Figure 5c with three features. These features have the values of the field activity as names and the values of the field grade as values. This query language creates a powerful and dynamic query method independent of the number of fields and only depends on a specific field’s values. In the EWS, this query option helps to generate a dynamic model for each course (i.e., slicing by the identifier of the course) and retrieving data from all the activities from a unique model specification.

### MongoDB

**activity_assessed:**

```
[
  { Activity: 'AA1', Grade: 'B' },
  { Activity: 'AA2', Grade: 'A' },
  { Activity: 'AA3', Grade: 'D' }
]
```

**Query**

```
{
  "View": "activity_assessed",
  "CollectBy": ["Activity"],
  "Field": ["Grade"],
  "Default": ["N"]
}
```

**Dataset**

```
..., AA1, AA2, AA3, ...
..., 'B', 'A', 'D', ...
```

Figure 5. Example of creation of a dataset from a query where (a) shows the data stored in the MongoDB, (b) defines the query to create the dataset for training a predictive model, and (c) shows the obtained dataset from data in MongoDB.

This specification file is managed by the computational service responsible for training, testing, and performing predictions from the available models. It is worth noting that the query operation accepts multiple arguments to create a specific model: the range of semesters where the data should be gathered and the value for each sliced field. Following the example in Figure 5, a dataset can be created for a course by setting up the identifier of the course and the ranges of semesters from where the data are gathered.

### 4.2. The Early Warning System

#### 4.2.1. Profiled Gradual At-Risk Model

This section presents the predictive model developed for the EWS to detect at-risk learners on individual courses. The model, denoted as the Profiled Gradual At-Risk model (PGAR), aims to detect the likelihood to fail a course on the basis of the grades of the continuous assessment activities and some profiling information of the learners. This model is an evolution of the Gradual At-Risk model described in [77] by adding the
learner’s profile. These last features improve the identification of at-risk situations by knowing the different kinds of enrolled learners. The profiling information in the UOC data mart is currently quite limited, and only academic data can be used. For a learner, the model uses the number of enrolled courses in the current semester to estimate the learner’s workload, the number of times they have repeated the course where the prediction is performed to know their aptitude within the course, whether they are new in the university to evaluate self-regulation capacity, and the GPA (global point average) that provides an estimation about their average performance during previous semesters at the institution.

A PGAR model is built for each course, and it is composed of a set of predictive models defined as submodels. A course has a submodel for each assessment activity that features the profiling information previously defined and the grades of the activities until the current one. The model’s outcome is to fail the course, and it is a binary variable (i.e., fail or pass).

**Example 1.** Let us describe the PGAR model for a course with four assessment activities (AA). In such a case, the PGAR model contains four submodels:

\[
\begin{align*}
 Pr_{A_{A1}}(\text{Fail}? &= (\text{Profile}, \text{Grade}_{A_{A1}}) \\
 Pr_{A_{A2}}(\text{Fail}? &= (\text{Profile}, \text{Grade}_{A_{A1}}, \text{Grade}_{A_{A2}}) \\
 Pr_{A_{A3}}(\text{Fail}? &= (\text{Profile}, \text{Grade}_{A_{A1}}, \text{Grade}_{A_{A2}}, \text{Grade}_{A_{A3}}) \\
 Pr_{A_{A4}}(\text{Fail}? &= (\text{Profile}, \text{Grade}_{A_{A1}}, \text{Grade}_{A_{A2}}, \text{Grade}_{A_{A3}}, \text{Grade}_{A_{A4}})
\end{align*}
\]

where \( Pr_{A_{An}}(\text{Fail}? \) denotes the name of the submodel to predict whether the learner will fail the course after the activity \( A_{An} \). Each submodel \( Pr_{A_{An}}(\text{Fail}? \) uses the learner’s profile (Profile) and the grades (Grade\(_{A_{A1}}\), Grade\(_{A_{A2}}\), \ldots, Grade\(_{A_{A_n}}\)), that is, the grades from the first activity until the activity \( A_{An} \).

Each submodel is evaluated on the basis of different accuracy metrics. We use four metrics [8]:

\[
\begin{align*}
 \text{TNR} &= \frac{TN}{TN + FP} \quad \text{ACC} = \frac{TP + TN}{TP + FP + TN + FN} \\
 \text{TPR} &= \frac{TP}{TP + FN} \quad F_{1.5} = \frac{(1 + 1.5^2)TP}{(1 + 1.5^2)TP + 1.5^2FN + FP}
\end{align*}
\]

where \( TP \) denotes the number of at-risk learners correctly identified, \( TN \) the number of non-at-risk learners correctly identified, \( FP \) the number of at-risk learners not correctly identified, and \( FN \) the number of non-at-risk learners not correctly identified. These four metrics are used for evaluating the global accuracy of the model (ACC), the accuracy when detecting at-risk learners (true positive rate—TPR), the accuracy when distinguishing non-at-risk learners (true negative rate—TNR), and a harmonic mean of the true positive value (precision) and the TPR (recall) that weights correct at-risk identification (F score—F1.5). A non-balanced F score with a value higher than one gives more emphasis on correctly identify at-risk learners. In our case, this metric is more significant than ACC due to the class imbalance and the EWS’s aim, which focuses on detecting those at-risk learners.

The PGAR model is built from the query language for model creation and trained from MongoDB data, as depicted in Figure 6. The submodels for all the activities were built to select the best classification algorithm and training data. As stated in [77], the decision tree (DT), naïve Bayes (NB), support vector machine (SVM), and \( k \)-nearest neighbors (KNN) classification algorithms are used for training. Moreover, different training sets are considered on the basis of selecting different subsets of previous semesters (i.e., selecting one semester, two semesters, until the previous semester to the current one). Finally, the best classifier and training set is selected on the basis of the cost function that maximizes the TPR and TNR sum.
Each submodel is evaluated on the basis of different accuracy metrics. We use four metrics [8]:

\[
\text{TPR} = \frac{TP}{TP + FN}
\]

\[
\text{TNR} = \frac{TN}{TN + FP}
\]

\[
\text{F}^1 = \frac{2 \cdot \text{TPR} \cdot \text{TNR}}{\text{TPR} + \text{TNR}}
\]

where \(TP\) denotes the number of at-risk learners correctly identified, \(TN\) the number of non-at-risk learners correctly identified, \(FP\) the number of at-risk learners not correctly identified, and \(FN\) the number of non-at-risk learners not correctly identified. These four metrics are used for evaluating the global accuracy of the model (\(ACC\)), the accuracy when detecting at-risk learners (true positive rate—\(TPR\)), the accuracy when distinguishing non-at-risk learners (true negative rate—\(TNR\)), and a harmonic mean of the true positive value (precision) and the \(TPR\) (recall) that weights correct at-risk identification (\(F^1\) score—\(F^1\)). A non-balanced \(F^1\) score with a value higher than one gives more emphasis on correctly identifying at-risk learners. In our case, this metric is more significant than \(ACC\) due to the class imbalance and the EWS’s aim, which focuses on detecting those at-risk learners.

The PGAR model is built from the query language for model creation and trained from MongoDB data, as depicted in Figure 6. The submodels for all the activities were built to select the best classification algorithm and training data. As stated in [77], the decision tree (DT), naïve Bayes (NB), support vector machine (SVM), and \(k\)-nearest neighbors (KNN) classification algorithms are used for training. Moreover, different training sets are considered on the basis of selecting different subsets of previous semesters (i.e., selecting one semester, two semesters, until the previous semester to the current one). Finally, the best classifier and training set is selected on the basis of the cost function that maximizes the \(TPR\) and \(TNR\) sum.

**Figure 6. Procedure to obtain each submodel of the PGAR model.**

### 4.2.2. Next Activity At-Risk Simulation

The PGAR model provides a binary prediction about the likelihood of failing a course on the basis of the already assessed activities of the course. However, it lacks usefulness to impact a learner because they might already be at-risk when the prediction is computed. The EWS uses the PGAR to provide early and personalized information, feedback, and recommendations concerning the next assessment activity. We define the Next Activity At-Risk (NAAR) simulation as the process for detecting the minimum grade the learner must obtain in the next activity to be out of risk of failing. This simulation is performed by using the submodel associated with the activity we want to analyze. The submodel uses the previous activities’ grades and simulates all possible grades for the last activity. The simulation will identify the grade that changes the prediction from failing to pass the course.

**Example 2.** Let us take the submodel \(Pr_{AA1}(\text{Fail?})\) of Example 1. For the minimum grade to be known, six simulations are performed on the basis of the grades the learner can obtain in AA1 and the learner’s profile. Each simulation will predict the likelihood of failing on the basis of the assigned grade. An example is shown next on the basis of the first assessment activity of a hypothetical course.

\[
\begin{align*}
Pr_{AA1}(\text{Fail?}) &= (\text{Profile, N}) \quad \rightarrow \quad \text{Fail?} = \text{Yes} \\
Pr_{AA1}(\text{Fail?}) &= (\text{Profile, D}) \quad \rightarrow \quad \text{Fail?} = \text{Yes} \\
Pr_{AA1}(\text{Fail?}) &= (\text{Profile, C−}) \quad \rightarrow \quad \text{Fail?} = \text{Yes} \\
Pr_{AA1}(\text{Fail?}) &= (\text{Profile, C+}) \quad \rightarrow \quad \text{Fail?} = \text{Yes} \\
Pr_{AA1}(\text{Fail?}) &= (\text{Profile, B}) \quad \rightarrow \quad \text{Fail?} = \text{No} \\
Pr_{AA1}(\text{Fail?}) &= (\text{Profile, A}) \quad \rightarrow \quad \text{Fail?} = \text{No}
\end{align*}
\]
We can observe that learners with the profile “Profile” have a high probability of failing on grades below B. In other words, learners who get a grade from N to C+ in the first activity tend to fail the course. However, it is possible to pass the course with these grades but less frequently. Note that the simulation will consider previous activity grades on further activities. Thus, personalization increases while increasing the available data about learners.

4.2.3. Warning Level Classification and Intervention Mechanism

Compared to the PGAR, the NAAR simulation enhances the information the stakeholders receive about the likelihood of failing, and it can be used to enrich such predictions with recommendations. The NAAR is transformed to a warning level classification tree (shown in Figure 7) with different meanings depending on the prediction of the NAAR simulation (i.e., Pred), the accuracy (i.e., \( TNR \) and \( TPR \)) of the PGAR submodel used, and the grade of the learner for such activity (i.e., Grade).

![Decision tree for warning level classification.](image)

From this classification, the learner can be warned of their risk level when an activity starts. They are informed using two methods. The first one reports a personalized dashboard that we can see in Figure 8a. When an activity starts, the learner is informed by a bar-like visualization about the warning level distribution among the grades of such activity. The bar-like chart is built from the decision tree presented in Figure 7 by simulating all grades (i.e., the variable Grade in the decision tree) of the activity. When the activity is submitted and assessed, this chart is updated with the obtained grade (see Figure 8b), the risk level classification semaphore is also updated, and a new bar is generated for the next activity. The second method is the intervention mechanism associated with the risk level classification. Each level shown in the decision tree has associated a message giving recommendations about how to continue in the course. The message for the low-risk level congratulates the learner. In contrast, the messages for higher risk levels warn the learner and recommend activities, resources, and learning paths to continue successfully on the course.
Figure 7. Decision tree for warning level classification.

From this classification, the learner can be warned of their risk level when an activity starts. They are informed using two methods. The first one reports a personalized dashboard that we can see in Figure 8a. When an activity starts, the learner is informed by a bar-like visualization about the warning level distribution among the grades of such activity. The bar-like chart is built from the decision tree presented in Figure 7 by simulating all grades (i.e., the variable Grade in the decision tree) of the activity. When the activity is submitted and assessed, this chart is updated with the obtained grade (see Figure 8b), the risk level classification semaphore is also updated, and a new bar is generated for the next activity. The second method is the intervention mechanism associated with the risk level classification. Each level shown in the decision tree has associated a message giving recommendations about how to continue in the course. The message for the low-risk level congratulates the learner. In contrast, the messages for higher risk levels warn the learner and recommend activities, resources, and learning paths to continue successfully on the course.

(a)

Figure 8. Early warning dashboard for learners (a) before performing the AA1 and (b) after being graded the AA1.

Figure 8. Early warning dashboard for learners (a) before performing the AA1 and (b) after being graded the AA1.
5. Results and Discussion

In order to build a robust and trustworthy predictive analytics infrastructure, we used the self-assessment guidelines ALTAI [17], where seven requirements must be analyzed. The summarization of the self-assessment is detailed in Tables 1 and 2. The following subsections discuss in more detail the results of the self-assessment.

| Requirement                        | Self-Assessment                                                                 |
|------------------------------------|--------------------------------------------------------------------------------|
| Human agency and autonomy          | - The system focuses on detecting at-risk learners to pass courses.            |
|                                    | - The system provides recommendations to pass and warns the learner when there is no possibility to pass. |
|                                    | - Learner’s autonomy is not affected, but guidance may generate over-reliance on the system. |
|                                    | - No addictive behaviour is expected.                                          |
| Human agency and oversight         | - Human-in-the-loop oversees prediction results by simulation prior to it being sent to learners. |
|                                    | - Human-on-the-loop allows for configuring of the EWS before starting the course and monitoring the autonomous tasks. |
| Human oversight                    | - Recommender system can be set to manual, autonomous, or semi-autonomous modes. |
|                                    | - Administrators and teachers are informed of the autonomous tasks performed by the system by e-mail. |
| Security                           | - The system meets the institution security standards.                         |
|                                    | - Data (databases) are encrypted and secured by password.                       |
| Reliability and reproducibility    | - As a research project, the system is hold on a single server supported by institutional backup policy. |
|                                    | - The system is deployed by docker technology to easily reproduce the system on any server or cloud service. |
| Accuracy                           | - The system has a testing module to evaluate precision of predictive models.   |
|                                    | - Accuracy is transparently informed to all stakeholders.                      |
|                                    | - The recommender system takes into account model accuracy to avoid unprecise information. |
|                                    | - The system highly depends on the quality of the data in the data mart.       |
| Privacy                            | - The system gathers the minimum learners’ data to train models and handle them with an anonymous identifier. |
|                                    | - Anonymization is provided by the institution                                  |
| Data governance                    | - Learners give their consent to use the system and their data.               |
|                                    | - Consent can be withdrawn.                                                   |
|                                    | - Roles have been defined to control the access to the data.                  |
Table 2. Self-assessment of the ALTAI guidelines (continuation).

| Requirement                        | Self-Assessment                                                                 |
|------------------------------------|-------------------------------------------------------------------------------|
| **Transparency**                   |                                                                               |
| Transparency                        | - The system has a developer section to analyze models and predictions.        |
| Explainability                      | - Predictions come with an explanatory message to learners.                    |
|                                    | - Predictions can be evaluated by teachers prior to being sent to the learners.|
| Communication                      | - Learners are informed about the capabilities of the system.                  |
|                                    | - Learners are informed about the precision of the predictions.               |
|                                    | - Manuals are provided to learners and teachers to use the system.            |
| **Diversity, non-discrimination, and fairness** |                                                                               |
| Avoidance of unfair bias            | - Unclear predictions that may produce confusion to learners are replaced by deterministic outcome. |
| Accessibility                      | - The system allows accessibility to disabled people using web standards tested by WAVE tool. |
|                                    | - User experience has been analyzed by means of questionnaires.               |
| Stakeholder participation           | - The system has been tested on stakeholders (learners, teachers, administrators, developers) on each development cycle. |
| **Societal and environmental well-being** |                                                                               |
| Environmental well-being           | - Training is performed once per semester.                                    |
|                                    | - Predictions are triggered only when data are available.                     |
|                                    | - Only registered courses in the system and only learners who consented to participate are processed. |
|                                    | - MongoDB documents schema design minimizes read/write operations.            |
| Impact on work and skills          | - No new skills are needed to use the system.                                |
|                                    | - The system leverages the teachers’ work by providing recommendations and guidance to learners. |
| **Accountability**                 |                                                                               |
| Auditability                       | - All models, predictions, and recommendations triggered are logged in the system. |
| **Risk management**                |                                                                               |
|                                    | - The data risk level is high (security, role access, data policies).         |
|                                    | - The system availability risk level is high.                                |
|                                    | - Disaster recovery risk level is low.                                       |
|                                    | - Application performance risk level is high.                                |

5.1. Human Agency and Oversight

Human agency helps ensure that an AI system does not undermine human autonomy or causes other adverse effects. When the system’s aim is analyzed, it is crucial to see its impact on learners’ decisions. The developed EWS aims to provide relevant information to the learners while their autonomy is preserved. As a recommender system, suggestions focus on describing materials, learning activities, exercises, and learning paths that could lead to a successful outcome in the course. However, the system also warns when the learner is going to fail the course inevitably. The learner has the right to know this failing event to optimize their study time and planning. Thus, the system may impact learners’ decisions to select the best learning path or drop out from the course when there is no possibility to pass. Moreover, the system may generate an over-reliance [83]. Learners may feel that guidance is the best way to pass the course and waiting upcoming recommendations to continue. However, such dependency may negatively impact their self-regulation.

Trust is also affected by the feeling of having control over the system. AI systems tend to work autonomously and perform tasks without human intervention. However, this automation decreases the confidence, and oversight is needed on the tasks [84]. The EWS
has been designed on the basis of the human-in-the-loop (HITL) and human-on-the-loop (HOTL) governance mechanisms. The HITL allows for intervening on every decision the system has to perform; meanwhile, HOTL allows for human intervention during the design cycle. Three modes are supported: manual, autonomous, and semi-autonomous. Using the autonomous mode and the HOTL mechanism, the system requires minimal configuration from teachers and uses template recommendations based on the risk level classification. The recommendations are triggered when predictions are ready after activities are assessed on the basis of the best accuracy models trained by the system [77]. In manual mode using the HITL mechanism, teachers take complete control by selecting the best prediction model on the basis of simulations, filling the recommendations for each risk level using their expertise, and sending the messages when they decide. Finally, the semi-autonomous mode combines both mechanisms where the teacher can set the best model and the specific messages and allows the system to manage the sending process when data are available. Nevertheless, any task performed autonomously by the system is informed to the teachers to avoid untrustworthiness, regardless of the selected mode.

5.2. Technical Robustness and Safety

The infrastructure described in this paper is especially relevant with regard to this requirement. The infrastructure has been designed to secure data and to increase reproducibility. Data are secured in the persistence data tier encrypted by password, and Docker technology makes the system reproducible on any server. As a research project, the main drawback of the system is the use of one server to support the complete infrastructure. The system is currently deployed on a single server, and any failure on the hardware or software level may negatively impact the availability. However, the system can be easily deployed with Docker technology to a cloud infrastructure and benefits from its advantages such as high reliability and fault tolerance.

The accuracy is a crucial factor in the system because learners are classified on risk levels and receive corresponding recommendations based on such risk. Thus, stakeholders will mistrust the system if learners are classified erroneously. To prove the accuracy of the PGAR model, we tested the model in all the institution courses. The objective is to show the average accuracy on each point of the semester in the whole institution. The models were trained with data from the data mart from the 2016 Fall to the 2018 Fall semesters and tested on the 2019 Spring semester. Using the query language for the ETL process, we gathered the learners’ profile information, the grades from the assessment activities, and the pass/fail binary information about passing the course. Nine hundred and seventy-nine courses were analyzed, wherein the number of activities ranged from 3 to 11. In the end, 585,936 registries were used for training, and 138,746 were used for testing.

The PGAR model was built as described in Section 4.2.1. The accuracy is checked by the metrics defined in Equation (1). To distribute the metric results through the semester, we checked the accuracy for each submodel and course, and it was uniformly distributed among the semester timeline on the basis of the submission date of the respective assessment activity. For instance, for a course with four activities, the metrics values for each activity were distributed to the positions 20%, 40%, 60%, and 80% of the timeline. This distribution aims to identify the average quality on the basis of the activities submitted at each point of the timeline.

The average quality is evaluated on the basis of a LOESS regression [85] since this regression shows a better approximation than the linear one when there are many scattered values. The results are plotted in Figure 9, where the LOESS regression is shown, and the scattered plots are the metrics values for all the trained submodels. Table 3 summarizes the LOESS regression for each metric across the semester. As we can observe, the average accuracy (ACC) for the whole institution was higher than 75%. Specifically, the accuracy ranged from 77.56% to 95.25% from the beginning to the end of the semester, respectively. However, detecting at-risk and non-at-risk learners is quite different because learners at UOC tend to pass the courses and, therefore, classes are imbalanced. Detecting non-at-risk
learners (TNR) is easier. The accuracy of detecting non-at-risk learners ranged from 84.79% at the beginning of the semester to 96.91% at the end. Detecting at-risk learners started with a lower accuracy at 55.71%, but it reached 91.22% at the end of the course.

![Figure 9. Scatter plots of the PGAR model for (a) accuracy; (b) TNR; and (c) TPR, F1.5.](image)

Table 3. LOESS regression of the PGAR model for ACC, TNR, TPR, and F1.5 metrics.

| Semester Timeline | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
|-------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| ACC               | 77.56 | 80.86 | 83.72 | 86.20 | 88.43 | 90.46 | 92.39 | 93.99 | 95.25 |
| TNR               | 84.79 | 86.30 | 87.78 | 89.30 | 90.95 | 92.46 | 94.04 | 95.54 | 96.91 |
| TPR               | 55.71 | 64.51 | 71.70 | 77.40 | 81.88 | 85.42 | 88.53 | 90.47 | 91.22 |
| F1.5              | 53.50 | 61.00 | 67.50 | 73.10 | 77.97 | 81.75 | 85.36 | 88.14 | 90.05 |

As we can observe, there were many submodels with low accuracy in the first activities. Although the system tried to find the best classification algorithm for each submodel, the accuracy was highly impacted by the data mart. As aforesaid, data from the data mart were not curated, and inconsistencies could lead to imprecise models. Although the loader service has a module to find such discrepancies, some submodels cannot capture the learners’ behavior for the corresponding activity. As described in Section 4.2.3, submodels with a TPR or TNR lower than 75% generate a yellow (amber) risk level to avoid reporting unprecise information, wherein recommendations are adapted and stakeholders are notified that a low-accuracy model is used.

5.3. Privacy and Data Governance

Privacy is also a critical issue in AI systems [49], and it is intimately related to data governance. Privacy on the stored data or the extracted knowledge must be taken into account. The infrastructure handles data anonymously to impersonate the data and the obtained knowledge from learners. Thus, the learner’s identity is unknown in any place of the infrastructure. Additionally, the anonymization is provided by a secured service of the university, leveraging the responsibility of the infrastructure to manage the complete anonymization process.
The system aims to exclusively store the data needed to train the models. This is achieved by specifying the appropriate fields to be captured in the ETL process on the basis of the needed ones in the predictive models. Since selecting the minimal set of fields to train a model is not an easy task, the computational service provides a tool to minimize the number of features using the Feature Selection technique [86]. Such a technique helps to identify superfluous features already covered by other fields captured in the model.

Data governance, which is related to privacy, manages the availability, usability, integrity, and security of sensitive data. UOC has its own data policy [87], and the EWS adheres to such policy following the European GDPR [26]. Additionally, the university has a Research Ethical Committee that assesses that the utilization of learners’ data meets ethical codes and current privacy legislation. The system should protect the data’s ownership, the data to be displayed, and who has access to which kind of data. The system meets the privacy policy by asking users to consent before utilizing their data, informing them of used data and the objective of data processing, and securing access to information by roles. In the designed EWS, the data’s owner is the learner and, thus, they have the right to consent to their utilization, as well as to revoke it. Note that the data visualization is restricted by the user’s roles described in Section 4.1.3. Although the ALTAI self-assessment guidelines recommend aligning the system’s data governance with other AI-related standards [15,18], it is currently out of the scope of this paper. However, it will be done as future work in the following design cycle of the system.

5.4. Transparency

Transparency encompasses three elements: traceability to understand which data trigger each prediction and the decisions performed by the system, explainability as to the reason behind the decisions, and proper communication about the capabilities of the system.

AI tools are commonly conceived as black boxes [88] in the sense that users are not able to understand the underpinning algorithms that produce the predictions and recommendations. Without knowing such parts, it is difficult to trust the results of such systems. The presented EWS was designed to transparently present this information to the stakeholders and provide insights about the reason for a specific prediction. As aforesaid, the model’s accuracy can be tested before using the model. Developers can test them, play with multiple sets of features, see their accuracy, and finally present the best model for detecting at-risk learners. Additionally, models can be debugged in a way that the used dataset for training can be extracted. Such robust characteristics can lead developers to understand why an erroneous prediction arose or even detect incongruent data from the data mart.

However, such detail level or the model’s accuracy is not suitable for teachers who probably may not be proficient in AI tools. A teacher needs to understand the learners’ behavior and which circumstances may impact on failing or dropping out of a course [89]. The designed EWS can simulate the complete decision tree classification of the risk level for a complete course regarding the profile features and the activities grades. Teachers can gain insights as to which profile’s features and grades mainly impact the likelihood of failing a course with such a tree.

Finally, good communication is the only way to open and disclose the “black box”. Users must be informed about the systems’ capabilities and limitations. Teachers must be trained to use the system, and learners must be informed about the used data and their purposes. A manual is provided to each user's role with the full detail of the capabilities they can access. Moreover, all users have access to multimedia material explaining the information they receive via the dashboards and messages and an email address to share their thoughts and doubts.
5.5. Diversity, Non-Discrimination, and Fairness

A well-known problem of AI models is that they depend on data used for training [90]. Biased models may discriminate against some users or provide predictions that may prejudice them when impacting their autonomy. However, discrimination can also appear when the system supports limited accessibility. To avoid both types of discrimination, stakeholders should participate in the system’s design cycles by giving their opinions and clarifying their needs.

The PGAR model highly depends on the learners’ profile and the percentage of learners who passed the course in the past. We have shown in Section 5.2 how the average $TPR$ and $TNR$ are significantly different due to class imbalance. Models for courses with high performance will negatively detect at-risk learners, and learners with a profile not covered by a model will probably be mispredicted.

As aforesaid, the risk level classification considers the $TPR$ and $TNR$ to decide the risk level. A threshold of 75% was set to determine a high- or low-accuracy model. As shown in Figure 7, high-accuracy models can ascertain all risk levels, while low-accuracy models are always set to medium risk to avoid erroneous messages to learners. Note that classifying a learner at the at-risk level may impact their decision to drop out of the course. Thus, we defined the system’s unfairness as “a learner is classified as at-risk, but meanwhile there is some possibility of passing the course.”

Accessibility is an important factor to analyze on AI-powered systems. One of the claimed advantages of AI utilization is the help and enhanced access to people with disabilities [91]. However, we need to avoid excluding them from new technologies and even avoid unfairness on the predictions [92]. The EWS allows accessibility to disabled people using web standards tested by the WAVE web accessibility evaluation tool [93]. However, models do not take into account disabled learners. Note that profile information about being disabled should be known to distinguish them and create distinctive models. However, this is not possible due to institutional privacy and ethical policies. Thus, disabled learners are kept anonymized within the models.

Related to stakeholders’ participation, the system development follows a mixed research methodology. The action research methodology allows for investigation and improvement of own practices, guided by an iterative cycle plan-act-reflect, and also emphasizes collaboration with practitioners. The design and creation approach is especially suited when developing new IT artifacts. It is crucial to interact with the stakeholders during the development of the artifact of the next cycle to fulfill their requirements. On each cycle, the stakeholders’ opinions and criticisms are considered to enhance the system’s features in the next cycle. The IT artifacts have been tested in different pilots. The prediction model and the risk classification system can be reviewed in [77], user experience is analyzed in [94], dashboard design is described on [95], intervention mechanism is defined in [96], and the impact on performance on case studies can be checked in [97,98].

5.6. Societal and Environmental Well-Being

In line with the prevention of harm on individuals, social influence should be analyzed as well. Workforce impact can deteriorate people’s physical and mental well-being; meanwhile, environmental effects should be evaluated to reduce the carbon footprint on AI utilization.

AI systems are commonly untrusted when they directly impact the workforce [99,100]. Education can be equally affected if AI-powered LMS can be proved as a replacement for teachers. However, AI tools should enforce teachers’ work instead of thinking of substituting [101]. As authors in [102] (p. 1) claimed, AI should never remove “the role of the teacher because how it works and what it does is so profoundly different from human intelligence and it has the potential to transform education in ways that make education more human, not less.” AI can leverage teacher’s work on repetitive tasks or enhance the personalization. As aforesaid in Section 5.1, the designed EWS can be set up to automate the teacher’s recommendation task. The autonomous mode
can be configured with basic recommendations provided as templates by the system. However, the semi-autonomous mode is also available, wherein the teacher’s expertise is considered. This second option is the recommended one because the teacher knows which resources, exercises, and learning paths can help the learner pass the course with their expertise. Moreover, this option increases trustworthiness with the system because the system empowers the teacher with additional knowledge about the learners’ behavior and needs. Therefore, better personalization can be supplied.

Furthermore, environmental effects are considered. Training models require an increment of energy consumption due to the resource’s utilization and operation calculations. The leading research in this topic focuses on evaluating energy consumption in terms of software (i.e., language programming libraries for AI) [103] and hardware levels (i.e., identifying critical circuit paths during AI operations) [104]. In our infrastructure, we focus on the software level in the computational service. The service has been implemented to process only registered courses in the system and learners who consented to participate. Thus, the operations have been optimized and run only for active courses and learners.

Three main operations affect energy consumption: read/write operations, training, and predicting outcomes.

The read/write operations from MongoDB are a critical bottleneck when big data pipelines are implemented [105]. Here, it is critical to design a proper schema and query aggregation pipeline. Aforesaid, the document schema is created on the basis of the query language for the ETL process described in Section 4.1.4, storing each learner as a document and identified by the anonymized identifier. All the information referred to a learner is stored in subdocuments identified by the name of the data mart dataset. The query aggregations can quickly obtain data for a specific learner by querying by document identifier. The critical operation is the dataset creation for training because the collection associated with the view must be parsed slicing by course and activity identifiers using the query language described in Section 4.1.5. In such a case, only MongoDB cache and RAM size can help to get results faster. However, the training task is only performed during developer testing or when setting up the model for a course, and it is only performed once at the beginning of the semester for the registered course. Finally, the prediction is only performed when grade information for a specific activity is detected in the data mart. Thus, it implies only a read operation and a prediction for each course, activity, and learner.

5.7. Accountability

Like human oversight, the system needs to be auditable, and auditability can help identify and mitigate risks transparently. The EWS provides administrator reports about different logged data. Currently, the system allows for auditing (1) accuracy of PGAR models for all courses, (2) risk classification in terms of activities already assessed and predicted, (3) accuracy of the risk classification in terms of learners of the previous semesters with identical profile, (4) recommendations delivered to an individual learner and recommendations sent by a teacher, and (5) real-time feedback from learners through their dashboard about potential errors on received predictions and recommendations. Such logs can facilitate the analysis of erroneous behaviors.

Finally, risk management is a critical issue in systems that are in production. Several risks should be carefully analyzed and tried to minimize their impact in case of happening. The proposed infrastructure and EWS have some inherent risks. Data management is one of the critical issues. Security, data breaches, conform data policies, and ethical issues must be continuously evaluated. The system manages private and sensitive data from learners. Although they are anonymized, a data breach is possible with unauthorized access to the server, the persistence data tier, and the institutional deanonymize service.

Availability and performance are other risks. Currently, all systems are handled in a unique on-premise server. Any failure may impact the system globally without unable to restart the system on another server automatically. Although the system can be restored from backups, manual intervention is needed. Performance is also affected by the same
problem because all services are virtualized in the same on-premise server, and horizontal scalability is limited to the server resources.

6. Conclusions

TAI systems offer immense potential in different areas such as health, energy, transport, commerce, and also education by respecting individual values and avoiding harnessing citizens. However, there are some risks. Guidelines such as ALTAI focus on mitigating them. Quality dataset, human oversight, security, accuracy, robustness, and accountability are some of the requirements that such systems should meet before being offered to stakeholders. The future of TAI systems is uncertain. The concept may disappear in the void, but a path is being forged. Governments and institutions are investing efforts into defining the essential requirements. Currently, these requirements are only focused on educating engineers at high-tech companies to focus on citizens’ rights, but probably in the future will be the basis of a certification ensuring that TAI systems work for people and not for other interests.

Aligned with the need that AI-based systems can be TAI systems, the contribution of this paper is twofold. First, we presented the predictive analytics infrastructure to support a trustworthy EWS that, following the applied research methodology, has evolved due to a second iteration of the system enhancement (iterative cycle plan, act, reflect). Secondly, the system was built to fit into the ALTAI guidelines, and a self-assessment process was carried out to check all the requirements. This allowed for verification as to which extent they are covered in the developed system. Until now, design research on AI systems was mainly focused on security [49], privacy [51], or ethics [52] because of the access and utilization of sensitive data are crucial points in AI systems. However, we claim that other requirements such as accountability, oversight, or transparency are also relevant and cannot be underestimated. They impact the same way in trusting a system because they are related to how users feel about using such a system. Precisely, our research shows that dealing with all requirements is feasible and necessary to obtain a comprehensive view of the system’s trustworthiness and to detect weaknesses and threats. Furthermore, the self-assessment process needs to be conducted at every system iteration-cycle, and guidelines and standards can help to this end.

After this evaluation process, some aspects should still be refined to fully address all the requirements. Regarding the technical robustness and safety requirement, it is still needed to improve the models’ accuracy. As data mart might provide data not well-curated, it impacts the course’s models, generating inconsistencies that should be arranged. Some submodels cannot appropriately capture data from learners’ behavior for the related activities, which may lead to some discrimination or unfairness issues. Following this system’s weakness, the predictive models are not capturing information about disabled learners, and such models could lead to undesirable predictions for such learners. The system cannot distinguish them since institutional ethics and privacy policy do not allow it, and this may have some disadvantages in terms of unfairness. Finally, as a research project, the accountability requirement has some hindrances to be solved. Risk management requires continuous revision of the guidelines, privacy concerns, an ethics review board, and a process to analyze potential vulnerabilities of the system. Such requirements require full support from the institution to adhere the TAI system to the current established privacy and security policies and designated ethical committee.

On the basis of this self-assessment process to fit ALTAI guidelines and the results obtained on this research, as future work, we will perform a new iterative cycle plan to test the final system once the main weaknesses have been addressed. Additionally, this final system will be aligned with other standards guidelines such as ISO [18] or IEEE [15] to have a fully predictive analytics infrastructure to support a trustworthy EWS.
Author Contributions: Conceptualization, D.B., A.E.G.-R. and M.E.R.-G.; data curation, D.B.; investigation, D.B., A.E.G.-R., M.E.R.-G. and A.K.; methodology, A.E.G.-R. and M.E.R.-G.; project administration, D.B. and A.K.; software, D.B.; supervision, D.B., A.E.G.-R. and M.E.R.-G.; validation, D.B., A.E.G.-R., M.E.R.-G. and A.K.; visualization, D.B., A.E.G.-R. and M.E.R.-G.; writing—original draft, D.B., A.E.G.-R. and M.E.R.-G.; writing—review and editing, D.B., A.E.G.-R., M.E.R.-G. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the eLearn Center at Universitat Oberta de Catalunya through the project New Goals 2018NG001 “LIS: Learning Intelligent System”.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: The authors would express their gratitude for the technical support received by the eLearn Center staff in charge of the UOC data mart.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yu, T.; Jo, I.-H. Educational Technology Approach Toward Learning Analytics. In Proceedings of the Fourth International Conference on Learning Analytics and Knowledge—LAK ‘14, Indianapolis, IN, USA, 24–28 March 2014; ACM Press: New York, NY, USA, 2014; pp. 269–270.
2. Delgado Kloos, C.; Alario-Hoyos, C.; Estevez-Ayres, I.; Munoz-Merino, P.J.; Ibanez, M.B.; Crespo-Garcia, R.M. Boosting Interaction with Educational Technology. In Proceedings of the 2017 IEEE Global Engineering Education Conference, EDUCON, Athens, Greece, 25–28 April 2017; IEEE: New York, NY, USA, 2017; pp. 1763–1767.
3. Craig, S.D. Tutoring and Intelligent Tutoring Systems; Nova Science Publishers, Incorporated: New York, NY, USA, 2018.
4. Romero, C.; Ventura, S. Data mining in education. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2013, 3, 12–27. [CrossRef]
5. Ratnapala, I.P.; Ragel, R.G.; Deegalla, S. Students behavioural analysis in an online learning environment using data mining. In Proceedings of the 7th International Conference on Information and Automation for Sustainability, Colombo, Sri Lanka, 22–24 December 2014; IEEE: New York, NY, USA, 2014; pp. 1–7.
6. Romero, C.; Ventura, S.; Espejo, P.G.; Hervás, C. Data mining algorithms to classify students. In Proceedings of the Educational Data Mining 2008—1st International Conference on Educational Data Mining, Montreal, QC, Canada, 20–21 June 2008; Joaozeiro de Baker, R.S., Barnes, T., Beck, J.E., Eds.; International Educational Data Mining Society: Boston, MA, USA, 2008; pp. 8–17.
7. Wolff, A.; Zdrahal, Z.; Herrmannova, D.; Knoth, P. Predicting Student Performance from Combined Data Sources. In Educational Data Mining: Applications and Trends; Peña-Ayala, A., Ed.; Springer International Publisher: Cham, Switzerland, 2014; Volume 524, pp. 175–202.
8. Marbouti, F.; Diefes-Dux, H.A.; Madhavan, K. Models for early prediction of at-risk students in a course using standards-based grading. Comput. Educ. 2016, 103, 1–15. [CrossRef]
9. Xing, W.; Chen, X.; Stein, J.; Marcinkowski, M. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. Comput. Hum. Behav. 2016, 58, 119–129. [CrossRef]
10. Sun, H. Research on the Innovation and Reform of Art Education and Teaching in the Era of Big Data. In Cyber Security Intelligence and Analytics 2021 International Conference on Cyber Security Intelligence and Analytics (CSIA 2021), 19–20 March, Shengang, China; Xu, Z., Parizi, R.M., Loyola-González, O., Zhang, X., Eds.; Springer: Cham, Switzerland, 2021; Volume 1343, pp. 734–741.
11. Holmes, W.; Porayska-Pomsta, K.; Holstein, K.; Sutherland, E.; Baker, T.; Shum, S.B.; Santos, O.C.; Rodrigo, M.T.; Cukurova, M.; Bittencourt, I.; et al. Ethics of AI in Education: Towards a Community-Wide Framework. Int. J. Artif. Intell. Educ. 2021, 31, 1–23.
12. European Commission. Fostering a European Approach to Artificial Intelligence. Available online: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75790 (accessed on 31 May 2021).
13. ANSI Standarization Empowering AI-Enabled Systems in Healthcare: Workshop Report. Available online: https://share.ansi.org/Shared%20Documents/News%20and%20Publications/Links%20Within%20Stories/Empowering%20AI-Enabled%20Systems,%20Workshop%20Report.pdf (accessed on 31 May 2021).
14. Stanton, B.; Jensen, T. Trust and Artificial Intelligence; NIST Interagency/Internal Report (NISTIR); National Institute of Standards and Technology: Gaithersburg, MD, USA, 2021. Available online: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087 (accessed on 31 May 2021).
15. The IEEE Global Initiative. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems; IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: New York, NY, USA, 2017.
16. ICO. Guidance on the AI Auditing Framework: Draft Guidance for Consultation. Available online: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (accessed on 31 May 2021).
17. High-Level Expert Group on Artificial Intelligence (AI HLEG). Assessment List for Trustworthy AI (ALTAI). Available online: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (accessed on 31 May 2021).

18. International Organization for Standardization. Overview of Trustworthiness in Artificial Intelligence (ISO/IEC TR 24028:2020). Available online: https://www.iso.org/standard/77608.html (accessed on 31 May 2021).

19. Karadeniz, A.; Bañeres Besora, D.; Rodríguez González, M.E.; Guerrero Roldán, A.E. Enhancing ICT Personalized Education through a Learning Intelligent System. In Proceedings of the Online, Open and Flexible Higher Education Conference, Madrid, Spain, 16–18 October 2019; pp. 142–147.

20. Oates, B.J. Researching Information Systems and Computing; SAGE: London, UK, 2006.

21. Vaishnavi, V.; Kuechler, W. Design Research in Information Systems. Available online: http://www.desrist.org/design-research-in-information-systems/ (accessed on 31 May 2021).

22. Bañeres, D.; Karadeniz, A.; Guerrero-Roldán, A.E.; Rodriguez, M.E. A Predictive System for Supporting At-Risk Students’ Identification. In Proceedings of the Future Technologies Conference (FTC), Vancouver, BC, Canada, 5–6 November 2020; Arai, K., Kapoor, S., Bhatia, R., Eds.; Springer: Cham, Switzerland, 2021; Volume 1288, pp. 891–904.

23. Queiroga, E.M.; Lopes, J.L.; Kappel, K.; Aguiar, M.; Araújo, R.M.; Munoz, R.; Villarroel, R.; Cechinel, C. A Learning Analytics Approach to Identify Students at Risk of Dropout: A Case Study with a Technical Distance Education Course. Appl. Sci. 2020, 10, 3998. [CrossRef]

24. Wan, H.; Liu, K.; Yu, Q.; Gao, X. Pedagogical Intervention Practices: Improving Learning Engagement Based on Early Prediction. IEEE Trans. Learn. Technol. 2019, 12, 278–289. [CrossRef]

25. Kostopoulos, G.; Karlos, S.; Kotsiantis, S. Multiview Learning for Early Prognosis of Academic Performance: A Case Study. IEEE Trans. Learn. Technol. 2019, 12, 212–224. [CrossRef]

26. The European Parliament and the Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Off. J. Eur. Union 2016, 59, 1–88.

27. Drachsler, H.; Grelle, W. Privacy and analytics. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge—LA’K ’16, Edinburgh, UK, 25–29 April 2016; ACM Press: New York, NY, USA, 2016; pp. 89–98.

28. SEG Education Committee. Principles for the Use of University-held Student Personal Information for Learning Analytics at The University of Sydney. Available online: https://www.sydney.edu.au/education-portfolio/images/common/learning-analytics-principles-april-2016.pdf (accessed on 31 May 2021).

29. The Open University. Policy on Ethical Use of Student Data for Learning Analytics. Available online: https://help.open.ac.uk/documents/policies/ethical-use-of-student-data/files/22/ethical-use-of-student-data-policy.pdf (accessed on 31 May 2021).

30. Kazim, E.; Koshiyama, A. AI Assurance Processes. SSRN Electron. J. 2020. [CrossRef]

31. Floridi, L. Establishing the rules for building trustworthy AI. Nat. Mach. Intell. 2019, 1, 261–262. [CrossRef]

32. House of Lords. AI in the UK: Ready, Willing and Able? Available online: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf (accessed on 31 May 2021).

33. Future of Life Institute. Asilomar AI Principles. Available online: https://futureoflife.org/ai-principles/ (accessed on 31 May 2021).

34. Wiens, J.; Saria, S.; Sendak, M.; Ghassemi, M.; Liu, V.X.; Doshi-Velez, F.; Jung, K.; Heller, K.; Kale, D.; Saeed, M.; et al. Do no harm: A roadmap for responsible machine learning for health care. Nat. Med. 2019, 25, 1337–1340. [CrossRef]

35. Thiebes, S.; Lins, S.; Sunyaev, A. Trustworthy artificial intelligence. Electron. Mark. 2020, 31, 1–18. [CrossRef]

36. High-Level Independent Group on Artificial Intelligence (AI HLEG). Ethics Guidelines for Trustworthy AI. Available online: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (accessed on 31 May 2021).

37. Sundaramurthy, A.H.; Raviprakash, N.; Devarla, D.; Rathis, A. Machine Learning and Artificial Intelligence. XRDS Crossroads ACM Mag. Stud. 2019, 25, 93–103.

38. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach. 2018, 28, 689–707. [CrossRef]

39. Huo, C.-C. Factors affecting webpage’s visual interface design and style. Procedia Comput. Sci. 2011, 3, 1315–1320. [CrossRef]

40. Siijing, L.; Lan, W. Artificial Intelligence Education Ethical Problems and Solutions. In Proceedings of the 2018 13th International Conference on Computer Science & Education (ICCSE), Colombo, Sri Lanka, 8–11 August 2018; IEEE: New York, NY, USA, 2018; pp. 1–5.

41. Vincent-Lancrin, S.; van der Vlies, R. Trustworthy Artificial Intelligence (AI) in Education; In OECD Education Working Papers, No. 218; OECD Publishing: Paris, France, 2020.

42. Persico, D.; Pozzi, F. Informing learning design with learning analytics to improve teacher inquiry. Br. J. Educ. Technol. 2015, 46, 230–248. [CrossRef]

43. Franzoni, V.; Milani, A.; Mengoni, P.; Piccinato, F. Artificial Intelligence Visual Metaphors in E-Learning Interfaces for Learning Analytics. Appl. Sci. 2020, 10, 7195. [CrossRef]

44. Williamson, B. The hidden architecture of higher education: Building a big data infrastructure for the ‘smarter university. Int. J. Educ. Technol. High. Educ. 2018, 15, 1–26. [CrossRef]
101. Selwyn, N. *Should Robots Replace Teachers? AI and the Future of Education*; John Wiley: Hoboken, NJ, USA, 2019.

102. Cope, B.; Kalantzis, M.; Searsmith, D. Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educ. Philos. Theory* **2020**, *1*, 1–17. [CrossRef]

103. Mahmoud, A.A.; Elkatatny, S.; Ali, A.Z.; Abouelresh, M.; Abdulraheem, A. Evaluation of the total organic carbon (TOC) using different artificial intelligence techniques. *Sustainability* **2019**, *11*, 5643. [CrossRef]

104. García-Martín, E.; Rodrigues, C.F.; Riley, G.; Grahn, H. Estimation of energy consumption in machine learning. *J. Parallel Distrib. Comput.* **2019**, *134*, 75–88. [CrossRef]

105. Mahajan, D.; Zong, Z. Energy efficiency analysis of query optimizations on MongoDB and Cassandra. In Proceedings of the 2017 Eighth International Green and Sustainable Computing Conference (IGSC), Orlando, FL, USA, 23–25 October 2017; IEEE: New York, NY, USA, 2017; pp. 1–6.