Open Education Analytics

Data and AI Use Case Template

Template v3 Published: December 2021

*This is a preliminary document and may be changed substantially prior to final commercial release of the software described herein. The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions,* ***it should not be interpreted as a commitment on the part of Microsoft****, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.*

***This template is for informational purposes only. Microsoft makes no warranties, express or implied, in this document.***

***Complying with all applicable copyright laws is the responsibility of the user.***

*© 2021 Microsoft Corporation. All rights reserved.*

# 

# Introduction

The OEA Use Case Template is an example of how Microsoft Education defines and plans specific data and AI projects for education use cases.

To ensure the appropriate and ethical use of data in analytics projects, Microsoft Education applies Microsoft’s Responsible AI Principles. The OEA Use Case Template operationalizes these principles, which include fairness, reliability and safety, privacy and security, inclusion, transparency, and accountability. For more information on how Open Education Analytics operationalizes these principles, see: [Responsible AI | Open Education Analytics](https://openeducationanalytics.org/responsible-ai/).

The OEA Use Case Template can be used to plan a single use case or multiple times to develop an inventory of use cases in a comprehensive plan for data modernization.

Sections 1-3 of the OEA Use Case Template and the section on Privacy and Security in Section 4 is used to develop any type of data use case from simple reports to more complex AI models. Completing these sections can help prevent many common problems in data projects such as:

* Asking the wrong questions or not fully understanding the problem to be solved with data
* Using the wrong type of data or too much data to solve the problem of the use case
* Making incorrect assumptions about the data and how it maps to the problem
* Developing a data solution that is not utilized by key groups for its intended purpose (e.g., not used to make decisions by schools, educators, students, families).

Section 4 is used throughout the use case development process to operationalize and document decisions made for each of the principles of Responsible AI. This section is especially important when a use case involves the development of a machine learning model or a predictive algorithm, as these have the potential to cause unintentional harm to students.

The Use Case Template is generally managed by the Project Manager for any specific use case, with input and review by all roles and key groups involved in use case planning. *The sections in blue italics are where input for a specific use case should be provided.*

# 1) The Use Case Problem

**Defining the Problem: What problem does this use case seek to solve?**

Defining the Use Case Problem helps clarify the opportunities and limitations of data and analytics to help solve challenges.

Some common Use Case Problems in Education:

* **Digital Access**: How do we ensure all students in our system have access to our digital learning applications and platforms when they are learning remotely?
* **Digital Learning Insights**: How are students using and behaving in digital learning platforms and applications, and how does that behavior relate to key learning outcomes?
* **Program Evaluation and Monitoring:** Is a new education program or policy such as a professional learning program being implemented as planned, and does the program have measurable impact on learning or education outcomes?
* **State and National Education Reporting**: How do we make our city/state/country education data and reporting more real time and actionable?
* **Predicting At Risk and Vulnerable Students**: How do we use our data to prevent learning dis-engagement, drop-outs, or harm to students?
* **Learner Records:** How do we use all our data to develop more holistic insights into student learning and better support students in representing their learning beyond an individual school, such as in applications for other schools or for jobs?
* **Personalization:** How do we enable every student to develop a more personalized learning experience, driven by their stage of learning progress, interests, and goals?

Note: [OEA Modules and Packages](https://github.com/microsoft/OpenEduAnalytics), support many of these education use cases through accelerating the ingestion of key data sources needed and providing resources to set up these use cases.

--------------------------------------------------------------------------------------------------------------------------

*In a few sentences, describe the education or learning problem that needs a solution with this use case, and how it aligns with the needs of learners.*

# 2) The Use Case Stakeholders

Considering the benefits of a use case – and its potential harms - requires the consideration of different stakeholders and their points of view. Stakeholders typically include the people who are responsible for, will use, or will be affected by the use case. Stakeholders are defined by their role: their duties, contextual identity, or circumstances in relationship to the use case.

* **Direct stakeholders** interact with the data from a use case directly and make decisions or take actions based on results. They include data users, system developers, and even data system maintenance staff.
* **Indirect stakeholders** are affected by the use case data system but, unlike direct stakeholders, they do not have a role that requires them to use or maintain the system. Indirect stakeholders can include groups who may be affected by the downstream effects of the system, such as parents, students, or future employers.
* **Malicious actors** include hackers and others who may intentionally misuse the system. Considering malicious actors is important to supporting safe and reliable data systems.

**Who are the stakeholder groups for this use case, and how are they involved in its development?**

|  |  |  |
| --- | --- | --- |
| **Possible Stakeholder Groups** | **Relationship to Use Case** | **Involvement in Use Case** |
| Students | *For example, Indirect Stakeholders* | *For example, have students been asked permission to use their data?* |
| Parents or Guardians | *For example, Indirect Stakeholders* | *For example, parents or guardians do not interact within the systems, but are the intended beneficiaries of this use case.* |
| Educators (Faculty or Teachers) |  |  |
| School or Department Leaders |  |  |
| School System or Institutional Leaders | *For example, Direct Stakeholders* | *For example, they will take advantage of system wide insights, to inform policy decision and resourcing.* |
| Researchers |  |  |
| Future Employers of Learners |  |  |
| Potential Malicious Actors |  |  |

**Engaging stakeholders in the use case development process.**

Including stakeholders in the early thinking and conceptualization of a data use case is a good way to ensure that the use case output will be accepted, trusted, and used by key stakeholders. For example, conducting interviews or focus groups with representatives of each stakeholder group can provide early insights into the conceptual model framing the use case and the appropriateness of specific data sources to be used. At a later stage in the project, involving key stakeholder groups in designing the use case outputs (such as a dashboard or notification) can be essential to that product’s eventual effective use.

**Outline how stakeholders will be involved in the development in different stages of the use case development:**

*Early Stages: Defining the use case problem, developing the local theory or conceptual model of the problem, identifying key data sources to include in the use case in the local context:*

*Reviewing and Designing Outputs Stages: Testing validity of the use case results, developing dashboard designs or set of interventions based on the use case results:*

**What type of outputs are expected from this use case, such as AI models, dashboards, or notification systems?**

*Describe the expected outputs, and which stakeholders will have access to or use those outputs:*

# 3) Mapping Theory to Data

For most common education use cases, prior research has already been conducted or a theory of the problem developed. For example, extensive research identifies the key data elements that best predict at-risk students. Another example is a model for school improvement that defines different theoretical constructs or categories to consider when improving or evaluating a school. This type of research, theory or model can help identify the most relevant data sources for a specific use case. The actual data analysis may show that the data identified in the theoretical construct does not ‘fit’ well in the local use case context, or that the analytical patterns are not the same as in prior research, but it provides a starting place for identifying key data sources.

**For this use case, what prior research or conceptual model frames your theory of the problem?**

*List or cite the research, theories, or conceptual models that inform data selection for this use case, identifying the categories of data relevant to the use case problem.*

Sources of education research can include:

* Local or national education agencies and their research organizations
* Local, national, or international education policy documents
* Education research databases: [ERIC - Education Resources Information Center](https://eric.ed.gov/)
* Google Scholar: [Google Scholar](https://scholar.google.com/)

**Mapping Theory to Data. From prior research or conceptual models what are they key data categories expected to inform this use case? What local data sources are available or needed for each category? Please note where no data is available for a Data Category**

A key part of the use case development process is deciding which data to use and how it should be mapped to the theory of the problem. Identifying which data should be viewed as a “feature” and which data is the “target outcome” is at the core of this mapping.

|  |  |
| --- | --- |
| **Key Data Category** | **Local Data Source** |
| 1. *For example, Attendance Data* | *For example, Student Information System – “Days Absent”* |
| 2. *For example, Digital Learning Application Data* | *For example, Canvas and M365 app usage data* |
| 3. *For example, Student Well-Being* | *For example, Student Health Records* |
| 4. *For example, Student Engagement* | *For example, No Data Available* |

**Note: Mapping theory to data with a ‘data dictionary.’**

A “data dictionary” allows the data team to examine specific data tables and data entities in the available datasets, and then map specific items to the Key Data Category.

New data services like [Azure Purview](https://docs.microsoft.com/en-us/azure/purview/overview) can support this work through creating a holistic, up-to-date map of a data repository with automated data discovery, sensitive data classification, and end-to-end data lineage.

**Please see “Privacy and Security” section below for more ensuring that sensitive data is protected.**

4) Responsible AI Principles Applied

**In these next sections, please answer the questions under each of the headings describing how responsible AI principles will be applied to this use case.**

Fairness Principle

|  |
| --- |
| AI systems should treat everyone in a fair and balanced manner and not affect similarly situated groups of people in different ways. Human decision makers are susceptible to many forms of prejudice and bias, such as those rooted in gender and racial stereotypes. To ensure AI models are trained in a way that does not embed or re-enforce those biases, models must be tested for fairness. Microsoft has developed an open-source toolkit to support this called [Fairlearn](https://fairlearn.org/), which can be applied within the Azure analytical services used in the OEA reference architecture.  [Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6) on Fairness Principle.  **Who is most likely to be at risk of experiencing harms from this use case?**  *Define any groups (immigrants, rural students) or subpopulations (gender, language group) that might face adverse consequences from the AI. For each group, identify if these groups and subpopulations are clearly labelled in the dataset.* |
| |  |  |  | | --- | --- | --- | | **Group or Subpopulation** | **Clearly Labelled in Dataset? Y/N** | **Planned Mitigations** | | 1. *For example, Immigrants* | *Y* | *For example, Test model outputs via Fairlearn for false positives / negatives* | | 1. *For example, Rural Students* | *N* | *For example, Add geolabel based on address mapping for rural, then test outputs via Fairlearn* | |  |  |  | |

Reliability and Safety Principle

Systems should operate reliably and safely when they function in the world. AI systems must be designed with a view to the potential benefits and risks to different stakeholders and undergo rigorous testing to ensure they respond safely to unanticipated situations and do not evolve in ways that are inconsistent with the original shared purpose.

[Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6) on Reliability and Safety Principle.

**What are possible risks faced by learners or educators from the analytics of this use case?**

1. *Risk 1:*
2. *Risk 2:*
3. *Risk 3:*

**Planned Mitigations:**

Transparency Principle

Transparency requires visibility into all levels of decision-making and design of an AI system. Designers should clearly document their goals, definitions, and design choices, and any assumptions they have made. Those who build and use AI systems should be forthcoming about when, why, and how they choose to build and deploy them, as well as their data and systems' limitations. Information should be readily available on the quality of the predictions and recommendations the AI system makes. Transparency also encompasses intelligibility, which means that people (in this case, educators, parents, students, etc.) should be able to understand, monitor, and respond to the technical behavior or recommendations of AI systems.

[Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6) on Transparency Principle

**What steps will the analytics or AI process include?**

*Describe data exploration of key data sources for the use case, specific definitions that map theory to data, data science design and method choices, and any assumptions made to provide documentation for traceability and communication.*

**Who will develop the analytics or models?**

*List the individuals or teams developing the analytics or models.*

**How will the limitations of the analytics or AI model be communicated to stakeholders and users?**

*For example, as part of a professional learning program.*

**What means will be built into the system for correction and model feedback by those who provide data and who use its outputs?**

*For example, will the outputs be presented in a dashboard to specific stakeholders, or will the outputs automatically generate a notification or workflow that communicates insight to key stakeholders?*

Privacy and Security

Private or personal data should not be collected or incorporated in analytics or AI products for education unless all groups have agreed this data is necessary to achieve the shared purpose of a specific analytics or AI project. Additionally, the people providing the data need to give permission for the data to be used for this purpose, such as through school policy at enrollment. Ideally, data providers should directly understand the value that they will receive as a result of sharing their data. Finally, the security of that data must be protected, guidelines or policies developed for which roles can access which data, and the level of anonymization needed for specific use case purposes defined.

[Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6) on Privacy and Security principle

Identifying sensitive data, such as personal information, should be part of the use case process. In OEA modules for individual datasets, sensitive data is often pre-identified, and scripts are written to pseudonymize or anonymize specific data fields before they “land” in Stage 2 data lakes and are accessed by researchers or data scientists. For datasets that are not OEA modules, the process of identifying data for sensitivity classification should be conducted through a collaboration between the project’s data engineers and individuals who understand the local education context and datasets.

**How will access to sensitive data be secured and protected in the data environment?**

*For example, is role-based access control defined and operationalized through* [*Azure Active Directory*](https://docs.microsoft.com/en-us/azure/active-directory/roles/custom-overview)*?*

**Does the dataset contain any personally identifiable information (PII) and how will that data be protected and governed?**

*Please describe your process for identifying sensitive data, de-identifying it, and ensuring only the right individuals or roles in the system have access to it (for example, through role-based access control).*

*See* [*this resource*](https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/govern/policy-compliance/data-classification) *for guidance on data governance.* [*This course*](https://docs.microsoft.com/en-us/learn/modules/build-cloud-governance-strategy-azure/) *is also available.*

Accountability

Accountability requires that people who develop and deploy AI systems be held responsible for how they operate. AI systems should never be left to operate unchecked, irrespective of the degree to which they may be capable of acting autonomously. This is what is meant by the phrase “humans in the loop.” A part of this is ensuring documentation of the decisions made during the AI system development. This document can be used for that purpose.

[Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6) on Accountability principle

**Who is responsible for reviewing the Use Case documentation and ensuring that the implementation meets responsible AI principles?**

*Name the primary individuals responsible for the use case documentation, and the process for review with key stakeholders.*

**How will stakeholders and end users be trained on the appropriate use of the system?**

*What capacities are needed to ensure the appropriate use of the data insights, and how will that training be provided?*

**How will the analytics or AI system be monitored over time to ensure analytics and prediction perform reliably? Who will be responsible for this?**

*Describe system governance and retraining over time, and name individuals or teams responsible.*

Inclusion

The datasets used in learning analytics and AI determine the insights and predictions produced. If those datasets do not represent the whole population of learners, if the data quality is poor, or if certain types of data are not included in the models, it will decrease the accuracy, validity, and inclusiveness of the insights. Similarly, if the way the insights are acted upon by the system do not include all groups (e.g., students with disabilities), it can reinforce exclusion from learning opportunities.

[Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6) on Inclusion principle

**What are the constraints of these local data sources for this specific use case?**

*Please describe the limitations of these datasets. For example, are the datasets missing data for certain student populations? Is there bias in the data collection method? This will inform the sections below on Responsible AI principles.*

|  |  |
| --- | --- |
| **Dataset Name** | **Constraints or Limited Representativeness** |
| 1. *For example, number of student absences* | *For example, number of student absences data must be adjusted for days in the school year when schools were closed and no students physically attended school but they engaged in learning remotely.* |
| *2. For example, digital learning app data* | *For example, how does data collection ensure that data inputs are provided by all relevant populations, including diverse or traditionally marginalized groups? Do all students have digital access?* |
| 3. |  |
| 4. |  |

**How will the analytics or AI outputs from the system be provided to all relevant populations, including diverse or traditionally marginalized groups?**