Open Education Analytics

Use Case Template

Jan 2023

This is a preliminary document and may be changed substantially prior to final release of the software or solution 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.

This Use Case for a Predictive Model of Student Well-being was developed through a partnership between Microsoft Education, and the Department for Education, Children and Young People of Tasmania, Australia (“the Department” in short).

To ensure the appropriate and ethical use of data, OEA recommends applying Microsoft's Responsible AI principles. This document shows how these principles can be operationalized in the case of predicting student well-being. The principles include fairness, reliability and safety, privacy and security, inclusion, transparency, and accountability. For more information see: Responsible AI | Open Education Analytics.

1) The Use Case Problem

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

All students have different levels of need, and some require more personalized care and support, yet the process of identifying them sometimes tends to be a subjective and case-by-case process across schools. Often, interventions are put in place only after their context has become critical. In addition, the recent global pandemic has added even more challenges for school systems to provide timely support for students.

To address these challenges, the Predicting Student Well-Being package is designed to help education systems streamline their processes of nominating students in need of personalized support for further well-being assessments. We advocate a human-in-the-loop design for creating such a system for student nomination. Central to the design is the use of machine learning to empower human decision makers to better identify students who need specialized support, suggest the best interventions based on their specific needs, and ultimately improve students’ well-being.

The human-in-the-loop system aims to nominate students in need of personalized support and refer them to human experts for well-being assessment and support. There are 3 stages in the system design: Model Development, Model Consumption, and Model Calibration and Evaluation. This way, we can include human experts early in the loop, and subsequently the human evaluation can be used to augment existing datasets and further retrain the model. In addition, the model generates explanations for why a student should be nominated in terms of the indicators used to make such predictions. These suggestions help the human experts generate insights and actionable interventions for the student population and individual students. This knowledge can also spark larger discussions and investigations on how to best support students based on their specific needs during the iterative process.

We recognize that the model built in the package is optimized with respect to the Tasmania dataset, and therefore subject to the same limitation on data quality and knowledge specific to the dataset. As a result, when applying this package to build their own predictive systems, users should understand that the final performance in general is a function of the data quality, assumptions made, and the modeling quality, on top of other factors. To obtain the best results and advocate responsible AI practices, we strongly recommend users to follow and be inspired by the best practice shown in the package:

- investigating model errors,
- conducting data balance analysis and mitigating data balance issues,
- generating model explanations,
- assessing model fairness, and
- creating robust intervention suggestions with casual inference tools

These practices are supported by the full-suite Responsible AI Toolbox and explanation framework SHAP offered by Microsoft.

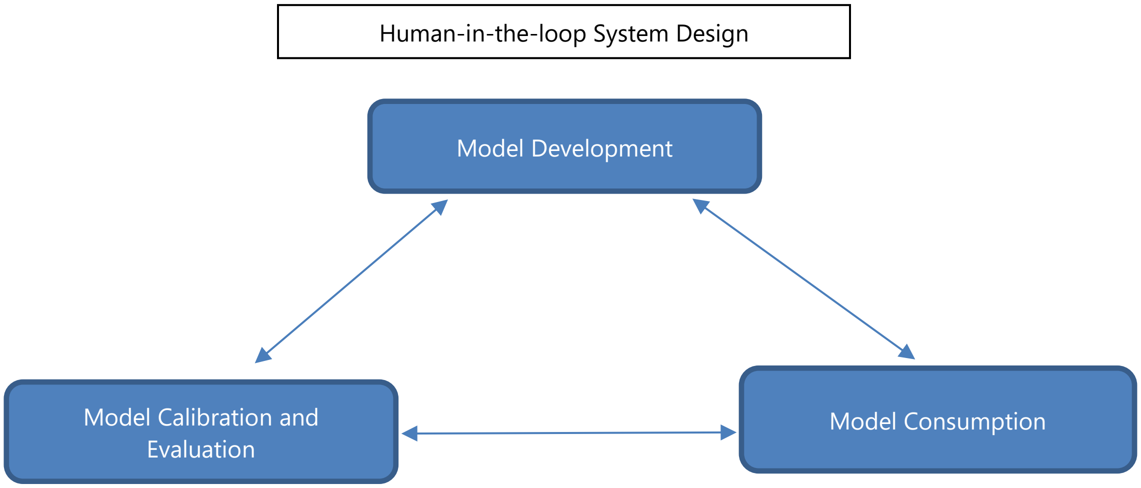
Recommendations on Responsible AI System Design

To develop a system to nominate students for well-being assessment, we recommend users of this package follow or be inspired by a three-stage human-in-the-loop approach described in detail below for Responsible AI best practice. Such a system should be iterative by nature and the three stages inform each other in a back-and-forth manner. We start with the Model Development stage but by no means a system should be deployed in a production setting until enough iterations of all three stages co-produce a mature model. For example, no model should be deployed without calibration and evaluation in a safe, non-production environment.

Model Development: This stage is where the package can be used to develop a predictive model that adapts to specific student datasets for an end-goal defined by the user. Through iterations of model design, we find the best use of the Tasmania dataset is to build a model to proactively nominate students in need of personalized support and then refer them to human experts for well-being assessment and support. Inside the package, the data engineering pipeline demonstrates how to go from pseudonymized dataset to features ready for predictive model consumption – a process called featurization in the data science world; the ML model pipeline demonstrates how to use the features to generate predictions and explanations. These predictions represent estimated probabilities that a student should be nominated for further well-being assessments, while the explanations represent what features make the model come to such predictions. In addition, we demonstrate model monitoring, error analysis, fairness assessment, and casual inference to enhance the robustness of the model. These analyses can gather useful insights into the model and surface important questions about the modeling process to human experts.

Model Consumption: This stage is where the model outputs aid different users in the educational system. Starting from the top level, school administrators can monitor trends and patterns in what categories of features drive students' need of personalized support at the aggregate level. They can also discuss how to guide resources more efficiently and create policies from this process. Subject matter experts on student well-being can assess the nominated students who potentially need personalized support and make recommendations for actions based on their specific needs. Well-being support team can use the expert evaluation results along with model insights to best support them. It is recommended that feedback from all types of users be documented, and relevant insights communicated back to the model developers in the next stage.

Model Calibration and Evaluation: This stage is where model developers calibrate the model with given human feedback. At this stage, data scientists can partner with subject matter experts and other users to incorporate their knowledge for model validation and debugging. Potential analyses can range from identifying biases in the dataset to validation of predictive performance results and top important features. Using the human feedback gathered, data scientists can then fine-tune the model by for example, addressing the biases in existing datasets by further data or feature engineering, and augmenting existing datasets or creating new ones based on the human ground truth to retrain or augment the model. If human interventions were made, changes in the outcome (e.g. nomination prediction) can also be observed and used to create more robust conclusions about the interventions using explicit casual inference approaches. These efforts can then power an improved model in the next iteration.


Human-in-the-loop System Design
Model Development
Model Calibration and Evaluation
Model Consumption

2) Stakeholder Involvement

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

Education system teams responsible for addressing students' well-being needs can collaborate with technology and data groups in the system and external education analytics companies to develop a predictive model that reliably nominates students for well-being assessment and support. To build an informed, ethical, and effective use case, many stakeholder groups should be involved in the design and development of the use case.

Stakeholder Groups	**Relationship to Use Case**	**Involvement in Use Case**
Students	***Indirect:*** *providers of data to the predictive model, who receive well-being assessment if they are so identified by the predictive system, and ultimately support if needed.* *They or their families or guardians should have awareness and give permission for their data to be used.*	*In the initial phases of model development and intervention designs, various types of students should be consulted in the model design process (review of data sources used, theory development). At a later stage, students identified as needing assessment for personalized support may receive or participate in intervention solutions and provide feedback to the system.*
Parents or Guardians	***Indirect:*** *May be providers of data to the predictive model, who may ultimately receive preventative solutions if their students assessed as needing personalized support.* *Families or guardians should have awareness and give permission for students' data to be used.*	*In the initial phases of model development and intervention designs, parents or guardians should be involved in the model design process (review of data sources used, theory development). At a later stage, families with students assessed as needing personalized support may receive or participate in intervention solutions and provide feedback to the system.*
Educators (Faculty or Teachers) and School Support Staff	***Indirect:*** *May participate in developing and implementing preventative solutions.*	*Educators and School Support Staff may directly provide data or feedback to the model and utilize data or insights if they are part of an intervention to provide well-being support.*
School or Department Leaders	***Indirect:*** *participate in developing and implementing preventative solutions.*	*Leaders would directly utilize a tool that predicts students likely to need personalized support.*
School System or Institutional Leaders	***Direct:*** *responsible for addressing students in need of well-being support in schools.*	*Will lead efforts to develop the model and implement preventative solutions; will integrate subject matter expertise with the predictive model in partnership with researchers.*
Researchers	***Direct:*** *research student well-being patterns in the system and be key partner in developing the model.*	*Responsible for maintaining and updating the system to ensure ongoing accuracy and interpretability of the predictive model in partnership with school systems.*
Potential Malicious Actors	***Indirect:*** *Student hackers, external hackers.*	*Corrupt data sources or modify predictive model so the model does not accurately nominate students for well-being assessment. Act to misuse intervention solutions.*

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:

Students, families or guardians, educators, school leaders, system leaders and researchers: In the initial phases of model development and intervention designs, these stakeholders will be involved in the model design process by providing their perspectives on the process for assessing and supporting student well-being needs in the local context (theory development), and in reviewing the data sources that are intended to be used in model development, for example to provide input on the quality and applicability of those data sources to the use case.

Focus group discussions should take place with these groups to assess their interest, concerns and ideas about this model development and the potential intervention solutions that might be valuable for the education system to provide to support students' need for personalized support.

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

As the predictive model is developed, these same stakeholders will again review the model and check for transparency, accountability, and ask other questions or concerns around how the model addresses responsible AI principles (see below). In addition, they will be asked for input on 1) how the model outputs should be communicated (e.g., dashboard designs) and 2) the set of interventions developed to address students' need for personalized support and whether some of these interventions can or should be automated. Finally, when the model starts to be used, they should provide continuous feedback on the system, correcting the model over time.

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

Stakeholder Group	Outputs
Students	*Students nominated by the predictive system and assessed by human experts as needing support may receive interventions such as support groups, medical or mental health supports, learning assistance, social services, and other forms of support, dependent upon their individual needs.*
Parents or Guardians	*Depending on the output results, families of students nominated and identified as needing support may receive interventions such as support groups, medical or mental health supports, learning assistance, social services, and other forms of support, dependent on their individual needs.*
Educators (Faculty or Teachers) and School Support Staff	*Depending on the output results, educators may have access to a tool, dashboard, or data set that nominate students for well-being assessment in their current classes, the reasons that may be causing the nomination, and based on the assessment results, recommends a specific intervention or provides intervention suggestions to the educator to choose among.*
School or Institution Leaders	*Depending on the output results, leaders may have access to a tool, dashboard, or data set that nominates students for well-being support in their schools, and based on the assessment results, recommends a specific intervention or provides intervention suggestions for those students.*

System Leaders	*Data analysis and exploration dashboards to understand patterns of need for personalized support, changes in the causes over time, and analysis of the impact of interventions that support individual students' well-being needs.*

3) Mapping Theory to Data

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

For most common education use cases, research has already been conducted or a theory of the problem developed. For example, extensive research has identified key data elements that are related to the well-being of students. This type of research, theory or model should help identify the most relevant data sources for a specific use case.

- Tasmanian Child and Youth Wellbeing Framework
    - Six domains of wellbeing:
        - Being loved and safe
        - Having material basics
        - Being healthy
        - Learning
        - Participating
        - Having a positive sense of culture and identity

- Vulnerable learners in the age of COVID-19: A scoping review | SpringerLink
    - Index of Community Socio-Educational Advantage (ICSEA) for each school
    - Attendance to video lectures – indicator of receiving proper cognitive/emotional interactions (regardless of grade)
    - Students from disadvantaged backgrounds reported as being more likely to experience markers of disengagement – ex. daily absence, disruptive behavior, poor school connectedness
    - NESCO highlights the importance of addressing the psychological challenges associated with the pandemic and recommends that this should take priority over teaching. Necessity to "ensure regular human interactions, enable social caring measures, and address possible psychosocial challenges that students may face when they are isolated"
    - Australian Digital Index (Thomas et al, 2019) measures digital inclusion in 3 discrete ways: access, affordability, digital ability. Digital divide between students from low and high socio-educational backgrounds

- Predictors of High School Graduation [36474].pdf
    - Demographic characteristics, family background, prior school performance, psychological characteristics, school or community characteristics (student vs institutional)
    - Individual factors – background (demographics, health, prior performance, past experiences); attitudes (goals, values, self-perceptions); behaviors (engagement, coursework, deviance, peers, employment); performance (achievement, persistence, attainment)
    - Institutional factors – families (structure, resources, practices); schools (composition, structure, resources, practices); communities (composition, resources)

Other related links

- Race, Gender, and Measures of Success in Engineering Education - Ohland - 2011 - Journal of Engineering Education - Wiley Online Library

- Models for early prediction of at-risk students in a course using standards-based grading - ScienceDirect

- Student Attendance and Educational Outcomes: Every Day Counts
    - Research shows that 'unauthorized absence' has a negative impact on Educational Outcome

Mapping theory to data and developing the 'data dictionary.'

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.

In the following, we will provide descriptions on a list of data features that we believe to be potentially predictive of the outcome that a student should be nominated for well-being assessment. We strongly recommend that such data should be collected and processed according to the local context and laws of the communities and based on consent and discussions of the stakeholders involved in the use case. For the purpose of illustration, we also provide examples of what features are used in the Tasmanian dataset and the relevant assumptions, processing, and limitations.

It is possible that sensitive information about students may be used in the predictive system, and it is important to use such data safely and properly. Please see "Privacy and Security" section of the Responsible AI Principles below for more ensuring that sensitive data is protected.

Note: OEA modules may provide data sources that support the student well-being use case through accelerating the ingestion of key data sources needed and providing resources to set up these use cases.

Theoretical Construct	**Local Data Source Mapped to Theoretical Construct**
Attendance – Research shows that unauthorized absence is an indicator of other risk factors	*This is where specific fields from available datasets should be mapped to the theoretical constructs important to student well-being.* *In the Tasmanian dataset, we create features indicating patterns of student attendance at school: the minimum, average and maximum number of consecutive days of absence and/or presence at school.*
Assessment scores – Assessment scores are reflective learning	*This is where specific fields from available datasets should be mapped to the theoretical constructs important to student well-being.* *In the Tasmanian dataset, we* - *select assessment scores of certain subjects based on data quality (proportion of missing value, whether an assessment of a subject is mandatory / optional, consistency of assessment criterion).*

	- *create features represented by the earliest assessment score and latest assessment score for each student.*
Medical Conditions – Represents the health context of a student	*This is where specific fields from available datasets should be mapped to the theoretical constructs important to student well-being.* *In the Tasmanian dataset, we* - *create features indicating seriousness of medical conditions;* - *create features indicating whether the student used to have medical condition that raised alert.*
Protection Order – Represents physical well-being	*This is where specific fields from available datasets should be mapped to the theoretical constructs important to student well-being.* *In the Tasmanian dataset, we count the number of protection orders of each individual.*
Disability – Represents additional health context	*This is where specific fields from available datasets should be mapped to the theoretical constructs important to student well-being.* *In the Tasmanian dataset, we create indicators of whether a student used to have certain disability registered in the system.*
Disciplinary – Represents disciplinary sanctions against a student	*This is where specific fields from available datasets should be mapped to the theoretical constructs important to student well-being.* *In the Tasmanian dataset, we* *flag on whether a student had certain disciplinary sanctions;* - *count the number of days a student was under a certain type of sanction.*
Observed Behaviors – Represents negative behaviors a student had in a specific timeframe	*This is where specific fields from available datasets should be mapped to the theoretical constructs important to student well-being.* *In the Tasmanian dataset, we* - *count number of negative behaviors recorded of a student during a specific timeframe;* - *count actions taken against the student during a specific timeframe.*

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. It is recommended users create data dictionaries for their use cases. New data services like Azure Purview 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.

What are the constraints of these datasets for this specific use case? In most systems, the data used for modelling is not of consistent quality or representativeness. It is important to clarify and describe these weaknesses in the data. The following provides a non-exhaustive list of examples of constraints from the predictive modelling in the specific dataset we analyzed.

Limitations on time stamps of features and labels: we have missing or incomplete information on the time elapse between feature and label data creation. Therefore, we are unable to assess the horizon for which the prediction should be used, and advise that this model demonstrated in the package should not be used in production before rigorous testing and iterations of the three stages of human-in-the-loop system. Consultation with subject matter experts is critical regarding questions such as how often features such as grades and medical conditions and nomination outcome (change between need to be nominated and otherwise) do change as students grow in transformative years. We strongly recommend that users should make their best efforts to collect such information and assess the horizon of the prediction.

Dataset Name	**Constraints**
Student Enrolment	- *Some students don't have enrolment records after grade 6. It might be because they transferred from public school to private school so there is no record in the system for them. It can also be that a student drops out and doesn't enroll in any school anymore. At current stage, it's impossible to tell if students who don't have enrolment records are just dropping out of school or transferring to private school.* - *Some students have duplicated records*
Student Assessment (e.g. exam grades)	- *The assessment records of some students are not complete – records are missing for certain reasons, for example, some students don't have KDC scores (an assessment that should be satisfied if she/he needs to enter grade 1) because they enter the system after graduating from kindergarten.* - *Complementary information needed – assessments of some subjects only need to be taken on certain student populations (e.g. TCE are required only for students at grade 11 and grade 12; ACF-MA is an optional assessment for students at grade 10, etc.) All these require efforts to clarify and clean.* - *It is not clear why grade 12 students do not have vulnerability labels and as such, any predictions made would not be accurate for this group of students.* - *Unclear and inconsistent categorization of assessments / subjects – when doing EDA we found the categorization of assessments is not consistent. Some assessments can be grouped by "AssessmentTierCode1" but some should be grouped by "AssessmentTierCode3".* - *Student attendance rates were significantly affected by the COVID-19 Lockdown in Tasmania from 16 March – 2020 to Tuesday 9 June 2020. The data from this period is not comparable with the same period in previous years.* - *If data is complete, we may also use the normalized percentile across the grade / class or other units.*
Student Attendance	- *There are missing attendance dates for some students for various years. Some students had very little attendance data while others had significantly more data.* - *Student attendance rates were significantly affected by COVID-19 lockdown in 2020. The data from this period is not comparable with the same period in previous years.* - *Tasmania moved from 3 Terms to 4 Terms – this occurred at the beginning of 2013. Previous to 2013 we had 3 Terms in a year – this Term 1 being between 15 to 16 weeks – though Term 1 incorporated a break of about 6 school days for Easter.* - *The dates of the Easter Holiday also effect the number of weeks in Term 1 and 2.*
Student Protection Order	- *Typos in protection order type of each record make it hard to categorize / aggregate each protection order type.*

	- *Erroneous information – multiple duplications / extremely long-lasting time of a protection order (e.g., till year 9999)* - *Complementary information needed – there are multiple protection orders issued at the same time on the same student.*
Student Vulnerability Indicators (Vulnerability and LevelofNeed columns)	- *The binary vulnerability indicator refers to whether a student needs personalized support and equivalently level 3 or 4 in the level of need indicator.* - *The current indicator in case management system contains only students who were assessed for well-being to some degree. Most students were not assessed and do not have data in the data set.* - *Time when students were assessed is missing (to be more specific, the time each level of need assessment was made or changed). Thus, we cannot determine the appropriate time range to be used for relevant features.* - *Since the exact time of when the vulnerability label was created was unknown and an approximate timeline of the labeling from April 2020 to August 2021 was assumed, efforts were made to ensure that features used in the model were dated before April 2020, the presumed first month of the labeling period, in order to avoid information leakage (i.e. using the future to predict the past).*
(Additional information needed) Organization Identifier / School information	- *Having the information about the organization of each student/ each instructor would be very helpful to tell the variousness of teachers' perspective of well-being.* - *It can help with generating the feature describing "distance from school" of each student.* - *It will enable building of more specific models, and allow for more careful examination of the model results (before being applied on unidentified students).* - *However, this information was not provided and therefore not used in modeling.*

4) Responsible AI Principles Applied

Fairness Principle

AI systems should treat everyone in a fair and balanced manner and not affect similarly situated groups of people in different ways. To ensure AI models are trained in a way that does not embed or re-enforce negative human biases, models must be tested for fairness. Microsoft has developed an open-source toolkit Fairlearn to support such effort, which can be applied within the Azure analytical services used in the OEA reference architecture.

Who is most likely to be at risk of experiencing harm from this use case?

Generally, when the predictive system will influence a decision or intervention which leads to allocation of valuable resources, such as providing well-being support of students, ***allocation harm*** may exist to extend or withhold certain resources. The accuracy performance of the predictive model, if it is poor, can lead to false positives and false negatives of the nomination, which relates to allocation harm. Too many false positives may overwhelm the assessment or support system, and dilute such resources available for those truly indeed. False negatives withhold or delay assessment or support for students truly in need.

Other forms of harm may include false or misleading information from the suggested explanations of the model or causes why a student should be nominated. Such information, if unexamined and applied directly to the interventions for a student in need, may provide the student with the wrong or inappropriate type of support, leading to an ineffective outcome.

More specifically, special attention should be paid to fairness assessment and other analyses regarding sensitive demographic groups, for example, children from low-income families and/or children in traditionally underserved student groups (student with disabilities, English language learners, youth involved in the juvenile justice system). We recognize that there are at least two aspects:

1. collection of sensitive demographic information may be controversial in certain local social context, or restricted or prohibited in certain local laws and regulations; and
2. analysis into or with the aid of such demographic information may also be controversial in certain local social context, or restricted or prohibited in certain local laws and regulations, even for the purpose of fairness assessment.

The reasons why the collection or analysis of such data is sensitive can be cultural, legal, political, and/or related to privacy protection. In the case of the Tasmania dataset, sensitive demographic information was collected under the consent of the stakeholders and analysis of such data including fairness assessment was conducted under the research agreement. In addition, efforts have been made to protect the privacy and security of the data. See "Privacy and Security" below for more details.

Finally, the processes of assessment by a human expert or experts and support for the students who were confirmed as those in need of it as a result of this process represent an intervention in the life of students. By analogy with medical field, any intervention may cause good, harm, or no effect, and needs to be studied in a rigorous manner. While this is not directly related to the ML model developed for this use case, it is one of the aspects that we see as important to highlight from data science perspective.

Planned Mitigations:

We generally deem that it is valuable to collect and analyze any sensitive demographic information, for the purpose of evaluating whether a predictive model is fair or not among different demographic groups. We also advise that users should be aware of the local context and comply with the local laws and regulations where the use case will be applied.

The package allows fairness assessment of the model across sensitive demographic subgroups in terms of multiple fairness metrics such as disparities in recall rates, accuracy rates, etc. If different demographics subgroup is found to generate disparate accuracy rates, we recommend users first discuss with subject matter experts to find out the underlying causes of the disparities, following the three-stage human-in-the-loop system design.

There can be many causes of such disparate results. One could come from the data generation or collection process, under the influence of human judgment or specific selection criteria for the purpose of educational interventions. For example, it may be the case where a higher percentage of students from a certain demographic minority indeed have a need of personalized well-being support according to a well-being framework, and the process selects those students more often naturally, as it is intended. Alternatively, it may also be the case that more students of a certain demographic minority are nominated intentionally and preemptively by school support teams.

In the case of Tasmania dataset, sensitive demographic information has been used to assess fairness of the predictive model under the research agreement. The research question specifically seeks to find differences between the predictive accuracy and other metrics among different demographic groups. We found that there are some patterns in the data generating process of the label data – whether a student is nominated by school support teams for further well-being assessment by experts. An example of the patterns in this data we found is that: a higher percentage of students from low-income backgrounds are nominated and identified as needing supports for well-being than students from high-income backgrounds. It is therefore not surprising to see that an unmitigated model has learned such a pattern from data and receives a higher recall rate for economically disadvantaged students than the counterpart. Similar patterns were observed in students with indigenous status and independent status, and female students to various degrees.

As mentioned, it may be possible that the nomination process selected one or more of the sensitive demographic groups for a purpose. Further discussion among education leaders, model developers, and subject matter experts is highly encouraged to discuss 1) whether this was indeed intentional and for what purpose, 2) whether such disparity should be corrected, and if so, 3) what kind of mitigation is needed. For example, mitigation may include a change in nomination procedure, or mitigate the disparity in the model as supported by the Mitigation section of the Fairlearn package.

Controlled experiments are the most reliable ways for estimating “treatment effects” in similar high-stake scenarios such as medicine. Even if controlled experiments are not possible, other causal inference approaches are possible. Causal inference dashboard is built as an important feature as part of the RAI dashboard. It is highly recommended to leverage relevant methods and experts to reduce the risks of causing harm to students through the framework-driven actions.

Are these groups and subpopulations clearly labelled in the dataset?

Demographic information is provided in the student dataset. Additionally, viewing the impact of the model across these subpopulations is easily visible with these fairness tools. As mentioned, demographic information such as gender, economic disadvantage status, indigenous status, independent status, and birth months, were included in the Tasmania dataset.

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.

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

1) Risk 1: Data drift in student populations caused by a changing current event landscape. Events such as the COVID pandemic can accelerate the change in model validity as new indicators for students' well-being become more relevant, or the population faces new risks.
2) Risk 2: Data drift in student populations caused by a change in the nomination or assessment procedure. The Tasmania dataset assumed a certain nomination and assessment procedure where students are nominated by school support teams and then assessed by subject matter experts according to specific well-being criteria. And models were built in the package based on this assumption. If there were, for example, no nomination process involved in a new dataset, the models should be re-developed by the user.
3) Risk 3: Complete automation of well-being nomination or labeling, or over reliance on model predictions with little or no human oversight in the system.
4) Risk 4: Direct application of the insights (such as model explanations or causal inferences) drawn from the model and translating them into immediate interventions, without examining the insights by human experts.
5) Risk 5: Data losses or incorrectness of entry may lead to incorrectness in the assessment of probabilities that students should be assessed in the direction either "false negatives" or "false positives"

Planned Mitigations:

We recommend mandatory human-in-the-loop decisions with the assistance of model output, as outlined in the Model Calibration stage of the recommended System Design. At no time should stakeholders rely solely on the output of the AI system to make decisions about students' well-being, broadly construed.

To aid human in decision-making in *monitoring, debugging, and understanding* model output, the RAI dashboard offers these key features:

1. Model statistics to *monitor* model performance on the population or across custom groups of students;
2. Error analysis to *discover* and *drill down* into specific groups of students with certain features (cohorts) in which most errors of the model are made;
3. Interpretability to *explain* what features are the most important to the model predictions;
4. Causal inference to *examine* whether certain features are causal in the model predictions; and
5. What-if analysis to help users *design* interventions for individual students.

It is worth emphasizing that users should not rely solely on a single feature but rather examine multiple or all aspects of the insights drawn from the model with experts. For example, it would generally be more convincing that an important feature is also causal and has statistically significant effect on overall student population or individual students. But human experts should be consulted before moving to designing interventions. Also, as mentioned in the fairness section, fairness assessment should be performed to facilitate discussion between model developers and experts.

As included in the package, the PowerBI dashboard template also contains high-level summary statistics of model performance and customer fairness metrics. It helps stakeholders such as education system leaders to gain a basic understanding of the model performance to facilitate discussions around providing students with well-being support.

We recommend periodic retraining of the model in intervals in which new data is prioritized over old data in order to capture more recent or relevant trends. Triggering retraining or recalibration can be considered based on different quantitative metrics such as:

- Based on regular scheduled intervals (say every 6 months)
- Based on degree in which new labels are updated and validated by humans
- Based on degree in which errors are detected by humans in the loop (regular validation of model decisions)

To address the risks of data loss and incorrectness, it is recommended that sound data engineering principles are established in the information systems where the data is collected and preprocessed, including but not limited to: detection of software and hardware failures, information handling policies that include failover mechanisms, quick detection and mitigation of outages and anomalies. In the case when it is known by system operators that data from a certain source within a particular period is not trustworthy, this information should be made transparent to the team in charge of the model so that appropriate actions can be taken during calibration process for mitigation, such as the assessment of impact of the issue, data cleaning actions, and potentially even changes to the model feature engineering.

Transparency Principle

Transparency requires visibility into all levels of decision-making and design of an AI system. Model developers including data scientists and software engineers 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.

What steps will the analytics or AI process include?

Tasmania and Microsoft provided documentation describing the capabilities and limitations of the AI system, including, but not limited to, warnings to the end-user about relying on outputs made by the system and descriptions of the intended uses of the system. Here, the goal is to encourage stakeholders to use the system only in the ways in which it was designed and intended to be used, by clarifying exactly what those use-cases are and how the system is designed to enable them. The analytics process should be integrated into all three stages in the human-in-the-loop System Design.

The analytics process demonstrated by in the package pipelines includes the following 4 steps:

1. *Data subsetting and aggregation:* Data identified in the above theory and data discussion will be located in the production data environment. Only the columns needed for model building will be used.

2. *Feature engineering and model table construction:* Data from step 1 will be combined into a single table for building the predictive model. Because each row in this table represents 1 student's data, certain columns such as attendance records will need to be aggregated into a single metric.
3. *Model building:* ML models will be used to build a best performing model. This process will record and catalog the model table and all models produced for future reference. The model will be used to make predictions on the model table and InterpretML (or SHAP) will be used to identify individual feature importance.
4. *RAI dashboard and PowerBI deployment*: Model results and other data (i.e., level of students' needs, and their time trends and demographics) can be visualized and interacted via the RAI and PowerBI dashboards. Users then can explore model findings such as fairness assessment and help decision making. See more details in the Reliability and Safety section above for dashboard features and the recommended practices of data handling, especially for incidents that may affect data trustworthiness.

Who will develop the analytics or models?

In general use case, software engineers and data scientists should form a team of model developers to develop the system based on the package with the help of stakeholders and subject matter experts. For the Tasmania dataset use case, a small group of data scientists from Microsoft and the Tasmania Department of Education worked together to develop the model iteratively in the packaged pipelines. The model was enhanced as more datasets representing more aspects of the well-being construct were added into the system.

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

At the model development stage, there should be ongoing conversations between all direct stakeholders and these conversations delved into the purposes and data limitations of the datasets used in the model building. In the Tasmania dataset use case, direct stakeholders had these conversations where Tasmania DoE also invited educators, support staff, and system leaders to develop the model jointly with Microsoft and identify limitations. At a later stage in the project, when the project Tasmania DoE will develop a training program to show model users how to use the system predications appropriately, how to give feedback to improve the system, and to clearly outline the model's limitations.

At no time should educators or support staff rely solely on the output of the AI system to make decisions about student well-being, broadly construed. The AI system will always be limited by both the data sources it incorporates and by the changing conditions of well-being in local school, family, and student contexts. The system will not have all the data for perfect predictions. Only subject matter experts, school support staff, and Tasmania system leaders will design actions with the aid of model outputs.

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

The package is intended to facilitate a three-stage system where stakeholders (especially direct stakeholders such as students and their families as well as researchers or model developers) are included in the design and decision-making process in every stage. We describe below what may happen in every stage as an example.

At the Model Development stage, students and their families who provide data should be engaged early, to provide inputs into what data should be collected to enhance students' well-being before model developers build the models. At the

Model Consumption stage, the model output after examination including consultation with experts should be provided to the relevant stakeholders including the students and their families for correction and model feedback. Feedback should be prioritized in the early iterations of the system building, so as to flag and finetune incorrect results or questionable recommendations. More specifically, the RAI and PowerBI dashboard are part of the tools built in the package to facilitate such process in this stage. At the Model Calibration and Evaluation stage, such feedback should be documented and used to calibrate the model. Again, the dashboards are meant to facilitate this process as they monitor model performance and generate insights into findings.

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. 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. This principle also applies to high-level data summary and insights from the model surfaced by the dashboards, including, but not limited to, sensitive information that educators and other stakeholders in the system should not see or otherwise have access to. The dashboard should present the analysis the educators and support staff need while limiting their control over using the AI system in unintended ways.

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. Common privacy and security risks may include:

1) Data about levels of students' well-being needs and whether a student is nominated for assessment, as well as the causes of that outcome (e.g. sensitive medical information) become publicly available or available to individuals other than those who should have role-based access to such information.

2) Data about the health and wellbeing of students, families, educators, or staff is unsecured and becomes associated with the students' personal information.

3) Data about interventions taken with students and families becomes available to individuals or the public who should not have access to such information.

4) Leakage of personal information through fields that are not pre-identified as sensitive and not scrubbed, such as free text notes on medical records

5) Certain ML inference attacks, such as membership inference attack or property inference attack, may be performed by adversaries if they get hold of the ML model to make .

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

Only pseudonymized student data will be used to build and assess models, so no students will be available or identified through this process. If the school system wishes to re-identify students (so that supports can be provided to the student), re-identifying the student data will be performed only by those inside the education system with appropriate role-based permissions governed through the above policies and implemented through Azure Active Directory, Azure Synapse, Azure ML and PowerBI.

At pre-processing stage, it is recommended to implement additional “sanity checks” for leaked PII that help increase confidence that all names have been scrubbed prior to modeling. This can be done by human reviewers, by automatic analysis of text and tabular data, or a combination of both.

To avoid the possibility of membership inference or property inference, it is important to protect the trained model from potential adversaries and only provide access to authorized personnel. The model and its outputs need to be as secured and protected as the data on which it was trained.

Accountability

Accountability requires that people should be held responsible for how AI systems operate. In other words, 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.

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

The decision makers in the education system who use the predictive model in practice to identify supports for students and schools will be responsible for continued implementation of the principles responsible AI described in this document. They should review this documentation thoroughly and update it if decisions or data changes.

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

Education system leaders will train key data stakeholders on interpretations of model accuracies via deployed RAI and PowerBI dashboards. Detailed technical documentation will be created as a reference.

Education system leaders and schools will be responsible for training schools, educators, and all support staff and stakeholders on how to understand the model, dashboards and other outputs from the model, and on the appropriate and intended use of the outputs to inform their decisions and actions at the school and student level.

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

Education system leaders and schools will be responsible for monitoring analytics and prediction performance. It is recommended that they establish a threshold for the accuracy performance (e.g. 80% accuracy or recall rate) of the model and use it to decide if model quality is sufficient for use in practice. Model accuracy should be checked at regular intervals with the use of new attendance record data. It is recommended that the model be retrained at least regularly by either the education system's data scientists or an external data science partner. It is also recommended that a process is established for ensuring data quality in the dataset that is used to train the model and run inference, incorporating practices for handling incidents that have potential to cause data loss or incorrectness, and transparency of information flow when and if such incidents occur so that appropriate actions can be taken.

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 in which the insights are acted upon by the system does not include all groups (e.g., students with disabilities), it can reinforce exclusion from timely well-being support. For more details, refer to the recommendations on allocation harm in the Fairness principle above.

How does data collection ensure that data inputs are provided by all relevant populations, including diverse or traditionally marginalized groups?

As a result of this modelling work in Tasmania, the Department of Education mandated that all students should be assessed for well-being by staff trained in the assessment's use to improve data quality. We recommend that other users of the package ensure all students are included as well in the data collection process.

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

The results of this model based on the Tasmania dataset will be applied equally to all students in the system. To ensure that results are served equally, the Tasmania DoE is ensuring equal training on the system for education stakeholders who would are responsible for applying this tool throughout the school system. We recommend that other school systems should make their best efforts to ensure that their model systems are as inclusive as possible and employ fairness assessment for sensitive demographic groups.

More on the OEA Use Case Template and Documentation:

The OEA Use Case Template can be used to plan a single use case or multiple times to develop an inventory of use cases, such as when an education system is developing 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 can be 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 should be 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 should generally be 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.

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 to be 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.