# SCENARIO-Predict student attrition

Data Factory

Machine Learning

SQL Database

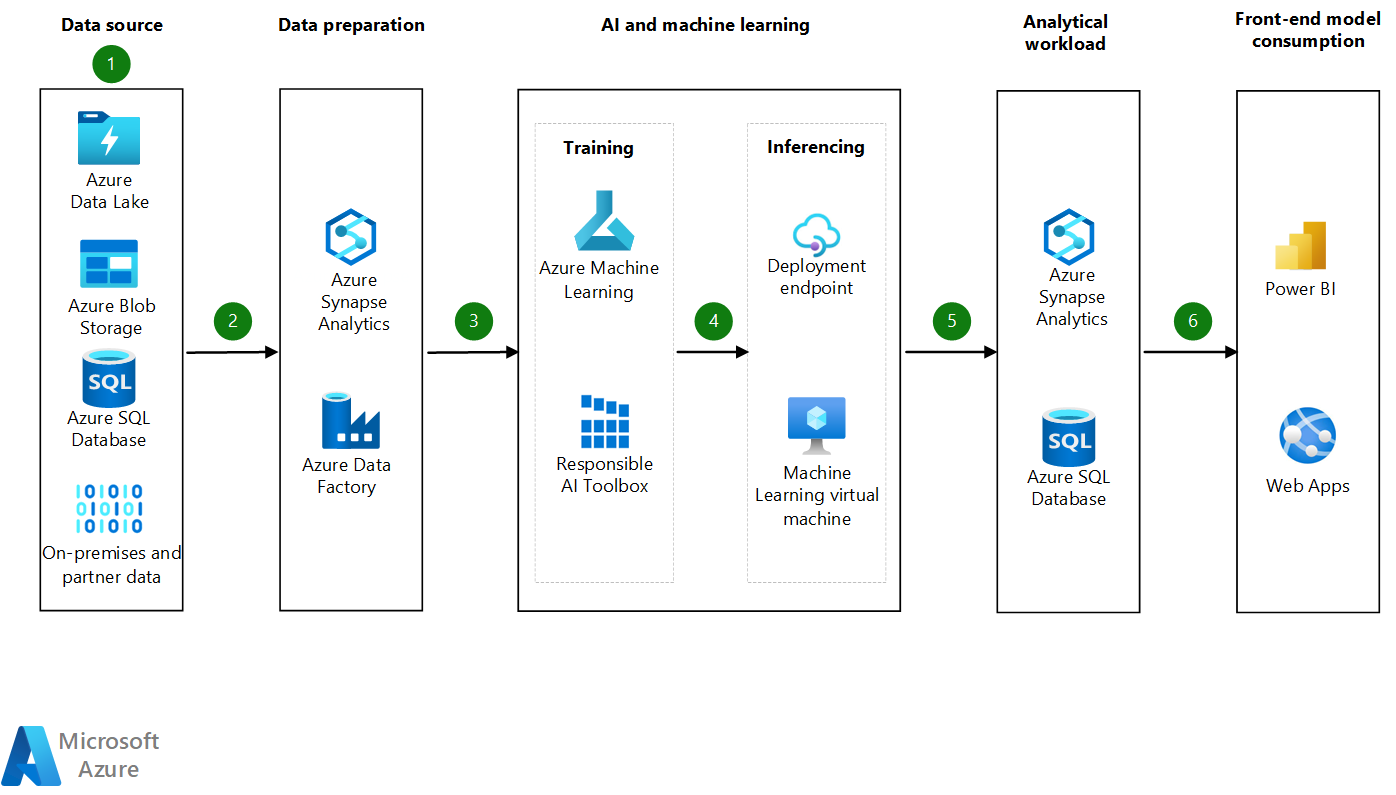
Synapse Analytics

Power BI

This article presents a solution for predicting student attrition. Core components include:

* Azure Machine Learning for training and deploying machine learning models.
* Responsible AI Toolbox for identifying contributing factors and detecting bias.
* Azure Synapse Analytics for data processing.

## Architecture

[](https://docs.microsoft.com/en-us/azure/architecture/example-scenario/ai/media/student-attrition-prediction-architecture.png#lightbox)

Download a [*Visio file*](https://arch-center.azureedge.net/US-1958830-student-attrition-prediction-architecture.vsdx) of this architecture.

### Dataflow

1. Data source. Educational data comes from various sources:
   * District records
   * Digital archives of instructional materials and grade books
   * Detailed information about financial grants
   * Student responses to course surveys

Educational data is retrieved from various data sources, including district databases and state databases. It's housed in an on-premises database.

1. Data preparation. During data preparation, the data is gathered, combined, structured, and organized. It's then ready to be used:
   * To build machine learning models.
   * For business intelligence purposes.
   * In analytics and data visualization applications.

The solution uses Azure Data Factory to orchestrate the process of transforming and loading the data. It uses Azure Synapse Analytics to process data and to trigger Machine Learning experiments.

1. AI machine learning training. The solution uses Azure Machine Learning studio to train a wide range of supervised learning algorithms and to find a model that accurately predicts student attrition. The following Responsible AI Toolbox tools help to implement responsible AI:
   * The interactive interpretability dashboard helps users understand the major factors that contribute to student attrition.
   * The fairness tool provides an interactive dashboard for detecting and mitigating bias that's related to student gender and race in the chosen model.
2. AI machine learning inferencing. During inferencing, previously unseen data points are fed into a machine learning model. The model calculates the probability of student attrition. In Machine Learning, a built-in model registry stores and provides version control for models in the Azure cloud. The model registry makes it easy to organize and keep track of trained models. Trained models are deployed to instances of Azure Data Science Virtual Machine or managed endpoints.
3. Analytical workload. The model scoring results are stored in Azure Synapse Analytics and Azure SQL Database. The results are then available for use in the front end and for monitoring and retraining the models.
4. Front-end model consumption. Power BI and the Web Apps feature of Azure App Service consume the scored results.

#### Student data schema

The information that's critical for the student attrition model consists of factors that influence student behavior. The following table lists data elements that affect student attrition and retention patterns. The list isn't complete.

| Feature | Subfeature | Description |
| --- | --- | --- |
| Gender |  | The gender that the student was assigned at birth or "Not reported." |
| Race |  | The student's reported race: either "Black," "White," "Pacific Islander," "Asian," "American Indian," or "Not reported." |
| First generation in college |  | Whether the student is a first-generation student. With a first-generation student, no parent or guardian has a four-year degree or higher from a college or university. |
| Total terms |  | The total number of terms that the student has been enrolled. |
| High school graduate or GED |  | Whether the student graduated from high school or has an equivalent education. |
| Cumulative grade point average (GPA) |  | The average over all the grades that the student has earned. |
| Cumulative credit hours earned |  | All the hours that the student has accumulated over the period of enrollment. |
| Transferred to another program |  | Whether the student has switched programs since starting. |
| Financial aid eligible |  | Whether the student is eligible for a grant. |
| Age | Minimum | The minimum age that's recorded for the period of enrollment. |
| Age | Maximum | The maximum age that's recorded for the period of enrollment. |
| Entry type | Dual enrollment | Whether the student is enrolled in a dual program. |
| Entry type | Early admission | Whether the student was admitted in advance of the usual notification date. |
| Entry type | First time in college | Whether the student is enrolled in college for the first time. |
| Entry type | Other | Other admission types, indicated as a numerical value. |
| Entry type | Re-entry | A numerical value that represents re-entry admission. |
| Entry type | Transfer | A numerical value that represents transfer admission. |
| Academic standing | Academic probation | The period in which the student must improve academic standing, expressed as a normalized value in proportion to the enrollment period. |
| Academic standing | Academic suspension | Whether the student has been suspended, expressed as a normalized value in proportion to the enrollment period. |
| Academic standing | Academic suspension for one year | The proportion of the enrollment period that the student has been suspended for one academic year. |
| Academic standing | Academic warning | Whether the student has been given an academic warning for failing grades. |
| Academic standing | Extended probation for low GPA | The period that the student's probation has been extended in proportion to the enrollment period. |
| Academic standing | Good academic standing | The proportion of the enrollment period that the student has been in good academic standing. |
| Academic standing | Probation after suspension or dismissal | Whether the student has been placed on probation after academic suspension. |
| Instruction type | Blended | The proportion of time that the student has received a combination of in-person and online instruction. |
| Instruction type | Fully online | The proportion of time that the student has received only online instruction. |
| Instruction type | Remote learning | Whether the student has enrolled in remote learning. |
| Instruction type | Remote learning blended | The proportion of time that the student has combined in-person instruction with remote learning. |
| Instruction type | Traditional | The proportion of time that the student has received conventional in-person instruction. |
| Faculty type | Faculty | The proportion of instruction that the student has received from academic staff at the institution. |
| Faculty type | Adjunct | The proportion of instruction that the student has received from contract professors. |
| Faculty type | Unknown instructor type | The proportion of instruction that the student has received of unknown type. |
| Attrition |  | Whether the student has dropped out of college. |

### Components

* [Azure Data Lake](https://azure.microsoft.com/solutions/data-lake) offers limitless storage for data in different shapes and formats. Besides enterprise-grade security and monitoring support, Azure Data Lake integrates easily with Azure analytics tools. Built on top of [Azure Blob Storage](https://azure.microsoft.com/services/storage/blobs), Azure Data Lake can manage large amounts of unstructured data, such as archives and data lakes. The service is a good fit for high-performance computing, machine learning, and cloud-native workloads. This solution provides a local data store for the machine learning data and a premium data cache for training the machine learning model.
* [SQL Database](https://azure.microsoft.com/products/azure-sql/database) is a fully managed database engine for modern cloud applications. This database service offers built-in intelligent optimization, global scalability and availability, advanced security options, and dynamic scalability with no downtime. SQL Database can automatically process relational data and non-relational structures such as graphs and JSON, spatial, and XML data. For this service's availability guarantee, see [SLA for Azure SQL Database](https://azure.microsoft.com/support/legal/sla/azure-sql-database/v1_8).
* [Data Factory](https://azure.microsoft.com/services/data-factory) is an orchestration and cloud extract-transform-load (ETL) tool. Besides offering over 90 built-in connectors across various data sources, Data Factory provides copy and transformation functionality in a no-code environment. You can use its diagram view to monitor and manage data integration processes.
* [Azure Synapse Analytics](https://azure.microsoft.com/services/synapse-analytics) is an analytics service for enterprise data warehouses. This service uses SQL and Spark technologies and offers dedicated or serverless options for querying data. Azure Synapse Analytics provides a unified experience for ingesting, exploring, preparing, transforming, and managing data. The service also makes data available for business intelligence and machine learning purposes.
* [Machine Learning](https://azure.microsoft.com/services/machine-learning) is an enterprise-grade machine learning service for easy model development and deployment to a wide range of machine learning targets. This service provides users at all skill levels with a low-code designer, automated machine learning, and a hosted Jupyter notebook environment that supports various integrated development environments (IDEs).
* [Azure Machine Learning studio](https://azure.microsoft.com/services/machine-learning/" \l "faq) is a cloud service that you can use to accelerate and manage the machine learning project lifecycle. This service handles training, model deployment, and the management of machine learning operations (MLOps).
* [Responsible AI Toolbox](https://responsibleaitoolbox.ai/) is a collection of integrated tools and functionalities that help you implement responsible AI principles. The toolbox integrates ideas from several open-source tools in the areas of error analysis, interpretability, fairness, counterfactual analysis, and causal decision-making. You can use this open-source framework to assess machine learning models quickly and easily.
* [Azure Data Science Virtual Machine](https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview) is a customized virtual machine (VM) image on the Azure cloud platform that's built specifically for data science. The image has many popular data science tools preinstalled and preconfigured to jump-start building intelligent applications for advanced analytics.
* [Machine Learning endpoints](https://docs.microsoft.com/en-us/azure/machine-learning/concept-endpoints) are HTTPS endpoints that clients can access to retrieve inferencing output, or scoring output, from a trained model. Each endpoint provides a stable scoring URI with key-token authentication.
* [App Service](https://azure.microsoft.com/en-us/services/app-service/) provides a framework for building, deploying, and scaling web apps. The [Web Apps](https://docs.microsoft.com/en-us/azure/app-service/overview) feature is an HTTP-based service for hosting web applications, REST APIs, and mobile back ends. With Web Apps, you can develop in .NET, .NET Core, Java, Ruby, Node.js, PHP, or Python. Applications easily run and scale in Windows and [Linux](https://docs.microsoft.com/en-us/azure/app-service/overview" \l "app-service-on-linux)-based environments.
* [Power BI](https://powerbi.microsoft.com/) is the Azure software as a service (SaaS) for business analytics and visually immersive and interactive insights. Power BI provides a rich set of connectors to various data sources, easy transformation capabilities, and sophisticated visualization capabilities.

### Alternatives

* This solution uses Machine Learning as a data modeling and deployment tool. Instead, you can use Azure Databricks with a code-first approach.
* The data sources in the solution are Azure components. You can also use training data from third-party sources.

## Scenario details

Student attrition is one of the most common challenges faced by educational institutions across the globe. Attrition is prevalent at all levels of both private and public education systems, leading to various consequences. For example, in the public system, student attrition results in ineffective use of tax revenue. At private institutions, attrition damages reputations and negatively affects school revenue. When student attrition levels are elevated, fewer graduating professionals enter the workforce, which threatens the global economy.

Given these consequences, educational institutions are heavily investing in ways to reduce student attrition. It has become increasingly important for these institutions to identify and address the root causes of attrition. Understanding the major contributing factors helps institutions support individual students and implement at-scale solutions to reduce attrition rates. Institutions also must be able to predict attrition rates to minimize attrition across a broad set of program offerings.

Predicting student attrition is a binary classification problem that predicts whether a student will leave school. This type of model is built on student-centric data that includes demographic data, financial data, the student's academic history, and the course delivery mode.

For interpretability and bias detection, this solution uses the Responsible AI Toolbox. When educational institutions understand the factors that impact continuous education, they can take appropriate actions to curtail student attrition. But institutions need to use models with no bias toward protected classes like gender and race. This point is important, as is the need to consider and implement responsible AI practices. Equitable models help provide unbiased support for student success.