Summary

The use of machine learning for astrobiological research has been rapidly growing over the past few years and is almost inevitable for future applications. Accordingly, it is imperative for the next generation of researchers and students to be equipped to understand and implement machine learning. One of the main aims for our application is to provide a foundation for teaching machine learning in astrobiology. Studies have shown over-and-over again the value in active learning. Thus, enabling educators to create interactive lessons, tutorials and introduce machine learning concepts in an accessible, easy-to-understand format can greatly benefit both learners and educators. Below, we demonstrate a simple example of how our application can be integrated into a lesson plan designed to teach machine learning fundamentals for astrobiology interactively.

Part 1: Teach & Discuss

Lesson 1: A Gentle Introduction to Machine Learning

Goal: Prior to exploring the application, it is important to lay down the fundamentals of important machine learning concepts including what machine learning is broadly, machine learning workflows, and machine learning terms and definitions. The goal of this lesson is to cover the basics of machine learning to lay down the foundation for its value in astrobiology research.

Learning Objectives

- What is machine learning
- Strengths and weaknesses of machine learning
- Data formats
- Machine learning workflows
    - Data acquisition
    - Exploratory data analysis and pre-processing
    - Unsupervised learning
    - Supervised learning
- Data quality and human-in-the-loop

Resources: https://vt-coga-ai2-astrobiology.hf.space/ and https://vt-coga-ai2-astrobiology.hf.space/defn_uses

Considering the large variety of machine learning approaches, we simplify our application to those most relevant to astrobiological research. We've divided the application into two parts, *General* and *Model Development*. The teaching and discussion can focus on the *General* section. To get started, the *About* page provides fundamental information for describing machine learning including its main components (i.e. exploratory data analysis, data pre-processing, model selection and tuning, model evaluation, and final model selection) (Figure 1).

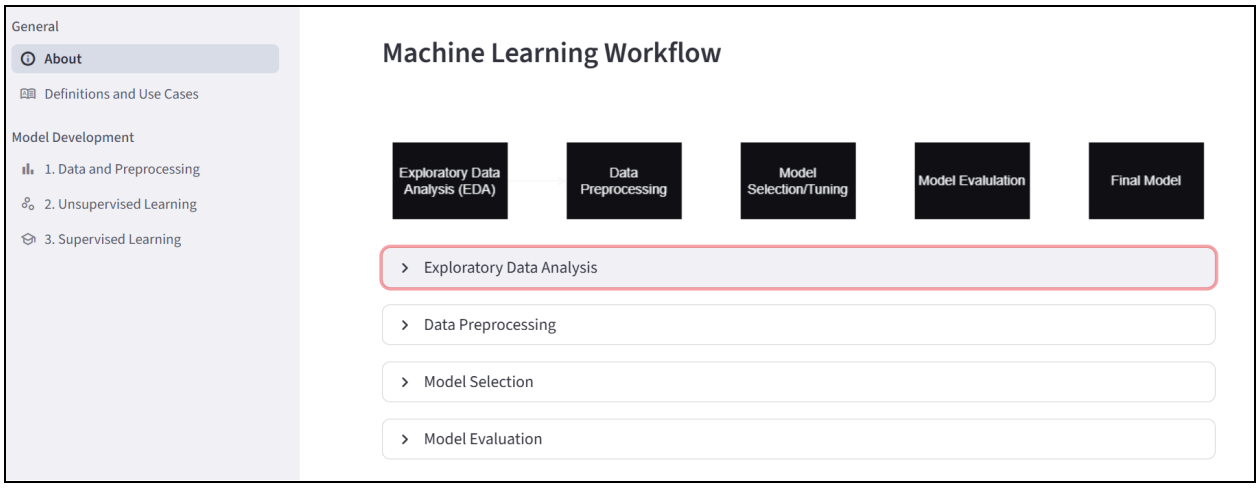


Figure 1. About Page. Here, the educator can find relevant information for a simplified linear machine learning workflow with associated definitions for each component.

Once the broad topics of machine learning are covered, we provide a *Definitions and Use Cases* page to provide more relevant information that will be needed to understand the subsequent part of the application for model development. This page covers a variety of definitions including basic machine learning terms, pre-processing terms, model terms, and model metric terms (Figure 2).

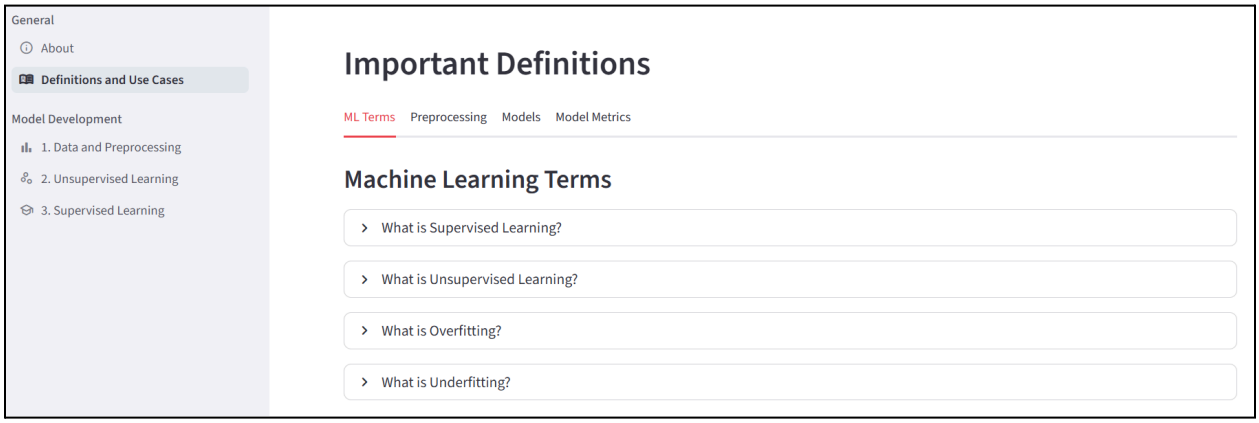


Figure 2. Definitions and Use Cases. Here, the educator can find definitions and use cases relevant to the machine learning algorithms present in our application.

Lesson 2: Data Pre-processing

Goal: Here, the concept of data pre-processing of tabular data will be explored.
Learning Objectives

- Feature engineering
- Scaling, standardization, and normalization
- Encoding of categorical variables
- Handling missing data
- Data augmentation

Resources: https://www.ibm.com/docs/hu/spss-modeler/saas?topic=values-handling-missing

Lesson 3: Fundamentals of Unsupervised Learning

Goal: Unsupervised learning is a powerful approach that can be used standalone to reveal structure and patterns in datasets; however, it can also be used as an embedded step in data pre-processing of a supervised machine learning workflow to reduce computational load. In this lesson, both aspects of unsupervised machine learning will be discussed.

Learning Objectives

- What is unsupervised learning?
- Unsupervised learning as a standalone approach
    - Dimensionality reduction vs. clustering
- Unsupervised learning as a pre-processing step for supervised learning

Resources: https://youtu.be/6TE1uDQAh0k?si=dCv4EQJpjpDtAB25

Lesson 4: Fundamentals of Supervised Learning

Goal: Teach the fundamentals of supervised machine learning.

Learning Objectives

- What is supervised learning?
- Training/Testing/Validation splits
- Classification vs. regression
- Statistical considerations of select supervised machine learning models
- Model Evaluation
- Post-processing of model
    - Hyperparameter tuning
    - Feature importance extraction

Resources: https://youtu.be/erQezPVG5ig?si=M4wfHgZxvQSQQjTk

Part 2: Show & Experiment

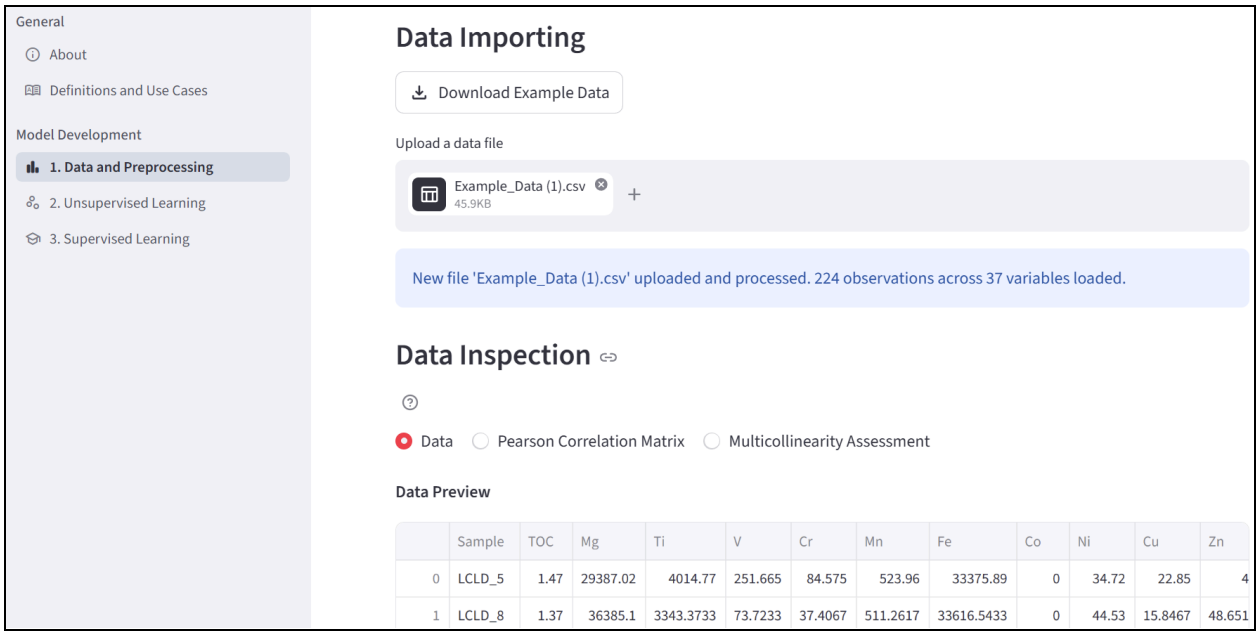
Lesson 5: What Does it Mean to Explore and Pre-process Data?

Goal: Using the application, the teacher will show and have the students follow along how to explore raw data statistically and pre-process a dataset using the example dataset provided for download in the application.

Learning Objectives

- Choosing correct model
    - Trade-off between performance and computational efficiency
    - Selection based on statistical criteria

Resources: https://vt-coga-ai2-astrobiology.hf.space/data_preprocessing



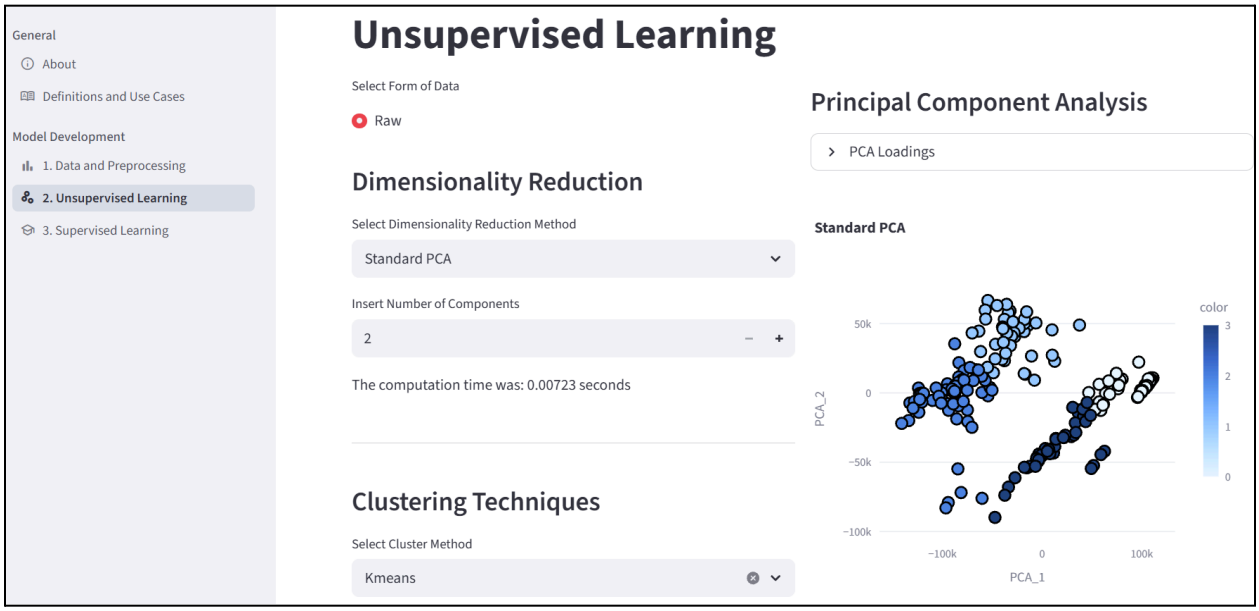
Lesson 6: How Can Unsupervised Learning Help Us Reveal Structure and Information in a Complex Dataset?

Goal: Using the application, the teacher will show and have the students follow along to build an unsupervised learning model.

Learning Objectives

- Understand the differences between dimensionality reduction and clustering
- Understand how to examine the structure of data from dimensionality reduction
- Compare the computational time between methods that are based on linear principles vs. methods that are based on non-linear principles.

Resources: https://vt-coga-ai2-astrobiology.hf.space/unsup_learn



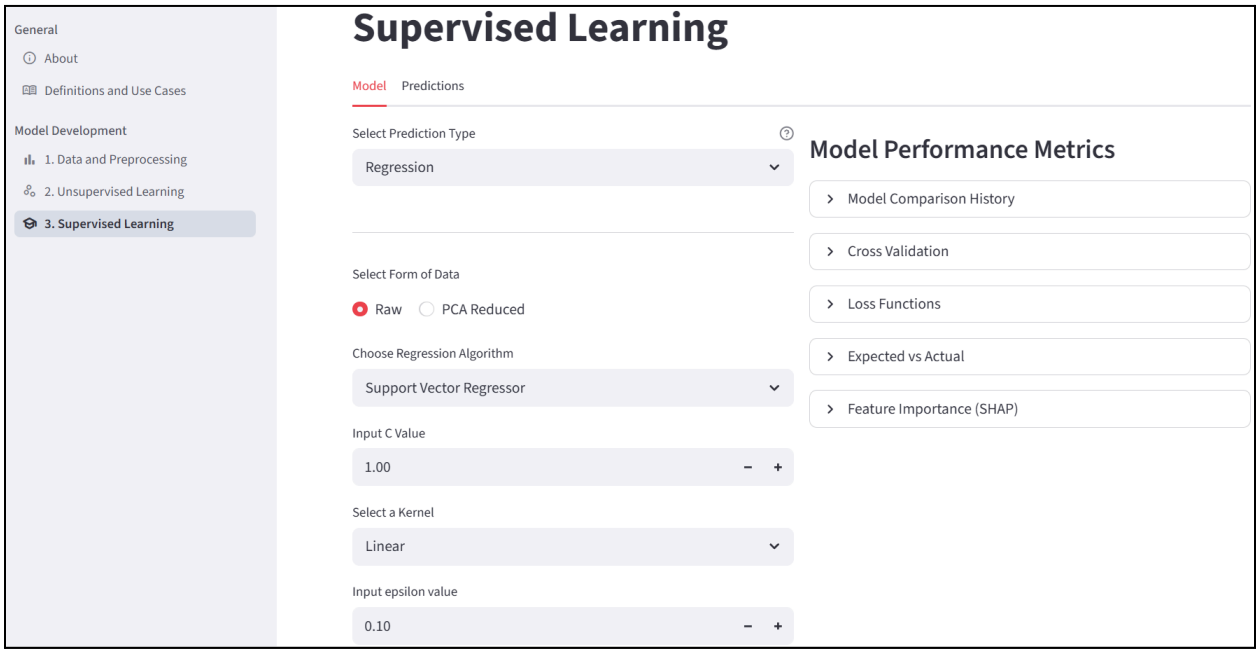
Lesson 7: How to Construct and Evaluate a Supervised Learning Model to Make Predictions from a Dataset

Goal: Using the application, the teacher will show and have the students follow along how to build a supervised machine learning model to make predictions.

Learning Objectives

- Understand the difference between classification and regression algorithms
- Examine the computational difference between linear and non-linear methods
- Understand how to evaluate both regression and classification models

Resources: https://vt-coga-ai2-astrobiology.hf.space/sup_learn



Part 3: Create and Test Own Machine Learning Model

Lesson 8: Choose Your Own Adventure

Goal: Enable the student to build their own machine learning model using data that they acquire themselves

Learning Objectives

- Know how to properly identify the correct pre-processing, unsupervised learning, and supervised learning model selection and evaluation
- Present dataset and machine learning model clearly to class