VINCI Open Source Multi-Persona POC

# The Challenge:

How can three programmers use their language of choice to collaborate on specialized aspects of the same business problem?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Language | Benefits | Editor | Key Technologies for Interfacing | User |
| SAS | A vast array of vetted statistical procedures with proven accuracy and support | SAS Viya | GitHub, CAS | Manuel Figallo |
| Python | A community-based repository of ML and AI libraries making use of the cutting edge of research | Microsoft Visual Studio Code | SWAT | Eli Lovelace |
| R | A beautiful set of statistics focused graphics and functions | RStudio | SWAT | John Cashy |

# Business Problem:

The Sashelp.Iris data set (Fisher 1936[[1]](#footnote-1)) is widely used for examples of discriminant analysis and cluster analysis. The data are measurements in millimeters of the sepal length, sepal width, petal length, and petal width of 50 iris specimens from each of three species: *Iris setosa, I. versicolor,* and *I. virginica*.

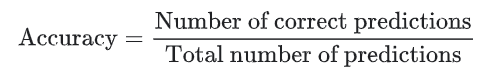
In this data we will be predicting the classes of the flowers based on these parameters. The data consists of continuous numeric values which describe the dimensions of the respective features. We will be training the model based on these features, and fulfill the following requirements of a robust ML workflow

1. **Understand** and define the problem
2. **Prepare** the data
3. **Explore** **and Analyze** the data
4. **Apply** the algorithms
5. **Reduce** the errors
6. **Predict** the result
7. **Understand** and define the problem

## **Understand** and define the problem

In this data we will be predicting the classes of the flowers based on these parameters – thus, we will aim to determine a champion model that will give us the highest accuracy in classifying Iris species in our set. We will use the following defition of Accuracy:

The number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives



## **Prepare** the data

**Python [sklearn, pandas, swat]**

The python user will be responsible for initiliazing the data of the project by connecting to the sklearn package and fetching the Iris dataset. They will add column names to the data as well, and then upload to the shared server in Viya

**SAS [DATA STEP]**

The SAS user can view the data. However – they notice that the target variable isn’t very clear. The SAS user will clean that up and push a new formatted table to the server.

## **Explore and Analyze** the data

**R [swat, ggplot2, Ggaly, gridextra, Rcpp]**

The R User will pull the formatted data and run EDA to investigate the data.

**Python [swat, pandas]**

The python User might have their own ideas of what data investigation looks like.

## **Apply** the algorithms

1. Fisher, R. A. (1936). “The Use of Multiple Measurements in Taxonomic Problems.” Annals of Eugenics 7:179–188. [↑](#footnote-ref-1)