## Project 2: Supervised Learning

Building a Student Intervention System

## Classification vs. Regression

**Your goal is to identify students who might need early intervention - which type of supervised machine learning problem is this, classification or regression? Why?**

This problem is categorized as classification. The deciding factor to decide this is that the prediction result will be discrete (‘yes’ or ’no’).

## Exploring the Data

**Can you find out the following facts about the dataset?**

* Total number of students: 395
* Number of students who passed: 265
* Number of students who failed: 130
* Graduation rate of the class (%): 67.09 %
* Number of features (excluding the label/target column): 31

## Preparing the Data

## Training and Evaluating Models

**Choose 3 supervised learning models that are available in scikit-learn, and appropriate for this problem.**

Decision Tree, SVM and Naive Bayes.

**What are the general applications of each model? What are their strengths and weaknesses?**

* Decision Tree:
* SVM:
* Naïve Bayes:

**Given what you know about the data so far, why did you choose this model to apply?**

* Decision Tree:
* SVM:
* Naïve Bayes:

**Fit this model to the training data, try to predict labels (for both training and test sets), and measure the F1 score. Repeat this process with different training set sizes (100, 200, 300), keeping test set constant.**

**Produce a table showing training time, prediction time, F1 score on training set and F1 score on test set, for each training set size.**

* Decision Tree:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training set size | | |
| 100 | 200 | 300 |
| Training time (s) | 0.002 | 0.001 | 0.003 |
| Prediction time (s) | 0.000 | 0.000 | 0.000 |
| F1 score for training set | 1 | 1 | 1 |
| F1 score for test set | 0.66 | 0.75 | 0.694 |

* SVM:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training set size | | |
| 100 | 200 | 300 |
| Training time (s) | 0.002 | 0.004 | 0.009 |
| Prediction time (s) | 0.001 | 0.003 | 0.006 |
| F1 score for training set | 0.842 | 0.875 | 0.874 |
| F1 score for test set | 0.818 | 0.784 | 0.786 |

* Naïve Bayes:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training set size | | |
| 100 | 200 | 300 |
| Training time (s) | 0.001 | 0.001 | 0.001 |
| Prediction time (s) | 0.001 | 0.000 | 0.000 |
| F1 score for training set | 0.581 | 0.812 | 0.783 |
| F1 score for test set | 0.297 | 0.677 | 0.713 |

## Choosing the Best Model

**Based on the experiments you performed earlier, in 2-3 paragraphs explain to the board of supervisors what single model you choose as the best model. Which model has the best test F1 score and time efficiency? Which model is generally the most appropriate based on the available data, limited resources, cost, and performance? Please directly compare and contrast the numerical values recorded to make your case.**

**In 1-3 paragraphs explain to the board of supervisors in layman’s terms how the final model chosen is supposed to work (for example if you chose a decision tree or support vector machine, how does it learn to make a prediction).**

**Fine-tune the model. What’s the model final F1 score?**