## 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: This model is usually used in supervised learning for classification problems. Decision trees are are simple to understand and explain, require little data preparation and are able to handle numerical and also categorical data. It’s weakness lie in creating over-complex trees. This occurs mostly from overfitting training data or trying to solve non-linearly separable problem such as XOR.
* SVM: This model is used in supervised learning for classification and regression problems. SVM is mathematically complex but by using the kernel trick non-linear problems can be transformed into linear problems. It works very well with smaller datasets with a small number of features giving high accuracy but is computationally intensive when training and predicting.
* Naïve Bayes: This model is used in supervised learning for classification. Naïve Bayes highly scalable but can be used with a small amount of training data. It is also computationally light-weight. Despite that, this model assumes that the features a independent so it doesn’t assume correlations between said features.

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

All chosen models benefit from the data size. Despite a large number of features…

**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.**

Based on the simple tests made in this project the model with the highest F1 score is SVM. But SVM is the model with the lowest time efficiency, this the probably due to the large number of features in the dataset. Considering limited resources, cost and performance I chose the decision tree model.

The decision tree model gives a perfect F1 score in the training test. This score probably comes from overfitting the tree to the training data. So, when this model is optimized the F1 score will increase considerably regarding the test data. Also, decision trees have a better performance Naïve Bayes when working with small datasets.

**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).**

Decision trees work by asking questions to the data and separating the answers in branches of a tree. Each level of the tree answer the question raised to a certain feature. In the process of training the tree will be created and at the end of each branch it’ll have a final answer regarding the feature we want to predict. The prediction is made by locating a branch that answer the questions in a similar way to the subject you want to test.

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

0.753846153846