1. 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?
Classification. Because it predicts binary label, which are yes and no.

2. Exploring the Data

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
30

4. Training and Evaluating Models

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

I choose Neighbours-Based, SVM, and Decision Trees models.

Neighbors-Based

What are the general applications of this model?

Neighbours-Bases classification is a instance-based learning. Which does not do learning but just store the training data. When classify it will compute k point that is near to input and vote for major label.

What are its strengths and weaknesses?

Strengths is it is very quick to learn and easy for adding new point because no need to train model.

Weakness is prediction time is slow because it have to search the nearest point and if there are more and more training data it will get slower and slower.

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

I choose Neighbours-Based classification because data point in this project is not large(395 students).

Fit model to the training data

Training set size	Training Time (secs)	F1 score on training set	F1 score on test set
100	0.001	0.825174825175	0.758620689655
200	0.001	0.809688581315	0.785714285714
300	0.002	0.85393258427	0.813793103448

SVM

What are the general applications of this model?

SVM model is a represent of data point in space that can separate class label as wide as possible. When predicting data it map to the space and predict label that belongs to side of class label.

What are its strengths and weaknesses?

The strengths are effective in high dimension spaces, memory efficient and it can use kernel function.

The weaknesses are if number of features is much greater that number of sample it is likely to give poor performance. And SVMs are expensive operation for large dataset.

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

I choose SVM because in this dataset there are 48 features after preprocess, which is considerably high compare to number of training set (395).

Fit SVM classification model to the training data

Training set size	Training Time	F1 score on training set	F1 score on test set
100	0.002	0.835443037975	0.802547770701
200	0.004	0.843137254902	0.81045751634
300	0.010	0.866379310345	0.805194805195

Decision Trees

What are the general applications of this model?

It is a graph-like model called decision tree. One node of a tree is a decision to go to one branch. And go deeper until reach leaf node which is output.

What are its strengths and weaknesses?

Strengths are simple to understand the model because model can be interpret.

Weakness are it is easy to get overfitted, and it is difficult to create optimal decision tree model.

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

I choose decision tree model because it is quite accurate and it is white box model which can be helpful when analyse how to deal with student.

Fit this model to the training data

Training set size	Training Time	F1 score on training set	F1 score on test set
100	0.001	1.0	0.677165354331
200	0.001	1.0	0.746031746032
300	0.002	1.0	0.704

5. Choosing the Best Model

Based on the experiments, SVM's F1-score are 0.80, 0.81, 0.80 in data size 100, 200, 300 and they are above 0.80 in both training set and test set in all training data size.

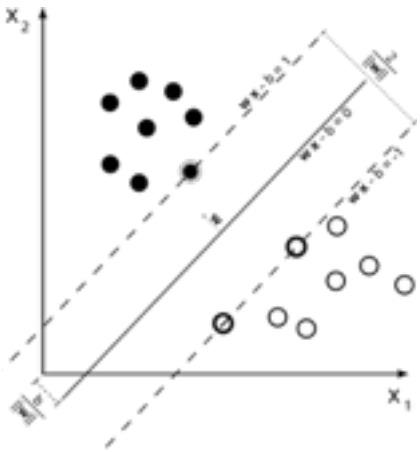
Neighbours-based model's F1-score is very near to SVM's when data size is 300 (0.81) but lower in 100 and 200 (0.75 and 0.78) since the data has high dimension (48 features after preprocess) and training data may not large enough to train neighbours-based model.

Decision Tree's F1-score are 1.0 on training set but has low F1-score on test set (0.68,0.75 and 0.70)

SVM use more training time (0.010 s) than neighbours-based (0.002 s) and decision tree (0.002 s) and SVM tends to increasing rapidly as training set size increase. But training data is not large and I assume that to help student we should focus on correct prediction than time performance so choosing SVM from others is the best model.

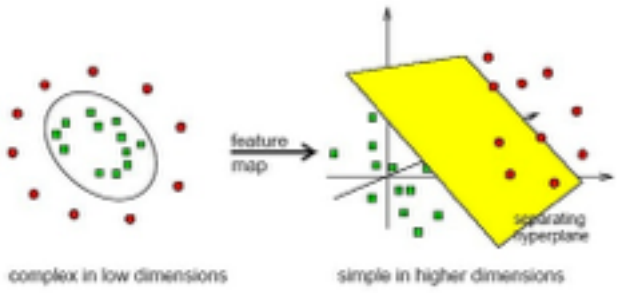
Compare from F1-score and training time, I choose SVM as best model.

SVM or Support Vector Machine create support vectors from representation data point in class, which is "passed" class and "failed" class in this project.The point are selected and make linear lines for each class's side. Those 2 lines are parallel and optimised to have as wide as possible gap while it still separating classes. When predict, point that fall on the side of centre line will be classified to be class in that side.



- https://en.wikipedia.org/wiki/Support_vector_machine

If data cannot be separated by line, we could apply function to map all point to higher dimension that can be linearly separable(see picture below). This method is called kernel trick.



- https://qph.is.quoracdn.net/main-qimg-ed058210b222061200a0a86ceb324b48

I use RBF kernel and fine tune C and gamma using gridSearchCV the model's. GridSearchCV function takes long time to process, so I change range and search from wide range and narrow in down to small range to find the best parameter in shorter time. Final F1 score is 0.825806451613.