PROJECT 2

1) 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 is a classification problem. In this problem we determine whether a student needs early intervention or not, so we are classifying the students into yes/no categories.

2) Important characteristics must include

   Total number of students: 395

   Number of students who passed: 265

   Number of students who failed: 130

   Number of features: 30

   Graduation rate of the class: 67.09%

3) What are the general applications of the models? What are its strengths and weaknesses?

   i) SVM
      a. Applications
         i. Anomaly detection (outlier analysis)
         ii. Dimensionality reduction
         iii. Classification of images, character recognition
      b. Strengths
         i. Can perform non linear classification
      c. Weaknesses
         i. Parameters of a solved model are difficult to interpret

   ii) Decision Tress
      a. Applications
         i. Decision analysis to identify strategies
         ii. Astronomy (star galaxy classification)
      b. Strengths
         i. Easy to interpret
         ii. Allows the addition of new scenarios
      c. Weaknesses
         i. Tend to overfit the training data

   iii) Naïve Bayes
      a. Applications
         i. Medical Diagnosis
         ii. Text Categorisation (spam filtering)
      b. Strengths
         i. Highly scalable
         ii. Requires small amount of data for estimating parameters for classification

c. Weaknesses

i. Oversimplified assumptions makes a naïve model

4) Given what you know about the data so far, why did you choose the models to apply?

i) SVM
Since SVMs divide the given examples in such a way that there is a clear gap (as wide as possible) between the closest 2 examples of different classes, when more data points turn up, it is less likely to be misclassified.

ii) Decision Trees
DT makes it much easier to understand how the algorithm arrived at that particular step. Also, for every scenario we may get to know what the result would be since DTs explore every option / path.

iii) Naïve Bayes
The prediction works on probability distributions and is used widely for dividing the data into 2 classes, as in the case of spam filtering or document classification.
[Reference: http://scikit-learn.org/stable/modules/naive_bayes.html]

5) Based on the experiments you performed earlier, in 1-2 paragraphs explain to the board of supervisors what single model you chose as the best model. Which model is generally the most appropriate based on the available data, limited resources, cost, and performance?

The algorithms used were SVM, Decision Trees and Naïve Bayes. The results are as as recorded below:

CLASSIFIER: SVM

	Test_f1	**Test_pred_time**	**Train_Size**	**Train_Time**	**Train_f1**	**Train_pred_time**
0	0.774648	0	100	0	0.877698	0
1	0.774648	0	200	0	0.841379	0
2	0.783784	0	300	0	0.876068	0

CLASSIFIER: DECISION TREE

	Test_f1	**Test_pred_time**	**Train_Size**	**Train_Time**	**Train_f1**	**Train_pred_time**
0	0.643478	0	100	0.000	1.000000	0

1	0.643478	0	200	0.000	0.848485	0
2	0.615385	0	300	0.016	1.000000	0

CLASSIFIER: NAIVE BAYES

	Test_f1	Test_pred_time	Train_Size	Train_Time	Train_f1	Train_pred_time
0	0.802920	0	100	0	0.846715	0
1	0.802920	0	200	0	0.829787	0
2	0.763359	0	300	0	0.803783	0

Firstly, comparing the time taken to predict on testing and training set, almost all the algorithms do it in no time, except when Decision Trees trains with the whole set.

Secondly, If we only compare the f1 score, DT seems to outperform the other algorithms while considering only testing set scores. But when we compare the scores for test sets, DT performs rather poorly because it overfits the training data very well, but fails to do it for new cases.

That leaves us with SVM and Naïve Bayes. Comparing their training set f1 scores, SVM seems to do better but isn't as good when it comes to classifying new sets. Keeping in mind the variance-bias trade-off, I preferred using NB as it classifies new examples (test set cases) better.

Thus, comparing all the training and test set f1 scores, I chose NB for this problem.

6) In 1-2 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 make a prediction).

Naïve Bayes, first the probabilities for both the classes for the 'passed' target column will be calculated (also called prior probability). Next the probability of the feature columns are calculated seeing the data like age, sex, family size, using the internet or not, etc, given the student passes or fails. For eg, probability of a student being female knowing that they passed, or probability of a student having a job knowing that they passed.

Then in the case of classifying training and testing data, all the probabilities that a particular feature that could affect the classification is calculated, also called as conditional probability or posterior probability using the probabilities calculated earlier. For eg, what is the probability of the student passing, knowing that the student is a female or the student has a job.

This can be expressed as: Given a class variable y and a dependent feature vector x_1 through x_n, Bayes' theorem states the following relationship:

$$P(y \mid x_1, \dots, x_n) = \frac{P(y) P(x_1, \dots x_n \mid y)}{P(x_1, \dots, x_n)}$$

Now this model can be deployed in the real world where we calculate the prior and the posterior probability to determine a class for the student.

7) What is the model's final F_1 score?

The final scores:

```
Test set f1 score:  0.763358778626
Training set f1 score:  0.80378250591
```