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?

To identify students who might need early intervention we will use classification type of supervised learning

In regression output is in continuous form while in classification output is in discrete form and here we need to know whether student need early intervention or not so output will be discrete.

2. Exploring the Data

Can you find out the following facts about the dataset?

1. **Total number of students:** 395
2. **Number of students who passed:** 265
3. **Number of students who failed:** 130
4. **Number of features:** 30
5. **Graduation rate of the class:** 67.09%

3. Training and Evaluating Models

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

- What are the general applications of this model? What are its strengths and weaknesses?

Model	General Application	Strengths	Weaknesses
Support Vector Machine	1.) Used in Face recognition 2.) Bioinformatics 3.) Protein Secondary Structure Prediction 4.) Signal Processing	1.) Works really well in complicated domains where there is clear margin of separation 2.) Using Kernel can project data to higher dimensions to perform Linear separation	1.) Don't Perform well in very large data sets, because the training time happens to be cubic in the size of data sets 2.) Don't work well with noise
Decision tree Classifier	1.) Used in Astronomy for Star galaxy classification. 2.) Used in Biomedical Engineering	1.) Really easy to use and they are beautiful to grow on 2.) Allow data to interpret data really well 3.) Can build bigger classifier out of decision tree in something called ensemble methods	1.) Prone to Overfitting 2.) We need to set min sample split to ensure it doesn't get over fit

Gaussian Naive Bayes	1.) To mark an email spam or not 2.) To determine whether a text shows positive or negative emotions	1.) Really easy to implement 2.) Works pretty good with large feature set. 3.) Highly efficient	1.) Sometimes when words in phrases have different meaning it doesn't take the whole phrase in account rather search for individual word 2.) Assumes independence of features

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

	Reason to Choose this Model
Support Vector Machine	Students are classified in two classes either pass or fail and therefore can easily be separated by fine tuning the parameters of an SVM and also data set is small
Decision Tree Classifier	It allows data to interpret really well and are simple to grow on and as are labels are only yes or no it will be easy to interpret
Gaussian Naive Bayes	It uses prior probability and test evidences to calculate posterior probability so in this data set it takes all the features and calculates the probability of a student whether it need early intervention or not

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

Support Vector Machine

	Training set size		
	100	200	300
Training time (secs)	0.008	0.010	0.047
Prediction time (secs)	0000	0.001	0.001
F1 score for training set	0.88059	0.86219	0.84210
F1 score for test set	0.746269	0.764705	0.7826

Decision Tree Classifier

	Training set size		
	100	200	300
Training time (secs)	0.001	0.001	0.003
Prediction time (secs)	0.000	0.000	0.000
F1 score for training set	1.0	1.0	1.0
F1 score for test set	0.699	0.7286	0.7

Gaussian Naive Bayes

	Training set size		
	100	200	300
Training time (secs)	0.000	0.001	0.001
Prediction time (secs)	0.000	0.000	0.000
F1 score for training set	0.8549	0.8320	0.8088
F1 score for test set	0.7480	0.7131	0.75

5. 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.

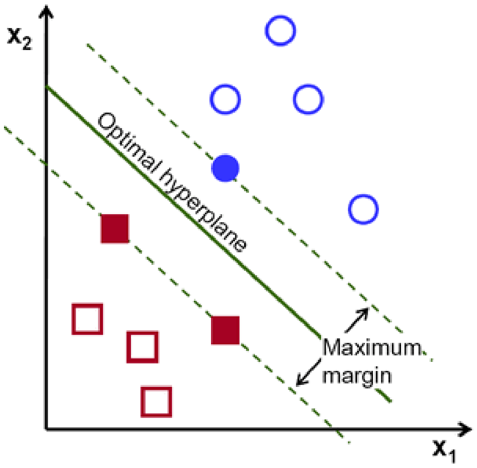
Based on the experiment performed I think Support vector machine is the best choice for our model. As it has the best F1 score in all the three. It's the only model whose F1 score increases with increase in data. Training time may be more than other two but space and prediction time is constant as it separates the students in two halves and make prediction based upon the relative positon of the student from decision boundary.

SVM also gives parameter to control whether we want a smoother boundary or we want more training points to be classified correctly

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).

I Choose Support Vector Machine because it predicts in constant time and can easily controls the trade-off between a simple model or a complex model which classifies more point correctly.

In this case we need to identify whether a student need early intervention or not. SVM uses all the past data (feature) to separate students into different classes in this case pass and fail and whenever a new data is given to the machine it uses it's training to predict whether the student need early intervention or not. SVM is particularly good for small datasets.



As Shown in the figure SVM creates a decision surface in this case a line (optimal hyperplane) to separates data into different classes (pass or fail). Decision surface is selected such that its distance (margins) from both classes is maximum.

Fine-tune the model. Use gridsearch with at least one important parameter tuned and with at least 3 settings. Use the entire training set for this.

I used Grid Search to fine tune SVM on kernel and C and gamma while keeping scoring to f1_score

What is the model's final F1 score?

```
F1 score for train set: 0.971698113208
```

```
F1 score for test set: 0.791946308725
```