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?

It is a classifcaiton - the process of taking input and mapping the discrete label weather it is true or false, where regession is continuous function.

2. 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
Number of features: 30
Graduation rate of the class: 67.00%
```

Use the code block provided in the template to compute these values.

3. Preparing the Data

Execute the following steps to prepare the data for modeling, training and testing:

- Identify feature and target columns
- Preprocess feature columns
- Split data into training and test sets

```
Feature column(s):- ['school', 'sex', 'age', 'address', 'famsize', '
Pstatus', 'Medu', 'Fedu', 'Mjob', 'Fjob', 'reason', 'guardian', 'tra
veltime', 'studytime', 'failures', 'schoolsup', 'famsup', 'paid', 'a
ctivities', 'nursery', 'higher', 'internet', 'romantic', 'famrel', '
freetime', 'goout', 'Dalc', 'Walc', 'health', 'absences'] Target col
umn: passed
```

4. 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?

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

Gaussian Navie Bayes is a simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features.

The reason using naïve bases is because it is easy to describe and calculate using ratios and also gives robust results.

GaussianNB

Training set size	Training Time	Prediction time	F1 score (training)	F1 score (test)
100	0.001	0.000	0.4	0.396
200	0.001	0.000	0.811	0.75
300	0.001	0.000	0.786	0.791

Random forest operates by constructing a multitude of decision tress at training time and outputting the class that is mode of the classes or mean prediction of the individual tree.

The reason using that, because it is invariant under scaling and various other transformations of features values, is robust to inclusion of irrelevant features, and produces inspectable models.

RandomForestClassifier

Training set size	Training Time	Prediction time	F1 score (training)	F1 score (test)
100	0.006	0.001	0.976	0.744
200	0.007	0.001	0.992	0.721
300	0.010	0.001	0.997	0.785

Support vector machine analyze data and recognize patterns used for classification and regression analysis. Also it is a non-probabilistic binary linear classifier.

I picked SVM, because it is separate categories are divided by a clear gap. Also it can efficiently perform a non-linear classification.

SVC

Training set size	Training Time	Prediction time	F1 score (training)	F1 score (test)
100	0.001	0.001	0.8823	0.7746
200	0.001	0.003	0.8814	0.7839
300	0.001	0.007	0.7855	0.8460

5. Choosing the Best Model

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?

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).

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.

What is the model's final F1 score?