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: 67.09%
```

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

Gaussian Naive Bayes is a simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. A disadvantage of Naive-Bayes is that if you have no occurrences of a class label and a certain attribute value together (e.g. class="nice", shape="sphere") then the frequency-based probability estimate will be zero. Given Naive-Bayes' conditional independence assumption, when all the probabilities are multiplied you will get zero and this will affect the posterior probability estimate.

The strength of Naïve Bayes can explicitly handles class priors, “normalizes” across observations: outliers comparable, and assumption: all dependence is “explained” by class label. Naïve Bayes can be use greatly on spam detections or hand writing input recognition.

The reason using naïve bases is because it is easy to describe and calculate using ratios, also gives robust results. By its quick classification, that saves up lots of computation time.

GaussianNB

Training set size	Training Time	Prediction time	F1 score (training)	F1 score (test)
100	0.001	0.001	0.8467	0.8029
200	0.001	0.001	0.8406	0.7244
300	0.001	0.001	0.8038	0.7634

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.

Pros: an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

Cons: For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable for this type of data. Methods such as partial permutations were used to solve the problem.

A good use of random forest was tuned with a large dataset, and validated by an independent data set. In Italy, an ecological modeling are based on this algorithm.

RandomForestClassifier

Training set size	Training Time	Prediction time	F1 score (training)	F1 score (test)
100	0.020	0.001	0.9846	0.6667
200	0.062	0.001	0.9928	0.8028
300	0.028	0.001	0.9831	0.7205

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.

Pros: The key features of SVMs are the use of kernels, the absence of local minima, the sparseness of the solution and the capacity control obtained by optimizing the margin.
Cons: "Besides the advantages of SVMs - from a practical point of view - they have some drawbacks. An important practical question that is not entirely solved, is the selection of the kernel function parameters - for Gaussian kernels the width parameter [sigma] - and the value of [epsilon] in the [epsilon]-insensitive loss function...[more]

SVM can be widely use on large datasets and suitable for scaling. Typically good application can be adopt in the solution of consulting and industrial data analysis problem.

SVM

Training set size	Training Time	Prediction time	F1 score (training)	F1 score (test)
100	0.002	0.002	0.8777	0.7746
200	0.010	0.004	0.8879	0.7815
300	0.010	0.009	0.8761	0.7838

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?

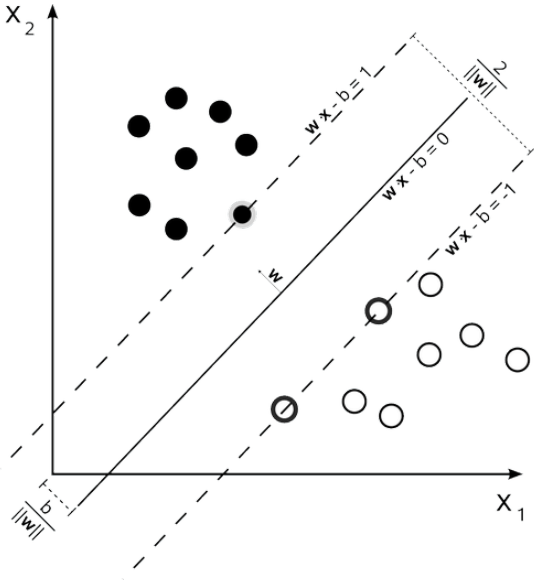
Though Support Vector Machine (SVM) does not performs as good as Naïve Base with a much higher training time and prediction time, yet it produces a better F1 score. Therefore, the cost of the additional time compare with the result is justified. Not to mention, the available size of the data, unbalanced data, and high graduation rate of 95% would require a higher scores.

Not to mention SVM is great on high dimensional space and high performance on its memory efficiency.
In contrast to what we have earlier on the SVM without fine-tuned

The F1 score from the grid search performs slightly better than the default one, also slightly faster.

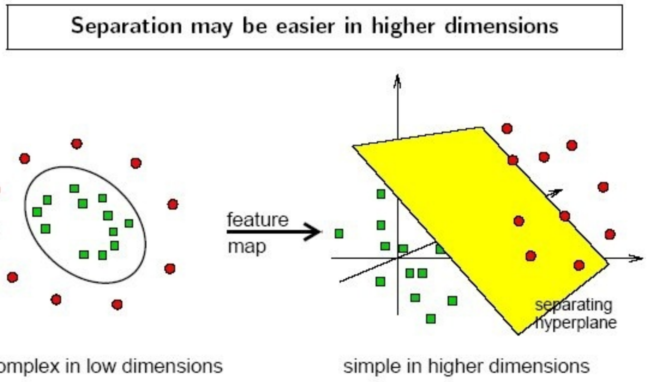
	Training set size	Training Time	Prediction time	F1 score (training)	F1 score (test)
SVM	300	0.010	0.009	0.8761	0.7838
SVM w/ grid search	300	0.008	0.007	0.9762	0.7895

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).



SVM classifier separates students who pass and did not pass by a line. The model determines weather the new student is able to graduate or not, thus intervention will be needed. SVM searches for the closest point where the best line depends or supported by the closest points.

The SVM can also translate the students' data into higher dimensions in order to create a more robust boundary. In a multidimensional problem, The graph on the right shows how higher dimensions able to separate on hyperplane better than 2d low dimensions. SVM is going to surface the different features instead on a single line. Therefore, though that the best outcome comes from maximizes distance between different points.


Separation may be easier in higher dimensions
feature
map
separating
hyperplane
complex in low dimensions
simple in higher dimensions

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

C = 100, size =200, degree = 3, gamma = 0.1, kernel = ‘rbf’.

What is the model’s final F1 score?

The F1 final score is 0.7947

Reference :

https://www.quora.com/What-does-support-vector-machine-SVM-mean-in-laymans-terms
https://www.reddit.com/r/Machin1eLearning/comments/15zrpp/please_explain_support_vector_machines_svm_like_i
https://www.cs.cmu.edu/~ggordon/SVMs/new-svms-and-kernels.pdf
https://www.quora.com/What-is-AdaBoost