You now have a fairly substantial starting toolbox of supervised learning methods that you can use to tackle a host of exciting problems. To make sure all of these ideas are organized in your mind, please go through the list of problems below. For each, identify which supervised learning method(s) would be best for addressing that particular problem. Explain your reasoning and discuss your answers with your mentor.

1. Predict the running times of prospective Olympic sprinters using data from the last 20 Olympics.

1.     SVR

2. Linear Regression

3.Random Forest Regression

4. KNN Regression

All 4 of these models are capable of predicting continuous values (our target will be continuous) and should be tested and tweaked to compare accuracies, and to see which one demonstrates the best performance on this dataset.

2.     You have more features (columns) than rows in your dataset.

1.     Lasso (setting lambda to reduce # of predictors)

2.     Random Forest (we can keep track of feature importance with one of the module attributes)

3.     GBM (we can keep track of feature importance with one of the module attributes)

4.     Feature engineering (PCA, Partial Least Squares Regression)

3.     Identify the most important characteristic predicting likelihood of being jailed before age 20.

1.     Random Forest (we can keep track of feature importance with one of the module attributes)

2.     GBM (we can keep track of feature importance with one of the module attributes)

3.     Feature engineering (PCA, Partial Least Squares Regression)

4.     Implement a filter to “highlight” emails that might be important to the recipient

1.     Naïve Bayes – conditional probability is intuitive, the model is quick to set up, and is known for performance on text classification problems

2. SVC – performs well on high-dimensional datasets like text classifiers where much of the data is linearly separable

5.     You have 1000+ features.

1.     Lasso (setting lambda to reduce # of predictors)

2.     Random Forest (we can keep track of feature importance with one of the module attributes)

3.     GBM (we can keep track of feature importance with one of the module attributes)

4.     Feature engineering (PCA, Partial Least Squares Regression)

6.     Predict whether someone who adds items to their cart on a website will purchase the items.

1.     Naïve Bayes (conditional probability)

2.     Logistic Regression

3.     SVC

4.     Random Forest

5.     KNN Classification

Approaches 2-5 are included because they can be configured to give a binary output, and all should be tested (and cross-validated) to determine which model works best on the given data.

7.     Your dataset dimensions are 982400 x 500

1.     Could use Lasso to reduce the feature space

2.     Random Forest (bagging / sub-sampling)

3.     Gradient Boosting (sub-sampling)

8.     Identify faces in an image.

1.     Random Forest (do we detect two black circles, are they a certain distance apart, do we detect a ‘nose’, mouth, jaw-line)

2.     SVM

9.     Predict which of three flavors of ice cream will be most popular with boys vs girls.

1.     SVC

2.     Random Forest Classification

3.     KNN classification

4.     Some advanced boosting model combining all 3

All of these approaches can classify into multiple groups (not just binary yes/no predictions) and all should be tested and cross-validated to determine which model works best with the dataset.