ANMOL SINGH (2013021), SATYAM KUMAR (2014096)

MINOR PROJECT I – PROPOSAL

- Dataset: http://archive.ics.uci.edu/ml/datasets/Student+Performance

- Problem Statement: The goal of this project is to predict the academic performance of students in two Portuguese high schools. The input dataset is a mix of about 30 numeric, binary (yes-no) and nominal (few fixed classes) attributes. A few examples are - number of past class failures, extra paid classes (binary: yes or no), free time after school. A full description of the attributes will be provided in the report.

  We will predict the final grade (G3) of a student as a numeric score between 0-20 corresponding to the 20 point grading scale used in Portugal.

- Learning Techniques:
  1. Linear Regression with regularisation (Ridge and LASSO)
  2. Random Forests

- Strategy For Model Selection and tuning hyper parameters: We will test various models (linear, polynomial, gaussian, etc.) along with regularisation and use cross-validation to select the one that performs best. We will be using k-fold cross-validation where 'k' will be decided later .

- Training approaches to be explored: We intend to try out both batch and stochastic gradient descent algorithms for linear regression. The data will be normalised and in general we will use gradient descent, however, Newton's method will be preferred for polynomial kernels of degree greater than 1.

- Ensemble approaches: Our second learning technique (random forests) is an ensemble approach for decision trees.

- Evaluation metrics: The dataset includes the true values of the final grade for each student. Thus, our primary evaluation metric will be the Root Mean Square Error (RMSE).