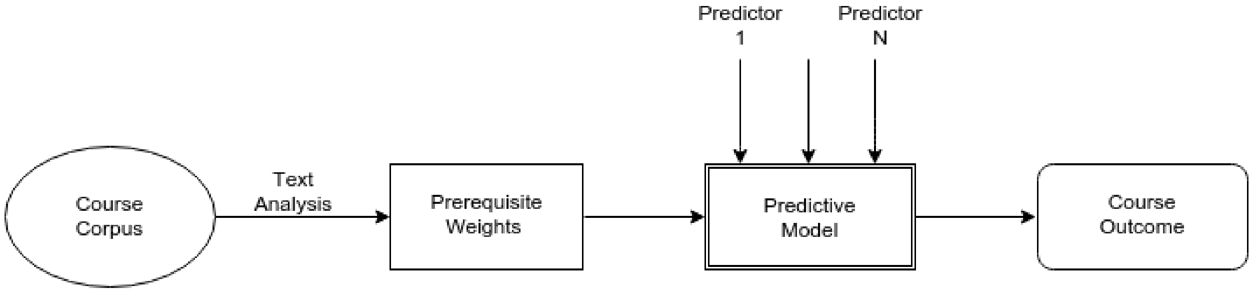
Proposed System

The proposed system is a predictive model that takes into account variables such as performance in previous tests such as marks obtained in each internal test as well as demographic variables such as health of the student and habits(both good and bad) of the student as predictors. We also plan to add weights based on course dependency to improve the system(as shown in the figure), however this integration is a future enhancement we look to add once we get an opportunity to work on data with respect to PES University students.

The system finally looks to predict the grade the student is most likely to obtain using each of the predictors.

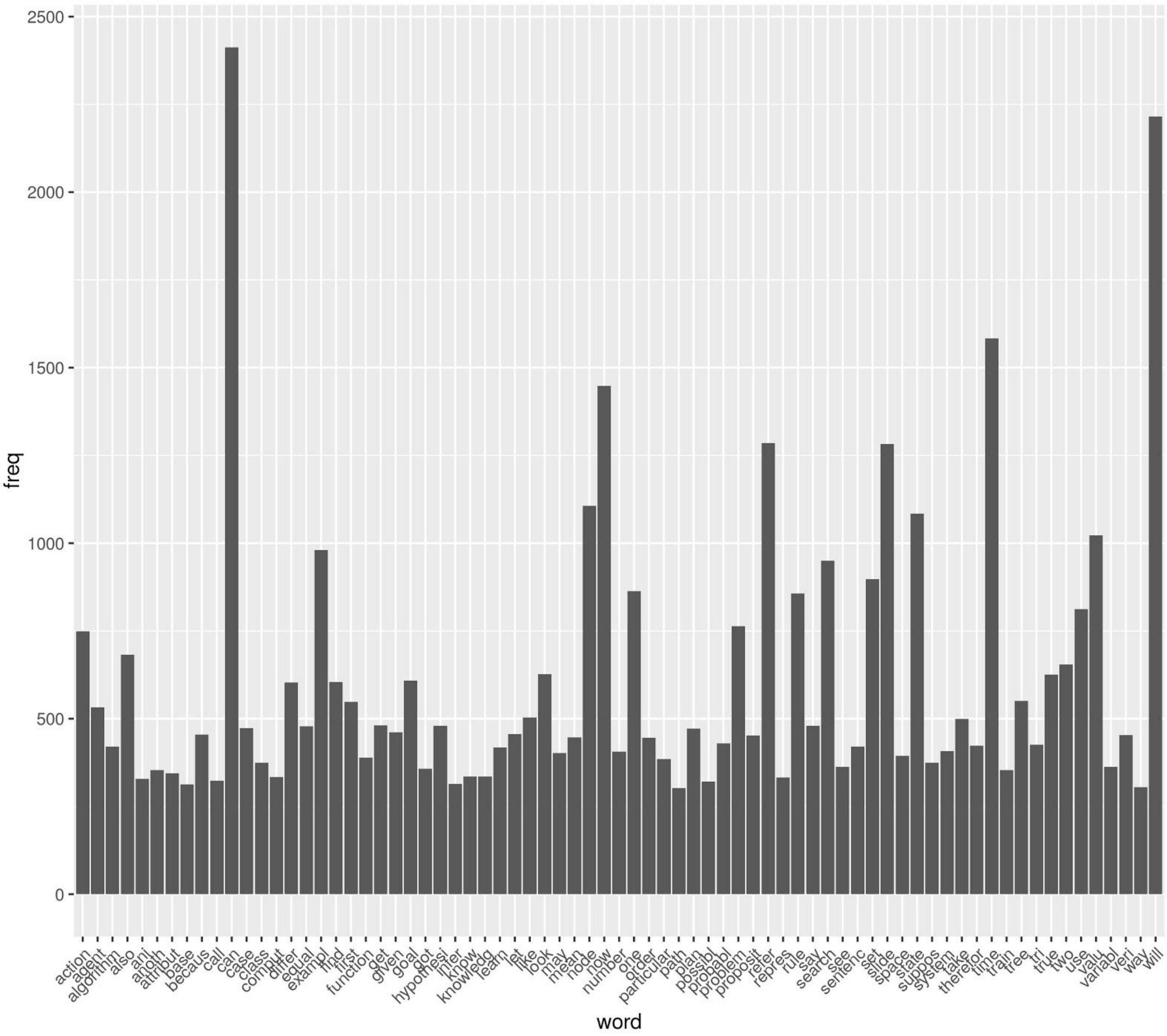
We have taken two distinct predictive models to build the system. However,before diving into each of these models that are very important components of the system, we would like to first illustrate the basic workflow of the system. [Figure 1]



Components of the system

The first component we would like to explain is that of extraction of prerequisite weights given a course corpus using text analysis.
The first step in this section was to web scrape NPTEL course transcripts from the NPTEL website. A python script was written through which transcripts of each of the courses was obtained. We then used R and applied text analysis concepts such as stop-word removal,stemming and term document matrix on each of the transcripts.
The next step involved defining a threshold N(say 100) and extracting all the words that occur greater than or equal to the value of the threshold defined.

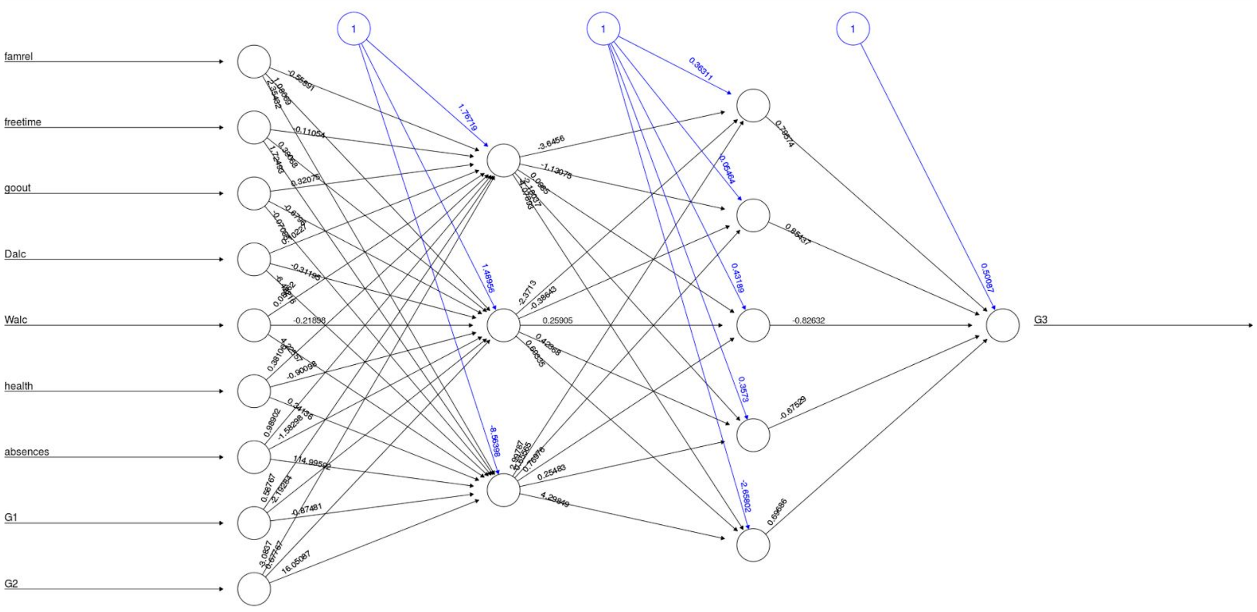
This was done for each course and its corresponding pre requisites and set intersection was applied to quantify percentage of pre requisite in each course. Visualisations for term frequency for the courses Artificial Intelligence is shown in FIGURE X. However due to unavailability of PES University student performance data we were unable to integrate the findings to our model. Thus this component is a future enhancement to our system. The core of the system for this semester’s course project lies in the predictive models built, namely, a linear model and a deep neural network model for predicting student performance using the UCI Exam Performance dataset. However the component involving text analysis allowed us to apply concepts of text analysis learnt in this course and the work done will allow us to improve our model once we get PES University Student Performance Data.



The second component we would like to discuss is the predictive models used to predict student performance. We built a linear model as well as a deep neural network and predicted student performance in Math and Portuguese based on the UCI Exam Performance Dataset. [Insert Reference Here]

We would like to first explain the neural network model used. The initial model was built using R. The first layer took 9 input variables to the first layer. These variables are the predictors which include previous grades (G1 and G2),the family relationship of the student(famrel), the amount of free time the student gets, the amount the student goes out,the alcohol consumption of the student(Dalc and Walc),the number of time the student was about during the course(absences) as well as the overall health of the student. The final prediction is that of predicted course outcome denoted by G3.
The neural network also has two hidden layers with 3 and 5 neurons respectively. A visualisation of the neural network built and trained with respect to the math course present in the UCI Exam Performance dataset is shown in FIGURE X.
The black lines show the connections between each layer and the weights on each connection while the blue lines show the bias term added in each step.

We further improved this model and built a deep neural network using Tensor Flow with 3 hidden layers with 10,20 and 10 neurons. The models were validated with that present in the paper “Using Data Mining to Predict Secondary School Student Performance “ by P Cortex and A Silva and exciting results were obtained which are discussed in the Experiments and Results section.



We also built a linear model as for this particular problem of prediction, a linear model will work well. The model was built with the same 9 predictors as that of the neural network with the final course performance outcome(G3) as the response variable.

Comparisons in terms of RMSE values of the models used as well as the models used in the paper [have a reference here] are discussed in the Experiments and Results section.