**Performance Prediction**

by

Atharva Wandile

**Abstract**

The aim of the project is to predict the performance of students of a particular class, given their performance in the previous years as well as the performance of the previous batch.

The method used to achieve the result is linear regression algorithm trained by gradient descent optimizer. The project further evaluates and compares different learning rates and regularization parameters to obtain the optimal solution.

**Key Words:** <List some keywords here>

**Acknowledgements**

I’d like to extend my thanks to Edumerge Pvt. Ltd. For giving me the opportunity to work as their intern and do this project. My special thanks to Mr Rajat and Mr Krishna Mohan for mentoring and guiding me through the internship.

**Content**

[Error: Reference source not found](#_Toc16465)

[Error: Reference source not found](#_Toc13268)

[Error: Reference source not found](#_Toc18264)

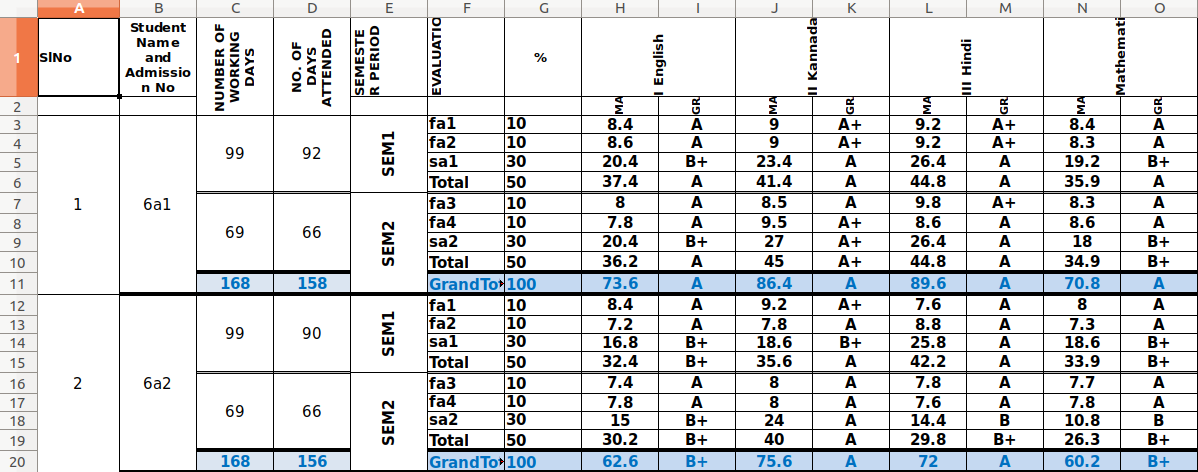
[Error: Reference source not found](#_Toc9302)

[Error: Reference source not found](#_Toc25006)

# Visualizing and Manipulating the Data

# Data Extraction

First we need to see what the given data is and convert it into a suitable form for easy manipulation and calculations. The data is given in excel format.



Pandas is used to extract the data from excel into a pandas dataframe. Pandas allows for easy manipulation of the data. Later the pandas dataframe is converted to a numpy array which is accepted for calculations.

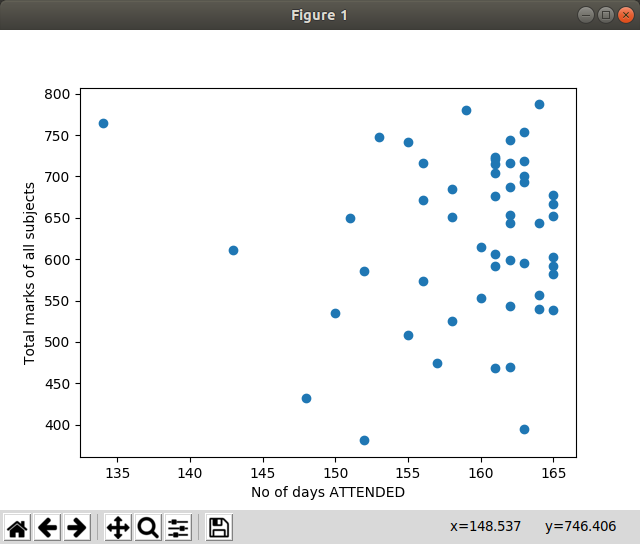
# Removing Unwanted Columns

The columns that are not required are removed. For eg. Student Id, Total, Semester, etc. have no consequence on the performance of the student.

These columns cannot be used for any analysis or prediction.

The grade column is removed as well as it is non-numeric and is based on the marks of the student which are already, given hence removing this does not have any effect on the performance of the model.

Next we check if the attendance has any correlation with the marks of the students



As seen from the plot above, the attendance has little consequence on the marks of the students. Hence the attendance is also removed.

# Removing Data Redundancy

It is important to remove redundant data to speed up the learning process and also to most importantly prevent overfitting in some cases.

The ‘total’ and ‘grand total’ for each student is the sum of all the individual tests only. It provides no new information and is therefore redundant.

Hence these rows are removed from the dataframe.

# Converting Dataframe Object into Numpy Matrix

Initially we used the pandas dataframe object to store the data as it is easy to manipulate and trasform. However for the machine learning we use tensorflow library which does not accept pandas dataframe as input. Therefore we need to convert the pandas dataframe into a numpy array. Pandas has an inbuilt function to convert the dataframe into a python matrix which is then converted to numpy array.

# Splitting Data and Feature Scaling

## Splitting Data

Now that the data has been converted into proper matrix format we need to split the data into the training and test sets.

Training set : The data that is used to train the model. The x as well as y data is given and the model trains the parameters according to the training set. This should be the main chunk of given data.

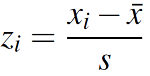
Test/Validation set : This data is used to test that accuracy of the model after it has been trained. This data is not seen by the model during training, hence it is similar to the real world data that the model needs to work on. This data is used to tune the hyperparameters of the model like learning rate, regularization parameter etc.

Before we split the data we need to make sure we are not picking the same data for training and testing every time as this can lead to bias. Hence we need to randomize the data set first. In order to randomize the data we first shuffle it and then split it into train and test sets. The extra students from each class are also dropped off at this stage ie. the students that are not present consistently in all the classes.

The splitting is done 80-20 that is 80% of the data is used for training and 20% of the data is used for testing.

## Feature Scaling

Once the data has been split we need to standardize the data. This brings all the different values of the features in the data to have mean 0 and standard deviation 1. This is important in order to achieve stable and fast convergence. Mean normalization has been used. In mean normalization we subtract each value with the mean and divide by the standard deviation.



Now the data is ready to be trained.

# Training the Data

## Declaring Variables

We first need to declare the different variables that need to be trained.

Wi  will be the weights of shape (1,no of features)

b is the bias variable which has same shape as W

Both these variables are initialized to zero at first.

## Hypothesis Function

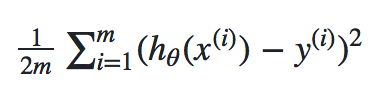
Now we need to form the hypothesis or prediction function.

The hypothesis is hθ = (Xi \*Wi) + b

This is our predicted outcome.

## Cost Function

The cost function is a measure of the error in the prediction compared to the actual values. The cost is calculated by the mean squared error.



## Training Parameters

Now that we have the hypothesis and the cost function we need to train the parameters to fit our training data and minimize the cost function.

Gradient decent is the training algorithm that has been used here.