THE FEDERAL UNIVERSITY OF TECHNOLOGY, AKURE
SCHOOL OF ENGINEERING AND ENGINEERING TECHNOLOGY
COMPUTER ENGINEERING DEPARTMENT

EEE415- SYSTEM MODELLING AND ANALYSIS

PROJECT REPORT

ON

LINEAR REGRESSION MODEL FOR 25TH AND 35TH CPE CANDIDATES.

COURSE LECTURER: **DR BUSARI SHERIFF**

DATE: **4TH DECEMBER 2022**

1. GROUP MEMBERS

S/N	MATRIC NUMBER	NAME
1	CPE/17/3117	HAMMED TAOFEEK
2	CPE/17/3087	ACHIAGA SAMSON
3	CPE/17/3097	AJIBOYE ABIODUN
4	CPE/17/3107	BOWOTO AYOMIDE
5	CPE/17/3126	OGUNDIYA DANIEL
6	CPE/17/3136	OLUSHINA AYOMIDE
7	CPE/17/3146	QOZIM IDRIS
8	CPE/18/6675	OLAOLUWA MARVELLOUS
9	CPE/16/7867	OMOSULE BABATUNDE

1. INTRODUCTION

Simple linear regression is used to estimate the relationship between **two quantitative variables**.

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Assumptions

Simple linear regression is a parametric test, meaning that we make certain assumptions about the data. These assumptions are:

1. The size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. The observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3. The data follows a normal distribution.
4. The relationship between the independent and dependent variable is linear
5. Students cannot retake failed courses until the final year
6. FUTA grading system is used for the student grade estimation as shown below

Table 1: Student Grade Computation

Grade	Point	Percentage	Student Grade (point* Course Unit)
A	5	70-100	5*3=15
B	4	60-69	4*2=8
C	3	50-59	3*3=9
D	2	45-49	2*1=2
F	0	0-44	0*3=0

Simple linear regression Model is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

where

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- $\mathbf{B_0}$ is the **intercept**, the predicted value of **y** when the **x** is 0.
- $\mathbf{B_1}$ is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

2. DATA PRE-PROCESSING TECHNIQUES

Machine learning model is as good as the quality of the dataset used for the project. Whenever a dataset is selected for machine learning task, it has to be relevant, meet the requirements and it must be readily available. The data processing techniques used on the raw dataset for this project dataset Linear regression task are briefly discussed below.

i. **Data Selection:** The modified dataset was extracted from the EEE/CPE dataset and the bottom half of the dataset were selected for CPE students.
ii. **Data Extraction**: The data for the 25^{th} and 35^{th} students were selected for regression task.
iii. **Data cleaning and Normalization:** The columns containing courses offered by EEE were removed and the remaining columns reflects courses offered by only CPE students. The empty cells were randomly filled with mean and mode score the remaining data.
iv. **Data Transformation:** There were more rows in the data which makes it unfit for our Linear regression task. The rows and columns were transposed and identify our dependent and independent variables.
v. **Data Balancing:** We added extra columns for course unit, grade points (student score point* course unit) for 25^{th} and 35^{th} students

3. TOOLS, FRAMEWORKS AND PROGRAMMING LANGUAGES USED

We applied some open source and proprietary technology tools for the project. The are enumerated below for reference:

- Microsoft Excel for data manipulation and linear regression analysis.
- Google Colab for python modules and scripts
- Framework: Sci-kit learn for building Linear regression model
- Scientific computing Libraries: numpy, pandas
- Visualization: Microsoft Excel, Seaborn, Matplotlib
- Programming language: Python

4. LINEAR REGRESSION MODEL AND ANALYSIS FOR THE 25TH AND 35TH CPE STUDENT

25th Student Linear Regression Analysis

The regression equation can be written as Y= 65.71 + 0.58*grade point which is the interpretation of the regression analysis gotten from Excel. Visualization of the regression analysis is shown below:

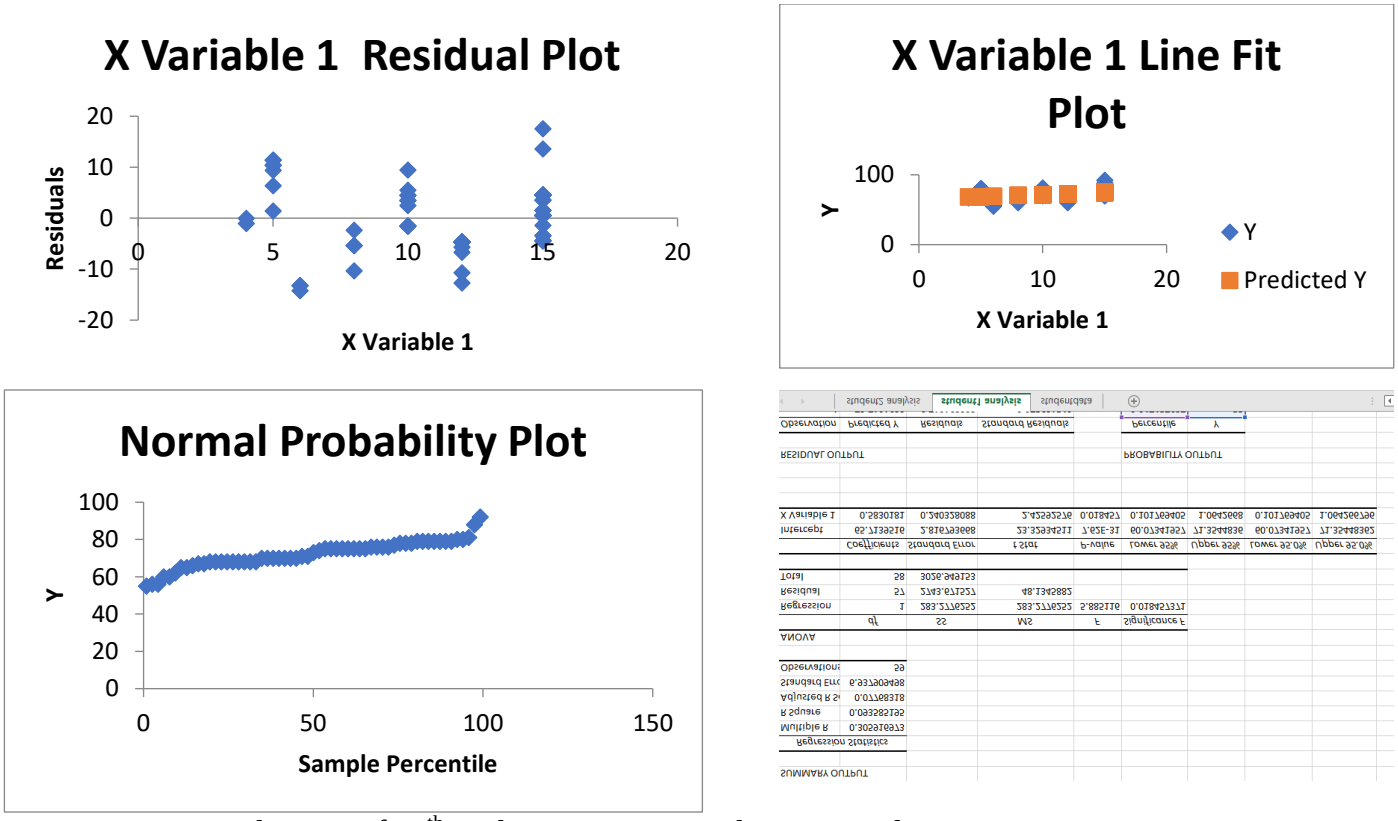


Fig 1: Visualization of 25th student regression analysis in excel

The below scatter plot is gnerated to visualize the 25th student regression before we build regression model to predict the student score in any new course based on the previous results in the dataset. The regression model was tested and we got a prediction that the student will get a score of approximately 70 marks in their next course result.

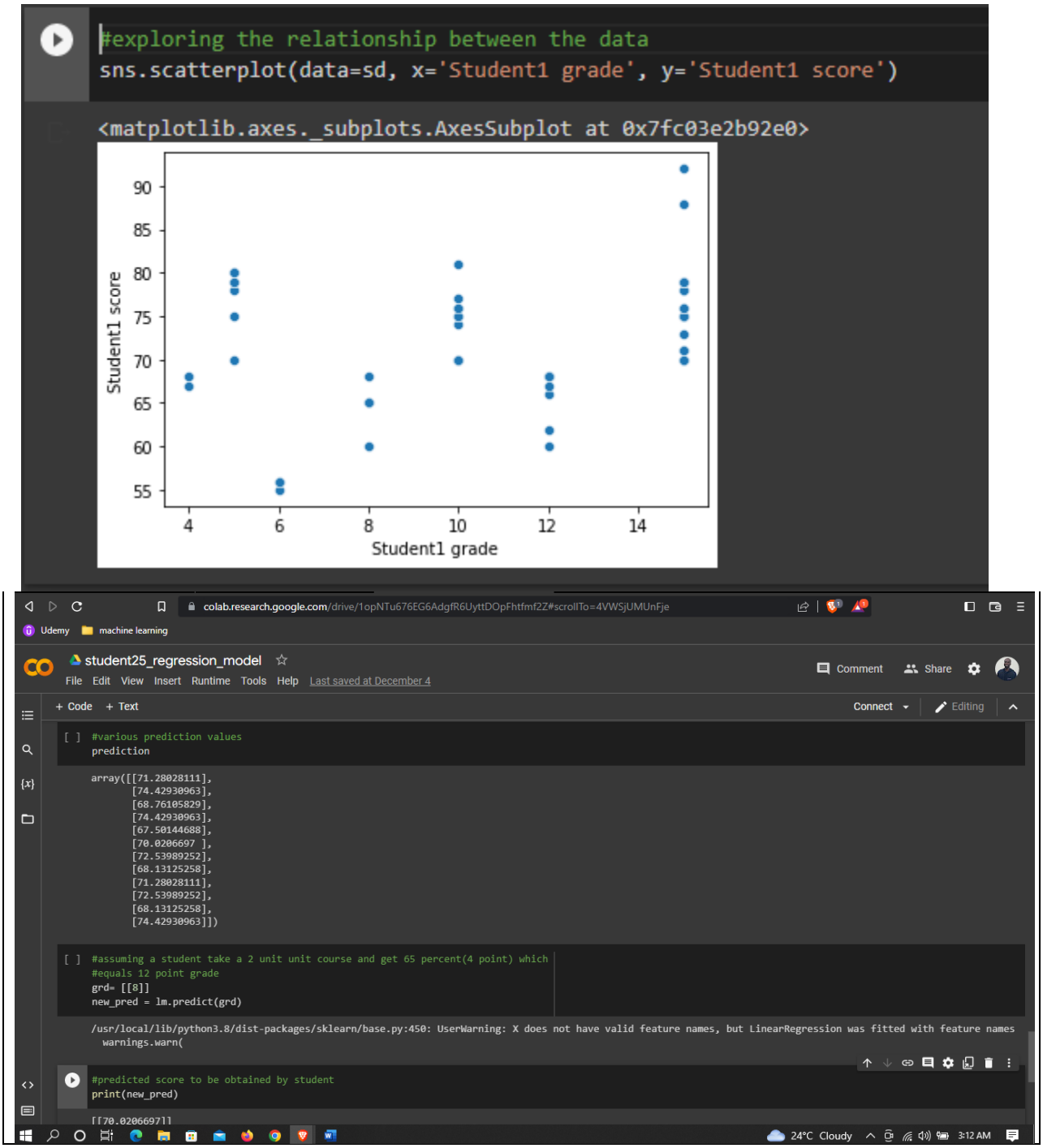
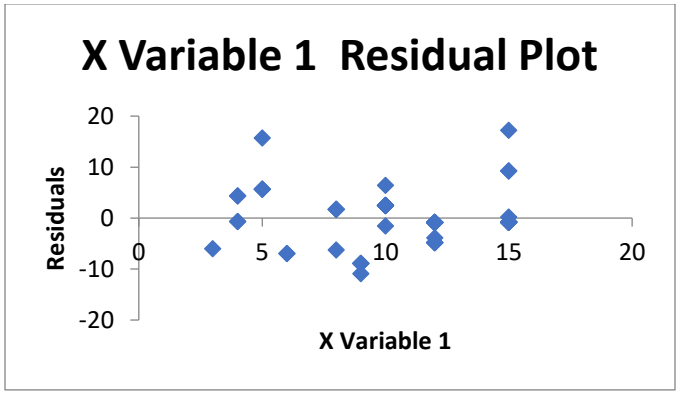
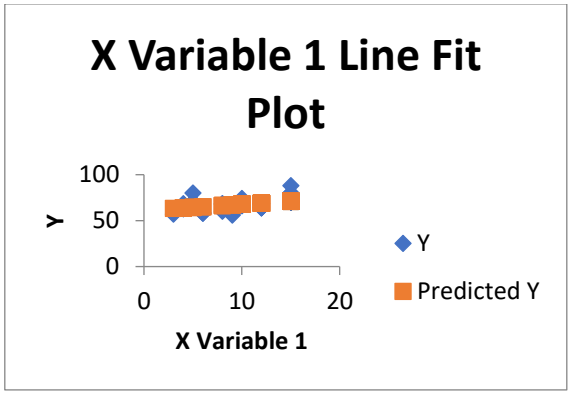


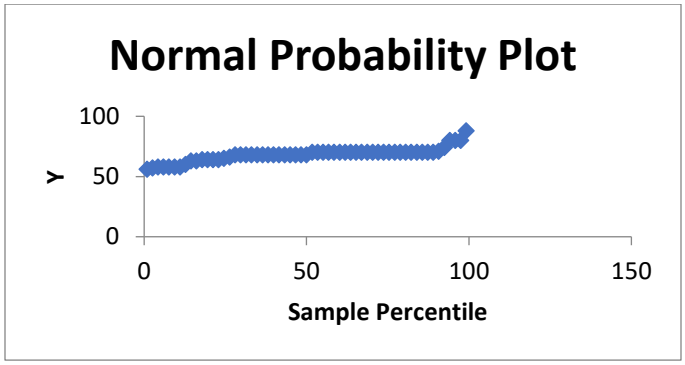
Figure 2: Scatter plot of the student score and prediction of the Linear Regression Model

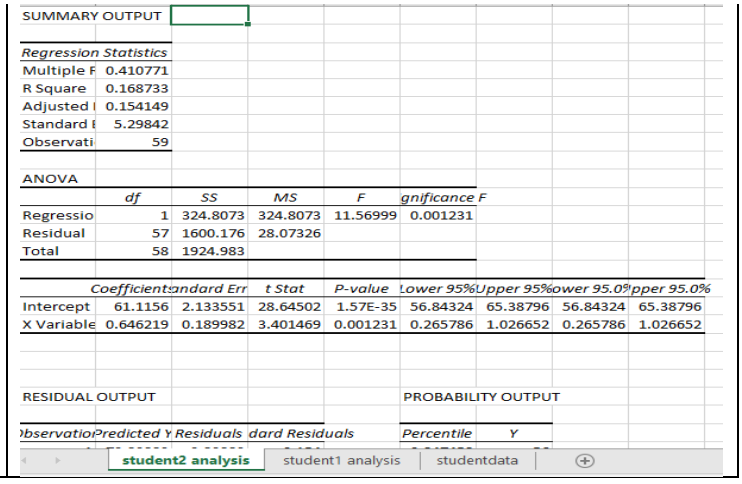
35th Student Linear Regression Analysis

The regression equation for the 35th candidate can be written as "**Y=61.12 + 0.65* grade point**" which is the interpretation of the regression analysis gotten from Excel. Visualization of the regression analysis is shown below:









SUMMARY OUTPUT

Regression Statistics	
Multiple F	0.410771
R Square	0.168733
Adjusted I	0.154149
Standard E	5.29842
Observati	59

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	324.8073	324.8073	11.56999	0.001231
Residual	57	1600.176	28.07326		
Total	58	1924.983			

	Coefficients	*ndard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	61.1156	2.133551	28.64502	1.57E-35	56.84324	65.38796	56.84324	65.38796
X Variable	0.646219	0.189982	3.401469	0.001231	0.265786	1.026652	0.265786	1.026652

RESIDUAL OUTPUT

)bservatio	*Predicted Y*	*Residuals*	*dard Residuals*

PROBABILITY OUTPUT

Percentile	*Y*

student2 analysis | student1 analysis | studentdata

Fig3: Visualization of 35th student regression analysis in excel

The below scatter plot is gnerated to visualize the 25th student score relationship with their grade point. We build regression model to predict the student score in any new course based on the previous results in the dataset. The regression model was tested and we got a prediction that the student will get a score of approximately 68 marks in their next course result.

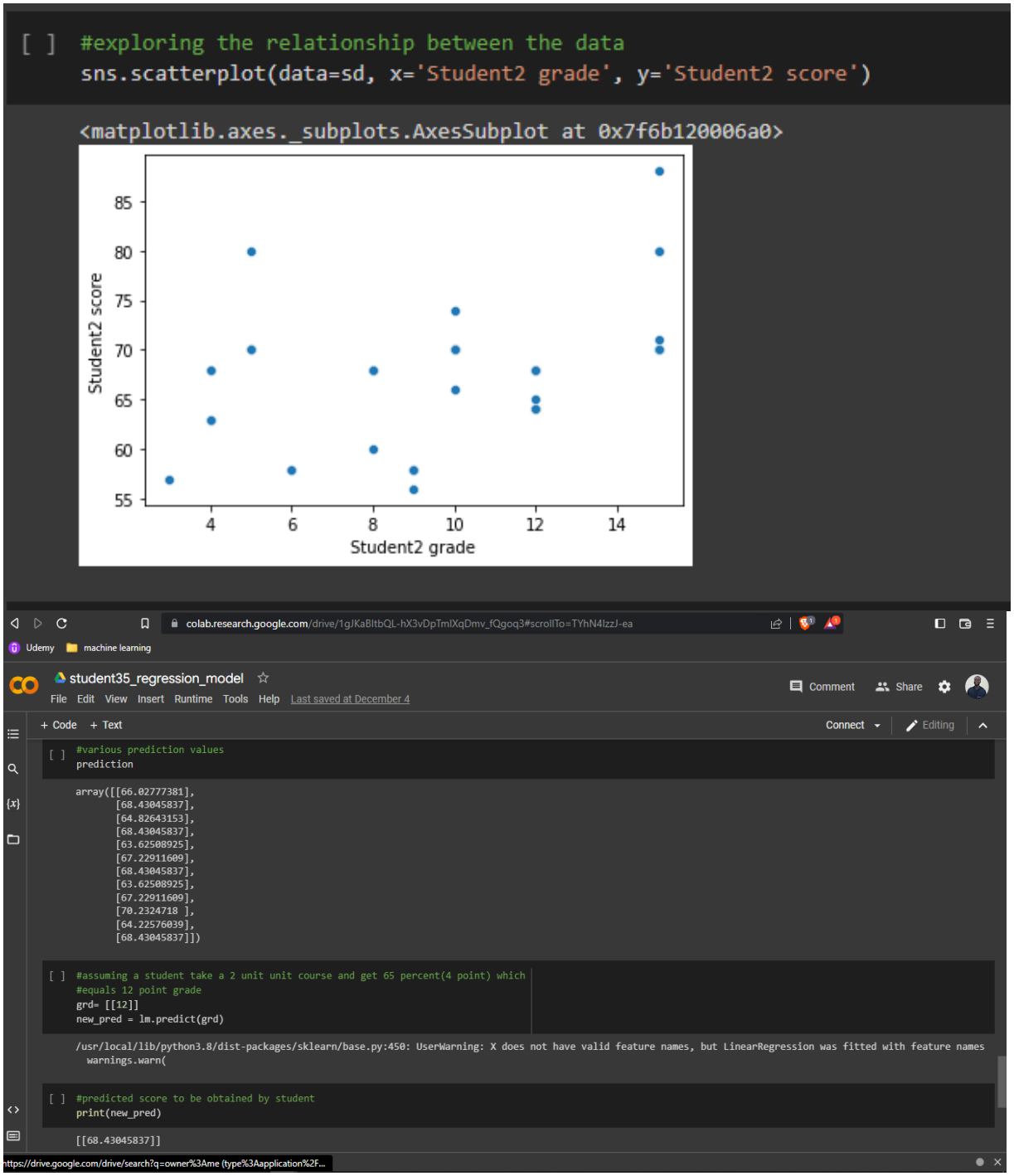


Figure 4: Scatter plot of the 35th student score and prediction out of the Linear Regression Model

5. CONCLUSION

The grade points of the students were used to predict their score in another course. The 25^{th} student has probability of scoring 70 marks on their next course while 35^{th} student has probability of 68 marks in their next course. The major assumption is that course unit determines student grade. Students put effort in courses with three units than two units or one-unit courses. The data was properly preprocessed and we built a Linear Regression model to predict future scores of the two students individually in their respective new courses.