Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Role Of Data Science And Machine Learning In Education: A Predictive Analysis For Students' Performance In Sultanate Of Oman

Rana Tarannum Ansari[1] , Kishore Kumar P. K[2]

[1]IT Department –UTASA University of Technology and Applied Sciences Al Musannah , Oman

[2]IT Department –UTASA University of Technology and Applied Sciences Al Musannah , Oman

---

Abstract—The comprehensive production of data in education has led to the practice of analyzing the data and find out the patterns and useful knowledge out of it These can be used to formulate policy and practices for improving results. A data science and Machine Learning research methodology is becoming more important in educational field, especially related to result prediction. A Machine Learning model i.e. Multiple Linear Regression is used in this research work to for prediction of students results. A case study of Post Foundation students is presented to predict the performance in final exam using multiple linear regression methods. So that we can identify the weaker students and take necessary actions to improve the results.

Index Terms—Data Science, Machine learning, Linear Regression methods.

I. INTRODUCTION

In developing countries, poverty can be reduced by education. It can create increase employment opportunities. With education, one can develop problem solving skills and communities, nations and human life can be changed .Education is the responsibility of the stakeholders which include government officials, teachers , parents and policy makers. Hence, education is important for social and economic growth in society.

To achieve this mission, quality of education should be improved by higher education institutions. Quality of education can be reflected from the high level of student success and low failure rate of students. There are several factors that affect the students' performance. Data mining in education is a latest area of research which uses algorithms to convert large volumes of academic data into valuable information to take decision for improving the educational processes.

A Machine Learning algorithm is a mechanism that runs on data to create a "model". These algorithms recognize "patterns" hidden in the dataset. Algorithms learn from data or are "fit" on a dataset. Emerging Machine Learning and

Artificial Intelligence technologies are capable of finding out patterns in the data. The purpose of this paper is to find out patterns in previous results of some math courses and predict the current semester results for the same. So that policies can be set accordingly to improve the upcoming results.

In this research, the data of around 2000 students has been collected from last 3 years in XYZ University in Oman (name is kept hidden because of confidentiality) and Machine Learning algorithm like single and multiple regression modelsare implemented.

With this model we tried to predict the students' grade in final exam from different class activities (e.g., tests and quizzes) marks. In foundation, they study 2 math courses. Also, in post foundation, they study one math course. So their complete split of marks is present with marks in different assessment, like Test 1, Test 2, Quizzes, Assignment, self-study and Final exam marks etc.

The main content of this paper include the following .
1)The algorithm/model used here can predict the performance of students in final examination of different Math courses. By studying this, the factors(different assessments marks) that determine final result are identified.
(2) The system also provides solution to improve the results.

II. LITERATURE REVIEW

In this section we reviewed the existing literatures on the educational data mining and the academic successes in the institution.
Educational Data Mining deals with the use of data mining techniques by, tapping into the data stored therein, in order to extract meaningful information that can support the decision-making processes by enabling a better understanding of the students and their learning environments [3]. EDM easily identifies those factors responsible for students to either graduate or not graduate [4]. However, the results are dependent on the selected dataset. Student model is defined as the representation of their characteristics, state, intelligent quotient, motivation, meta-cognition and behaviors [3]. This model allows the educational software systems to adapt to the responses of students. Baker et al. focused on identifying frustrated [5]. Personalized learning environments are systems flexible to students' characteristics. They are closely related to the recommendation systems, and allows students achieve their educational goals [6]. The resource management systems can be improved by integrating the Data Mining tools and all works in the EDM. The essence is to allow the average user to be able to make use of such tools [7]. Pedagogical support revolves around identifying the most effective type of support for a given situation and group of students. Beck and Mosto was sociated a student's performance to the type of pedagogical support received [8].

Educational theories deal with its empirical analysis and phenomenon. In order to enable a deeper comprehension of the key aspects of these theories, Gong et al stated that there was a relationship between an individuals' self-discipline and the number of mistakes made by that person [9] Prediction is defined as the determination of the value of an unknown variable using the values of known variables. The known variables are called predictors. Problems associated with prediction can be either be

classified as unknown variable belonging to several pre-established classes, or as a regression whose objective is to predict the value of a continuous numerical variable [10, 11]. There are several other algorithms predictions that can be used to predict the students' performances in their work, such as decision trees and Bayesian classifiers. [12, 13] detection of outliers in the EDM, is an inconsistent process. This helps to identify students who have slow learning processes and those students who are gifted. [14].

III. DESIGN AND METHODOLOGIES

Python3.6 is an object-oriented language.As it is object-oriented, platform independent and has efficient execution, it is used for the large-scale software development.In Python3.6 you can write codes and also run the code simultaneouslywhich improves the engineers' efficiency.In this paper Python 3.6 is used to establish a linear regression model to find out the effect of variation in class assessments marks on the Final exam marks.

The paper herein will predict marks in Final Examination of math courses affected by the variation of class assessments, like Test 1, Test 2, quizzes marks etc.

After Test 1 and Test 2, the addition of two tests taken and the final exam results are predicted, based on the previous records. The addition of Test1 and Test2 in Math course 1 and Math course 2 dataset is taken as x and Final exam marks is taken as y and then submitted to a Regression algorithm For the third model, the post foundation Math Course 1 results are taken, and Final exam marks are predicted from the Course Work marks based on the pattern available in previous records.

Regression Model: A linear regression model gives the relationship between variables. A simple linear regression models determines how an independent variable predicts the values of dependent variable. It can be expressed by the following equation.

$$y = \beta_0 x + \beta_1 + \varepsilon$$

Where x is the feature orindependent variable and y is the response or dependent variable. β_1 represents y - intercept of the regression line and β_0 is the slope of the regression line or a regression coefficient . ε represents the random error

In machine learning,the dataset is partitioned into two parts. model is trained for 70% of data and the remaining 30% is test data . The training and test data both consist of x values as independent variable and y values as independent or response variable. After training the model, the model will predict the y values for test data. The results are compared. If y - actual = y- predicted, it means our model is accurate. Generally, it cannot predict exactly same, so root mean squared error is found out in terms of R^2. R^2 is the 'Goodness of Fit' of a regression model that is used to determine the total variance.

For this research, we used a dataset from the XYZ University' in Oman (The name is hidden for confidentiality). The dataset spans 9 semesters and it contains 3 courses around 2000 students' marks records.

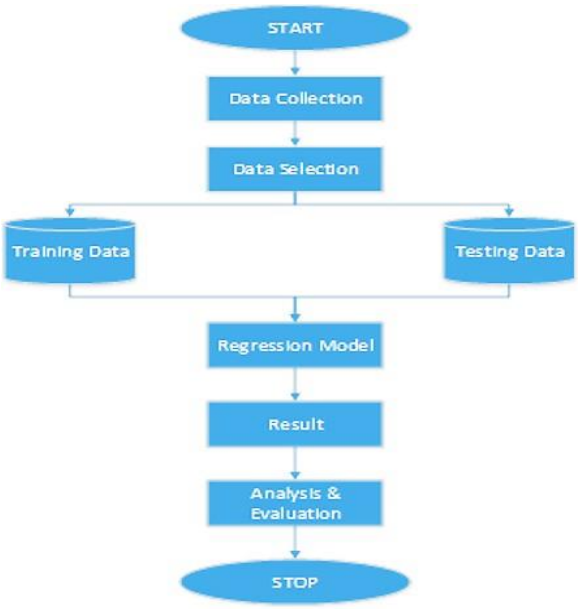

Fig 1: Machine Learning Algorithm Flow diagram

IV. Discussion and Analysis

We used here 999 same students records in all 3 courses.We analyze the concepts of linear regression equations and the R^2 score which is the coefficient of determination. The following parameters were analyzed in this context namely the y-intercept and the regression coefficient.

1) The following table the course 1 marks with test 1 and test 2 added together and total marks a student score out of 100.The following table shows the count, mean, standard deviation, minimum marks, quartile deviations and maximum marks of course 1 studied at foundation level.

	T1+T2(30)	Final(55)
count	999.00000	999.000000
mean	23.14965	28.806807
std	6.21092	11.392862
min	0.00000	0.000000
25%	18.50000	21.000000
50%	23.50000	27.500000
75%	27.50000	36.000000
max	40.00000	55.000000

Fig 2.Summary of data with course 1 Test 1 and Test2 marks added together with Final marks (55)

The following SNS graph shows the comparison between test1 and test2 marks added up together out of 30 with the final marks out of 55.

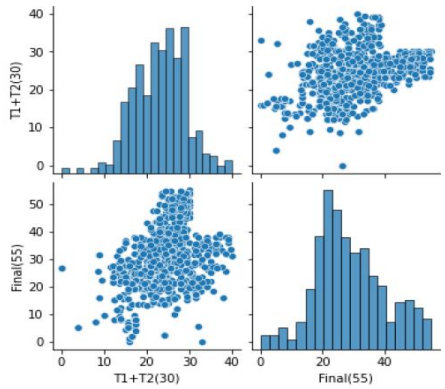

Fig 3.Scatter diagram matrix for Test1+Test 2(30) and Final Exam (55) for math course 1

The regression equation $y = 0.7895x + 10.31$ has been found from the workings on the record collected. Also the comparison study on predicted data were analyzed. Its R^2 score is 0.1852. The following figure shows the scatter diagram between predicted data and test data.

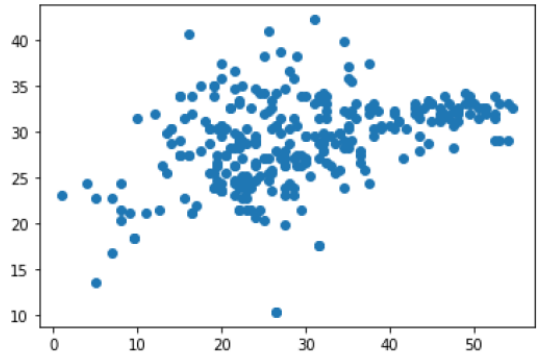

Fig 4.Scatter diagram between predicted data(y-predicted) and test data(y-actual) for math course 1 with Test 1 + Test 2 (30) as input variables and Final Exam (55) as output variable.

2)The following table shows the count, mean, standard deviation, minimum marks, quartile deviations and maximum marks of course 1 studied at foundation level.

	T1+T2(30)	Total(100)
count	999.00000	999.000000
mean	23.14965	65.266266
std	6.21092	15.540276
min	0.00000	20.000000
25%	18.50000	53.000000
50%	23.50000	64.000000
75%	27.50000	77.000000
max	40.00000	100.000000

Fig 5. Summary of data with course 1 Test 1 and Test2 marks added together with Final marks (100)

The following SNS graph shows the comparison between test1 and test2 marks added up together out of 30 with the total marks out of 100.

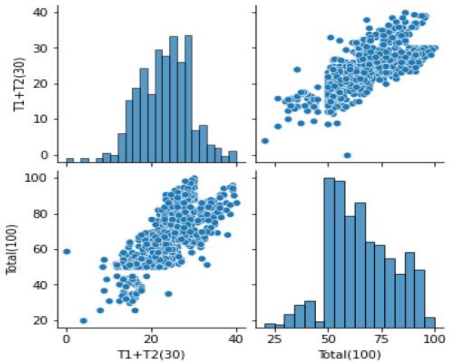

Fig 6. Scatter diagram matrix for Test1+Test 2(30) and Total(100) for math course 1

The regression equation $y = 1.895x + 21.30$ has been found from the workings on the record collected. Also the comparison study on predicted data were analyzed. Its R^2 score is 0.5604. The following figure shows the scatter diagram between the predicted data and test data.

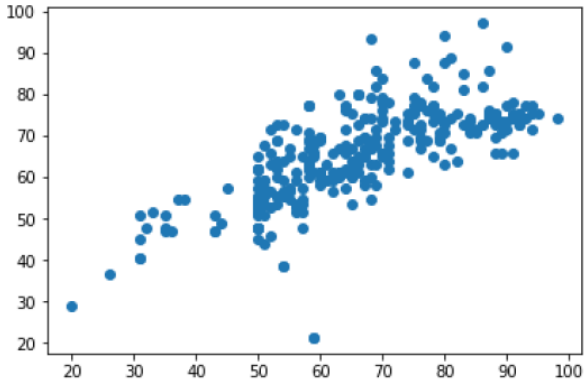

Fig 7.Scatter diagram between predicted data(y-predicted) and test data(y-actual) for math course 1 with Test 1 + Test 2 (30) as input variables and Final Exam (100) as output variable.

3)The following table shows the count, mean, standard deviation, minimum marks, quartile deviations and maximum marks of course 2 studied at foundation level with test1 and test2 out of 30 and total marks out of 100.

	T1+T2(30)	Final(55)
count	999.000000	999.000000
mean	20.813814	27.919279
std	6.518770	11.006022
min	0.000000	0.000000
25%	16.500000	21.000000
50%	21.500000	27.500000
75%	25.500000	35.000000
max	39.500000	54.000000

Fig 8.Summary of data with course 2 Test 1 and Test2 marks added together with Final marks (55)

The following SNS graph shows the comparison between test1 and test2 marks added up together out of 30 with the final marks out of 55.

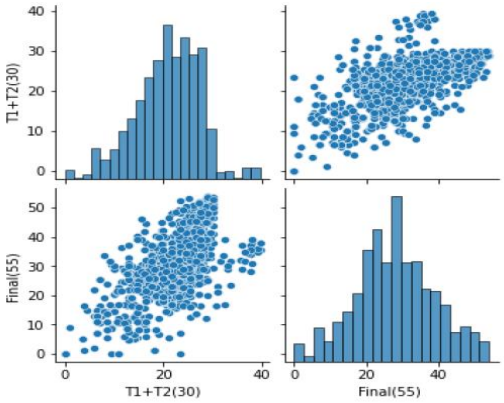

Fig 9.Scatter diagram matrix for Test1+Test 2(30) and Final Exam (55) for math course 2.

The regression equation $y = 1.0648x + 5.756$ has been found from the workings on the record collected.Also, the comparison study on predicted data were analyzed.Its R^2 score is 0.3978. The following figure shows the scatter diagram between predicted data and test data.

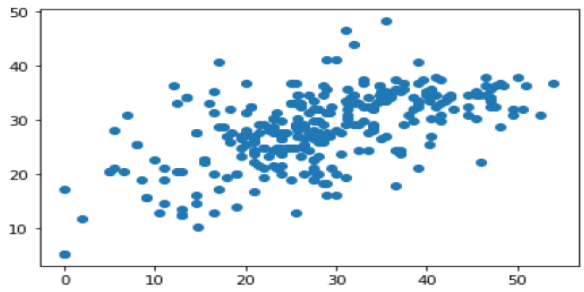

Fig 10.Scatter diagram between predicted data (y-predicted) and test data(y-actual) for math course 2 with Test 1 + Test 2 (30) as input variables and Final Exam (55) as output variable.

4)The following table shows the count, mean, standard deviation, minimum marks, quartile deviations and maximum marks of course 2 studied at foundation level with test1 and test2 out of 30 and final marks out of 55.

	T1+T2(30)	Total(100)
count	999.000000	999.000000
mean	20.813814	60.476476
std	6.518770	16.963275
min	0.000000	0.000000
25%	16.500000	51.000000
50%	21.500000	60.000000
75%	25.500000	73.000000
max	39.500000	98.000000

Fig 11. Summary of data with course 2 Test 1 and Test2 marks added together with Final marks (100)

The following SNS graph shows the comparison between test1 and test2 marks added up together out of 30 with the total marks out of 100.

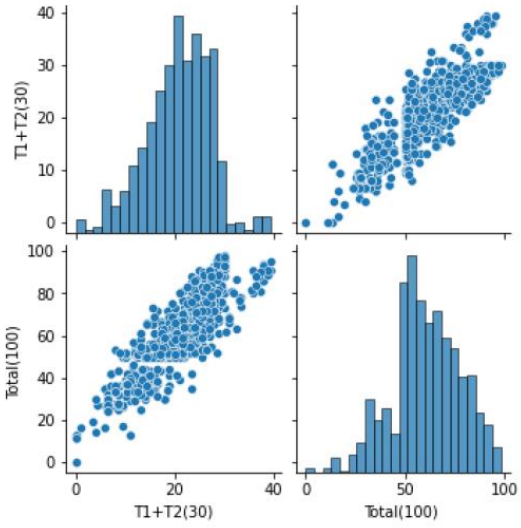

Fig 12.Scatter diagram matrix for Test1+Test 2(30) and Total (100) for math course 2.

The regression equation $y = 2.262x + 13.47$ has been found from the workings on the record collected. Its R^2 score is 0.7331.Also the comparison study on predicted data were analyzed. The following figure shows the scatter diagram between predicted data and test data.

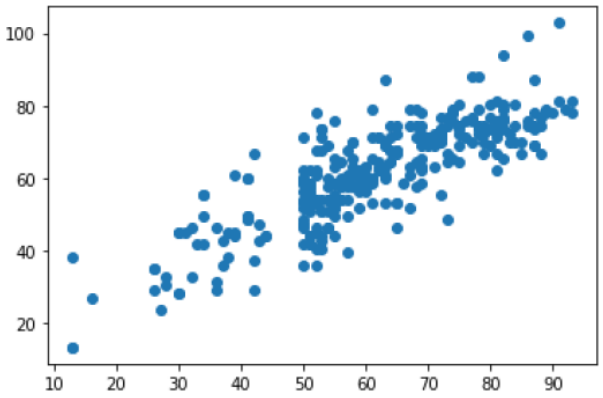

Fig 13. Scatter diagram between predicted data(y-predicted) and test data(y-actual) for math course 2 with Test 1 + Test 2 (30) as input variables and total (100) as output variable.

5)The following table shows the count, mean, standard deviation, minimum marks, quartile deviations and maximum marks of a course studied at post foundation level with course work marks out of 50 and final marks out of 50

	CW(50)	Final(50)
count	999.000000	999.000000
mean	36.776276	31.910911
std	6.869872	10.897779
min	0.000000	0.000000
25%	33.000000	26.500000
50%	37.000000	34.000000
75%	42.000000	40.000000
max	49.000000	50.000000

Fig 14. Summary of data with post foundation math course, with course work marks(50) added together with Final marks (50)

The following SNS graph shows the comparison between course work marks out of 50 with the final marks out of 50.

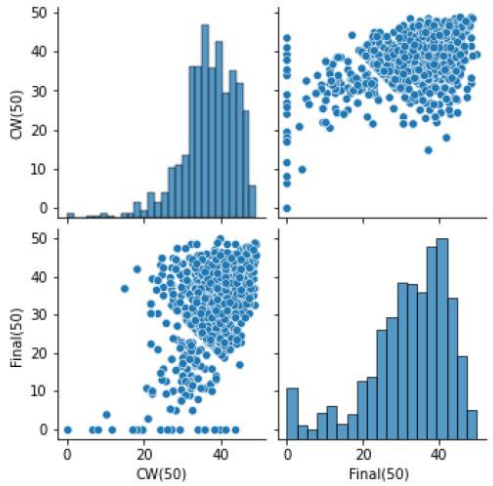

Fig 15.Scatter diagram matrix for Course Work marks (50) and Final Exam (50) for post foundation math course.

The regression equation $y = 0.8071x + 2.23$ has been found from the workings on the record collected. Its R^2 score is 0.2588.Also, the comparison study on predicted data were analyzed. The following figure shows the scatter diagram between predicted data and test data.

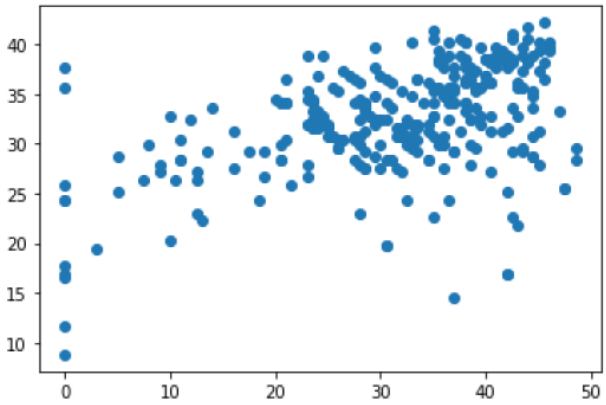

Fig 16Scatter diagram between predicted data(y-predicted) and test data(y-actual) for post foundation math course with Course Work Marks (50) as input variables and Final Exam marks (50) as output variable

6)The following table shows the count, mean, standard deviation, minimum marks, quartile deviations and maximum marks of a course studied at post foundation level with course work marks out of 50 and total marks out of 100.

	CW(50)	Total(100)
count	999.000000	999.000000
mean	36.776276	68.968969
std	6.869872	15.581193
min	0.000000	0.000000
25%	33.000000	61.500000
50%	37.000000	70.000000
75%	42.000000	80.000000
max	49.000000	98.000000

Fig 17. Summary of data with post foundation math course, with course work marks(50) added together with Final marks (100)

The following SNS graph shows the comparison between course work marks out of 50 with the total marks out of 100.

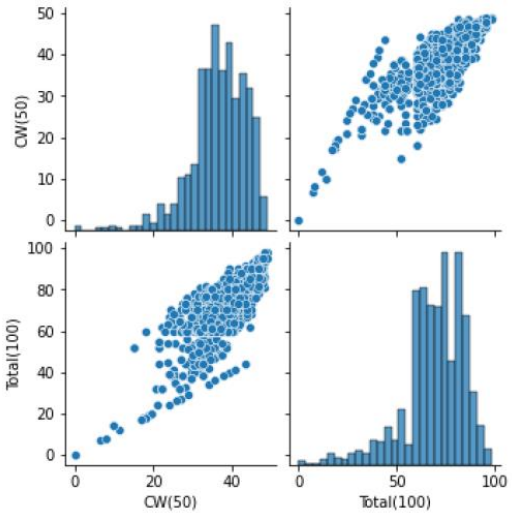

Fig 18. Scatter diagram matrix for Course Work marks (50) and Total (100) for post foundation math course.

The regression equation $y = 1.81x + 2.52$ has been found from the workings on the record collected. Its R^2 score is 0.6366.Also the comparison study on predicted data were analyzed. The following figure shows the scatter diagram between predicted data and test data.

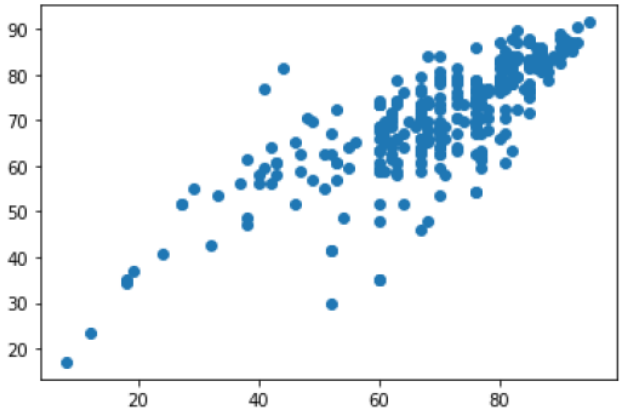

Fig 19.Scatter diagram between predicted data(y-predicted) and test data(y-actual) for post foundation math course with Course Work Marks (50) as input variables and Total (100) as output variable.

V. HYPOTHESIS TESTING

The following hypothesessare formulated for the study.

H1: There is a significant difference in performance of students in class assessment (CW marks) and in final exam.

H0:There is no significant difference in performance of students in class assessment (CW marks) and in Final exam

	CWM	Final Exam
Mean	37.70497763	31.62891499
Variance	35.75328802	101.6597846
Observations	1788	1788
Pearson Correlation	0.535559101	
Hypothesized Mean Difference	0	
df	1787	
t Stat	30.10434945	
P(T<=t) one-tail	1.0685E-161	
t Critical one-tail	1.645706769	
P(T<=t) two-tail	2.1369E-161	
t Critical two-tail	1.961292385	

Fig 20.Paired sample T-test results, performed on Course Work Marks and Final Exam Marks of post foundation Math course.

T–test results are given above in table 1. Here we observe that P(T<t) is less than the level of significance (0.05). Hence, Null Hypothesis is rejected. It means that the performance in Final exam is reduced compared to Course Work Marks. So the effort can be taken to improve the performance in final exam after Course work marks finalized.

H2 : There is no significant difference between the observed values and expected values for frequencies of male and female students in different grade.(i.e The grades obtained by students are independent of the gender).i.e., $O_{ij} = E_{ij}$

H0 : There is a significant difference between the observed values and expected values for frequencies of male and female students in different grade.(i.e The grades obtained by students are dependent on gender) i.e.,$O_{ij} \neq E_{ij}$

By performing the chi square test, the following observations are obtained.

Observed Values

Count of Sno	lumn Labels										
Row Labels	A	A-	B+	B	B-	C+	C	C-	D	F	Grand Total
Male	31	59	89	100	59	4	72	190	60	156	190
Female	71	87	121	91	70	63	72	163	45	109	294
Grand Total	102	16	210	191	129	137	144	353	105	265	484

Expected Values

Count of Sno	lumn Labels										
Row Labels	A	A-	B+	B	B-	C+	C	C-	D	F	Grand Total
Male	40.0	6.3	82.4	75.0	50.6	53.8	56.5	138.6	41.2	104.0	190
Female	62.0	9.7	127.6	116.0	78.4	83.2	87.5	214.4	63.8	161.0	294
Grand Total	102	16	210	191	129	137	144	353	105	265	484

LOS	0.05
DF	9
ChiTest	1E-257
Chi Calculated	253.6
Chi Critical	16.919

Fig 21. Chi Squared Test, performed on Final Exam Grades obtained by male and female students of post foundation Math course.

As we observe in the above table that Chi calculated is much greater than Chi critical. It means that the Null hypothesis is rejected. Which means that there is some association of grade obtained by the students and their gender.Now, if we observe the following graph,some facts are uncovered.

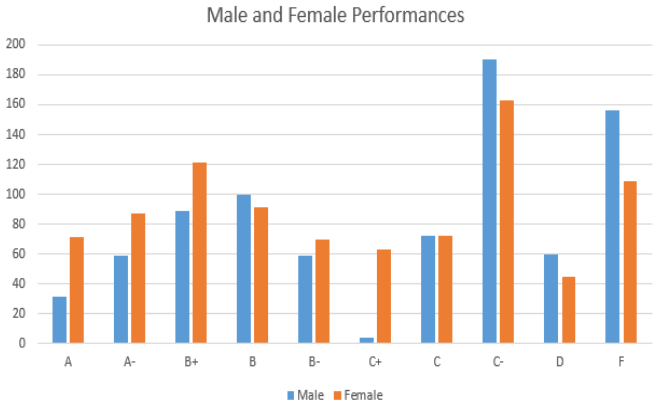

Fig 22.Population of male and female students in different grades.

From the Graph it is very clear that female population is better in higher grades. Where as in lower grades, F, D and C-, male population is more than females. It means we have to focus more on male students.

VI. CONCLUSION AND FUTURE WORK

An objective of this paper is to provide approximate guidelines that may be used by the organization to improve the final results of Math courses is achieved.

The paper herein uses the Regression model in the field of Machine learningusing programming language Python 3.6. From the results it is quite obvious that Linear Regression model can be used to predict the final exam results based on course work marks or Test 1 and Test 2 marks. Most of the places , we get significant R^2 values to accept the model. One thing is also clear that Programming language like Python can also be used for as it makes data mining easier.

The hypotheses testing explains the different factors affecting the Final Marks for post foundation Math course. T-test explains that there is a significant variation between course work marks and final exam marks.

Similarly, Chi squared test results represent that the female students' performance is significantly better than male students.

So the organization can set the policies to improve the Final exam results.

ACKNOWLEDGMENT

The authors would like to express thanks and gratitude to the Department of Information Technology(Mathematics section), University of Technology and Applied Sciences, Al Musannah, Sultanate of Oman for supporting this research.

REFERENCES

[1] L.W. Santoso and Yulia, "Analysis of the impact of information technology investments - a survey of Indonesian universities", ARPN JEAS, vol. 9, no. 12, pp. 2404-2410, Dec 2014.

[2] L.W. Santoso and Yulia, "Data warehouse with big Data technology for higher education", Procedia Computer Science, vol. 124, no. 1, pp. 93-99, 2017.

[3] R.S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions", JEDM-Journal of Educational Data Mining, vol. 1, no. 1, pp. 3-17, 2009.

[4] P. Strecht, JM Moreira and C. Soares, "Educational data mining: preliminary results at university of porto", 2014.

[5] R.S. Baker, A.T. Corbett and A.Z. Wagner, "Human classification of low-fidelity replays of student actions", Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems, pp. 29-36, 2006.

[6] R.A. Huebner, A survey of educational data mining research, 2013.

[7] E. García, C. Romero, S. Ventura and C. de Castro, "A collaborative educational association rule mining tool", The Internet and Higher Education, vol. 14, no. 2, pp. 77-88, 2011.

[8] J.E. Beck and J. Mostow, "How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students", In Intelligent tutoring systems, pp. 353-362, 2008.

[9] Y. Gong, D. Rai, J.E. Beck and N.T. Heffernan, "Does self-discipline impact students' knowledge and learning?", International Working Group on Educational Data Mining, 2009.

[10]S.B. Kotsiantis, I. Zaharakis and P. Pintelas, "Supervised machine learning: A review of classification techniques", 2007.

[11]L.W. Santoso and Yulia, "Predicting student performance using data mining", In the Proceedings of 5th International Conference on Communication and Computer Engineering (ICOCOE), 2018.

[12]N.T. Nghe, P. Janecek and P. Haddawy, "A comparative analysis of techniques for predicting academic performance", In Frontiers in education conference - global engineering: Knowledge without borders opportunities without passports, Oct 2007.

[13]D. Kabakchieva, "Predicting student performance by using data mining methods for classification", Cybernetics and Information Technologies, vol. 13, no. 1, pp. 61-72, 2013.

[14]V. Hodge and J. Austin, "A survey of outlier detection methodologies", Artificial Intelligence Review, vol. 22, no. 2, pp. 85-126, 2004.

[15][15] Barber R., and Sharkey M., "Course correction: Using analytics to predict course success," 2nd International Conference on Learning Analytics and Knowledge, pp. 259-262, 2012.