Journal of Engineering Education Transformations,
Volume No. 37, January 2024 Special Issue, eISSN 2394-1707

Use of Microsoft Excel for Data Collection and Processing to Predict Students' Performance in EDM

Sunita M. Dol[1], Dr. P. M. Jawandhiya[2], Dr. Pravin R. Satav[3]
[1]Department of Computer Science and Engineering, Walchand Institute of Technology, Solapur - 413006, Maharashtra, India
[2] Department of Computer Science and Engineering, Pankaj Laddhad Institute of Technology and Management Studies, Buldhna - 443001, Maharashtra, India
[3] Department of Computer Engineering, Government Polytechnic, Murtijapur – 444107, Maharashtra, India
[1]sunita_aher@yahoo.com,[2] pmjawandhiya@gmail.com , [3]prsatav@gmail.com

***Abstract*— Educational Data Mining (EDM) referes to the research designed to classify, analyze, and predict the students' academic performance from the data collected from educational setting. Data collection and data processing are an important task in any research such as EDM. In this research article, data collection and data processing task are explained in detailed to build the model for predicting students' performance and provide the recommendation in Educational Data Mining.**

In data collection step, we have collected the result ledgers in PDF form related to Four Year Computer Science and Engineering (CSE) course from university. The PDF ledgers for two academic years 2014-15 and 2015-16 of Four Years - First Year, Second Year, Third Year, and Final Year are downloaded from site http://www.sus.ac.in/examination/Online-Result-(Ledger) or https://su.digitaluniversity.ac/Content.aspx?ID=29445 to prepare the dataset to predict students' performance in Educational Data Mining (EDM). In current study, Syllabus structure of Four Year course of Computer Science and Engineering, Credit system pattern, attributes required for preparing dataset, and types of assessment methods such as Types of assessment methods such as Theory + Practical, Theory + Practical + Practical Oral Exam (POE), Practical + POE, Practical + OE, Practical, Term Work, and Theory are explained in detailed. So original data downloaded from university site for two academic years 2014-15 and 2015-16 of Four Years CSE course from Sem-I to Sem-VIII is prepared with the help of Excel and contain approximately 10,616 students data with 544 number of attributes.

For data processing, Microsoft Excel is used. Microsoft Excel features such as Text to Column – Delimited, Text to Column - Fixed width, Filter, and Conditional Formatting – Highlight Cells Rules – Text that contains – are considered for preparation of dataset. Also various functions such as SUM, IF, COUNTIF, MOD and % are employed for processing the data.

After data processing step, final dataset for two academic years 2014-15 and 2015-16 from Sem-I to Sem-VIII consists of 6906 students data with 970 number of attributes.

In addition to the data collection and processing, research gaps related to the dataset size, etc. are also identified and mentioned the same in this article. These two steps - data collection and processing discussed in detailed in this research article will help the researcher working in EDM to prepare the dataset to build the model so that more work in education sector related to students' performance can be carried out to improve the teaching-learning process.

Keywords—Data Collection, Data Processing, Microsoft Excel, Educational Data Mining (EDM)

JEET Category—Research

This paper was submitted for review on July, 12, 2023. It was accepted on November, 15, 2023.

Corresponding author: Sunita M. Dol[1], Dr. P. M. Jawandhiya[2], Dr. Pravin R. Satav[3]
[1]Computer Science and Engineering, Walchand Institute of Technology, Solapur, Maharashtra, India
[2]Computer Science and Engineering, Pankaj Laddhad Institute of Technology and Management Studies, Buldhna, Maharashtra, India
[3]Computer Engineering, Government Polytechnic, Murtijapur, Maharashtra, India
Address: Sunita M. Dol, Assistant Professor, Computer Science and Engineering , Walchand Institute of Technology, P.B.No.634, Walchand Hirachand Marg, Ashok Chowk, Solapur - 413006 Maharashtra, India (e-mail: sunita_aher@yahoo.com).

I. INTRODUCTION

In EDM, predicting students' performance from the educational setting data is of paramount importance. So data collection and data processing plays an important role in collecting data for students' prediction system.

Data collection is the process of gathering and analyzing data collected from relevant sources while data processing is the translation of collected data into useful and valuable data so that it can be used to build the model for the prediction system. During data collection, researchers need to identify the sources of data, data required to build dataset, and data types depending upon the purpose of research. So for data collection, university PDF ledgers related to result of four year engineering course of CSE stream is considered. These PDF ledgers consist of result of each students of CSE stream for First Year, Second Year, Third Year and Final Year. So from these leadgers, the original dataset in the form of result analysis is prepared with the help of Excel for First Year, Second Year, Third Year and Final Year of two Academic Year 2014-15 and 2015-16. Result analysis of original data is prepared based on Permanent Registration Number (PRN) of students using Microsoft Excel.

For data processing, as useful and valuble information needs to be extracted from data collected from various relevnt resources, various tools are also required for processing of that data. There are various tools for processing the data such as Microsoft Excel, Python, Apache Spark, R, Microsoft Power BI, Jupyter Notebook, SAS, Tableau, KNIME, etc. In this research article. For data processing, we used the Microsoft Excel. Data processing is done with the help of various features of Excel such as Filter, Sort, etc. and functions such as SUM, IF, COUNTIF, etc.

In this research article, how data collection step is performed and how data processing is done with the help of Microsoft Excel features are explained in detailed.

The article is arranged in the following manner: Section 2 discusses about literature review related dataset collection technique, dataset size, etc. Section 3 elucidates dataset description and collection while Section 4 describe the use of Microsoft Excel for data processing. Section 5 explains the data collection and processing method while Section 6 talks about the analysis of prepared dataset for student performance prediction system followed by Conclusion in Section 7.

II. LITERATURE SURVEY

There are various research articles which mke the use of dataset to predict the performance of students using data mining technique/ machine learning / deep learning in Educational data mining. In such prediction model, dataset plays an important role. This section discusses about the data collection techniques, dataset size and number of years considered to collect the data by various researcher in Eductional Data Mining and identifies the research gap.

Table 1 discusses about various data collection techniques used by various researcher such as UCI Dataset, Online/ e-learning/ Distance education/ courses data, Programming courses data, Learning Management System (LMS) data, and Students' placement data. From Table 1, it is observed that Online/ e-learning/ Distance education/ courses data is considered in maximum research articles having reference (El Aissaoui, 2020), (Abyaa, 2018), (Ayub, 2017), (Buenaño-Fernández, 2017), (Figueira, 2017), (Kularbphettong, 2017), (Costa, 2017), (Kassak, 2106), (Salinas, 2015), (Shukor, 2015), (Sen, 2012), (Bodea, 2010), (Cocea, 2010), (Wang, 2011), and (Nasiri, 2012) while the research articles (Lagus, 2018), (Leppänen, 2017), (Costa, 2017), (Amornsinlaphachai, 2016), (Badr, 2016), (Ahadi, 2015), (Sisovic, 2015), (Pathan, 2014), and (Márquez-Vera C., Cano A., Romero C. & Ventura S., 2013) use the programming language data to predict the srudents' performance. The research articles such as (Teoh, 2022), (Malini, 2021), (Zaffar, 2018), (Chaudhury, 2016) and (Ahmed, 2016) make the use of UCI Machine Learning Repository Dataset dataset also while Learning Management System (LMS) such as MOODLE is also considered to collect the dta in research articles (Altaf, 2019), (Dimić, 2019), (Ajibade, 2018), (Chellatamilan, 2011) and(Umer, 2019)

TABLE I
ANALYSIS BASED ON DATA COLLECTION TECHNIQUE

Dataset collection technique	Reference Number
UCI Machine Learning Repository Dataset	(Teoh, 2022), (Malini, 2021), (Zaffar, 2018), (Chaudhury, 2016), (Ahmed, 2016)
Online/ e-learning/ Distance education/ courses data	(El Aissaoui, 2020), (Abyaa, 2018), (Ayub, 2017), (Buenaño-Fernández, 2017), (Figueira, 2017), (Kularbphettong, 2017), (Costa, 2017), (Kassak, 2016), (Salinas, 2015), (Shukor, 2015), (Sen B. & Ucar E., 2012), (Bodea, 2010), (Cocea, 2010), (Wang, 2011), (Nasiri, 2012)
Programming courses data	(Lagus, 2018), (Leppänen, 2017), (Costa, 2017), (Amornsinlaphachai, 2016), (Badr, 2016), (Ahadi, 2015), (Sisovic, 2015), (Anh, 2014), (Márquez-Vera C., Cano A., Romero C. & Ventura S., 2013)
Learning Management System (LMS) data	(Altaf, 2019), (Dimić, 2019), (Ajibade, 2018), (Chellatamilan, 2011), (Umer, 2019)
Students' placement data	(Pruthi, 2015)

Table 2 describes the analysis based on number of years to collect the data. So the number of years considered are < 1 Year, 1 Year, 1 & ½ Year, 2 Years, 2 & ½ Years, 3 Years, and 4 Years. From Table 2, it is noted that maximum research articles such as (Karthikeyan, 2020), (Altaf, 2019), (Kularbphettong, 2017), (Costa, 2017), (Lehr, 2016), (Sisovic, 2015), (Guarín, 2015), (Pratiwi,2013), (Márquez-Vera C., Morales C. R. & Soto S. V., 2013), (Dejaeger, 2012), (Şen B., Uçar E. & Delen, D., 2012), and (Chuan, 2011), (Meedech, 2016) consider 1 year period for collecting the data while the research articles (Mengash, 2020), (Crivei, 2019), (Amazona, 2019), (Rustia, 2018), (Ayub, 2017), (Devasia, 2016), (Natek, 2014), (Chau, 2013), and (Palazuelos, 2013) uses 3 years data. The research articles (Lottering, 2020), (Utari, 2020), (Tasnim, 2019), (Santoso, 2019), (Ibrahim, 2018), (Vila, 2018), and (Jung, 2018) collect four years data.

TABLE II
ANALYSIS BASED ON NUMBER OF YEARS TO COLLECT THE DATA

Data collection for number of years	Reference Number
< 1 Year	(Márquez-Vera C., Cano A., Romero C. & Ventura S., 2013), (Zengin, 2011)
1 Year	(Karthikeyan, 2020), (Altaf, 2019), (Kularbphettong, 2017), (Costa, 2017), (Lehr, 2016), (Sisovic, 2015), (Guarín, 2015), (Pratiwi,2013), (Márquez-Vera C., Morales C. R. & Soto S. V., 2013), (Dejaeger, 2012), (Şen B., Uçar E. & Delen, D., 2012), (Chuan, 2011), (Meedech, 2016)
1 and ½ Year	(Castro-Wunsch, 2017), (Chen, 2014)
2 Years	(Al Breiki, 2019), (Martins, 2019), (Sanchez-Santillan, 2016), (Mashiloane, 2013)
2 and ½ Years	(Jishan, 2015)
3 Years	(Mengash, 2020), (Crivei, 2019), (Amazona, 2019), (Rustia, 2018), (Ayub, 2017), (Devasia, 2016), (Natek, 2014), (Chau, 2013), (Palazuelos, 2013)
4 Years	(Lottering, 2020), (Utari, 2020), (Tasnim, 2019), (Santoso, 2019), (Ibrahim, 2018), (Vila, 2018), (Jung, 2018)

Table 3 shows the analysis based on dtaset size. Large dataset gives more accurate result as compared to small dataset. The dataset size ranges considered are <100, 100 - 150, 151-200, 201-250, 251-300, 301-350, 351-400, 401-450, 451-500, 601-700, 701-800, 901-1000, 1001-1500, 1501–2000, 2001-2500, 2501-3000, 3001-3500, 4001-5000, 5001-6000, 6001-8000, and 8001-10000. From Table 3, it is seen that the research articles (Ma, 2021), (Dabhade, 2021), (Rawat, 2019), (Abyaa, 2018), (Pise, 2017), (Ramanathan, 2016), (Bakaric, 2015), (Shukor, 2015), (Pathan, 2014), (Chellatamilan, 2011), (Cocea, 2010), and (Wang, 2011) use the dataset size less than 100. The dataset size range 100-150 is noted in the research articles (Ashraf, 2020), (Al Breiki, 2019), (Akram, 2019), (Rojanavasu, 2019), (Martínez-Abad, 2018), (Burgos, 2018), (Sorour, 2015), (Natek, 2014), and (Kan, 2010) while

the dataset size range 1001-1500 is observed in the research (El Aissaoui, 2020), (Lagman, 2019), (Crivei, 2019), (Zaffar, 2018), (Castro-Wunsch, 2017), (Costa, 2017), (Guo,2015), (Anh, 2014), and (Abaya, 2013). There are very few research articles (Hussain, 2022), (Dejaeger, 2012), and (Chuan, 2011) which use the dataset size greater than 6000.

TABLE III
ANALYSIS BASED ON DATASET SIZE

Size of dataset	Reference Number
<100	(Ma, 2021), (Dabhade, 2021), (Rawat, 2019), (Abyaa, 2018), (Pise, 2017), (Ramanathan, 2016), (Bakaric, 2015), (Shukor, 2015), (Pathan, 2014), (Chellatamilan, 2011), (Cocea, 2010), (Wang, 2011)
100 - 150	(Ashraf, 2020), (Al Breiki, 2019), (Akram, 2019), (Rojanavasu, 2019), (Martínez-Abad, 2018), (Burgos, 2018), (Sorour, 2015), (Natek, 2014), (Kan, 2010)
151-200	(Figueira, 2017), (Sanchez-Santillan, 2016), (Hamsa, 2016), (Kassak, 2016), (Sisovic, 2015), (Jishan, 2015), (Kaur, 2015), (Mayilvaganan, 2015), (Dangi, 2014), (Palazuelos, 2013), (Bodea, 2010), (Göker, 2013)
201-250	(Kaunang, 2018), (Kasthuriarachchi, 2018), (Badr, 2016)
251-300	(Amazona, 2019), (Dimić, 2019), (Leppänen, 2017), (Ahadi, 2015)
301-350	(Lagus, 2018), (Ayub, 2017), (Pratiwi,2013)
351-400	(Rahman, 2020), (Agrawal, 2020), (Tasnim, 2019), (Athani, 2017), (Mashiloane, 2013), (Umer, 2019)
401-450	(Rustia, 2018), (Pruthi, 2015), (Barbosa Manhães, 2015)
451-500	(Malini, 2021), (Almutairi, 2019), (Jalota, 2019), (Ajibade, 2018), (Amornsinlaphachai, 2016), (Ketui, 2019), (Kamal, 2019)
601-700	(Teoh, 2022), (Injadat M., Moubayed A., Nassif A. B. & Shami A., 2020), (Devasia, 2016), (Chaudhury, 2016), (Stahovich, 2016), (Márquez-Vera C., Morales C. R. & Soto S. V., 2013), (Márquez-Vera C., Cano A., Romero C. & Ventura S., 2013)
701-800	(Ibrahim, 2018), (Figueira, 2017)
901-1000	(Altaf, 2019), (Maitra, 2018), (Buenaño-Fernández, 2017), (Lehr, 2016), (Hassan, 2016)
1001-1500	(El Aissaoui, 2020), (Lagman, 2019), (Crivei, 2019), (Zaffar, 2018), (Castro-Wunsch, 2017), (Costa, 2017), (Guo,2015), (Anh, 2014), (Abaya, 2013)
1501–2000	(Adekitan, 2019), (Santoso, 2019), (Guarín, 2015), (Blagojević, 2013)
2001-2500	(Utari, 2020), (Mengash, 2020), (Adekitan, 2020), (Miguéis, 2018), (Hoe, 2013)
2501-3000	(Spatiotis, 2018), (Srivastava, 2018), (Jung, 2018), (Agaoglu, 2016), (Salinas, 2015)
3001-3500	(Jung, 2016), (Sen B. & Ucar E., 2012)
4001-5000	(Lottering, 2020), (Martins, 2019), (Şen B., Uçar E. & Delen, D., 2012), (El-Halees, 2011)
5001-6000	(Kularbphettong, 2017), (Ahmed, 2016), (Ragab, 2014), (Chau, 2013)
6001-8000	(Hussain, 2022)
8001-10000	(Dejaeger, 2012), (Chuan, 2011)

As from Table 1, 2 and 3, it is observed that research articles

- Make the use of UCI Machine Learning Repository Dataset, Online/ e-learning/ Distance education/ courses data, Programming courses data, Learning Management System (LMS) data, Students' placement data, etc.
- uses less number of years for collecting the data.
- generally make the use of small dataset to build the model to predict the students' performance in Eductional Data Mining.

So to bridge this gap, we build the dataset of five years with dataset size 10,616. The dataset is created with the help of various features of Microsoft Excel.

III. DATASET DESCRIPTION AND COLLECTION

For preparing the dataset for predicting students' performance in EDM, data of four years engineering stream related to the Computer Science and Engineering (CSE) in the form of PDF ledgers are collected from University and the link is http://www.sus.ac.in/examination/Online-Result-(Ledger). The PDF ledgers for two academic years 2014-15 and 2015-16 of First Year, Second Year, Third Year, and Final Year CSE were downloaded to prepare the dataset for prediction of students' performance.

A. Data for Academic Year 2014-15 and 2015-16

Two academic years 2014-15 and 2015-16 are considered for collection of dataset because Pandemic period started in March – 2020 and Multiple Choice Questions as assessment method was adopted to evaluate the course of each semester of each stream. In order to correctly predict the students' performance, these two academic years university result ledegers are examined as the university had conducted three hours written paper as assessment method for all courses of all semesters of all years that First, Second, Third and Final Year. So four years data of these two academic years are studied for preparation of dataset. Also as students' performance prediction is for CSE stream, First Year data is ignored to prepare the final dataset to build the prediction model in EDM.s So total five years data from 2014-15 to 2018-19 is collected for data collection as shown below in the Table 4.

TABLE IV
ACADEMIC YEARS CONSIDERED FOR DATA COLLECTION

Year→ Academic Year ↓	First Year	Second Year	Third Year	Final Year
2014-15	2014-15	2015-16	2016-17	2017-18
2015-16	2015-16	2016-17	2017-18	2018-19

B. Syllabus Structure of CSE

The syllabus structure of CSE for Second Year, Third Year, and Final Year is given in Table 5. Each Year that is Second, Third and Final consist of two semesters. Table 5 consist of Course Name of Semester, abbreviation considered for the course, credits assigned to that course and assessment method. Assessment method are Theory + Practical**,** Theory + Practical + Practical Oral Examination (POE), Practical + Practical Oral Examination, Practical + Oral Examination (OE), Practical, Term Work and Theory.

TABLE V
SYLLABUS STRUCTURE OF CSE

Paper Name	Abbr-eviation	Cre-dits	Assessment Method

Applied Mathematics-I	AM-I	4	Theory+Practical
Discrete Mathematical Structures	DMS	4	Theory+Practical
Advanced C Concepts	ACC	5	Theory+Practical+POE
Digital Techniques	DT	5	Theory+Practical+POE
Computer Graphics	CG	4	Theory+Practical
Lab-Visual Basic	VB	3	Practical+POE
Total		25	
Applied Mathematics-II	AM-II	4	Theory+Practical
Theory of Computation	TOC	4	Theory+Practical
Microprocessors	MP	5	Theory+Practical+POE
Data Communication	DC	4	Theory+Practical
Data Structures	DS	5	Theory+Practical+POE
Lab-Object Oriented Design & Programming Through C++	OODP	3	Practical+POE
Total		25	
Operating System Concepts	OSC	5	Theory+Practical+ POE
Computer Networks	CN	5	Theory+Practical+ POE
System Programming	SP	4	Theory+Practical
Design and Analysis of Algorithm	DAA	4	Theory+Practical
Computer Organization	CO	3	Theory+Practical
Lab-Java Programming	JP	4	Practical+POE
Self Learning-I	SL-I	2	Theory
		27	
Database Engineering	DBE	5	Theory+Practical+ POE
Compiler Construction	CC	4	Theory+Practical
Unix Operating System	UOS	4	Theory+Practical
Mobile Computing	MC	4	Theory+Practical
Software Engineering	SE	4	Theory+Practical
Lab-Programming in C#.net	C#	3	Practical+POE
Mini Project	Mproj	1	Practical+OE
Self Learning -II		2	Theory
Total		27	
Advanced Computer Architecture	ACA	3	Theory+Practical
Distributed Systems	DSys	4	Theory+Practical
Modern Database Systems	MDS	6	Theory+Practical+ POE
Lab-I (Project Phase-I)	PP-I	2	Practical+POE
Lab-II (Python)	Py	3	Practical
Vocational Training	VT	1	Term Work
Elective-I	Ele-I	3	Theory+Practical
Elective-II	Ele-II	3	Theory+Practical
Total		25	
Management Information System	MIS	3	Theory+Practical
Information & Cyber Security	ICS	4	Theory+Practical+ POE
Lab-I (Web Technology)	WT	3	Practical+POE
Lab-II (Project Phase-II)	PP-II	3	Practical+POE
Lab-III (Open Source Technology)	OST	3	Term Work
Elective-III	Ele-III	3	Theory+Practical
Elective-IV	Ele-IV	3	Theory+Practical
Total		22	

The Patten used for two Academic Years 2014-15 and 2015-16 is the Credit System - Ten Point Scale. Table 6 shows the Credit System Pattern. This Table contains Grade Abbreviation, From marks, To marks, From Grade Point Average (GPA), To GPA, Status, and Description. Grade abbreviations - O, A+, A, B+, B, C+, C, and F are for the marks range 80-100, 70-79.99, 60-69.99, 55-59.99, 50-54.99, 45-49.99, 40-44.99, and 0-39.99 respectively while GPA ranges 9.5-10, 8.5-9.49, 7.5-8.49, 6.5-7.49, 5.5-6.49, 4.5-5.49, 4-4.49, and 0-3.99 are for the Grade abbreviations - O, A+, A, B+, B, C+, C, and F respectively. The Status is 'Pass' for Grade O, A+, A, B+, B, C+, and C while it is 'Fail' for Grade F. Description for Grade O, A+, A, B+, B, C+, C, and F are - Excellent/Outstanding, Very Good, Good, Fair, Above Average, Average, Below Average, and Fail respectively.

TABLE VI
CREDIT SYSTEM PATTERN FOR ACADEMIC YEARS 2014-15 AND 2015-16

Sr. No.	Grade Abbreviation	From (Marks)	To (Marks)	Grade Point	From (GPA)	To (GPA)	Status	Description
1	O	80	100	10	9.5	10	Pass	Excellent/Outstanding
2	A+	70	79.99	9	8.5	9.49	Pass	Very Good
3	A	60	69.99	8	7.5	8.49	Pass	Good
4	B+	55	59.99	7	6.5	7.49	Pass	Fair
5	B	50	54.99	6	5.5	6.49	Pass	Above Average
6	C+	45	49.99	5	4.5	5.49	Pass	Average
7	C	40	44.99	4	4	4.49	Pass	Below Average
8	F	0	39.99	0	0	3.99	Fail	Fail

Types of assessment methods such as Theory + Practical, Theory + Practical + Practical Oral Exam (POE), Practical + POE, Practical + OE, Practical, Term Work, and Theoryare considered in Table 7. Table 7 also specify the maximum marks and minimum marks required to pass the End Semester Exam (ESE), In Semester Exam (ISE), Internal Contineous Assessment (ICA) and POE. In Table 7, ESE of 70 marks, ISE of 30 marks and ICA of 25 marks are considered for assessment method - Theory + Practical while for the assessment method - Theory + Practical + POE, total 175 marks are there with ESE of 70 marks + ISE of 30 marks + ICA of 25 marks + POE of 50 marks. For the assessment type - Practical + POE/ Practical + OE, ICA and POE is considered. So total 75 marks are given for Practical + POE and 50 marks are given for Practical + OE. In both cases, 25 marks are for ICA.

In Practical / Term Work assessment method, only ICA is considered. In Table 4, ICA of 50 marks is for Practical assessment method while ICA of 25 marks is for Term Work assessment method. For Theory assessment method, only ESE is specified for maximum marks 50.

TABLE VII
TYPES OF ASSESSMENT METHODS

Assessment Method	ESE (End Semester Exam)		ISE (In Semester Exam)		ICA (Internal Contineous Assessment)		POE (Practical Oral Exam)		Total
	Mix. marks	Min Marks	Mix. marks	Min Marks	Mix. marks	Min Marks	Mix. marks	Min Marks	
Theory + Practical	70	28	30	12	25	10	--	--	125
Theory + Practical + POE	70	28	30	12	25	10	50	20	175
Practical + POE	--	--	--	--	25	10	50	20	75
Practical + OE	--	--	--	--	25	10	25	10	50
Practical	--	--	--	--	50	20	--	--	50
Term Work	--	--	--	--	25	10	--	--	25
Theory	50	20	--	--	--	--	--	--	50

Following attributes are considered for each course type of assessment method to prepare original data -

- For assessment method - Theory + Practical - End Semester Exam ESE (70 Marks), In Semester Exam ISE (30 Marks), Internal Contineous Assessment ICA (25 Marks), Condol Marks (1, 2 3, 4, 5, 6, 7,), Total (125 Marks), Grade (O, A+, A, B+, B, C+, C, and F), Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0), Earned Grade Point EGP (Credits assigned to the Course * GP), Status (Pass/ Fail), and Number of Attempts required to pass the course.
- For assessment method - Theory + Practical + POE - End Semester Exam ESE (70 Marks), In Semester Exam ISE (30 Marks), Theory Status (Pass/ Fail), Theory Condol Marks (1, 2 3, 4, 5, 6, 7,), Internal Contineous Assessment ICA (25 Marks), Internal Contineous Assessment POE (50 Marks), Practical Status (Pass/ Fail), Practical Condol Marks (1, 2 3, 4, 5, 6, 7,), Total (175 Marks), Grade (O, A+, A, B+, B, C+, C, and F), Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0), Earned Grade Point EGP (Credits assigned to the Course * GP), Status (Pass/ Fail), Number of Theory Attempts required to pass the course, and Number of Practical Attempts required to pass the POE.
- For assessment method - Practical + POE/ Practical + OE – Internal Contineous Assessment ICA (25 Marks), Internal Contineous Assessment POE (50 Marks for Practical + POE and 25 Marks for Practical + OE), Condol Marks (1, 2 3, 4, 5, 6, 7,), Total (75 Marks for Practical + POE and 50 Marks for Practical + OE), Grade (O, A+, A, B+, B, C+, C, and F), Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0), Earned Grade Point EGP (Credits assigned to the Course * GP), Status (Pass/ Fail), Number of Attempts required to pass the POE/ OE
- For assessment method - Practical / Term Work - Internal Contineous Assessment ICA, Grade (O, A+, A, B+, B, C+, C, and F), Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0), Earned Grade Point EGP (Credits assigned to the Course * GP), Status (Pass/ Fail), Number of Attempts required to pass the course
- For assessment method - Theory - End Semester ExamESE (50), Condol Marks (1, 2 3, 4, 5, 6, 7,), Total (50), Grade (O, A+, A, B+, B, C+, C, and F), Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0), Earned Grade Point EGP (Credits assigned to the Course * GP), Status (Pass/ Fail), Number of Theory Attempts required to pass the course.

Following attributes are considered for each course type of assessment method to prepare final dataset. If the student is passed in the course then values of subbulleted point related to the marks will be the same as that of bulleted point. If the student is failed in the course then values of sub-bulleted point related to the marks will be the values after passing the course and will have different vlues than that of bulleted points.

- For assessment method - Theory + Practical - End Semester Exam ESE (70 Marks), In Semester Exam ISE (30 Marks), Internal Contineous Assessment ICA (25 Marks), Condol Marks (1, 2 3, 4, 5, 6, 7,), Total (125 Marks), Grade (O, A+, A, B+, B, C+, C, and F), Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0), Earned Grade Point EGP (Credits assigned to the Course * GP), Status (Pass/ Fail), and Number of Attempts required to pass the course.
  - End Semester Exam ESE (70 Marks) Final, In Semester Exam ISE (30 Marks) Final, Internal Continuous Assessment ICA (25 Marks) Final, Condol Marks (1, 2 3, 4, 5, 6, 7,) Final, Total (125 Marks) Final, Grade (O, A+, A, B+, B, C+, C, and F) Final, Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0) Final, Earned Grade Point EGP (Credits assigned to the Course * GP) Final, Status (Pass/ Fail) Final.
- For assessment method - Theory + Practical + POE - End Semester Exam ESE (70 Marks), In Semester Exam ISE (30 Marks), Theory Status (Pass/ Fail), Theory Condol Marks (1, 2 3, 4, 5, 6, 7,), Internal Contineous Assessment ICA (25 Marks), Internal Contineous Assessment POE (50 Marks), Practical Status (Pass/ Fail), Practical Condol Marks (1, 2 3, 4, 5, 6, 7,), Total (175 Marks), Grade (O, A+, A, B+, B, C+, C, and F), Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0), Earned Grade Point EGP (Credits assigned to the Course * GP), Status (Pass/ Fail), Number of Theory Attempts required to pass the course, and Number of Practical Attempts required to pass the POE.
  - End Semester Exam ESE (70 Marks) Final, In Semester Exam ISE (30 Marks) Final, Theory Condol Marks (1, 2 3, 4, 5, 6, 7,) Final, Internal Continuous Assessment ICA (25 Marks) Final, Internal Continuous Assessment POE (50 Marks) Final, Practical Condol Marks (1, 2 3, 4, 5, 6, 7,) Final, Total (175 Marks) Final, Grade (O, A+, A, B+, B, C+, C, and F) Final, Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0) Final, Earned Grade Point EGP (Credits assigned to the Course * GP) Final, Status (Pass/ Fail) Final
- For assessment method - Practical + POE/ Practical + OE – Internal Contineous Assessment ICA (25 Marks), Internal Contineous Assessment POE (50 Marks for Practical + POE and 25 Marks for Practical + OE), Condol Marks (1, 2 3, 4, 5, 6, 7,), Total (75 Marks for Practical + POE and 50 Marks for Practical + OE), Grade (O, A+, A, B+, B, C+, C, and F), Grade Point GP (10, 9,

8, 7, 6, 5, 4, and 0), Earned Grade Point EGP (Credits assigned to the Course * GP), Status (Pass/ Fail), Number of Attempts required to pass the POE/ OE.
    - Internal Continuous Assessment ICA (25 Marks) Final, Internal Continuous Assessment POE (50 Marks for Practical + POE and 25 Marks for Practical + OE) Final, Condol Marks (1, 2 3, 4, 5, 6, 7,) Final, Total (75 Marks for Practical + POE and 50 Marks for Practical + OE) Final, Grade (O, A+, A, B+, B, C+, C, and F) Final, Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0) Final, Earned Grade Point EGP (Credits assigned to the Course * GP) Final, Status (Pass/ Fail) Final
- For assessment method - Practical / Term Work - Internal Contineous Assessment ICA, Grade (O, A+, A, B+, B, C+, C, and F), Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0), Earned Grade Point EGP (Credits assigned to the Course * GP), Status (Pass/ Fail), Number of Attempts required to pass the course
    - Internal Continuous Assessment ICA Final, Grade (O, A+, A, B+, B, C+, C, and F) Final, Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0) Final, Earned Grade Point EGP (Credits assigned to the Course * GP) Final, Status (Pass/ Fail) Final
- For assessment method - Theory - End Semester ExamESE (50), Condol Marks (1, 2 3, 4, 5, 6, 7,), Total (50), Grade (O, A+, A, B+, B, C+, C, and F), Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0), Earned Grade Point EGP (Credits assigned to the Course * GP), Status (Pass/ Fail), Number of Theory Attempts required to pass the course
    - End Semester Exam ESE (50) Final, Condol Marks (1, 2 3, 4, 5, 6, 7,) Final, Total (50) Final, Grade (O, A+, A, B+, B, C+, C, and F) Final, Grade Point GP (10, 9, 8, 7, 6, 5, 4, and 0)Final, Earned Grade Point EGP (Credits assigned to the Course * GP) Final, Status (Pass/ Fail) Final

Table 8 shows the number of attributes for preparing original data from university ledgers and attributes required for final dataset considered for each course as per assessment method. So number of attributes 544 and 970 are considered for preparing the original data and final dataset respectively.

TABLE VIII
NUMBER OF ATTRIBUTES FOR PREPARING ORIGINAL DATA AND ATTRIBUTES REQUIRED FOR FINAL DATASET CONSIDERED FOR EACH COURSE

Abbreviation	Assessment Method	Attributes for preparing original data	Attributes for final dataset
AM-I	Theory+Practical	10	19
DMS	Theory+Practical	10	19
ACC	Theory+Practical+POE	15	26
DT	Theory+Practical+POE	15	26
CG	Theory+Practical	10	19
VB	Practical+POE	9	17
	Additional Attributes of SY I result	13	16
AM-II	Theory+Practical	10	19
TOC	Theory+Practical	10	19
MP	Theory+Practical+POE	15	26
DC	Theory+Practical	10	19
DS	Theory+Practical+POE	15	26
OODP	Practical+POE	9	17
	Additional Attributes of SY II result	21	32
OSC	Theory+Practical+POE	15	26
CN	Theory+Practical+POE	15	26
SP	Theory+Practical	10	19
DAA	Theory+Practical	10	19
CO	Theory+Practical	10	19
JP	Practical+POE	9	17
SL-I	Theory	8	15
	Additional Attributes of TY I result	13	16
DBE	Theory+Practical+POE	15	26
CC	Theory+Practical	10	19
UOS	Theory+Practical	10	19
MC	Theory+Practical	10	19
SE	Theory+Practical	10	19
C#	Practical+POE	9	17
Mproj	Practical+OE	9	17
SL-II	Theory	8	15
	Additional Attributes of TY II result	21	32
ACA	Theory+Practical	10	19
DSys	Theory+Practical	10	19
MDS	Theory+Practical+POE	15	26
PP-I	Practical+POE	9	17
Py	Practical	6	11
VT	Term Work	6	11
Ele-I	Theory+Practical	10	19
Ele-II	Theory+Practical	10	19
	Additional Attributes of BE I result	13	16
MIS	Theory+Practical	10	19
ICS	Theory+Practical+POE	15	26
WT	Practical+POE	9	17
PP-II	Practical+POE	9	17
OST	Term Work	6	11
Ele-III	Theory+Practical	10	19
Ele-IV	Theory+Practical	10	19
	Additional Attributes of BE II result	22	32
Additional Attributes for overall evaluation		-	13
	Graand Total	544	970

IV. USE OF EXCEL FEATURES USED FOR DATA PROESSING

Following Excel features are used to prepare the dataset for Students' performance prediction system.

A. Text to Column – Delimited

The result ledgers collected from university site were in PDF file, so to prepare the datset to build the model for predicting students' performance using data mining techniques, Excel feature – Text to column is used. For the feature 'Text to column', there are two options –

- Delimited – Character such as commas or tabs separate each field
- Fixed Width – Fields are aligned in the columns with spaces in each field.

To use 'Text to Column – Delimited'**,** create the Excel sheet and copy the content of PDF ledger of particular semester of CSE in the sheet. Figure 1 shows the content of PDF ledger Second Year Bachelor of Engineering (COMPUTER SCIENCE & ENGINEERING) Semester-I (Credit System - Ten Point Scale) copied to Excel sheet. As shown in the Figure 1, it is not possible to map the fields mentioned in row numbers 4 and 5 to the values given in row numbers 9 to 17. The feature 'Text to Column' required three steps to convert text to column in sheet. So steps used for 'Text to Column - Delimited' feature are as follows –

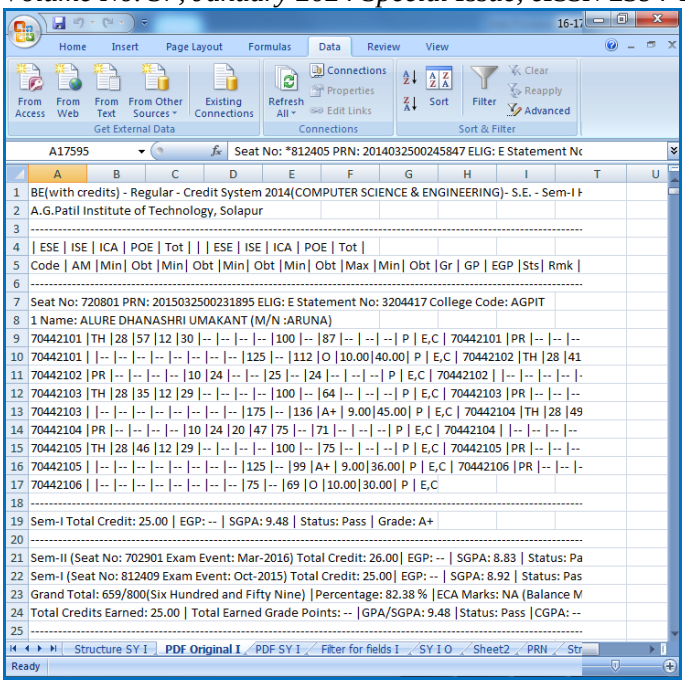

Fig.1. Copy of University leadger PDF file to Excel

Step 1: Click on 'Data' tab on menu bar and select 'Text to Columns'. After selecting 'Text to Columns', there are two options – Delimited and Fixed Width. So click on 'Delimited' option as the fields of PDF ledger copied to Excel sheet are required to be separated for preparing the dataset. Click on 'Next' to go to the next step in 'Text to Columns' option as shown in Figure 2.

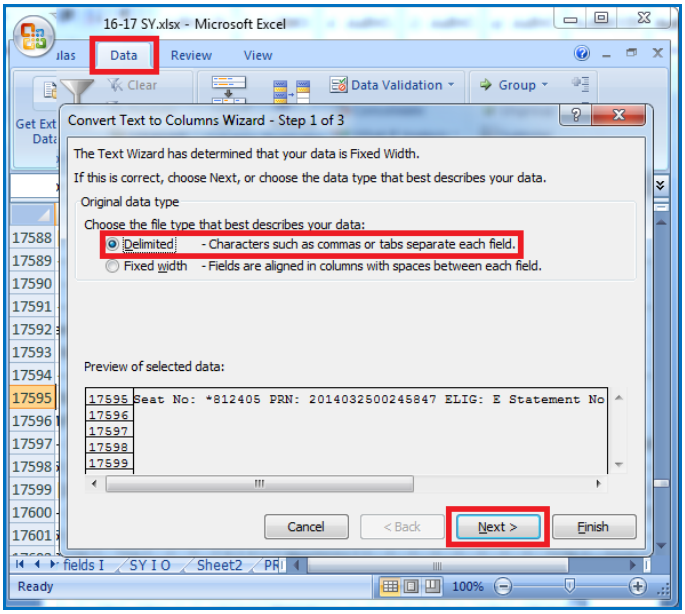

Fig.2. Seleting 'Delimited' in 'Convert Text to Column' window of Excel

Step 2: The next step 'Convert Text to Columns Wizard - Step 2 of 3' is shown in Figure 3. This step set the delimiters that the data contains. So here in this Figure, delimiter that the data contains, has to mention. So here delimiter 'Other' as vertical line (|) is selected and clicked on 'Next' to go to next step.

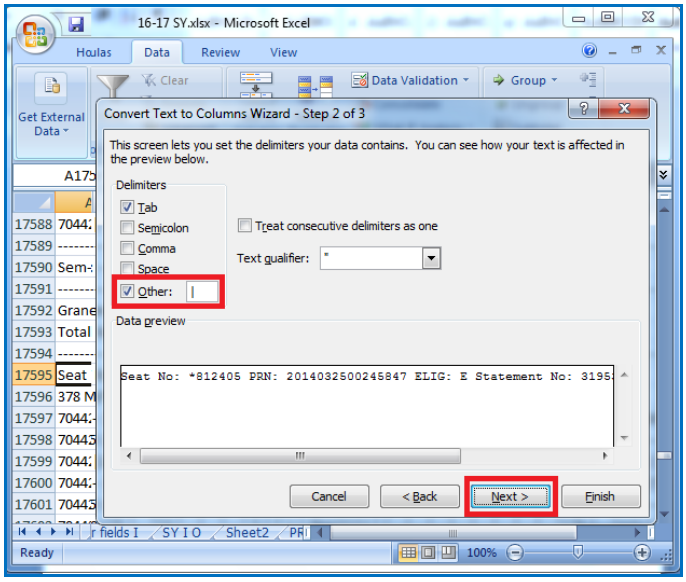

Fig.3. Selecting Other as | in 'Convert Text to Column' window of Excel

Step 3: Figure 4 shows the next step 'Convert Text to Columns Wizard - Step 3 of 3'. This step select each column and set the data format. So here in this step, 'Column data format' selected is 'General'. So 'General' converts numeric values in sheet to number and date values to date and remaining values to text. Here destination cell considered is A1 which allows neither the column nor the rows reference to change.

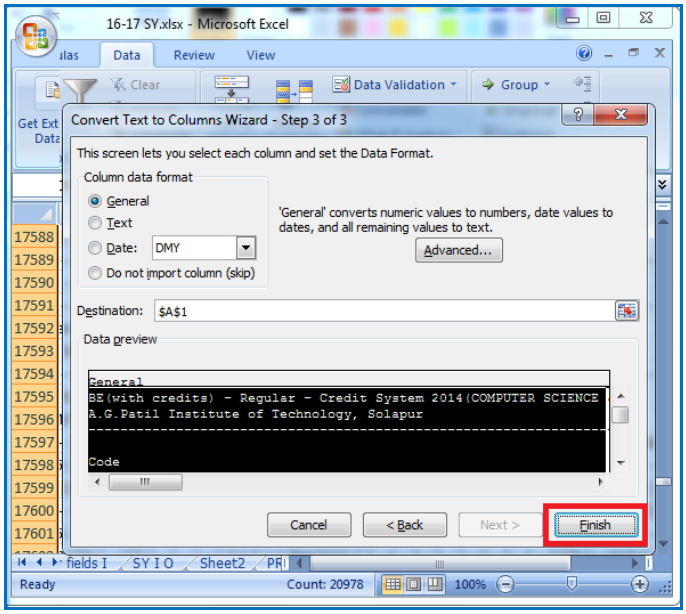

Fig.4. Clicking on 'Finish' in 'Convert Text to Column' window of Excel

Figure 5 shows the data of PDF ledgers after converting Text to Column – Delimited in Excel. This data is now easy to handle and easy to prepare the result analysis of that semester of CSE. So in this sheet shown in Figure 5, it is possible to map the fields given in row numbwrs 4 and 5 with row numbers 9 to 17 of that particular student e.g. column A of row number 5 contains the filed 'Code' and course code of various courses of that semester is given in column A of row number 9 to 17. Similarly other field of data can be mapped to

particular row and column. In this way, it is easy to prepare the result analysis of that particular semester.

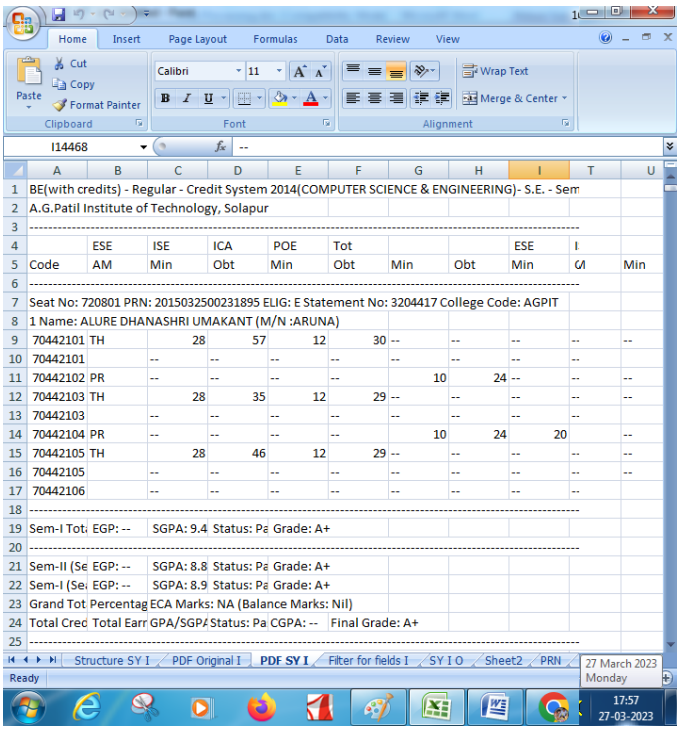

Fig.5. Data after Converting Text to Column in Excel

B. Text to Column - Fixed width

After the university result ledgers copied to Excel sheet, it may contain the data in the form of statement which we need to separate in the various columns e.g. 'Seat No: *812405 PRN: 2014032500245847 ELIG: E Statement No: 3195300 College Code: WIT' which consist of Seat No., PRN, Eligibility, Statement Number and College Name. The feature of Excel 'Text to Column - Fixed width' is used for this purpose in which fields are aligned in the columns with spaces in each field. So such data for separating the fields is described in Figure 6.

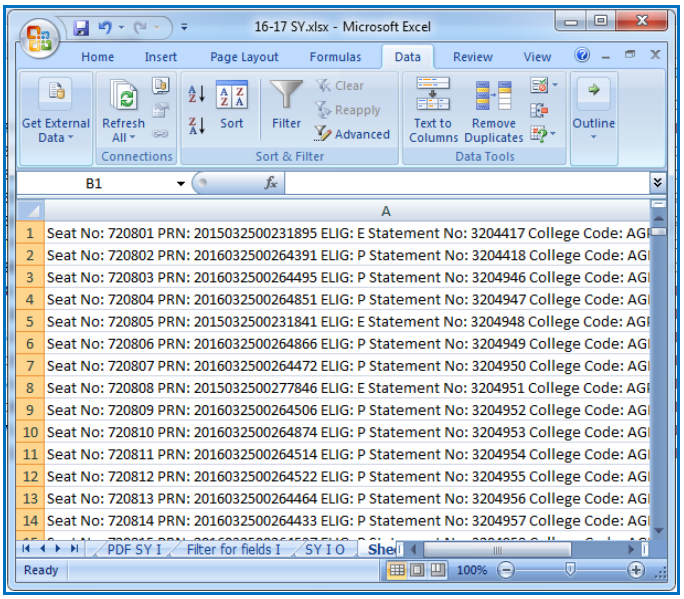

Fig.6. Data for separating fields in Excel

Following are the steps to convert 'Text to Column - Fixed width':

Step 1: Click on 'Data' tab on menu bar and select 'Text to Columns'. After selecting 'Text to Columns', click on 'Fixed Width' option as the fields of statement in Excel sheet are required to be separated. Click on 'Next' to go to the next step in 'Text to Columns' option as shown in Figure 7.

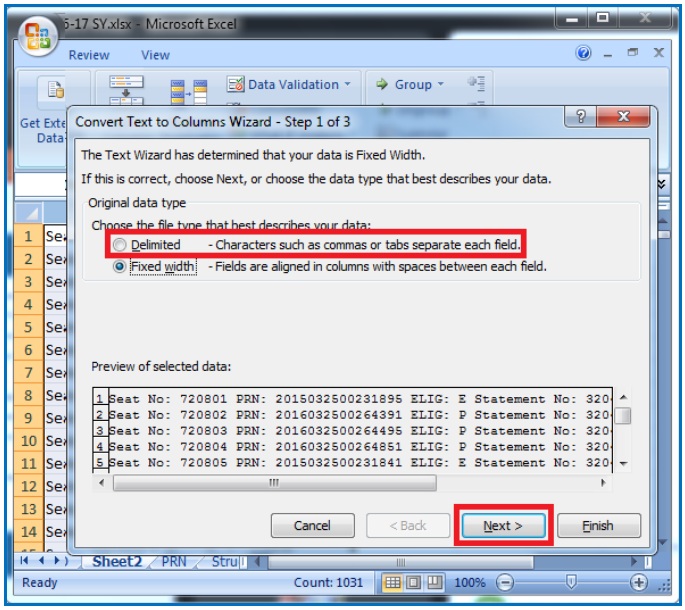

Fig.7. Seleting 'Delimited' in 'Convert Text to Column' - Fixed Width window of Excel

Step 2: The next step 'Convert Text to Columns Wizard - Step 2 of 3' is shown in Figure 8. Using this screen, field width can be set. Here, lines with arrows are used to specify the column break. The break lines are created by clicking at the desired column while these lines can be deleted by double click on line. The break lines can be moved to desired position by clicking and dragging it to that position. Click on 'Next' to go the next step. Using these break lines, the fields Seat No., PRN, Eligibility, Statement Number and College Name are separated.

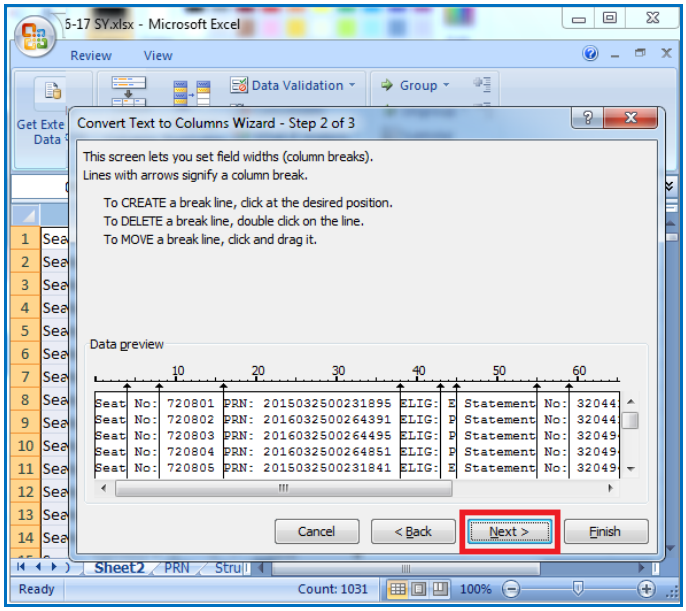

Fig.8. Seleting 'Next' in 'Convert Text to Column' - Fixed Width window of Excel

Step 3: Figure 9 shows the next step 'Convert Text to Columns Wizard - Step 3 of 3'. This step select each column and set the data format. So here in this step, 'Column data format' selected is 'General'. So 'General' converts numeric values in sheet to number and date values to date and remaining values to text. Here destination cell considered is A1 which allows neither the column nor the rows reference to change.

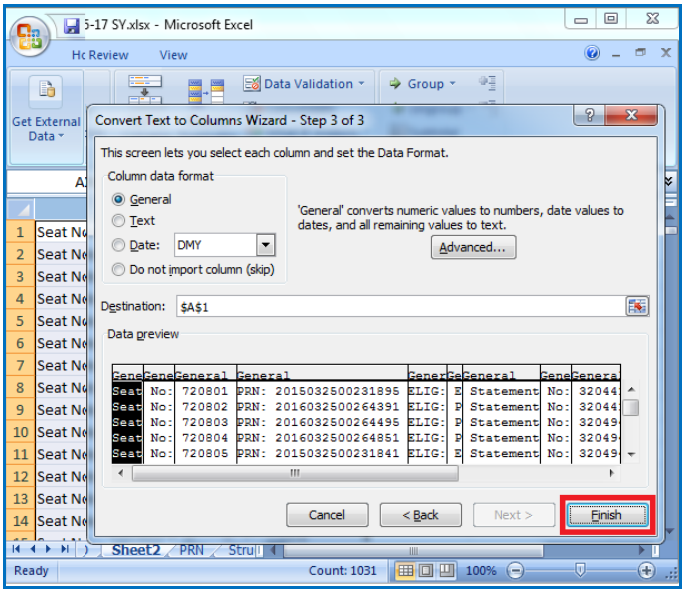

Fig.9. Clicking 'Finish' in 'Convert Text to Column' - Fixed Width window of Excel

After clicking on 'Finish' in Figure 9, the fields in statement get separated in columns as shown in Figure 10. The fields Seat No., PRN, Eligibility, Statement Number and College Name now are in separate columns.

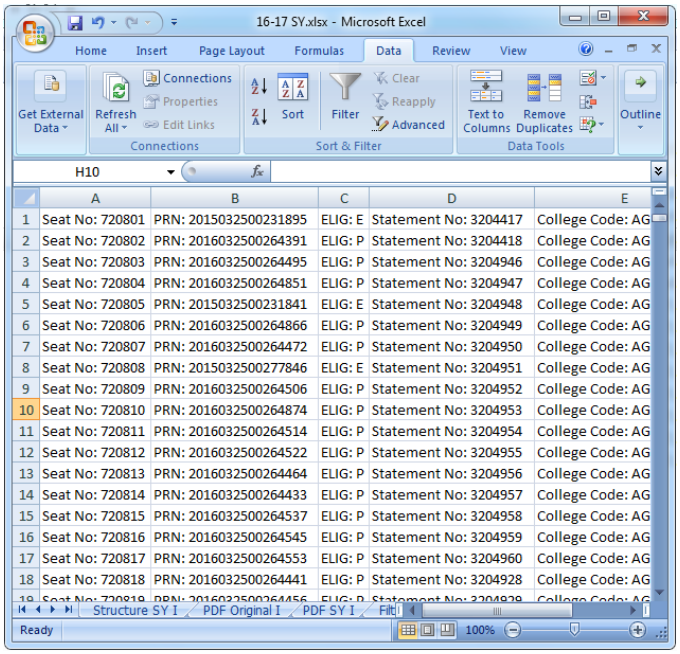

Fig.10. Fileds in statement separated to columns

C. Filter

The 'Filter' function in Excel is used to filter the range of data based on the criteria defined. How Filter option is used for data processing is explained in this section as below.

Step 1: As all fields mentioned in row number 5 are needed to add filters, hence select row number 5 as shown in Figure 11. Click on 'Data' and select 'Filter'.

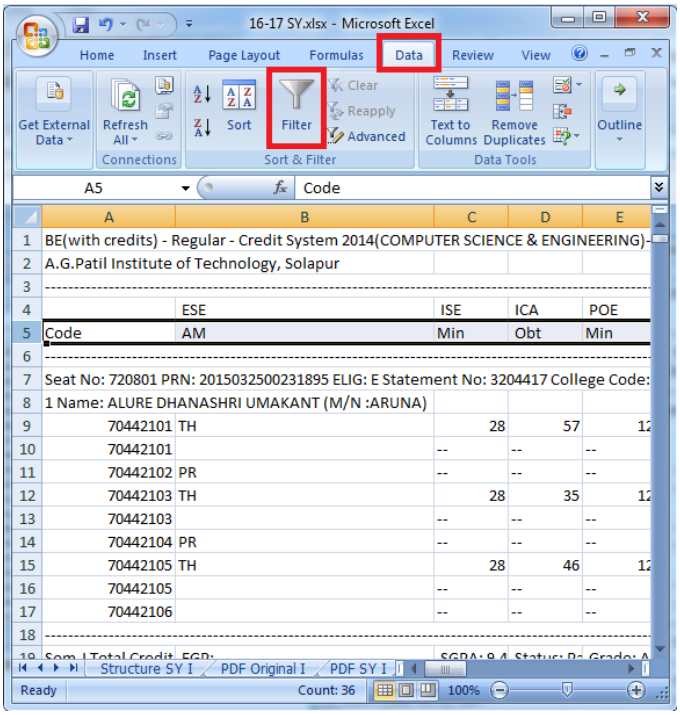

Fig.11. 'Filter' option in Excel

Step 2: After selecting 'Filter' as shown in Figure 11, Figure 12 is displayed. Click on 'Filter'.

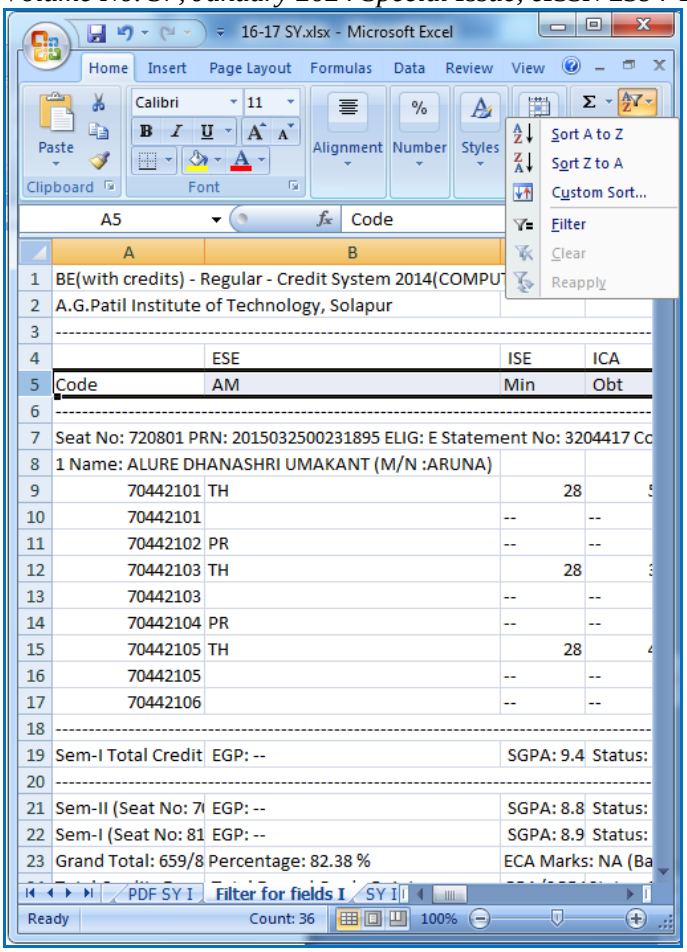

Fig.12. Screen after clicking on 'Filter' option in Excel

Step 3: As shown in Figure 13, 'Filter' option is applied to all fields mentioned in the row number 5. Now it is possible to apply the filter to any field mentioned in row number 5.

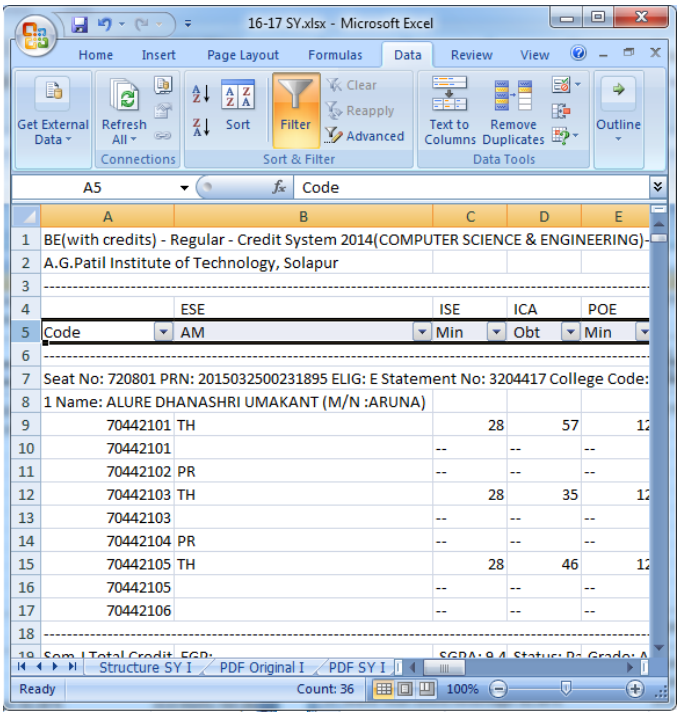

Fig.13. 'Filter' option at row number 5 in Excel

Step 4: In result analysis, each course marks such as ESE marks, ISE marks, ICA marks, POE marks, Total Marks, etc. are required to be entered in the result analysis of that particular semester, so here 'Filter' on course code is required to copy the marks of that course in Excel sheet. So Figure 14 shows how the Filter is applied to course code '70442101'.

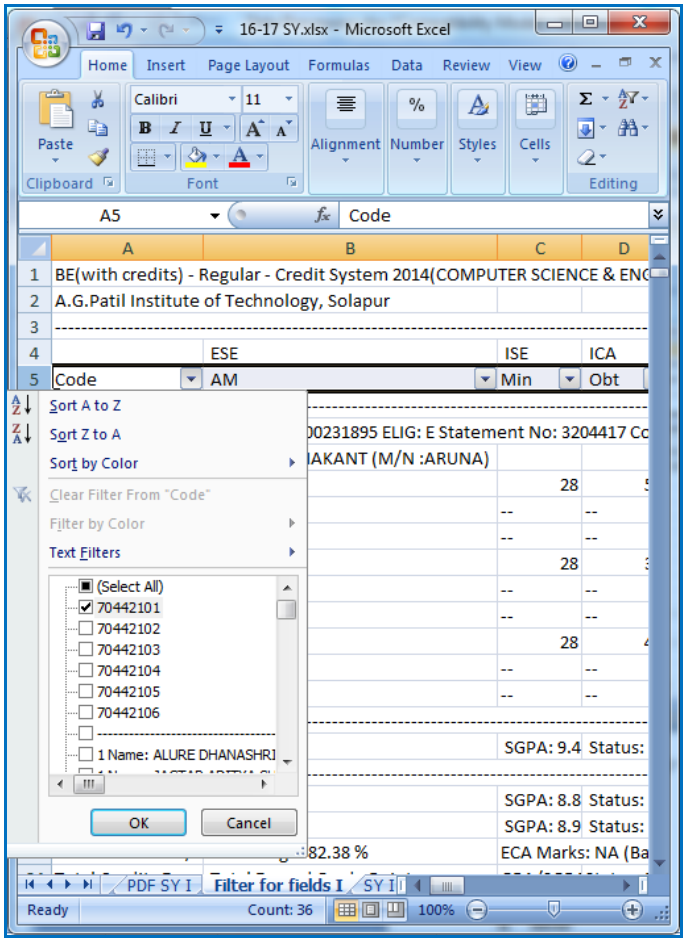

Fig.14. Selecting particular course code marks using 'Filter' option

Figure 15 shows the all marks related to the course code '70442101' that is ESE minimum marks & obtained marks, ISE minimum marks and obtained marks, ICA minimum marks & obtained marks, etc.

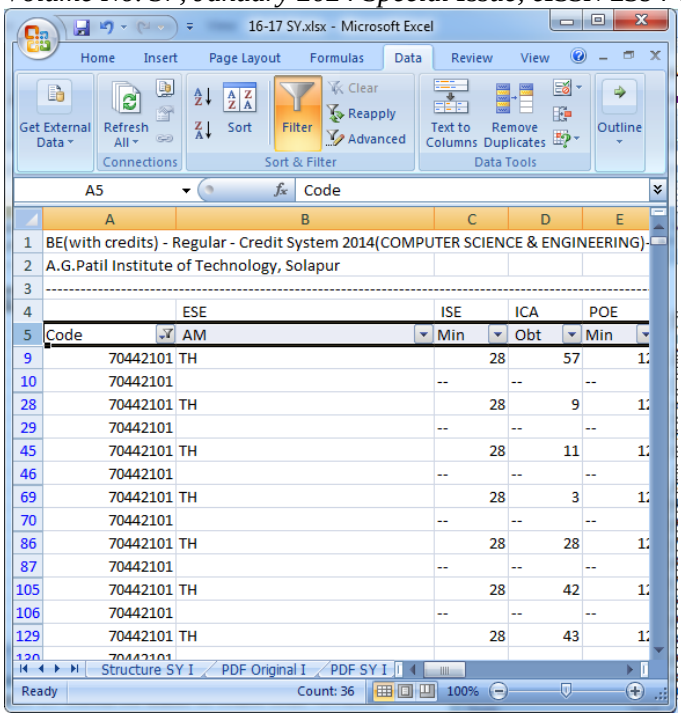

Fig.15. Screen for marks of course code 70442101

In this way, the 'Filter' option in Excel sheet is used to filter the marks of each course and prepare the data for dataset in EDM. So data filtering is used to rearrange, exclude or divide the data according to certain criteria.

D. Conditional Formatting – Highlight Cells Rules – Text that contains

Conditional formatting is used to apply particular format to a cell or range of cell and have that formatting change based on the value of formula. Suppose in column 'Sem III Status' as shown in Figure 16, the status 'ATKT' is to be highlighted. To highlight the 'ATKT' status in Red color, following steps are used.

Step 1: Click on 'Home' and select 'Styles'. In 'Styles' option, click on 'Conditional Formatting'. In 'Conditional Formatting', there are various options such as 'Highlight Cells Rules', 'Top/Bottom Rules', 'Data Bars', 'Color Scales' and 'Icon Sets'. Out of these options, click on 'Highlight Cells Rules'. In 'Highlight Cells Rules', select 'Text that conatins'.

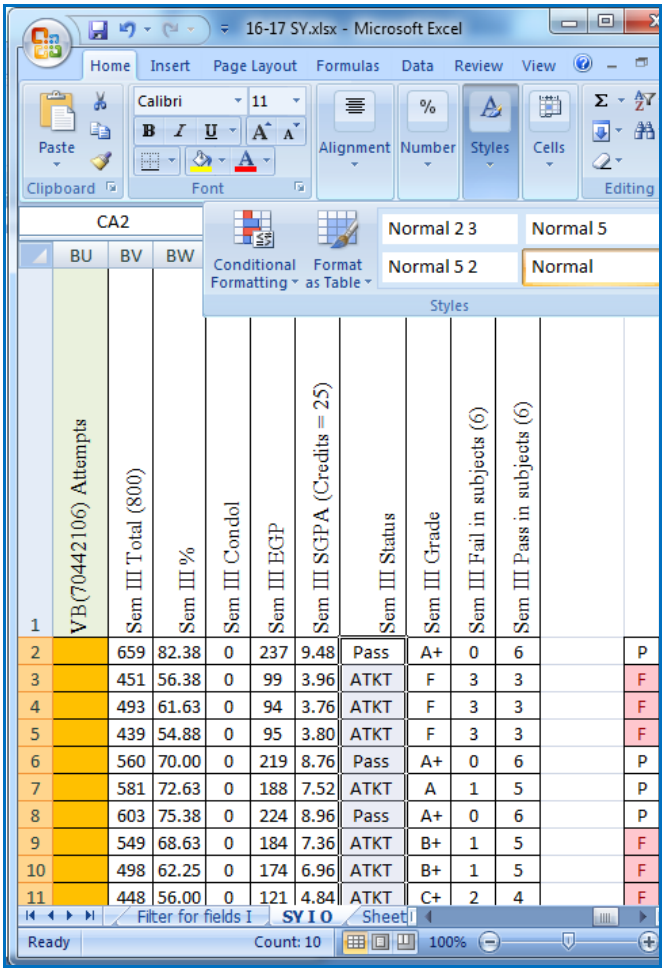

Fig.16. 'Conditional Formatting' option in Excel

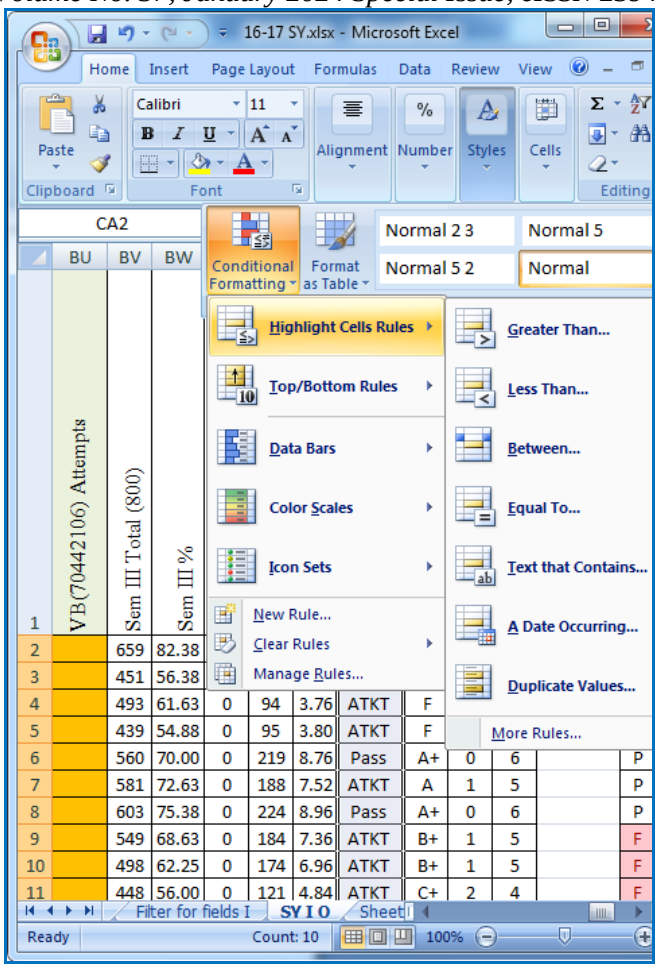

Fig.17. Selecting 'Text that Contains' in 'Conditional Formatting' option in Excel

Step 2: After clicking on 'Text that conatins', a small window as shown in Figure 18 will appear on the screen. So enter 'ATKT' in the and the color of cell in 'Format cells that contains the text'

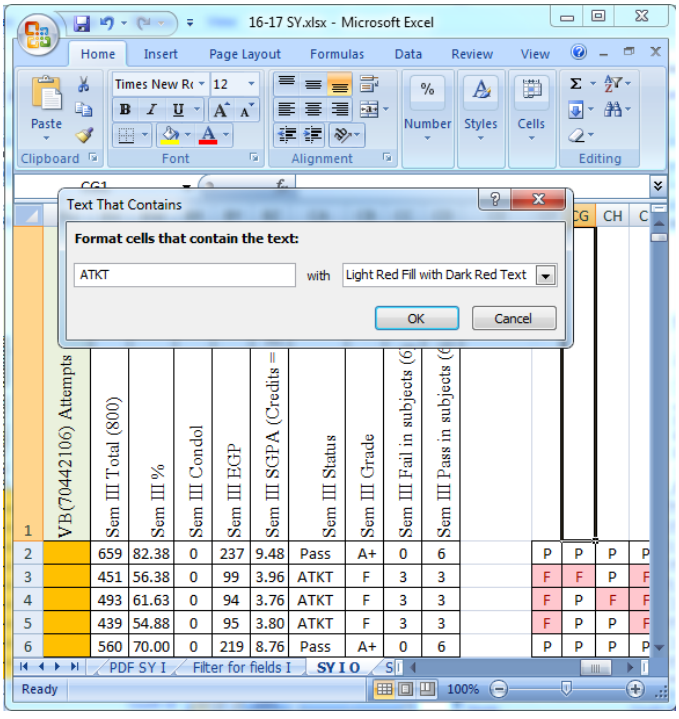

Fig.18. Selecting 'ATKT' in 'Text that Contains'

Figure 19 shows the Red color for ATKT cells in that particular column. This 'Conditional Formatting – Highlight Cells Rules – Text that contains' is used to 'Fail' cell color in particular course status (Pass/ Fail).

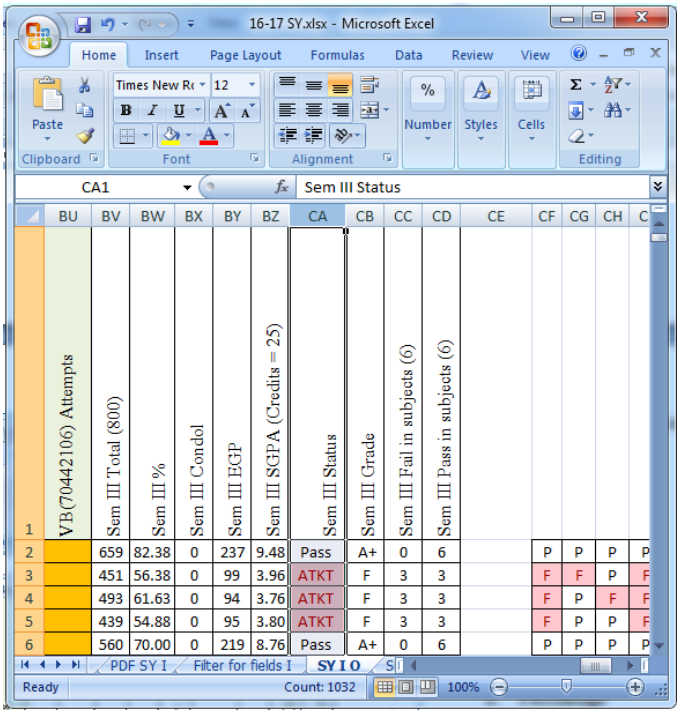

Fig.19. Column containing highlighted color for ATKT

E. Formulae used

Following formulae are used for various purpose

1) SUM

The SUM formula adds values such as individual values, cell references or ranges or mix of these three. Sum formula is used for most of attributes of dataset such as semester-wise total, yearwise total, grand total, etc.

- The formula
  '=SUM(AJS2,XN2,KL2)'
  is used to calculate the grand total marks obtained by students.
- The formula
  '=SUM(I2,S2,AG2,AV2,BG2,BP2)'
  is used to calculate the total marks obtained by the student in particular cell.

2) IF

The IF function returns one value if a condition is true otherwise return another value. This IF function is used to calculkate the grade of particular semester or year or overall grade.

Example: To calculate the grade based on the grade point average obtained by students as per Table 2, following formula is used.

'=IF(FK2>=9.5,"O",IF(FK2>=8.5,"A+",IF(FK2>=7.5,"A",IF(FK2>=6.5,"B+",IF(FK2>=5.5,"B",IF(FK2>=4.5,"C+",IF(FK2>=4,"C","F")))))))'

3) COUNTIF

The COUNTIF function counts the cells in the range that satisfy certain criteria. While preparing the dataset, this function is used to count number of passed courses and number of failed courses of particular students.

The formula =COUNTIF(FY2:GD2, " F ") and =COUNTIF(FY2:GD2, " P ") are used to check the number of failed and number of passed course respectively by particular student.

4) MOD

Sometimes in Excel sheet, after using Filter option on particular course code, result may contains marks of two courses in in even and odd rowssuch as one course code marks in odd numbered rows and other course code marks in even numbered rows. So in such case, the MOD function is used. To use MOD fuction, create one helper column and enter number from 1 to number of rows having data. To sort the odd and even rows, use the formula =MOD(A$, 2) in one more new column and drag it to number of rows having data in the sheet.

- If =MOD(A$, 2)=0 then then it will sort all even numbered rows and
- If =MOD(A$, 2)=1 then it will sort all odd numbered rows

So using Filter option, the data of two courses can be easity separate out to prepare the data for dataset in EDM

5) PERCENTAGE (%)

The PERCENTAGE (%) function is used to calculate the percentage obtained by students semester-wise, and yearwise.

V. DATASET COLLECTION AND PROCESSING METHOD

In this section, data processing method used for preparing the dataset for students' performance prediction system is explained. For data processing, Microsoft Excel sheet is considered. The data processing method is explained in Figure 1. So this method consist of following steps –

- Collection of University Result Ledgers related to CSE
- Preparation of semester-wise result analysis
- Mapping of PRN numbers of all semester result analysis
- Prepare all semester result analysis as per mapping
- Combining the semester-wise result analysis into Yearwise result analysis
- Checking yearwise result of all students for backlog (failed) courses if any
- Combining the result of all years for final dataset

A. Collection of University Result Ledgers related to CSE–

CSE result ledgers for two Academic Years 2014-15 and 2015-16 were collected from Punyashlok Ahilyadevi Holkar University, Solapur and URL is http://www.sus.ac.in/examination/Online-Result-(Ledger) or https://su.digitaluniversity.ac/Content.aspx?ID=29445

B. Preparation of semester-wise result analysis –

Uniersity result ledger of odd sesmester contains of result of all courses of odd semester only but Uniersity result ledger of even sesmester comprise of result of all courses related to odd as well as even semester. Hence semester-wise result analysis such as Second Year CSE SY Sem-III & SY Sem-III IV, Third Year CSE TY Sem-V & TY Sem-VI and Final Year CSE Sem VII & Final Year Sem-VII VIII is prepared using various features of Microsoft Excel sheet. First year result analysis is not considered as this students' performance prediction system is related to CSE courses. For example, Second Year CSE Sem III Sem-III & SY Sem-III IV Excel sheet consist of various sheet such as

1. Structure of SY III sheet which contains the information about course code of each course, minimum and maximum marks required in ESE, ISE, ICA, and POE, grade information, grade point average information, etc.
2. Original PDF SY III sheet which contains copied content of university result ledger of SY CSE Semester III to Excel sheet.
3. PDF SY III which contain the data obtained after applying the 'Text to Column – Delimited' feature of Excel to 'Original PDF SY III' sheet.
4. Filter for Fields III sheet which is used to copy the data after applying 'Filter' option course codewise.
5. SY III Original sheet which contains the result analysis of all courses related to SY Semester III after applying the 'Filter' option to each course in 'Filter for Fields III' sheet and copying the data to this sheet. Result analysis is prepared based on Permanent Registration Number (PRN) of students.
6. Structure of SY IV sheet which contains the information about course code of Semester III & IV courses, minimum and maximum marks required in ESE, ISE, ICA, and POE, grade information, grade point average information, etc.
7. Original PDF SY IV sheet which contains the copied content of university result ledger of SY CSE Semester III & IV to Excel sheet.
8. PDF SY IV which contain the data obtained after applying the 'Text to Column – Delimited' feature of Excel to 'Original PDF SY IV' sheet.
9. Filter for Fields IV sheet which is used to copy the data after applying 'Filter' option course codewise.
10. SY IV Original sheet which contains the result analysis of all courses related to SY Semester III & sIV after applying the 'Filter' option to each course in 'Filter for Fields IV' sheet and copying the data to this sheet.

All these 10 sheets will be for Third Year CSE TY Sem-V & TY Sem-VI and Final Year CSE Sem VII & Final Year Sem-VII VIII

C. Mapping of PRN numbers of all semester result analysis –

While preparing the result analysis of SY Sem-III, SY Sem-III IV, TY Sem-V, TY Sem-V VI, Final Year Sem-VII, and Final Year Sem-VII VIII, PRN number order is not same, so mapping of PRN of all result analysis is required before proceeding further. So serial number and PRN number from result analysis of SY Sem-III, SY Sem-III IV, TY Sem-V, TY Sem-V VI, Final Year Sem-VII, and Final Year Sem-VII VIII are copied in the Excel sheet. For mapping of PRN numbers, particular PRN number is taken, and checked whether that PRN number is there in all semesters PRN number. If it is available then number 1 is replaced in all such columns. So same procedure is followed for all PRN numbers. If particular

PRN number is not available in one or more columns but less than six column then such PRN number is not replaced by the number in sequence. In this way mapping of PRN number to new sequence number is required to collect the result analysis of SY Sem-III, SY Sem-III IV, TY Sem-V, TY Sem-V VI, Final Year Sem-VII, and Final Year Sem-VII VIII of each student.

D. Prepare all semester result analysis as per mapping –

After mapping PRN, all result analysis is sorted as per the new sequence number so that result analysis of particular student for all six semesters must be available in particular cell of each result analysis sheet. For example if PRN number 1234567890 is available in row number 10 of result analysis of SY Sem-III, then this PRN number should be available in row number 10 of all other result analysis of SY Sem-III IV, TY Sem-V, TY Sem-V VI, Final Year Sem-VII, and Final Year Sem-VII VIII.

E. Combining the semester-wise result analysis into Yearwise result analysis –

In this step, result analysis of SY Sem-III, and SY Sem-III IV are combined into one as SY Sem-III IV while TY Sem-V and TY Sem-V VI as TY Sem-V VI. Combining of Final Year Sem-VII, and Final Year Sem-VII VIII results in Final Year Sem-VII VIII.

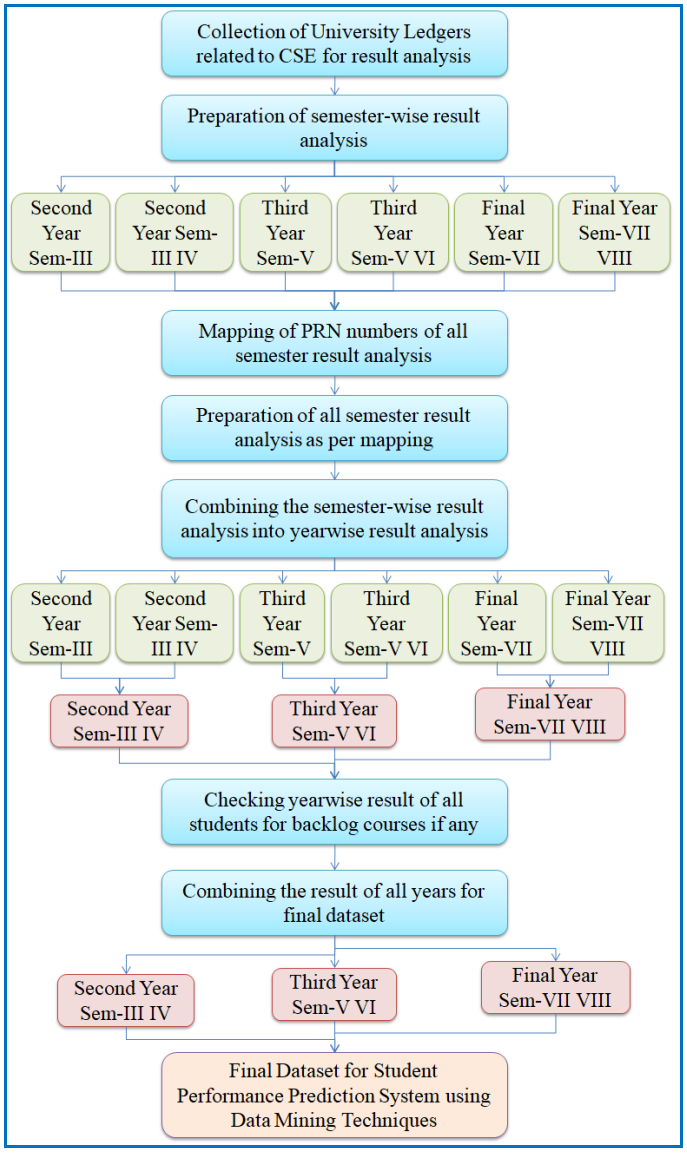

Fig.19. Data processing method for students' performance prediction system

F. Checking yearwise result of all students for backlog (failed) courses if any –

If student is failed in particular course of particular semester then that student's result for failed course of that semester is checked in the respective University Result Ledger and marks are entered for that particular course in particular semester result analysis sheet along with number of attempts required to pass that particular course. If student passes the course or POE in first attempt then number of attempts required to pass the course or POE will be one otherwise number of attempts for the course or POE will be depend upon number of attempts taken by student to pass the course or POE .

For example,

- If particular student A is failed in the course Theory of Computation of SY Sem-II held in May 2017 university exam then the result of that student A for that course needs to be checked in the university result ledger – 'SY Sem-II held in October 2017'.
- If that student is passed that course in October 2017 then number of attempts required to pass the course Theory of Computation will be two.

- If that student A is failed to clear the course Theory of Computation in university exame 'SY Sem-II held in October 2017' then the result of that student A need to be checked in university result ledger 'SY Sem-II held in May 2018'.
- If that student A is passed this course in 'SY Sem-II held in May 2018' then number of attempts required to pass the course Theory of Computation will be three.

In this way, the result of failed students need to be checked in respective university result ledgers and accordingly, number of attempts required to pass that course require to be entered in the respective sheet. Same procedure is considered if students gets failed in Practical Oral Examination.

G. Combining the result of all years for final dataset –

Finally all result analysis sheets SY Sem-III IV, TY Sem-V VI, and Final Year Sem-VII VIII are combined into one sheet. This sheet contains the result of all semester III, IV, V, VI, VII and VIII of each student.

VI. ANALYSIS OF PREPARED DATASET FOR STUDENT PERFORMANCE PREDICTION SYSTEM

Table 9 shows the data analysis at each stage of data processing. This Table contains the analysis of First Year Sem-I to Final Year Sem-VIII of Academic Year 2014-15 and 2015-16 original data collected from university result ledger.

TABLE IX
ORIGINAL DATA COLLECTION ANALYSIS

Year and Semester	2014-15 Number of Students	2015-16 Number of Students	Total Number of Students	Number of Attributes
FY-I	248	457	705	90
FY-I-II	235	439	674	100
SY-I	737	1031	1768	82
SY-I-II	722	1002	1724	90
TY-I	664	774	1438	90
TY-I-II	661	800	1461	102
BE-I	650	758	1408	89
BE-I-II	678	760	1438	91
Total	4595	6021	10616	544

Total of 4595 and 6021 student data of Academic Year 2014-15 and 2015-16 are considered for preparing the dataset. Out of 10616 total data, 1379 data of First Year Semester I and II is ignored as the students' performance prediction system is for CSE data. The graphical representation and analysis of original data collection is shown in Figure 20. This Figure shows the total number of students for academic year 2014-15, total number of students for academic year 2015-16 and number of attributes considered for the First Year Semester-I, First Year Semester-I II, Second Year Semester-III, Second Year Semester-III IV, Third Year Semester-V, Third Year Semester-V VI, Final Year Semester-VII, and Final Year Semester-VII VIII.

After preparing the result analysis of Second Year Semester-I, Second Year Semester-III IV, Third Year Semester-V, Third Year Semester-V VI, Final Year Semester-VII, and Final Year Semester-VII VIII, PNR number in all result analysis are not in the same order. So mapping of PRN numbers of all these result analysis needs to be done before proceeding further to prepare the dataset. So Table 10 shows the data analysis after initial processing and mapping of PRN number. So total 10100 students data is obtained in this step.

TABLE X
DATA ANALYSIS AFTER INITIAL PROCESSING AND MAPPING OF PRN NUMBERS

Year and Semester	2014-15 Number of Students	2015-16 Number of Students	Total Number of Students	Number of Attributes
FY-I	-	-	-	-
FY-I-II	-	-	-	-
SY-I	720	963	1683	82
SY-I-II	720	963	1683	90
TY-I	720	963	1683	90
TY-I-II	720	963	1683	102
BE-I	720	965	1685	89
BE-I-II	720	963	1683	91
Total	4320	5780	10100	544

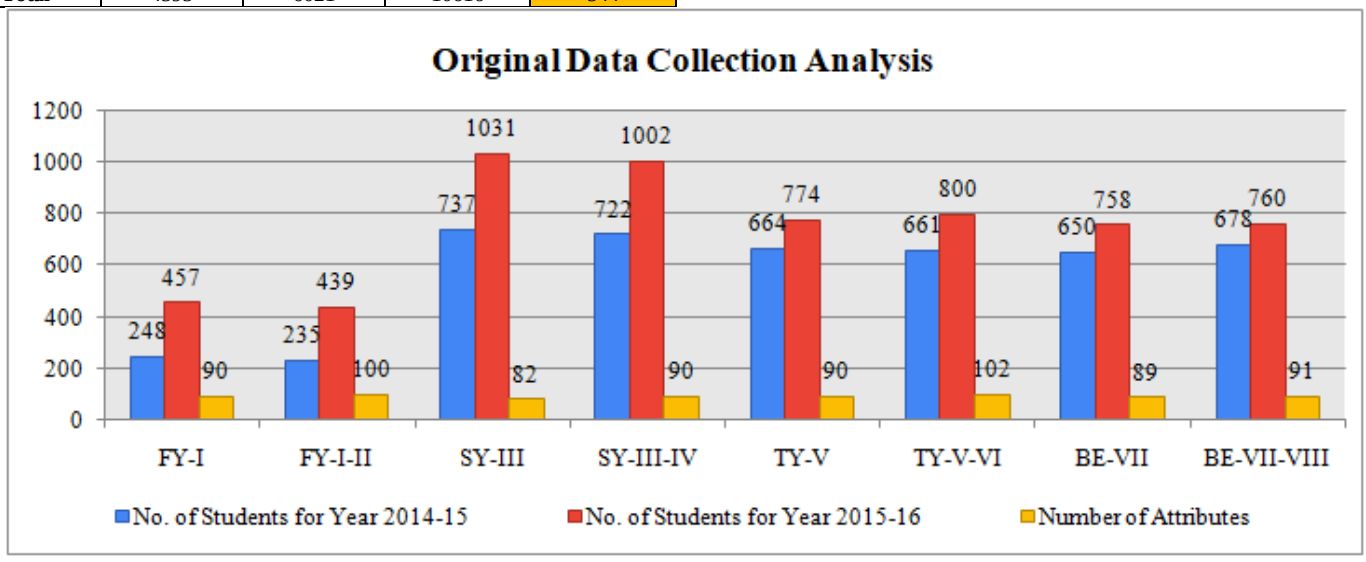

Fig.20. Analysis of Original data

So 720 and 963 students data for academic year 2014-15 and 2015-16 respectively are obtained for all semester Second Year Semester-III, Second Year Semester-III IV, Third Year Semester-V, Third Year Semester-V VI, Final Year Semester-VII, and Final Year Semester-VII VIII result analysis. Figure 21 shows graphically the data analysis after initial processing and mapping of PRN numbers.

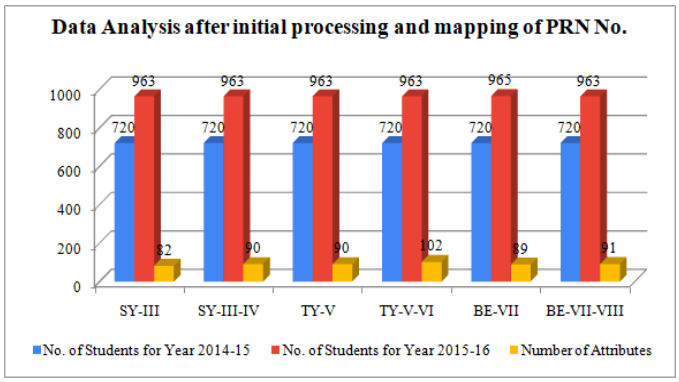

Fig.21. Data analysis after initial processing and mapping of PRN number.

After initial processing and mapping of PRN numbers of all semester result analysis, all result analysis in Excel sheet are sorted as per PRN mapping. Semester I and II result analysis of Second Year, Third Year and Final Year are combined to form yearwise result analysis Second Year III IV, Third Year V VI and Final Year VII VIII.

TABLE XI
DATA ANALYSIS AFTER COMBINING SEMESTER-I AND II RESULT ANALYSIS OF EACH YEAR

Year and Semester	2014-15 Number of Students	2015-16 Number of Students	Total Number of Students	Number of Attributes
FY-I	-	-	-	-
FY-I-II	-	-	-	-
SY-I	571	671	1242	303
SY-I-II	571	671	1242	
TY-I	571	671	1242	343
TY-I-II	571	671	1242	
BE-I	571	671	1242	333
BE-I-II	571	671	1242	
Total	3426	4026	7452	978

As per Table 11, total 7452 students data (total 3426 students data of academic year 2014-15 and total 4026 students data of academic year 2015-16) is derived after combining Semester I and II of Second Year, Third Year and Final Year. Figure 22 illustrates graphically the data analysis after combining Semester I and II result analysis into yearly result analysis. This Figure also indicate the total number of students for academic year 2014-15, total number of students for academic year 2015-16 and number of attributes considered for the Second Year Semester-III IV, Third Year Semester-V VI, and Final Year Semester-VII VIII result analysis.

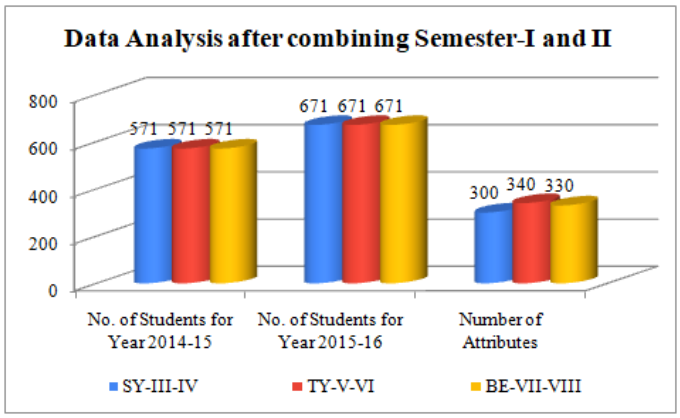

Fig.22. Data analysis after combining Semester I and II of each year of Academic Year 2014-15 and 2015-16

After combining the result analysis of Semester I and II of each year into yearly result analysis, if particular students is failed in the particular course of particular semester (Second Year III & IV, Third Year V & VI and Final Year VII & VIII) of particular year (Second, Third, or Final Year) then that student's result is checked in the respective university result ledger. If that students is passed then updated marks are entered in respective result analysis Excel sheet along with number of attempts required to pass that course by that particular student. Table 12 describe the data analysis after checking the result of the failed courses.

TABLE XII
DATA ANALYSIS AFTER CHECKING THE RESULT OF FAILED COURSES

Year and Semester	2014-15 Number of Students	20s15-16 Number of Students	Total Number of Students	Number of Attributes
FY-I	-	-	-	-
FY-I-II	-	-	-	-
SY-I	549	602	1151	303
SY-I-II	549	602	1151	
TY-I	549	602	1151	343
TY-I-II	549	602	1151	
BE-I	549	602	1151	333
BE-I-II	549	602	1151	
Total	3294	3612	6906	970

Total 6906 student data of academic year 2014-15 and academic year 2015-16 are obtained with number of attributes 303, 343 and 333 related to Second Year, Third Year and Final Year respectively. Figure 23 shows the data analysis graphically after checking the result of failed student in particular courses. Even if the result if single failed course for particular student is not found then that particular student data is deleted from result analysis.

Figure 24 gives the summary of final dataset analysis. From Figure 24, it is observed that original students data of 4594 and 6021 for academic year 2014-15 and academic year 2015-16 is collected with number of attributes 544. It is also noted from Figure 24 that final dataset conisist of 3294 and 3612 student data for academic year 2014-15 and academic year 2015-16 is collected with number of attributes 970.

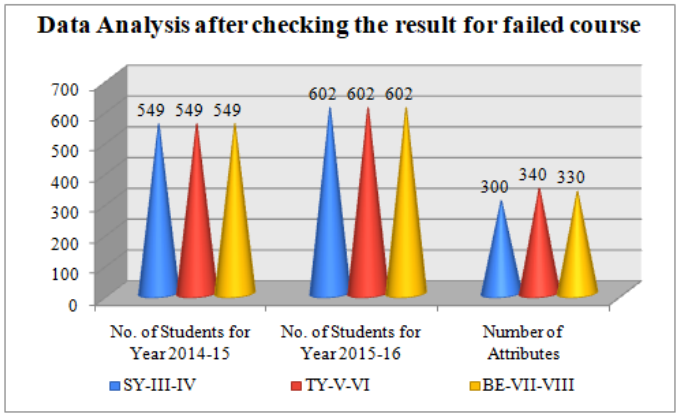

Fig.23. Data analysis after checking the result of failed course

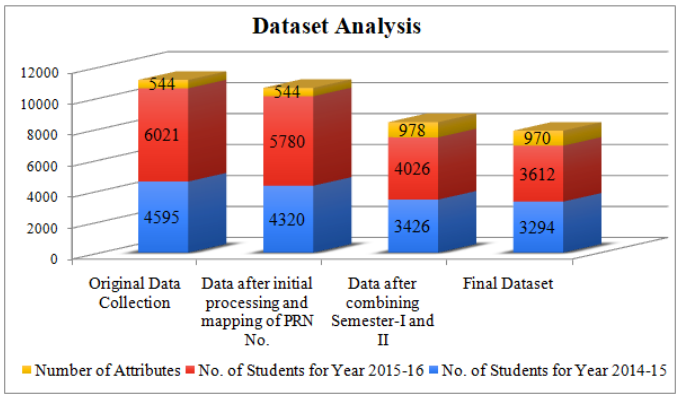

Fig.24. Final dataset analysis

VII. CONCLUSION

In this research article, two important steps of research methodology that is data collection and data processing are explained in detailed. These two steps are described related to data collected from University site related to four years engineering course of CSE stream to predict students' performance in Educational Data Mining. For data collection and processing, two Academic Year data 2014-15 and 2015-16 of Four Year CSE course that is First Year, Second Year, Third Year and Final Year are considered from university site. In data collection step, the detailed information about Syllabus structure, credit system pattern, types of assessment methods and attributes required to prepare the dataset are descrtibed. Result analysis of First Year, Second Year, Third Year and Final Year CSE stream is prepared based on Permanent Registration Number (PRN) of students. So approximately 10600 students data from Sem-I to Sem-VIII is preaperd in data collection step with 544 attributes. Data collection and processing method is also discussed in detailed which consist of seven steps - Collection of University Result Ledgers related to CSE, Preparation of semester-wise result analysis, Mapping of PRN numbers of all semester result analysis, Prepare all semester result analysis as per mapping, Combining the semester-wise result analysis into Yearwise result analysis, Checking yearwise result of all students for backlog (failed) courses if any, and Combining the result of all years for final dataset. Analysis of preapered dataset is also considered for

- Original Data Collection
- Data after initial processing and mapping of PRN No.
- Data after combining Semester-I and II
- Final Dataset

So final dataset prepared using Excel consist of five years data of two Academic Years 2014-15 and 2015-16, and 6906 students record from Semester-III to VIII with 970 number of attributes.

In future, we are plnning to use Tableau – data visualization tool for more data analysis and graphical representation.

REFERENCESS

Hussain, S., Ayoub, M., Jilani, G., Yu, Y., Khan, A., Wahid, J. A., ... & Weiyan, H. (2022). Aspect2Labels: A novelistic decision support system for higher educational institutions by using multi-layer topic modelling approach. Expert Systems with Applications, 209, 118119.

Teoh, C. W., Ho, S. B., Dollmat, K. S., & Chai, I. (2022, February). An Evolutionary Algorithm-Based Optimization Ensemble Learning Model for Predicting Academic Performance. In 2022 11th International Conference on Software and Computer Applications (pp. 102-107).

Ma, X., Qu, J. H., Xu, H. M., & Ling, Y. T. (2021, May). E-learning performance prediction based on attention mechanism. In 2021 the 6th International Conference on Distance Education and Learning (pp. 152-156).

Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. Materials Today: Proceedings, 47, 5260-5267.

Malini, J., & Kalpana, Y. (2021). Investigation of factors affecting student performance evaluation using education materials data mining technique. Materials Today: Proceedings, 47, 6105-6110.

Lottering, R., Hans, R., & Lall, M. (2020, August). A model for the identification of students at risk of dropout at a university of technology. In 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD) (pp. 1-8). IEEE.

Utari, M., Warsito, B., & Kusumaningrum, R. (2020, June). Implementation of Data Mining for Drop-Out Prediction using Random Forest Method. In 2020 8th International Conference on Information and Communication Technology (ICoICT) (pp. 1-5). IEEE.

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. IEEE Access, 8, 55462-55470.

Rahman, M., & Mahmud, A. (2020, January). Classification on Educational Performance Evaluation Dataset using Feature Extraction Approach. In Proceedings of the International Conference on Computing Advancements (pp. 1-6)

Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Multi-split optimized bagging ensemble model selection for multi-class educational data mining. Applied Intelligence, 50(12), 4506-4528.

Karthikeyan, V. G., Thangaraj, P., & Karthik, S. (2020). Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation. Soft Computing, 24(24), 18477-18487.

Adekitan, A. I., & Salau, O. (2020). Toward an improved learning process: the relevance of ethnicity to data mining prediction of students' performance. SN Applied Sciences, 2(1), 1-15.

El Aissaoui, O., El Madani, Y. E. A., Oughdir, L., Dakkak, A., & El Allioui, Y. (2020). Mining Learners' Behaviors: An Approach Based on Educational Data Mining Techniques. In Embedded Systems and Artificial Intelligence (pp. 655-670). Springer, Singapore.

Agrawal, R., Singh, J., & Ghosh, S. M. (2020). Performance Appraisal of an Educational Institute Using Data Mining Techniques. In Computing in Engineering and Technology (pp. 733-745). Springer, Singapore.

Ashraf, M., Zaman, M., & Ahmed, M. (2020). An intelligent prediction system for educational data mining based on ensemble and filtering approaches. Procedia Computer Science, 167, 1471-1483.

Almutairi, S., Shaiba, H., & Bezbradica, M. (2019, December). Predicting Students' Academic Performance and Main Behavioral Features Using Data Mining Techniques. In International Conference on Computing (pp. 245-259). Springer, Cham.

Al Breiki, B., Zaki, N., & Mohamed, E. A. (2019, November). Using Educational Data Mining Techniques to Predict Student Performance. In 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA) (pp. 1-5). IEEE.

Lagman, A. C., Calleja, J. Q., Fernando, C. G., Gonzales, J. G., Legaspi, J. B., Ortega, J. H. J. C., ... & Santos, R. C. (2019, November). Embedding naïve Bayes algorithm data model in predicting student graduation. In Proceedings of the 3rd International Conference on Telecommunications and Communication Engineering (pp. 51-56).

Crivei, L. M., Czibula, G., & Mihai, A. (2019, August). A Study on Applying Relational Association Rule Mining Based Classification for Predicting the Academic Performance of Students. In International Conference on Knowledge Science, Engineering and Management (pp. 287-300). Springer, Cham.

Amazona, M. V., & Hernandez, A. A. (2019, May). Modelling Student Performance Using Data Mining Techniques: Inputs for Academic Program Development. In Proceedings of the 2019 5th International Conference on Computing and Data Engineering (pp. 36-40).

Altaf, S., Soomro, W., & Rawi, M. I. M. (2019, April). Student Performance Prediction using Multi-Layers Artificial Neural Networks: A Case Study on Educational Data Mining. In Proceedings of the 2019 3rd International Conference on Information System and Data Mining (pp. 59-64).

Martins, M. P., Miguéis, V. L., Fonseca, D. S. B., & Alves, A. (2019, February). A data mining approach for predicting academic success–A case study. In International Conference on Information Technology & Systems (pp. 45-56). Springer, Cham

Jalota, C., & Agrawal, R. (2019, February). Analysis of educational data mining using classification. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (pp. 243-247). IEEE.

Tasnim, N., Paul, M. K., & Sattar, A. S. (2019, February). Identification of drop out students using educational data mining. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.

Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. Heliyon, 5(2), e01250.

Dimić, G., Rančić, D., Pronić-Rančić, O., & Milošević, D. (2019). An approach to educational data mining model accuracy improvement using histogram discretization and combining classifiers into an ensemble. In Smart Education and e-Learning 2019 (pp. 267-280). Springer, Singapore.

Akram, A., Fu, C., Li, Y., Javed, M. Y., Lin, R., Jiang, Y., & Tang, Y. (2019). Predicting students' academic procrastination in blended learning course using homework submission data. IEEE Access, 7, 102487-102498.

Santoso, L. W. (2019). The analysis of student performance using data mining. In Advances in Computer Communication and Computational Sciences (pp. 559-573). Springer, Singapore.

Rawat, K. S., & Malhan, I. V. (2019). A hybrid classification method based on machine learning classifiers to predict performance in educational data mining. In Proceedings of 2nd International Conference on Communication, Computing and Networking (pp. 677-684). Springer, Singapore.

Rojanavasu, P. (2019). Educational data analytics using association rule mining and classification. In 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON) (pp. 142-145). IEEE.

Ajibade, S. S. M., Ahmad, N. B., & Shamsuddin, S. M. (2018, December). A data mining approach to predict academic performance of students using ensemble techniques. In International Conference on Intelligent Systems Design and Applications (pp. 749-760). Springer, Cham.

Abyaa, A., Idrissi, M. K., & Bennani, S. (2018, October). Predicting the learner's personality from educational data using supervised learning. In Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications (pp. 1-7).

Kaunang, F. J., & Rotikan, R. (2018, October). Students' Academic Performance Prediction using Data Mining. In 2018 Third International Conference on Informatics and Computing (ICIC) (pp. 1-5). IEEE.

Martínez-Abad, F., Gamazo, A., & Rodríguez-Conde, M. J. (2018, October). Big data in education: detection of ICT factors associated with school effectiveness with data mining techniques. In Proceedings of the sixth

international conference on technological ecosystems for enhancing multiculturality (pp. 145-150).

Ibrahim, Z. M., Bader-El-Den, M., & Cocea, M. (2018, September). Mining unit feedback to explore students' learning experiences. In UK Workshop on Computational Intelligence (pp. 339-350). Springer, Cham.

Vila, D., Cisneros, S., Granda, P., Ortega, C., Posso-Yépez, M., & García-Santillán, I. (2018, August). Detection of desertion patterns in university students using data mining techniques: A case study. In International Conference on Technology Trends (pp. 420-429). Springer, Cham.

Kasthuriarachchi, K. T. S., & Liyanage, S. R. (2018, July). Predicting Students' Academic Performance Using Utility Based Educational Data Mining. In International Conference on Frontier Computing (pp. 29-39). Springer, Singapore.

Spatiotis, N., Perikos, I., Mporas, I., & Paraskevas, M. (2018, July). Evaluation of an educational training platform using text mining. In Proceedings of the 10th Hellenic Conference on Artificial Intelligence (pp. 1-5).

Srivastava, S., Karigar, S., Khanna, R., & Agarwal, R. (2018, July). Educational Data Mining: Classifier Comparison for the Course Selection Process. In 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE) (pp. 1-5). IEEE.

Zaffar, M., Hashmani, M. A., & Savita, K. S. (2018, June). Comparing the performance of FCBF, Chi-Square and relief-F filter feature selection algorithms in educational data mining. In International Conference of Reliable Information and Communication Technology (pp. 151-160). Springer, Cham.

Rustia, R. A., Cruz, M. M. A., Burac, M. A. P., & Palaoag, T. D. (2018, February). Predicting Student's Board Examination Performance using Classification Algorithms. In Proceedings of the 2018 7th International Conference on Software and Computer Applications (pp. 233-237).

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Computers & Electrical Engineering, 66, 541-556.

Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. Decision Support Systems, 115, 36-51.

Jung, E. (2018). An Educational Data Mining with Bayesian Networks for Analyzing Variables Affecting Parental Attachment. In Advances in Computer Science and Ubiquitous Computing (pp. 557-563). Springer, Singapore.

Maitra, S., Madan, S., Kandwal, R., & Mahajan, P. (2018). Mining authentic student feedback for faculty using Naïve Bayes classifier. Procedia computer science, 132, 1171-1183.

Lagus, J., Longi, K., Klami, A., & Hellas, A. (2018). Transfer-learning methods in programming course outcome prediction. ACM Transactions on Computing Education (TOCE), 18(4), 1-18.

Ayub, M., Toba, H., Wijanto, M. C., & Yong, S. (2017, November). Modelling online assessment in management subjects through educational data mining. In 2017 International Conference on Data and Software Engineering (ICoDSE) (pp. 1-6). IEEE.

Buenaño-Fernández, D., Luján-Mora, S., & Villegas-Ch, W. (2017, October). Improvement of massive open online courses by text mining of students' emails: a case study. In Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 1-7).

Figueira, Á. (2017, October). Mining Moodle logs for grade prediction: a methodology walk-through. In Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 1-8).

Leppänen, L., Leinonen, J., Ihantola, P., & Hellas, A. (2017, September). Predicting academic success based on learning material usage. In Proceedings of the 18th Annual Conference on Information Technology Education (pp. 13-18).

Kularbphettong, K. (2017, September). Analysis of students' behavior based on educational data mining. In Proceedings of the Computational Methods in Systems and Software (pp. 167-172). Springer, Cham.

Athani, S. S., Kodli, S. A., Banavasi, M. N., & Hiremath, P. S. (2017, May). Student academic performance and social behavior predictor using data mining techniques. In 2017 International Conference on Computing, Communication and Automation (ICCCA) (pp. 170-174). IEEE.

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017, April). Predicting student performance using advanced learning analytics. In Proceedings of the 26th international conference on world wide web companion (pp. 415-421).

Castro-Wunsch, K., Ahadi, A., & Petersen, A. (2017, March). Evaluating neural networks as a method for identifying students in need of assistance. In Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education (pp. 111-116).

Pise, N., & Kulkarni, P. (2017). Evolving learners' behavior in data mining. Evolving Systems, 8(4), 243-259.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior, 73, 247-256.

Lehr, S., Liu, H., Kinglesmith, S., Konyha, A., Robaszewska, N., & Medinilla, J. (2016, July). Use educational data mining to predict undergraduate retention. In 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT) (pp. 428-430). IEEE.

Sanchez-Santillan, M., Paule-Ruiz, M., Cerezo, R., & Nuñez, J. (2016, April). Predicting students' performance: Incremental interaction classifiers. In Proceedings of the Third (2016) ACM Conference on Learning@ Scale (pp. 217-220).

Devasia, T., Vinushree, T. P., & Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) (pp. 91-95). IEEE.

Chaudhury, P., Mishra, S., Tripathy, H. K., & Kishore, B. (2016, March). Enhancing the capabilities of student result prediction system. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (pp. 1-6).

Amornsinlaphachai, P. (2016, February). Efficiency of data mining models to predict academic performance and a cooperative learning model. In 2016 8th International Conference on Knowledge and Smart Technology (KST) (pp. 66-71). IEEE.

Agaoglu, M. (2016). Predicting instructor performance using data mining techniques in higher education. IEEE Access, 4, 2379-2387.

Ramanathan, L., Geetha, A., Khalid, M., & Swarnalatha, P. (2016). Apply of sum of difference method to predict placement of students' using educational data mining. In Information Systems Design and Intelligent Applications (pp. 367-377). Springer, New Delhi.

Stahovich, T. F., & Lin, H. (2016). Enabling data mining of handwritten coursework. Computers & Graphics, 57, 31-45.

Badr, G., Algobail, A., Almutairi, H., & Almutery, M. (2016). Predicting students' performance in university courses: a case study and tool in KSU mathematics department. Procedia Computer Science, 82, 80-89.

Hassan, S. M., & Al-Razgan, M. S. (2016). Pre-university exams effect on students GPA: a case study in IT department. Procedia Computer Science, 82, 127-131.

Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. J. (2016). Student academic performance prediction model using decision tree and fuzzy genetic algorithm. Procedia Technology, 25, 326-332.

Ahmed, A. M., Rizaner, A., & Ulusoy, A. H. (2016). Using data mining to predict instructor performance. Procedia Computer Science, 102, 137-142.

Kassak, O., Kompan, M., & Bielikova, M. (2016). Student behavior in a web-based educational system: Exit intent prediction. Engineering Applications of Artificial Intelligence, 51, 136-149.

Jung, E. (2016). A comparison of data mining methods in analyzing educational data. In Advances in Computer Science and Ubiquitous Computing (pp. 173-178). Springer, Singapore

Salinas, J. G. M., & Stephens, C. R. (2015, October). Applying data mining techniques to identify success factors in students enrolled in distance learning: a case study. In Mexican International Conference on Artificial Intelligence (pp. 208-219). Springer, Cham.

Pruthi, K., & Bhatia, P. (2015, October). Application of Data Mining in predicting placement of students. In 2015 International Conference on Green Computing and Internet of Things (ICGCIoT) (pp. 528-533). IEEE.

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015, July). Predicting students performance in educational data mining. In 2015 International Symposium on Educational Technology (ISET) (pp. 125-128). IEEE.

Ahadi, A., Lister, R., Haapala, H., & Vihavainen, A. (2015, August). Exploring machine learning methods to automatically identify students in need of assistance. In Proceedings of the eleventh annual international conference on international computing education research (pp. 121-130).

Sisovic, S., Matetic, M., & Bakaric, M. B. (2015, June). Mining student data to assess the impact of moodle activities and prior knowledge on programming course success. In Proceedings of the 16th International Conference on Computer Systems and Technologies (pp. 366-373).

Bakaric, M. B., Matetic, M., & Sisovic, S. (2015, June). Text mining student reports. In Proceedings of the 16th International Conference on Computer Systems and Technologies (pp. 382-389).

Barbosa Manhães, L. M., da Cruz, S. M. S., & Zimbrão, G. (2015, April). Towards automatic prediction of student performance in STEM undergraduate degree programs. In Proceedings of the 30th Annual ACM Symposium on Applied Computing (pp. 247-253).

Sorour, S., Goda, K., & Mine, T. (2015, February). Correlation of topic model and student grades using comment data mining. In Proceedings of the 46th ACM Technical Symposium on Computer Science Education (pp. 441-446).

Guarín, C. E. L., Guzmán, E. L., & González, F. A. (2015). A model to predict low academic performance at a specific enrollment using data mining. IEEE Revista Iberoamericana de tecnologias del Aprendizaje, 10(3), 119-125.

Jishan, S. T., Rashu, R. I., Mahmood, A., Billah, F., & Rahman, R. M. (2015). Application of optimum binning technique in data mining approaches to predict students' final grade in a course. In Computational Intelligence in Information Systems (pp. 159-170). Springer, Cham.

Shukor, N. A., Tasir, Z., & Van der Meijden, H. (2015). An examination of online learning effectiveness using data mining. Procedia-Social and Behavioral Sciences, 172, 555-562.

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Computer Science, 57, 500-508.

Mayilvaganan, M., & Kalpanadevi, D. (2015). Cognitive skill analysis for students through problem solving based on data mining techniques. Procedia Computer Science, 47, 62-75.

Dangi, A., & Srivastava, S. (2014, December). Educational data classification using selective Naïve Bayes for quota categorization. In 2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE) (pp. 118-121). IEEE.

Pathan, A. A., Hasan, M., Ahmed, M. F., & Farid, D. M. (2014, December). Educational data mining: A mining model for developing students' programming skills. In The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014) (pp. 1-5). IEEE.

Anh, N. T. M., Chau, V. T. N., & Phung, N. H. (2014, September). Towards a robust incomplete data

handling approach to effective educational data classification in an academic credit system. In 2014 International Conference on Data Mining and Intelligent Computing (ICDMIC) (pp. 1-7). IEEE.

Ragab, A. H. M., Noaman, A. Y., Al-Ghamdi, A. S., & Madbouly, A. I. (2014, June). A comparative analysis of classification algorithms for students college enrollment approval using data mining. In Proceedings of the 2014 Workshop on Interaction Design in Educational Environments (pp. 106-113)

Chen, X., Vorvoreanu, M., & Madhavan, K. (2014). Mining social media data for understanding students' learning experiences. IEEE Transactions on learning technologies, 7(3), 246-259.

Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. Expert systems with applications, 41(14), 6400-6407.

Hoe, A. C. K., Ahmad, M. S., Hooi, T. C., Shanmugam, M., Gunasekaran, S. S., Cob, Z. C., & Ramasamy, A. (2013, November). Analyzing students records to identify patterns of students' performance. In 2013 International Conference on Research and Innovation in Information Systems (ICRIIS) (pp. 544-547). IEEE.

Chau, V. T. N., & Phung, N. H. (2013, November). Imbalanced educational data classification: An effective approach with resampling and random forest. In The 2013 RIVF International Conference on Computing & Communication Technologies-Research, Innovation, and Vision for Future (RIVF) (pp. 135-140). IEEE.

Palazuelos, C., García-Saiz, D., & Zorrilla, M. (2013, September). Social network analysis and data mining: An application to the e-learning context. In International Conference on Computational Collective Intelligence (pp. 651-660). Springer, Berlin, Heidelberg.

Pratiwi, O. N. (2013, August). Predicting student placement class using data mining. In Proceedings of 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE) (pp. 618-621). IEEE.

Márquez-Vera, C., Morales, C. R., & Soto, S. V. (2013). Predicting school failure and dropout by using data mining techniques. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 8(1), 7-14.

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Applied intelligence, 38(3), 315-330.

Mashiloane, L., & Mchunu, M. (2013). Mining for marks: a comparison of classification algorithms when predicting academic performance to identify "students at risk". In Mining Intelligence and Knowledge Exploration (pp. 541-552). Springer, Cham

Blagojević, M., & Micić, Ž. (2013). A web-based intelligent report e-learning system using data mining techniques. Computers & Electrical Engineering, 39(2), 465-474.

Dejaeger, K., Goethals, F., Giangreco, A., Mola, L., & Baesens, B. (2012). Gaining insight into student satisfaction using comprehensible data mining techniques. European Journal of Operational Research, 218(2), 548-562.

Şen, B., Uçar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Systems with Applications, 39(10), 9468-9476.

Sen, B., & Ucar, E. (2012). Evaluating the achievements of computer engineering department of distance education students with data mining methods. Procedia Technology, 1, 262-267.

Chuan, H., Ruifan, L., & Yixin, Z. (2011, August). Combining Different Classifiers in Educational Data Mining. In International Conference on Applied Informatics and Communication (pp. 467-473). Springer, Berlin, Heidelberg.

El-Halees, A. (2011, June). Mining opinions in user-generated contents to improve course evaluation. In International conference on software engineering and computer systems (pp. 107-115). Springer, Berlin, Heidelberg.

Zengin, K., Esgi, N., Erginer, E., & Aksoy, M. E. (2011). A sample study on applying data mining research techniques in educational science: Developing a more meaning of data. Procedia-Social and Behavioral Sciences, 15, 4028-4032.

Bodea, C. N., Bodea, V., & Mogos, R. (2010, September). Student Performance in Online Project Management Courses: A Data Mining Approach. In World Summit on Knowledge Society (pp. 470-479). Springer, Berlin, Heidelberg.

Kan, L., Xingyuan, X., & Ping, L. (2010, March). DMCMS: A Data Mining Based Course Management System. In 2010 Second International Workshop on Education Technology and Computer Science (Vol. 3, pp. 145-148). IEEE.

Chellatamilan, T., Ravichandran, M., Suresh, R. M., & Kulanthaivel, G. (2011). Effect of mining educational data to improve adaptation of learning in e-learning system.

Cocea, M., & Weibelzahl, S. (2010). Disengagement detection in online learning: Validation studies and perspectives. IEEE transactions on learning technologies, 4(2), 114-124.

Wang, Y. H., & Liao, H. C. (2011). Data mining for adaptive learning in a TESL-based e-learning system. Expert Systems with Applications, 38(6), 6480-6485.

Meedech, P., Iam-On, N., & Boongoen, T. (2016). Prediction of student dropout using personal profile and data mining approach. In Intelligent and Evolutionary Systems (pp. 143-155). Springer, Cham.

Göker, H., Bülbül, H. I., & Irmak, E. (2013, December). The estimation of students' academic success by data mining methods. In 2013 12th International Conference on Machine Learning and Applications (Vol. 2, pp. 535-539). IEEE.

Umer, R., Mathrani, A., Susnjak, T., & Lim, S. (2019, March). Mining Activity Log Data to Predict Student's Outcome in a Course. In Proceedings of the 2019

International Conference on Big Data and Education (pp. 52-58).

Ketui, N., Wisomka, W., & Homjun, K. (2019). Using classification data mining techniques for students performance prediction. In 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON) (pp. 359-363). IEEE.

Kamal, P., & Ahuja, S. (2019). Academic performance prediction using data mining techniques: Identification of influential factors effecting the academic performance in undergrad professional course. In Harmony Search and Nature Inspired Optimization Algorithms (pp. 835-843). Springer, Singapore.

Abaya, S. A., & Gerardo, B. D. (2013, September). An education data mining tool for marketing based on C4. 5 classification technique. In 2013 Second International Conference on E-Learning and E-Technologies in Education (ICEEE) (pp. 289-293). IEEE.