# Enhancing Student Performance Using Data Pre-Processing Techniques

[1] Kapil Saxena, [2] Shailesh Jaloree, [3] R.S.Thakur, [4] Sachin Kamley
[1] Research Scholar
[1,2,4] S.A.T.I., Vidisha, [3] M.A.N.I.T., Bhopal

***Abstract :*** In present scenario, education data mining is an important concern and develops methods to find hidden and useful patterns from large data set. In this way, Knowledge Discovery in Databases (KDD) provides important steps to convert raw data into useful information. However, quality results always depend on the collected data and feature. For further data mining task, data collection and pre-processing is an important step in order to get the data in correct form. The four years (2012-2015) Government Girls College (GGC), Vidisha, real data is obtained for study purpose. This study mainly highlights important data preprocessing steps like data cleaning, integration and normalization etc. for education perspective. Finally, it is shown by study that after preprocessing task data can be used for further mining task and will be helpful to achieve quality results.

***IndexTerms*** **- Education Data Mining, KDD, GGC, Data-Preprocessing.**

## I. INTRODUCTION

Data collection is one kind of the loosely controlled method of gathering the data. Typically, mostly data are out of range, noisy, missing values, impossible data combinations and many more [1]. The data will generate misleading results if data have not been properly screened. To managerial decision making and quality results, it is very essential to pre-process the raw data. However, the data pre-processing steps converts raw data into an appropriate and understandable form for further processing [2].

Mostly, the real world data are inconsistent, incomplete, missing, uncertain, and contains many errors. The phrase "Garbage in Garbage Out (GIGO)" is well suited for machine learning and data mining applications [3]. Hence, data pre-processing plays an important role to make quality and effective decision making. In this study, different data collections used for implementation, their pre-processing and implementation have been discussed in detail. Figure 1 shows important steps of data pre-processing [4] [5].

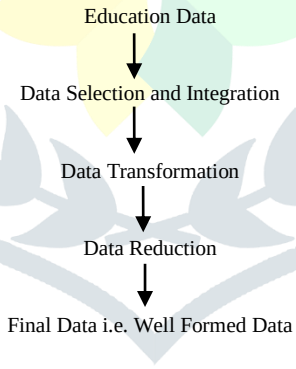


Figure 1. Data Pre-processing Steps

Figure 1 shows important data pre-processing steps.

II. LITERATURE REVIEW

This section describes the brief literature of significant researchers.

Han and Kamber (2006) [6] have discussed Various knowledge discovery process like data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation have been discussed.

Tomar and Agarwal (2014) [7] have used pre and post data pre-processing techniques to enhance data quality. Moreover, they have used various visualization tools and data mining software in order to handle the data. They also discussed research opportunities and challenges of KDD process.

Manisha (2014) [8] conducted data cleaning and distinct user identification technique which enhance the preprocessing steps of web log usage data. Using user identification they found out the distinct user based on their attended session time. This will help in personalizing the websites.

Suad and Wesam (2017) [9] have used various data preprocessing techniques kike cleaning, integration, transformation and reduction in order to improve data quality.

Danubianu (2015) [10] has designed data preprocessing framework for student performance prediction based on data mining techniques. He applied a case study regarding the preprocessing operations which transformed the raw data collected by Moodle system in "Stefan cel Mare" University of Suceava, in datasets.

Christa et al. (2012) [11] have discussed the role of data mining in decision making, but inconsistencies and noise present in it may result in poor and erratic decisions. It is important to consider that the real world data is prone to inconsistencies, duplicate values and can use dissimilar units for same attribute. Mining with that kind of data may give unreliable knowledge model.

Gonçalves et al. (2012) [12] have discussed about data cleaning step which makes the data suitable as input for certain algorithms and leads to correct and usable knowledge. The data pre-processing phase typically requires a significant amount of manual work, which may take up 60– 90 % of the time, efforts and resources employed in the whole knowledge discovery process.

## III. DATA COLLECTION

Data collection is the organized process of collecting and measuring knowledge on variables of concerns that enables us to evaluate outcomes. There are various data collection methods available, but in this research study real student data sets of Government Girls College (GGC), Vidisha (M.P.) for the academic year (2012-2015) have been used [13]. However, the data set contains approximately 248 tuples which is used to build the model. The data set employs more than 20 attributes but only relevant attributes have been considered for study purpose. Therefore, the pre-processing steps have been applied to make them quality data for further data mining process. Here the sample of student data set as showed in Table I and Table II [13].

Table I. Sample of Student Data Set 1

| SSC | HSC | S1TH | SIPR | S1CCE |
|---|---|---|---|---|
| 255 | 241 | 211 | 20 | 46 |
| 286 | 262 | 275 | 17 | 50 |
| 308 | 337 | 309 | 18 | 46 |
| 274G | 238 | 220 | 18 | 44 |
| 400 | 347 | 330 | 20 | 50 |
| 270 | 254 | 252 | 19 | 55 |
| 309 | 229 | 163 | 15 | 47 |
| 513 | 365 | 297 | 18 | 58 |
| 353 | 292 | 225 | 15 | 55 |
| 347 | 347 | 275 | 18 | 55 |
| 487 | 364 | 355 | 62 | 61 |
| 525 | 417 | 365 | 65 | 65 |
| 275 | 320 | 312 | 22 | 62 |
| 288 | 276 | 321 | 70 | 52 |
| 382 | 315 | 282 | 20 | 55 |
| 249 | 290 | 291 | 65 | 55 |
| 452 | 397 | 376 | 72 | 65 |
| 290 | 302 | 235 | 19 | 55 |
| 376 | 355 | 266 | 20 | 55 |
| 455 | 333 | 344 | 74 | 55 |
| 318 | 304 | 232 | 17 | 56 |
| 298 | 316 | 294 | 19 | 59 |
| 255 | 241 | 211 | 20 | 46 |
| 286 | 262 | 275 | 17 | 50 |
| 308 | 337 | 309 | 18 | 46 |

Table II. Sample of Student Data Set 2

| S4CCE | ATTD. | INCOME | SGPA | CGPA |
|---|---|---|---|---|
| 56 | 74% | 50000 | 48.89 | 47.222 |
| 57 | 78% | 35000 | 47.78 | 57.5 |
| 61 | 61% | 30000 | 56.44 | 61.333 |
| 55 | 48% | 45000 | 56.67 | 54.389 |
| 57 | 47% | 50000 | 62.44 | 66.389 |
| 65 | 58% | 45000 | 52 | 55.444 |
| 56 | 55% | 40000 | 50 | 44.833 |

Table III describes attributes in an abbreviated form.

Table III. Attributes in Abbreviated Form

| S.No. | Attributes | Description |
|---|---|---|
| 1 | SSC | Matriculation Marks |
| 2 | HSC | Higher Secondary Marks |
| 3 | S1-S4TH | Sem1 to Sem 4 Theory Marks |
| 4 | S1-S4PR | Sem1 to Sem 4 Practical Marks |
| 5 | S1-S4CCE | (Sem 1 to Sem 4 Continuous and Comprehensive Evaluation Marks |
| 6 | SGPA | Semester Grade Point Average Marks |
| 7 | CGPA | Cumulative Grade Point Average Marks |
| 8 | Attendance | Class Attendance Marks |
| 9 | Income | Family Income |
| 10 | WST | Weekly Studying Time |
| 11 | IAH | Internet Access at Home |
| 12 | STAS | Study Time After School |
| 13 | MTU | Mother's Education |
| 14 | FTU | Father's Education |
| 15 | SA | Student Address |
| 16 | ECA | Extra-Curricular Activities |
| | Total Attributes=25 | |

After attribute description, we have classified attribute in 2 forms i.e. academic and non-academic performance indicator which is exhibited by Table IV and Table V correspondingly.

Table IV. Academic Performance Indicators

| Attributes | Description | Values |
|---|---|---|
| SSC | Per (%) Marks in Matriculation | Very Good, Good, Avg, Poor |
| HSC | Per (%) Marks in Higher Secondary | Very Good, Good, Avg, Poor |
| S1-S4TH | Theory Marks Sem1 to Sem 4 | Very Good, Good, Avg, Poor |
| S1-S4PR | Practical Marks Sem1 to Sem 4 | Very Good, Good, Avg, Poor |
| SGPA | Semester Grade Point Average | Very Good, Good, Avg, Poor |
| CGPA | Cumulative Grade Point Average | Very Good, Good, Avg, Poor |

Table V. Non-Academic Performance Indicator

| Attributes | Description | Values |
|---|---|---|
| FJ | Father's Job | Government, Private, Business, Farmer, Labor |
| ME | Mother's Education | Primary, Middle, SSC, HSC, UG, PG |
| FE | Father's Education | Primary, Middle, SSC, HSC, UG, PG |
| WST | Weekly Study Time in Hours | 14-32 |
| IAH | Internet Access at Home | Yes, No |
| ECA | Extra-Curricular Activities | Yes, No |
| STAS | Study Time After School in Hours | 1-4 |
| RAC | Resident Address Category | Urban, Rural |

Table VI shows value classification of attributes

Table VI. Value Classification of Attributes

| Attributes | Range |
|---|---|
| SSC | IF Marks>=75%=Very Good, Marks>=65% and <75%=Good, Marks>=55% and <65%=AVG, Marks>=40% and <55%=Poor |
| HSC | IF Marks>=75%=Very Good, Marks>=65% and <75%=Good, Marks>=55% and <65%=AVG, Marks>=40% and <55%=Poor |
| S1-S4TH | IF THM>=75%= Very Good, THM>=60% and <75%=Good, THM>=50% and <60%=AVG, THM>=40% and <50%= Poor |
| S1-S4PR | IF PRM >=75%= Very Good, PRM>=60% and <75%=Good, PRM>=50% and <60%=AVG, PRM>=40% and <50%= Poor |
| S1-S4CCE | IF CCEM>=75%= Very Good, CCEM>=60% and <75%=Good, CCEM>=50% and <60%=AVG, CCEM>=40% and <50%= Poor |
| SGPA | IF SGPA>=75%=Very Good, SGPA>=60 and <75=Good, SGPA>=50 and <59=AVG, SGPA>=40 and <=50=Poor |
| CGPA | IF CGPA>=75%=Very Good, CGPA>=60% and <75=Good, CGPA>=50% and <60%=AVG, CGPA>=40% and <50%=Poor |
| Attendance | IF ATTD>=75% Very Good, ATTD>=60% and <75%= Good, ATTD>=50% and <60%= AVG, ATTD>=40% and <50%=Poor |

## IV. DATA PRE-PROCESSING

In order to make quality data, firstly data need to be pre-processed that acquire the quality analysis and information to make quality decision [14]. But, quantity of data also plays measure role for data mining process same as the relevance of the data. The quantity of the data is observed based on:

1) Number of tuples (records): it is very good if data set contains more records. If less then results are not optimum.

2) Number of attributes (fields): if numbers of attributes are more then use feature reduction and selection.

3) Number of targets: decide the target class. It can be greater than 1.

Next, the data preprocessing task are described below.

1) Data cleaning: it includes smooth noisy data, identify or remove outliers, fill in missing values and resolve inconsistencies [14] [15].

2) Data integration: data are integrated from multiple databases, data cubes, or files.

3) Data transformation: data are transformed or normalized in an appropriate form as well as aggregated as per requirement [15].

4) Data reduction: reducing the volume, there should not be effect on analytical results i.e. creating the identical or related analytical outcomes.

5) Data discretization: it is the division of data reduction step i.e. replacing numerical attributes with nominal ones.

During this study, mean (average) is calculated in order to fill missing values as well as data normalization formula is used to transform the value in [0, 1]. Table VII shows preprocessed data.

Table VII. Preprocessed Data

| SSC | HSC | S1TH | SIPR | S1CCE |
|---|---|---|---|---|
| poor | verygood | good | good | verygood |
| poor | good | avg | avg | avg |
| avg | avg | avg | good | verygood |
| poor | avg | avg | good | poor |
| verygood | poor | poor | good | poor |
| poor | good | poor | good | poor |
| avg | avg | avg | verygood | Avg |
| | | **Table VII Coni…..** | | |
| **S4CCE** | **ATTD.** | **INCOME** | **SGPA** | **CGPA** |
| verygood | good | good | verygood | good |
| avg | avg | poor | avg | avg |
| verygood | avg | good | verygood | poor |
| good | avg | good | avg | good |
| good | poor | poor | avg | good |
| good | poor | poor | good | poor |
| avg | avg | verygood | good | poor |
| avg | poor | poor | good | good |
| poor | poor | poor | avg | poor |
| avg | avg | poor | avg | good |
| verygood | good | good | verygood | good |
| avg | avg | poor | avg | avg |
| verygood | avg | good | verygood | poor |
| good | avg | good | avg | good |
| good | poor | poor | avg | good |
| good | poor | poor | good | poor |
| avg | avg | verygood | good | poor |

## V. CONCLUSION AND FUTURE SCOPE

In data mining process, data pre-processing is a very important step. Usually, data cleaning techniques are not applicable to all kinds of data. Deduplication and data linkage are important tasks in the pre-processing step for many data mining projects. Before data is loaded into data warehouse, it is very essential to improve data quality. For large databases finding duplicates is an important part of data management and plays a critical role in the data cleaning process. In this research wok, a framework is designed to clean duplicate data for improving data quality and also to support any subject oriented data. The main benefits of the data pre-processing framework are as follows:

(1) Duplicate data is easily handled.

(2) Data cleaning process complexity is reduced using attribute selection algorithm.

(3) The speed of the data cleaning process is improved

(4). Identical records are grouped together as cluster by applying some methods i.e. binning to reduce its computational cost.

(5) Increasing database accuracy by reducing false mismatches.

In future, big education data set will be considered and more data preprocessing techniques will be applied for student performance enhancement.

**REFERENCES**

[1] C. Romero, S. Ventura, and, E. Garcia, "Data mining in course management systems: Moodle case study and tutorial", Computers & Education, Vol. 51, Issue (1), pp. 368–384, 2008.

[2] P. Long and G. Siemens, "Penetrating the fog: analytics in learning education", Educause Review Online, Vol. 46, No. 5, pp.31-40, 2011.

[3] A. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview", In Proceedings of the IADIS European Conference on Data Mining, pp 182-185, 2008.

[4] S. E., Sorour, T. Mine, K. Goda, and, S. Hirokawa, "A predictive model to evaluate student performance", Journal of Information Processing, Vol. 23, Issue (2), pp.192–201, 2015.

[5] B.K. Pal and Bharadwaj, "Data mining: A prediction for performance", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, 2011.

[6] J.H. Kamber and Micheline, "Data Mining: Concepts and Techniques", second edition. San Francisco: Morgan Kaufmann, 2006.

[7] D. Tomar and S. Agarwal, "A survey on pre-processing and post-processing techniques in data mining", International Journal Database Theory Application, Vol. 7, pp. 99-128, 2014.

[8] Manisha V. "A Step up in Data Cleaning and User identification of Preprocessing on Web Usage data". International Journal of Advanced Research in Computer Engineering and Technology IJARCET, 2014.

[9] A. Alasadi Suad and S. Bhaya Wesam, "Review of Data Preprocessing Techniques in Data Mining", Journal of Engineering and Applied Sciences, 1212: 4102-4107, 2017.

[10] M. Danubianu, "A data preprocessing framework for students' outcome prediction by data mining techniques", 19th International Conference on System Theory, Control and Computing (ICSTCC), October 14-16, pp. 836-841, Cheile Gradistei, Romania, 2015.

[11] S. Christa, L. Madhuri, and V. Suma, "An effective data preprocessing technique for improved data management in a distributed environment", In International Conference on Advanced Computing and Communication Technologies for High Performance Applications, International Journal of Computer Applications, Cochin, pp. 52-57, 2012.

[12] P. M. Gonçalves, R. S. Barros and D. C. Vieira, "On the use of data mining tools for data preparation in classification problems", In Computer and Information Science (ICIS), IEEE/ACIS 11th International Conference on (pp. 173-178), IEEE, 2012.

[13] Data obtained from Government Girls College (GGC), Vidisha.

[14] C. Romero and S. Ventura, "Data mining in education", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27, 2013.

[15] R. Thakur, and A. R. Mahajan, "Preprocessing and classification of data analysis in institutional system using WEKA". International Journal of Computer Applications, 112(6), 2015.