Chapter 3: Research Methodology in Educational Data Mining

Research methodology (RM) is a way to systematically study and solve a research problem [89, 90]. It is necessary for the researcher to know not only the research methods/techniques which are available but also the methodology. It is a search for knowledge, an investigation, and a voyage of discovery. The purpose of research is to find the truth and to understand the unknown. Research contributes to the existing knowledge. In view of this, following sections are about definition and nature of research, then a discussion on evaluation of research. This is followed by sections on scientific approach and nature of educational data mining research.

3.1 Introduction

‘Research’ refers to the systematic method consisting of enunciating the problem, formulating a hypothesis, collecting the facts or data, analysing the facts and reaching certain conclusions either in the form of solutions(s) towards the concerned problem or in certain generalisations for some theoretical formulation.

A researcher should be able to identify and select appropriate & relevant research methods/techniques to achieve required outcome from a study[90]. In research methodology, the researcher selects what tools should be used for the analysis and why, taking into account all the underlying assumptions and all the criteria under consideration. This implies that the design of research methodology might differ from problem to problem. To meet out learning and research objectives we need relevant methods.

Research methodology has many dimensions and research methods do constitute an important part of the research methodology but the scope of research methodology is wider than that of research methods. Having a sound research methodology also helps in cross-verification and validation of different research studies conducted.
3.2 Nature of Research

The purpose of research is to discover answers to questions through the application of scientific procedures. This research has turned out to be Descriptive with a hint of all other types too – Exploratory, Hypothesis testing & Diagnostic. Data provided, is used for the purpose of determining characteristics of students, which in turn, help the administrators and teachers in understanding the potential of the students and enabling them to take corrective measures.

3.3 Research Methodology for Educational Data Mining

Currently, in EDM, a wide variety of DM methods are available such as prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment. An EDM researcher has to capture patterns in the data. Understanding problems of students from an educational psychology point of view is important here. An EDM research can help us in the following areas –

**Exploration:** Exploring hidden causes behind a phenomenon i.e. examining evidences. In EDM there is a lot of scope for exploring and identifying the factors affecting educator and educand. Human behaviour is complex and difficult to capture. We can identify the reasons for wide disparities in performance of students in a class. How does the environment of a student affect their concentration, study methods, liking/disliking of a particular method or subject? Such answers can be partially predicted on the basis of student’s profiles. We can identify the influential factors and contributing factors from this study.

**Description:** Defining or differentiating a phenomenon from others, e.g. describing characteristics of a population or its subsets. Mining of students’ characteristics can be useful to both, administrator and teacher, in many ways. The word ‘characteristics’ here means
distinguishing properties or attributes[91]. One such important attribute is ‘finance‘ or ‘parental income‘ which determines health, food/nutrition a family can buy and also the influence of other luxuries and exposure to technology and information. Another attribute is ‘disability‘, which covers categories like physical, mental or learning disability.

**Prediction:** Identifying relationships $P \rightarrow Q$. $P$ is the cause of $Q$ and $Q$ happens after (follows) $P$. So $P$ aids in $Q$’s occurrence, and hence $Q$ can be used to say that probably $P$ occurred. But if $Q$ did not occur then, $P$ did not occur as well. For example very poor families mostly can’t afford nutritional food and necessary health care facilities. Students from such families are very likely to suffer from disability or serious ailments. This family income, history & tags/keywords selected for life style can be used to predict health conditions.

**Explanation:** Comparing two or more theories in an unbiased manner, free from authoritarianism. While dealing with enrolment data of students it can be seen that some columns contain sensitive information, like: religion, caste and creed. A researcher and the authorities under whom he or she is working should know that research being done for the betterment of mankind and not for their personal reasons or grudges. Researcher and authorities should themselves remain unbiased towards each other and the students. Similar controversial attributes are gender, marital status, and children. At some level, all these attributes have to be considered to accurately separate the most influential factors or variables. Culture and religion have a great impact on lives of students especially in countries or communities with diversity.

**Action:** Finding a solution, applying and verifying it as well. This is about usefulness of a research. An action based on the suggestive results and evaluation of its effectiveness, is the
best evidence to prove the vitality of such a research. In the field of education, the action based on such studies is usually lacking or delayed.

### 3.4 Evaluation of research

Evaluation of research involves focusing on both sides of the coin – positive and negative, good and bad. Relevance of the research is a key decisive factor. There are following questions to be answered –

**Who:** Researchers, participants and consumers of the research. Competences and biases of the researcher play a role. EDM requires a research scholar to be unbiased like a teacher.

**What:** Topic and theory on which the research is based and how it is viewed by the world. EDM is a new field. There are only a handful of journals and books in this area. There are some commercial or business concerns with a focus on EDM. Most research is coming from academics. SPSS is the most popular software in this field. Books and papers with a focus on particular problems and issues are not there. EDM is considered as an important R&D area by computer science, education, social science, management and many related fields. It is also important to tie in the final objective of the research with these fields and understand the impact in each area.

**Where:** In what kind of environment research can be conducted. EDM research can be conducted in an unbiased environment. Data sets to be used here are so huge; a researcher can’t collect them individually. These databases are mostly provided by some educational body. These authorities should respond to researcher’s request if they can. Researcher should work as per rules laid down by authorities.
**When:** What is the best possible time to conduct a research. The best time is when the problem and its solution are relevant. A true research in DM is relevant at that time, even if it becomes irrelevant later.

**Why:** Other than the above, what are other motivational reasons behind doing a research. A research is important for the organizations involved it. However, we should try to fit the conclusions to the data and not the other way around even if that does not agree with the overall picture being created by the organization.

**How:** What are the methods of conducting a research i.e. research methodology (RM). DM research can be conducted using empirical, simulated or real world data. It depends on the demand of the organization providing data and sponsorship.

### 3.5 Scientific approach

There are no sacred truths. A researcher must be given some freedom under some systematic guidelines and approach. For example, there should not be too many rules in a Fuzzy Systems. Data is analysed to state a conclusion, to give a theory which fits the facts (rules or models) that have come out of a data. If more than one theory is possible, best fit theory is the conclusion. Tomorrow, even this final best fit theory may become invalid. Then it has to be replaced by some other theory, after a new research. Conditions, under which a research is conducted, are important e.g. today there is stress on female education which was not there in the past centuries in India. Education budget has increased but at the same time globalization has made educational organizations look like markets. There’s lot of competition between different organizations which is helping shape the current system but can also lead to negative outcomes if not managed carefully.
Other examples of this research are in the field of EDM including issues about time of assessment, use of drop out and retention rate is used to predict the future pathway of a student. Similarly data of leave of students, teacher and other staff can also be used for medical DM i.e. health issues of students and staff. Data of salary and perks could also be used for analysis.

Scientific research is more about observations followed by logical analysis to generalize inductively. But over generalization makes it non-scientific. Rules generated are acted upon and tested in the form of a pilot study. The hypothesis can then be proved that such a functional dependency of attributes is there.

**Serendipity**

Serendipity is discovery by chance. For examples: invention of dynamite, discovery of x-rays, saccharine, penicillin etc. were serendipities in science. Their inventors and discoverers were established scientists and chemists who were not even looking for these products, but for something else. Still, we need to use our basic research methods to find the unknown: analysis, surveys, questionnaire, qualitative & quantitative techniques. In the present attempt we are using enrolment form for the analysis which is similar to surveys and questionnaires based research. This research is quantitative as it is based on statistics & measurement. This study is qualitative as well as it is based on known fundamentals and also matches with the existing findings of various disciplines. The dataset used is a random sample.

### 3.6 Profile of the Student

Data mining technique used in this study have been applied for the purpose of determining students’ characteristics that can help an administrator or a teacher’s understanding about background information of students and its impact on students’ learning.
For the purpose of data mining, the dataset, used in the research consists of IGNOU. The word ‘characteristics’ here means distinguishing properties or attributes. Various students’ attributes can be:

1. Enrolment number
2. Roll number
3. Date of admission
4. Date of birth
5. Religion
6. Gender
7. Marital status
8. Children
9. Height
10. Weight
11. Identifying mark
12. Previous marks, division or rank
13. Year of pass
14. Employment status
15. Job experience
16. Parent’s (mother and father’s)
   a. Name
   b. Age
   c. Qualification
   d. Job title
   e. Income
17. Number of siblings
a. Siblings’ age
b. Qualifications
c. Job title
d. Experience

18. Rank or number in the family (eldest, 2\textsuperscript{nd} … youngest etc.)

19. Physical condition
   a. Physical handicap
      i. Type
      ii. Degree
   b. Serious ailment
   c. Learning disability

20. Address
   a. House number
   b. Block/colony
   c. City

21. Phone number

22. E-mail id

23. Course opted
   a. Code
   b. Medium
   c. Lab course choices
   d. Elective choices

24. College etc. preference

25. Fee
   a. Mode of payment
b. Bank

c. DD/receipt number

26. Photograph of the student

27. Signature of student and guardian

   a. Date

   b. Place

   …and so on.

All these attributes can have various values e.g. attribute Employment Status can have two values – Employed or Unemployed. Then details can be provided by the pupil about the same, if asked. The research methodology for analysing these students’ attributes can be selected based on the specific problem [92].

3.7 Nature of EDM research

Educational Data is available in an institution from various sources. The data file can be in the form of students’ profile, log files, usage data, progress records, lock box or portfolio etc. The hierarchical or nested structure is shown below (Figure 20) as an inverted pyramid. Since the attributes values from a pupil are going to vary, the variations can have very important implications for the way in which we design the research studies [93]. Data can be separated into components:

\[
DATA = FIT + RESIDUAL, \text{ or } \quad DATA = SIGNAL + NOISE
\]

The better the fit, or the stronger the signal; the smaller the residue or the weaker the noise, the closer our model comes to the real world. If a good fit is found, then variables have explained or accounted for the variability in a good way.
User profile can vary as shown below [17]:

1. Disability (physical, sensory, learning, developmental, situational)
2. Skill level (novice, expert, designer)
3. Social issues (globalization, socioeconomic factors)
4. Cognitive factors (memory, intelligence)
5. Cultural interpretation & linguistic issues
6. Age (children, older)
7. Gender (Stereotype, societal)

All the points above mentioned change their meaning with context and form various studies in many streams. A national & now international concern today is – how the skill level varies and social issues arise due to differences in curriculum, infrastructure, budget, language problem, differences in local needs & culture; amongst students, employees, their spouses and wards who migrate to other institution in states & countries. Bridge course is a possible solution.
It is further assumed that analogous to Nyquist Frequency, patterns (good or bad) found are likely to be repeated in future samples [94]. Education can’t be denied in distance education system of a democratic country. Educational data has a non-independence nature, due to interactions and influences in students’ lives. Therefore, psychometric methods are required to analyse learning. One such example is formative evaluation where a program is evaluated while it is being developed. EDM can always address a few questions important for improving educational designs and decision making.

3.8 Research Methodology used for this study

Data mining research falls in the category of experimental research. In this study, enrolment data of students has been analysed. The data analysis done in this study confirmed the existence of various known divides in the given population of students (as presented in chapter 4). The variables in the datasets indicate performance and other behaviours. Based on the divides and trends observed in student enrolment data, we used an ID3 based analysis to design a decision tree for the current data set. Also based on the population division observed, better interfaces are suggested for special students under Adaptive Educational Hypermedia section. Scalability, feasibility and test of mandate were also studied and tried to some extent in this research.

3.9 Summary

Educational data mining is a complex field and hence research in this area needs development of a research methodology to tackle problems presented to the researcher. The field is especially challenging since it typically involves a lot of sensitive data based on student attributes. A complete and thorough research requires considering a large number of variables into account. Such studies are harder to conduct as Student data is usually not easily available. Some of these limitations are discussed in later chapters as well. In this study, we
used a methodology where existing data has been analysed statistically and also using decision tree to propose tools to enable better learning for the group of target students in this study.