Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.98549371

SIGNIFICANCE OF DATA MINING TECHNIQUES IN SOFTWARE DEVELOPMENT

[1] Vikas S. Chomal, [2]Ms. Mubashshirahbanu M. Shekh, [3]Dr. Kavita Venkatachari
[4]Dr. Anita Venaik
[1]Assistant Professor, Faculty of Information Technology & Computer Science, Parul University, Vadodara, India
[2]Assistant Professor, Shrimad Rajchandra Institute of Management and Computer Application, Uka Tarsadia University, Bardoli, India
[3]HoD - Business Analytics Business Analytics Deptt. Universal Business School Karjat.
[4]Professsor Amity Businesss School,Noida

Abstract
Data mining is a rapidly expanding field in many industries. Data mining techniques have been used in a number of industries, including intrusion detection, advertisement, entertainment, manufacturing, process control, fraud detection, security and network administration. Data mining is often seen by modern corporations as an increasingly important method for turning data into business intelligence that gives them a competitive advantage. Further, the expanding multidisciplinary subject of data mining includes many different disciplines, including statistical database technology, artificial intelligence, machine learning, pattern recognition, data visualization, and high-performance parallel computing. There are numerous potential uses for it. Apart from above stated disciplines, one of the most useful research fields for data mining is software engineering. This study illustrates the systematic literature review carried out to analysis and describes the utilization of data mining techniques in various phases of software development life cycle.
Keywords: *Data Mining, Data mining techniques, Software development life cycle (SDLC), Software Engineering (SE),*

INTRODUCTION

The term "data mining" has become widely used by 1995, the year the First International Conference on Knowledge Discovery and Data Mining was held in Montreal. Data mining is a technique used to sift large data sets in order to find patterns and relationships that could be used to solve business challenges. Businesses can forecast future trends and make more informed business decisions by utilizing data mining techniques and technologies. One of the potent new technologies that have arisen is data mining. It makes it easier for both individual and business users to locate data inside a collection or huge amount of data. Data mining techniques require a lot of processing power. We employ data mining tools, processes, and theories to find trends in the data. Over the past ten years, the adoption of data mining techniques has rapidly increased as well as enhanced due to the development of big data analysis, data warehousing technologies, and other related fields. Although data processing technology is constantly improving, researchers still have issues with automation. Decision-making skills are improved by data mining. It uses machine learning techniques to predict

outcomes from a target dataset [1]. Software engineering is a discipline that deals with the economic, design, maintenance, and implementation of software as well as its development. A number of tasks necessary for quality software development are carried out by software engineering. These tasks include requirement analysis, software scoping and boundary setting, implementation, testing, documentation, resource and effort estimation, risk and change management, and project management [13].

The aim of the research is as follows:

- To provide with a summary of data mining techniques and how they can be used in the context of software engineering.
- Mapping of the appropriate data mining technique that will be most useful at each stage of the software development process.
- Represent the impact of use of data mining technique in software development life cycle.

The organization of the chapter is as follows: A comparative literature review is discussed in section 2. The research methodology is presented in section 3. The experiment and result of the research are presented in section 4. Lastly, conclusions and future enhancements drawn from this research are highlighted in section 5.

I. LITERATURE REVIEW

Thummalapenta et al [12] states that, software engineers are increasingly using data mining methods to complete various SE jobs in an effort to increase software productivity and quality. Further in their research study authors focuses on increasing trend of mining SE data and the reason behind this trend is (a) the growing amount of such data and its proven value in resolving a variety of practical issues and (b) mining SE data has lately gained popularity as a potential method for achieving this objective. Hong [15] examines the correlation analysis of numerous bug repair source code update data and bug defect reports in the software engineering project development process using the version control system SVN and the bug tracking system Bugzilla, and attempts to categorize the bug report by data mining technology: defect changes and potential defects change. Fengxian [4] asserts that, the use of data mining technology in software engineering has grown significantly in recent years, and doing so can raise the software system's maintenance effectiveness and, to a lesser extent, its stability. The author concludes that - data mining technology is frequently used in code analysis, software defect detection, software project management, etc., and it can successfully enhance the management of software engineering and control abilities.

Thomas et al [10] measured the effect of data mining techniques in each phases of software development. Further authors articulate that using data mining, software engineers may foresee, plan, and understand the many complexities of the project with the insights derived from the extracted knowledge patterns, which will enable them to maximize subsequent software development operations. In many areas, data mining is a rapidly increasing field. It is getting more and more important to locate data mining software that is suitable for a particular investigation. Several fields, including intrusion detection, manufacturing, process control, fraud detection, marketing, and network administration, have benefited from the use of data

mining techniques [8]. Bondyopadhyay [2] urge the use of data mining techniques in testing. According to the author's analysis, data Mining algorithms can be efficiently used for automated modeling of tested systems. Induced Data Mining models can be utilized for recovering system requirements, identifying equivalence classes in system inputs, designing a minimal set of regression tests, and evaluating the correctness of software outputs. Kiyak [5] demonstrated in the research work that data mining techniques namely - classification, clustering, text mining, and association rule mining can be applied to five SE tasks: defect prediction, effort estimation, vulnerability analysis, design pattern mining, and refactoring. Husain et al [14] discussed in their research study regarding the numerous forms of SE data that may be mined using data mining techniques in order to address the difficulties presented by tasks like programming, issue spotting, debugging, and maintenance. Further the research also demonstrates that data mining approaches are useful for enhancing SE by raising software reliability and quality. According to Tholuchuri [11] data mining in a nutshell, a set of procedures to extract knowledge from relevant patterns and correlations in massive quantities of datasets and utilize that knowledge to enhance the software engineering process. To examine huge digital collections, or data sets, it combines database management with technologies from artificial intelligence and statics.

Anupama et al [3] investigated the possibilities of using data mining techniques in order to better manage software development stages and create high-quality software systems that can be delivered on schedule and within the allotted budget. Furthermore Kumar and Durga [6], examined at how information mining enhances product development metrics related to timing, expense, availability of resources, reliability, and viability. Lovedeep and Atri [7] states that, in software engineering, a wide variety of data formats are available, including graphs, text, facts, and figures. Using well-known data mining techniques like association, classification, clustering, etc., meaningful information can be extracted from this complex data. Data mining software engineering makes data actionable by revealing hidden patterns. In software engineering, there are many different objectives, including cost estimation, documentation, and optimization. Throughout each stage of the software development life cycle, choosing the optimum mining method aids in efficiently fulfilling these objectives and raising software success rates.

II. RESEARCH METHDOLOGY

In this research, phase-wise research methodology is followed. The research methodology is highlighted in figure 1.

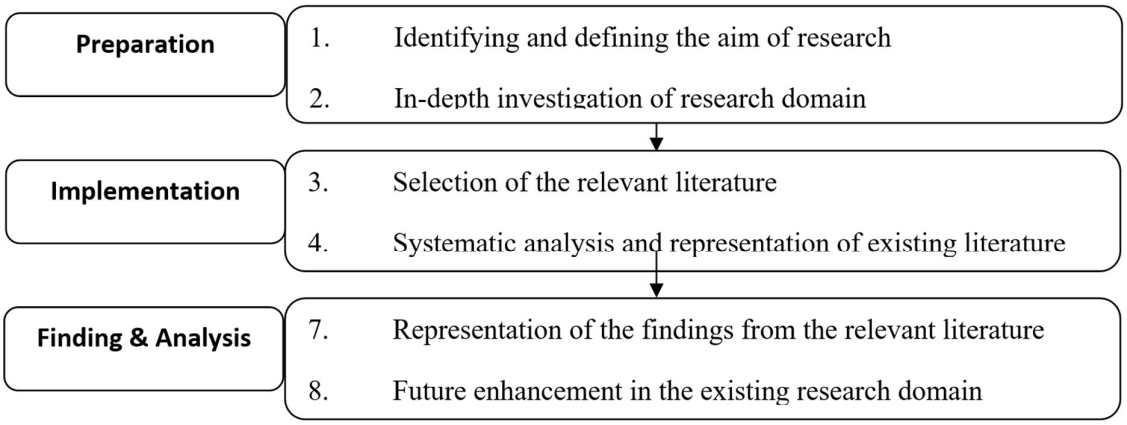

Figure 1 Phase-wise Research Methodology

An exhaustive analysis of related literature to study the utilization of data mining in software development was executed and is represented in tabular format in Table – 1.

Table – 1 Data Mining Techniques Applied in Software Development

Sr. No.	Data Mining Techniques
1	Classification
2	Association discovery
3	Clustering
4	Optimized set reduction
5	Visual data mining
6	Text matching
7	Sequence association rule mining
8	Regression
9	Outer detection
10	Sequential patterns
11	Prediction
12	FL-M-Gspan algorithm performance analysis
13	TS-M-GSpan algorithm
14	Metaheuristic
15	Frequent Pattern Mining
16	Case-based reasoning

Furthermore, after listing the widely used data mining techniques in SDLC, we represent the tasks and activities were these data mining techniques are applied in Table – 2.

Table – 2 Task-wise utilization of Data Mining Techniques

Sr. No.	Tasks & Activities

1	Software reliability
2	Defect prediction
3	Maintenance
4	Bug detection, debugging
5	Source code extraction
6	Software project management
7	Software documentation
8	Requirement Analysis
9	Testing
10	Effort estimation
11	Vulnerability analysis
12	Design
13	Refactoring
14	Scheduling
15	Implementation

III. FINDINGS & ANALYSIS

Finding & analysis is a crucial aspect of the study. After identification and listing of significant data mining techniques and its utilization in SDLC and various tasks and activities in Table 1 and Table 2, we represent the mapping and association of data mining techniques with relevant tasks and activities of software development in Table – 3.

Table – 3 Mapping of Data Mining Techniques with Tasks of SDLC

Sr. No.	Data Mining Techniques	Tasks & Activities
1	Classification	Requirement Analysis, Implementation, Testing, Design, Refactoring, Bug detection, debugging, Maintenance, Defect prediction, Software reliability
2	Association rule	Implementation, Source code, Software reliability
3	Clustering	Design, Implementation, Testing, Design, Refactoring, Bug detection, debugging, Maintenance, Defect prediction, Software reliability
4	Text matching	Vulnerability analysis, Refactoring, Maintenance
5	Regression	Design
6	FL-M-Gspan algorithm performance analysis	Source code
7	TS-M-GSpan algorithm	Source code
8	Metaheuristic	Requirement Analysis, Effort estimation, Scheduling
9	Frequent Pattern Mining	Implementation, Source code

10	Case-based reasoning	Effort estimation

From Table – 3 it can be observed that data mining techniques namely – classification, clustering, association rule and text mining are the most widely used and recommended techniques that are applied in SDLC. Further, software development activities such as requirement analysis, project management and documentation require more improvement and have ample chances and opportunities to take improvement of these data mining techniques. Further, the overall impact and benefit of using data mining and its technique in software development are mainly – (a) Better utilization of time & resources, (b) Improvement in estimation of resources & budget, (c) Quality improvement, (d) Efficient project scheduling, (e) Improve the productivity & development process.

IV. CONCLUSION & FUTURE ENHANCEMENT

Many diverse fields have effectively used data mining techniques. The purpose of the current study is to emphasize on the use of data mining and its techniques in software development process. The primary contribution of the study is the specification of the data mining method best suited for a specific stage of the development process. We would like to conclude that the most widely used and applied data mining techniques in software development process are classification, clustering, text mining, and association rule mining. The suggested research also offers a platform for others to study and investigate the advantages, disadvantages, and impact of data mining and its techniques on requirement analysis, project management and software documentation and other areas of software engineering.

References:

[1] Amitava Bondyopadhyay , Dr. A.C Mandal, Improvement of Software Reliability using Data Mining Technique, International Journal of Scientific and Research Publications, Volume 12, Issue 6, , ISSN 2250-3153, June 2022.

[2] Amitava Bondyopadhyay, Study of improving software testability by Data Mining Techniques, International Journal of Research in Engineering and Science (IJRES) ISSN (Online): 2320-9364, ISSN (Print): 2320-9356 , Volume 10 Issue 5 ǁ 2022 ǁ PP. 21-27.

[3] Anupama Das, Kaberi Das, B.Puthal, Improving Software Development Process through Data Mining Techniques Embedding Alitheia Core Tool, International Journal of Computer Science and Information Technologies, Vol. 2 (2) , 629-632, 2011.

[4] Deng Fengxian, Research Progress on Software Engineering Data Mining Technology, International Conference on Education Technology, Management and Humanities Science (ETMHS 2015).

[5] Elife Ozturk Kiyak, Data Mining and Machine Learning for Software Engineering, DOI: http://dx.doi.org/10.5772/intechopen.91448.

[6] Kishore Kumar , R. Durga, Estimation of Software Defects Use Data Mining-Techniques of Classification Algorithm, International Journal of Engineering Research & Technology (IJERT) , ISSN: 2278-0181 , Vol. 10 Issue 12, December-2021.

[7] Lovedeep, Varinder Kaur Atri, International Journal of Electrical, Electronics and Computer Systems (IJEECS) Applications of Data Mining Techniques in Software Engineering, ISSN (Online): 2347-2820, Volume -2, Issue-5,6, 2014.

[8] Manoj, Jatinder Singh, Study and Analysis of Data Mining Techniques, International Journal of Educational Planning & Administration, Volume 1, Number 1 (2011), pp. 29-35.

[9] Mustafa Abdalrassual Jassim, Sarah N. Abdulwahid, Data Mining preparation: Process, Techniques and Major Issues in Data Analysis, doi:10.1088/1757-899X/1090/1/012053, IOP Conf. Series: Materials Science and Engineering 1090 (2021) 012053.

[10] Nidhin Thomas, Atharva Joshi, Rishikesh Misal, Dr Manjula R, Data Mining Techniques used in Software Engineering: A Survey, International Journal of Computer Sciences and Engineering, Volume-4, Issue-3 E-ISSN: 2347-2693, 2015.

[11] Sreenivasulu Tholuchuri, A Study On Data Mining In Software, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 08 August 2018.

[12] Thummalapenta, Suresh; LO, David; and LIU, Chao. Data Mining for Software Engineering. (2009). Computer. 42, (8), 55-62. Research Collection School Of Information Systems. Available at: https://ink.library.smu.edu.sg/sis_research/763.

[13] Vikas S Chomal, Jatinderkumar R Saini, Significance of Software Documentation in Software Development Process, International Journal of Engineering Innovation & Research Volume 3, Issue 4, ISSN: 2277 – 5668, 2014.

[14] Wahidah Husain, Pey Ven Low, Lee Koon Ng, Zhen Li Ong, Application of Data Mining Techniques for Improving Software Engineering, ICIT 2011 The 5th International Conference on Information Technology.

[15] Xiaobin Hong, Application of Data Mining Technology in Software Engineering, Journal of Physics: Conference Series 2066 (2021) 012013, doi:10.1088/1742-6596/2066/1/012013.