**DATA MINING REVIEW PROJECT**

Data mining is the process that uncovers meaningful patterns, trends, and relationships hidden in large datasets. Data mining combines statistical analysis, machine learning, and database technologies to turn raw data into actionable intelligence. By applying various techniques, data mining helps organizations make informed decisions, optimize the existing operations, and predict future trends. Data mining and data warehousing are complementary technologies designed to transform raw data into strategic insights. The data warehouses are collections of information that support decisions, whereas data mining is the process of extracting meaningful, previously unknown patterns from either aggregated or operational datasets. The best practice involves integrating data mining tools in the early design of a warehouse, especially for big databases in terabytes or petabytes.

Data mining is one of the sub- processes of the broader Knowledge Discovery in Databases (KDD) process. KDD process involves six essential steps:

* **Data Selection:** The process of finding and selecting relevant data subsets for analysis based on the specific objectives of the data mining task.
* **Data Cleansing:** The process of correcting or removing inaccuracies, inconsistencies, and missing values to ensure high-quality data.
* **Enrichment:** Increasing the value of a dataset by adding valuable external data or derived features that will enhance the analysis.
* **Transformation:** Converting data into a suitable format, scale, or structure for efficient processing and mining.
* **Mining:** The application of algorithms to the transformed data to discover patterns, relationships, and insights.
* **Interpretation:** Analysing the mined results to extract meaningful conclusions and make them understandable for decision-making.

These stages make the data accurate, relevant, and ready for analysis. The mining phase includes complex techniques such as association rule mining, classification, clustering, regression, and many more. These methods are used to extract patterns and relationships from data that the traditional analysis cannot reveal. Data mining includes four primary strategic goals:

* **Prediction**: which forecasts future trends and outcomes.
* **Identification**: which detects unique patterns and anomalies.
* **Classification**: which systematically categorizes data into meaningful groups.
* **Optimization**: which enhances resource allocation and improves operational efficiency.

These objectives enable organizations to transform raw data into actionable insights, and to make strategic decision-making across diverse industries and domains.

Association rule mining identifies relationships between items in datasets, like discovering purchasing patterns in market-basket analysis. For example, customers who buy milk are more likely to purchase bread as well. Key metrics for assessing the strength of these relationships include support (the frequency of itemset occurrences) and confidence (the probability of purchasing one item given another). These types of algorithms utilize two key properties to reduce the computational complexity: **downward closure**, which indicates every subset of a frequent itemset is also frequent, and **antimonotonicity**, which denotes any superset of an infrequent itemset is also infrequent. These properties effectively prune the search space. Common algorithm used in association rule mining is the ***Apriori algorithm***, which is used for finding the frequent item sets by limiting the candidate generation. It generates frequent patterns by pairing the items into singletons, pairs and triplets. Another algorithm is ***FP-Growth algorithm***, which is an improvement to the Apriori. Here, a frequent pattern is generated without the need for candidate generation. This method builds an efficient data structure called an FP-tree for faster pattern search and discovery. The ***sampling algorithm***, is another algorithm that selects a small database sample to identify frequent itemsets, typically forming a superset of the actual frequent itemsets by applying algorithms like Apriori with reduced support thresholds. Few other algorithms include the ***partition algorithm***, that divides the database into non-overlapping subsets, treating each as a separate database to generate local frequent itemsets in a single pass. Local itemsets are then combined into global candidate itemsets, which are verified in a second scan to measure their support across the entire database. Association rules among hierarchies is also possible where relationships across item categories are explored, such as linking beverages to desserts, often uncovering insightful cross-category patterns. ***Multidimensional association rules*** extend this by incorporating attributes like time or location to discover richer patterns. ***Negative associations*** identify items that rarely co-occur, focusing on meaningful exceptions using domain hierarchies.

**Classification** is the process of finding a model that describes and distinguishes data classes and concepts. It is used to categorize data into predefined classes or labels. The two steps involved in data mining classification are Learning phase and Classification phase. For example, in banking systems, customers can be classified as "good risk" or "poor risk" for credit approval. ***Decision trees*** are a popular technique used in classification. A decision tree is a structure that has a root node, branches and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is called root node. Here the data is split based on the most informative attributes to maximize the separation between classes. The effectiveness of classification depends on measures like information gain, which evaluates the ability of a feature to reduce uncertainty in predictions.

**Clustering** technique is an unsupervised learning technique that groups similar data points together without predefined classes or labels. Cluster is a group of objects that belongs to the same class. It is particularly useful for discovering inherent structures in data, such as segmenting customers based on buying behaviour or grouping patients with similar medical conditions. Clustering relies on a similarity function to group data points, with **Euclidean distance** commonly used for numeric data. The Euclidean distance between two n-dimensional points measures similarity by calculating the square root of the sum of squared differences across all dimensions. A smaller distance indicates greater similarity between the points. One is the most frequently applied, the ***K-Means***, where the dataset gets divided into a fixed number of clusters by determining a central mean for every cluster. The other is ***BIRCH*** (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm, which was specially designed for large datasets. It works by creating a summary of the dataset that preserves as much information as possible, and then clustering that summary instead of the entire dataset.

Other approaches to data mining include discovering sequential patterns to predict itemset purchases, analysing time series for trends, and using regression for predicting continuous outcomes. A neural network is a technique, derived in artificial intelligence, which includes generalized regression and iterative learning to model complex patterns. It uses a curve-fitting approach to infer functions from sample data, enabling adaptive learning for classification and prediction tasks. Genetic algorithms, inspired by biological evolution, optimize solutions through random search and mutation but are computationally complex. These methods enhance problem-solving and pattern discovery process in large datasets.

Data mining is widely applied across industries to enhance decision-making.

* In **marketing**, it analyses consumer behaviour, optimizes advertising strategies, and aids for customer segmentation.
* **Finance** uses it for credit risk analysis, fraud detection, and investment performance evaluation.
* **Manufacturing** benefits from resource optimization, process design, and product customization.
* In **healthcare**, data mining discovers patterns in medical data, assesses drug effectiveness, optimizes hospital processes, and analyses links between patient health and treatment.

Commercial data mining tools utilize techniques like association rules, clustering, neural networks, and statistical analysis to extract insights. Advanced methods such as genetic algorithms, Bayesian networks, and fuzzy logic are also integrated into some products. Many tools support the ODBC interface, enabling access to popular databases like Oracle and SQL Server, and most operate in the Microsoft Windows environment.

Overall, the essential field of data mining, utilizes database technology to uncover knowledge and patterns within data. Ongoing research in data mining is expanding, and future developments in database technologies will significantly enhance the scope and functionality of data mining tools.

**Selected Paper Review**

The paper selected for this report is “Data Mining in Healthcare: Applying Strategic Intelligence Techniques to Depict 25 Years of Research Development” authored *by Maikel Luis Kolling, Leonardo B. Furstenau, Michele Kremer Sott, Bruna Rabaioli, Pedro Henrique Ulmi, Nicola Luigi Bragazzi, and Leonel Pablo Carvalho Tedesco.* It was published in the “International Journal of Environmental Research and Public Health”, the original paper can be found in this link: <https://doi.org/10.3390/ijerph18063099>. In the context of healthcare industry, large amounts of information associated with the organizational processes and patient care is being generated. As per the report, the development and use of wearable technologies transform healthcare through more personalized treatments, generating huge volumes of patient data. Global healthcare data is expected to exceed 1.2 exabytes annually, yet despite this massive information database, a significant portion of this still remains underutilized.

This paper was designed to identify the strategic topics and the thematic evolution structure of data mining in healthcare industry. It conducts a *Bibliometric Performance and Network Analysis* (BPNA), a method that combines science mapping with performance analysis, to the field of data mining in healthcare with the support of the *Science Mapping Analysis Software Tool* (SciMAT) software. SciMAT is particularly suitable for this kind of analysis because it combines performance analysis with science mapping, allowing researchers to visualize trends, identify key research themes, and measure the centrality and density of those themes. This is crucial for identifying emerging fields and understanding how data mining techniques have influenced healthcare research. The derived results form the basis for future research and facilitate decision-making by researchers and practitioners, institutions, and governments interested in exploring data mining techniques in the healthcare field. While the report was being drafted, there were no published works that provided a complete analysis of the field using a bibliometric performance and network analysis (BPNA). The following three research questions were defined and explored in this paper:

RQ1: What are the strategic themes of data mining in healthcare?

RQ2: How is the thematic evolution structure of data mining in healthcare?

RQ3: What are the trends and opportunities of data mining in healthcare for academics and practitioners?

The methodology of the study included Bibliometric Performance and Network Analysis (BPNA) combined with SciMAT software to investigate the field of data mining in healthcare. This methodology involved four primary steps. First**, research themes** **were discovered** by normalizing the keywords using Salton’s Cosine, which were further clustered through a simple centre algorithm, and then the thematic evolution structure was improved using the equivalence index. Second, **the identified themes were visualized** on a bi-dimensional diagram divided into the following four quadrants:

1. Motor Themes: Highly relevant and well-developed research areas
2. Basic And Transversal Themes: Research areas that have potential to grow
3. Emerging Or Declining Themes: Research areas that requires qualitative analysis
4. Highly Developed but Isolated Themes: Areas with restricted potential for further growth and research

Each of these categories reflect different stages of research relevance in the field. For example, Neural Networks (a motor theme) evolved from being just a theoretical tool to an important key technique in healthcare applications. As new machine learning methods emerged, such as Support Vector Machines and Random Forests, they became critical in comparing the effectiveness of Neural Networks in clinical applications.

Third, **the thematic network structure was analysed** to understand relationships and strengths between different themes. They were classified using data mining techniques and medical research concepts. The thematic evolution structure tracked the advancement of these themes across three time periods. Finally, in the **performance analysis** the h-index, sum of citations, centrality, and density were used to identify research themes and areas that were of greatest significance.

The dataset used in this study included 6138 non-duplicated articles and reviews gathered from the Web of Science (WoS) database using a query string that has terms such as "data mining," "health," "clinic," "medicine," and "disease." A preprocessing step was performed to eliminate duplicates and keywords with less than two occurrences, reducing the initial set of 21,838 keywords to 5310 for greater clarity and accuracy. The analysis covered three periods (1995–2003, 2004–2012, and 2013–2020), allowing for an understanding of the field's evolution. Network reduction techniques was used to eliminate irrelevant words and their co-occurrences. Simple centre algorithm enabled the mapping process. Metrics such as the h-index and citation counts were used to evaluate research contributions, resulting in a comprehensive analysis of thematic trends and relationships in the data mining and healthcare domain.  
The study highlighted that the mapping analysis of data mining in healthcare utilized strategic diagrams, thematic networks, and thematic evolution structures to identify key research themes, relationships, and trends. The strategic diagram revealed 19 clusters: eight motor themes (e.g., Neural Networks, Cancer, Electronic Health Records, Diabetes Mellitus), two basic and transversal themes, seven emerging or declining themes, and two highly specialized themes. Among these, Neural Networks ranked highest in density and centrality, while Cancer had the most citations.

Motor themes such as Neural Networks and Cancer were essential in this field, addressing areas like machine learning integration for diagnosis and bioinformatics for cancer mutation studies. Other motor themes included Electronic Health Records (enhancing decision-making), Diabetes Mellitus (risk factor prediction), Breast Cancer (metastasis analysis), Alzheimer’s Disease (gene identification and prediction), and Depression (symptom analysis via social media and electroencephalograms). Random Forest emerged as an impactful but less connected cluster, linked with applications like air pollution risk assessment.

The thematic evolution structure highlighted the development of themes over three periods (1995–2012). Early clusters like Knowledge Discovery shifted towards machine learning and artificial intelligence, suggesting technological disruption. Additionally, clusters such as Security gained relevance due to the sensitivity of healthcare data, showcasing de-identification and data protection measures.

The report indicated that the health concepts such as Gene Expression and Molecular Classification became essential in disease-based research, evolving to connect with major clusters like Cancer and Alzheimer's Disease. This was true. In recent years, clusters such as Pharmacovigilance, Depression, and Electronic Health Records have gained importance due to improvements in detecting drug reactions, diagnosing and diagnosing mental health conditions, and the use of large-scale health databases.

In the later section of the paper, the bibliometric performance of data mining in healthcare was assessed from 1995 to 2020, focusing on publications, citations, productive authors, journals, institutions, and countries. The field demonstrated steady growth, with 316 publications and 13,483 citations in 1995–2003, rising to 4250 publications and 41,821 citations in 2013–2020. Key contributions included Szolovits’ 1995 work on uncertainty in healthcare and Fawcett’s 2006 introduction of Receiver Operating Characteristics (ROC). Li Chien-Feng emerged as the most productive author, while Andrew C. Bate was the most cited, reflecting diverse contributions from cancer diagnosis to pharmacovigilance. Top journals included PLOS One and Expert Systems with Applications, while Columbia University, the U.S. FDA, and Harvard University were the prominent institutions.

The findings of this study have several practical implications for healthcare practitioners and institutions. For example, healthcare institutions could use the analysis of Neural Networks and Electronic Health Records to integrate machine learning models into clinical decision support systems (CDSS). These systems could assist doctors in diagnosing diseases more accurately and personalizing treatment based on individual patient data. The thematic network structure identified co-occurring clusters, revealing hidden patterns that could potentially promote the development of new scientific insights. The paper suggested that future research should explore emerging and declining themes to better understand the dynamics of the theme evolution. Technologies like blockchain could play a key role in ensuring the data integrity. Finally, while this study relied on the WoS database and SciMAT software, it was proposed that the future researchers could explore additional databases like Scopus, PubMed and consider using other bibliometric tools such as VOS Viewer or CiteSpace.

**Bibliography**  
Kolling, Maikel Luis, Leonardo B. Furstenau, Michele Kremer Sott, Bruna Rabaioli, Pedro Henrique Ulmi, Nicola Luigi Bragazzi, Leonel Pablo Carvalho Tedesco: Data Mining in Healthcare: Applying Strategic Intelligence Techniques to Depict 25 Years of Research Development. International Journal of Environmental Research and Public Health 18(6): 3099 (2021)