Kenzo C. Ragundiaz

BSIT 3-5

March 5, 2025

Prof. Severino M. Bedis Jr.

Midterm Assignment #1
Exploring the Landscape of Data Mining

Instructions: Kindly read the questions carefully and write your answer in the yellow pad paper together with name, course, subject and signature on it.

1. Compare and contrast data mining with traditional statistical analysis. What are the key differences and similarities?

Data Mining = the process of sorting large datasets to investigate and extract patterns, trends, and relationships between the data to discover mysteries, answer questions, and solve business problems. It also helps business owners to predict future outcomes and trends to help the company thrive.

Traditional Statistical Analysis = the process of analyzing, and summarizing data. A bit similar to data mining, but in this case, it is responsible for new scientific discoveries, improving human health, and guiding business decisions.

Differences:

— **Data Mining** is available on commercial systems and industries while **traditional statistical analysis** is used in academic research and scientific studies.

— **Data Mining** analyzes both quantitative data, and qualitative data, while **traditional statistical analysis** analyzes quantitative data only.

— **Data Mining** primarily focus on large datasets while **traditional statistical analysis** focuses on smaller datasets.

- **Data Mining** applications are financial data analysis, retail industry, and telecommunication industry while **traditional statistical analysis** applications are demography, and quality control.
- **Data Mining** undergoes in an inductive process where it is the generation of modern hypothesis from data while **traditional statistical analysis** undergoes in a deductive process where it does not include making forecasts.
- **Data Mining** is associated with prediction and classification while **traditional statistical analysis** is associated with testing hypotheses with model building.

Similarities:

- Both investigate patterns, trends, and relationships.
- Both learn from data.
- Both gather, prepare, and cleans the data.
- Both extract insights to support future decision-making.
- Both use charts, and graphs for data visualization.

2. Describe a specific data mining application in a field of your choice. Explain the problem being addressed, the data used, the techniques applied, and the benefits achieved.

- The field I chose that has applied data mining application is **retail/e-commerce**. The problems being addressed here are increasing profitability, reducing business risks, product preferences, and forecasting sales. The data used here are customer details, transactions, product sales data, seasonal sales, sales patterns, and historical sales. For understanding customer purchase, techniques are classification, and clustering. For detecting product sales preferences, the techniques are association rule mining, and clustering. For forecasting sales, the technique is regression models. For reducing business risks, the techniques are predictive modeling, and risk analysis. For increasing the profitability, the techniques are trend analysis, and optimization techniques. With those techniques, they can understand customer purchase behavior, identify which products are bought together, predict the future trends, and solve their business problems.

3. Explain the importance of data preparation in the data mining process. What are some common data preprocessing techniques, and why are they necessary?

Data Preparation (Data Preprocessing) = the preliminary step of data mining where you transform the data into a different format like .csv, and .json.

Data Preparation is important to do in data mining because you got to find or gather your data first for you to have something to manipulate.

Data Mining = the process of gathering, searching, examining a batch of raw data to figure out **patterns** and get insightful data. This assists companies to find what things are their customers interested at, helps in spam filtering, and in fraud detection.

Data preprocessing techniques:

— **Data Cleaning =** process of correcting errors, identifying problems and inconsistencies, removing duplicates, removing missing values, and making the data more reliable.

— **Data Integration =** merges multiple data into a single, unified data set.

— **Data Tranformation =** the normalization process for converting the data to a more usable format. From data cleaning to aggregation to coding. Also helps in scaling the data to a common range like average, totals, and converting xx data to discrete categories.

— **Data Reduction =** the process of reducing the dataset size while maintaining the important information.

These techniques are useful for us to extract meaningful, summarized, clean, and unbiased data to gather insightful data to help us discover patterns, trends, and solutions to a certain problem.

4. Discuss the ethical considerations surrounding data mining, particularly concerning privacy and security. What measures can be taken to mitigate these concerns?

Ethical Considerations = these are the guidelines that the people should follow to show transparency and avoid potential damages to users and other people. These make the data mining process fair and safe, away from potential data theft and malicious intents.

Examples of ethical considerations:

— **Consent** = we must ask our users or participants their consent to provide a safe, and ethical environment.

— **Transparency** = we should inform people with our intent with their data, asking them for their permission, and communicate openly about our data practices.

— **Data Breaches Preparedness** = focusing on keeping the data safe from unauthorized accesses from hackers. Protecting the data from cyber-attacks, and accidental loss.

— **Encryption** = transforms data into an unreadable format where the data will still remain safe even if hackers manage to gain unauthorized access in our database.

— **Purpose Limitation** = we must share our purpose on why we collect people's data to provide a safe environment to build trust with them.

— **Access Control** = safeguarding the protection of data and add data protection laws. Implementing privacy shield frameworks help too.

— **Data Minimization** = when collecting data, we should only collect necessary data. Nothing more than that.

5. How can biases in data affect the results of data mining models? What steps can be taken to identify and mitigate bias in data mining?

Bias = systematic error or deviation from the true/desired value of the data. It can be seen in data collection, preprocessing, analysis, and interpretation. If the data is biased, it may confuse the people looking at your visualizations, data report, affects the quality of the data, and will cause deception. Your company may be seen as an untrustworthy source or you, as a person if you provide biased data. This is called an "insidious threat" as it is a sneaky type of data that deceives the reader/people.

— To avoid data biases, we can remove sensitive variables (race, age, and gender), regularly evaluate the model, let humans see the processes and the data output, and conduct data preprocessing.

6. Choose one data mining algorithm (e.g., decision trees, clustering, association rules). Briefly explain how it works, what types of problems it is suited for, and its strengths and weaknesses.

— The data mining algorithm I chose is apriori algorithm.

Apriori Algorithm = unsupervised algorithm used for association rules in big data sets. It identifies the relationships between items in market basket analysis.

— **Suited for these problems:** frequent patterns, large datasets, recommendation systems, medical diagnosis

— **Strengths:** easy implementation, robust, scalability, versatility, defines support and confidence, and data preparation

— **Weaknesses:** takes a lot of memory, time consuming, and computational complexity.

References:

1. Compare and contrast data mining with traditional statistical analysis. What are the key differences and similarities?

Kiwop. (2019, February 16). Traditional Statistics versus Machine Learning. What's the Difference? | ToolsGroup. ToolsGroup. https://www.toolsgroup.com/blog/traditional-statistics-versus-machine-learning-whats-the-difference/

GeeksforGeeks. (2022, September 30). Difference between data mining and statistics. GeeksforGeeks. https://www.geeksforgeeks.org/difference-between-data-mining-and-statistics/

Gillis, A. S., Stedman, C., & Hughes, A. (2024, February 13). data mining. Search Business Analytics. https://www.techtarget.com/searchbusinessanalytics/definition/data-mining#:~:text=Data%20mining%20is%20the%20process,make%20more%20informed%20business%20decisions.

Priyadharshini. (2024, August 13). The difference between data mining and statistics. Simplilearn.com. https://www.simplilearn.com/data-mining-vs-statistics-article

Scribbr. (n.d.). The Beginner's Guide to Statistical Analysis | 5 Steps & Examples. Scribbr. Retrieved from https://www.scribbr.com/category/statistics/

Staff, C. (2025, February 4). What is statistical analysis? Definition, types, and jobs. Coursera. https://www.coursera.org/articles/statistical-analytics

2. Describe a specific data mining application in a field of your choice. Explain the problem being addressed, the data used, the techniques applied, and the benefits achieved.

Bhattacharyya, S. (n.d.). 8 Applications of Data Mining in Retail | Analytics Steps. https://analyticssteps.com/blogs/8-applications-data-mining-retail#google_vignette

Indeed. (n.d.). 15 Popular Data Mining Applications: A Complete Guide. *Indeed Career Guide*. Retrieved from https://in.indeed.com/career-advice/career-development/data-mining-applications

3. Explain the importance of data preparation in the data mining process. What are some common data preprocessing techniques, and why are they necessary?

GeeksforGeeks. (2025, January 28). *Data preprocessing in data mining*. GeeksforGeeks. https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

Lawton, G. (2022, January 31). *data preprocessing*. Search Data Management. https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing

Twin, A. (2024, February 23). *What is data mining? How it works, benefits, techniques, and examples*. Investopedia.
https://www.investopedia.com/terms/d/datamining.asp

4. Discuss the ethical considerations surrounding data mining, particularly concerning privacy and security. What measures can be taken to mitigate these concerns?

ForumCosmos. (2023, July 13). Ethical considerations in data privacy and security. *Medium*. https://medium.com/@armaanakhan91/ethical-considerations-in-data-privacy-and-security-1874a10061f0

5. How can biases in data affect the results of data mining models? What steps can be taken to identify and mitigate bias in data mining?

Blog, D. C. (2025, February 27). Reducing bias and ensuring fairness in machine learning | *DeepChecks. Deepchecks. https://www.deepchecks.com/reducing-bias-and-ensuring-fairness-in-machine-learning/*

How does bias impact your data mining results? (2023, August 15). www.linkedin.com. https://www.linkedin.com/advice/0/how-does-bias-impact-your-data-mining-results-skills-data-mining

Team, C. (2023, November 21). *Data-Mining bias*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/data-mining-bias/

6. Choose one data mining algorithm (e.g., decision trees, clustering, association rules). Briefly explain how it works, what types of problems it is suited for, and its strengths and weaknesses.

BotPenguin. (2024, December 6). *Apriori Algorithm: Advantages & Disadvantages* | *BotPenguin*. https://botpenguin.com/glossary/apriori-algorithm

Dharshinni, N. P., Mawengkang, H., & Nasution, M. K. M. (2018). Mapping of medicine data with k-means and apriori combinations based on patient diagnosis. *Journal of Physics Conference Series*, *978*, 012027. https://doi.org/10.1088/1742-6596/978/1/012027

GeeksforGeeks. (2025a, January 15). *Apriori algorithm*. GeeksforGeeks. https://www.geeksforgeeks.org/apriori-algorithm/

Overload, D. (2024, April 30). Apriori algorithm - data overload - medium. *Medium.* https://medium.com/@data-overload/apriori-algorithm-4b8d7f3bc26c#:~:text=of%20length%20k.-,The%20process%20is%20repeated%20until%20no%20more%20frequent%20itemsets%20can,performance%20with%20low%20support%20values.

Team, D. A. (2024, May 30). *10 Examples of data mining algorithms*. Digital Adoption. https://www.digital-adoption.com/data-mining-algorithms/