

Q1: Choose the most suitable answer for each of the following statements then fill out the table with your answers

[5 points, CLO 2]

1) ..... is often used to explore the relationship between two or three variables (in 2D or 3D visuals).

- A) line chart
- B) bar chart
- C) pie chart
- ☒ D) scatter plot

2) The combination of visualization and predictive analytics is called:

- A) visualization.
- B) predictive analytics.
- ☒ C) visual analytics.
- D) advanced histograms.

3) According to kdnuggets.com, which data mining process/methodology is the most comprehensive?

- A) SEMMA
- B) proprietary organizational methodologies
- ☒ C) KDD Process
- D) CRISP-DM

4) In the data mining process, data preparation is also known as .....

- ☒ A) data preprocessing
- B) data harvesting
- C) knowledge discovery
- D) pattern discovery

5) Because of its successful application to retail business problems, association rule mining is commonly called

- A) sequence analysis
- ☒ B) market basket analysis
- C) link analysis

[6 points, CLO 3]

Q2: Answer each of the following questions

- Explain the statement: "The first 3 steps in the data mining process consume around 85% of the total project time".

~~at~~ we should pay extra attention to the first three steps ~~to make sure~~ to make sure ~~to~~ not to put the whole study on an incorrect level of understanding. It's also been done by human. on the other hand, <sup>or path</sup> model building are done using tool ~~at~~ which will not take time.



- List and briefly describe the three main types of business reports.

1. Metric Management Reports: business performance is managed through outcome oriented metrics. for internal management KPIs indicators.
2. Dashboard-Type Reports: present a range of different performance indicator on one page i.e. a dashboard - dashboard vendor provide a set of predefined reports with allowing for customization.
3. Balanced Scorecard-Type Reports: attempts to present an integrated view of success in organization.

- Give one example of using NLP that is NOT discussed in class.

Using Alexa (Smart home), it will understand your natural language and fulfill your order.

### Q3: Data mining

(4) [4 points, CLO 3]

Choose the most suitable data mining tool (classification, clustering, regression, time series analysis, association, outlier analysis) for each of the following problems.

- Given the number of students who applied to join the university for the last 15 academic years, you are asked to build a predictive model to estimate the number of students for the next academic year.  
...Time Series.
- There is a plan to start a new student club for the business technology school. You are asked to study the student's profiles and find the common interests to define the new club's aim. Clustering
- PSUT is planning to study the factors affecting the students' performance in their exams. So that they can work on these factors to improve the academic files of its students. Association
- The admission and registration deanship is looking to build a predictive model using students' historical data that can be used to tell whether a new applicant (student) fits into the BIT program or not. Classification  
output  $\rightarrow$  yes fits  
                  no doesn't fit



#### Q4: Case study "Text mining for patent analysis"

[5 points, CLO 2]

A patent is a set of exclusive rights granted by a country to an inventor for a limited period of time in exchange for a disclosure of an invention (note that the procedure for granting patents, the requirements placed on the patentee, and the extent of the exclusive rights vary widely from country to country). The disclosure of these inventions is critical to future advancements in science and technology. If carefully analyzed, patent documents can help identify emerging technologies, inspire novel solutions, foster symbiotic partnerships, and enhance overall awareness of business' capabilities and limitations.

Patent analysis is the use of analytical techniques to extract valuable knowledge from patent databases. Countries or groups of countries that maintain patent databases (e.g., the United States, the European Union, Japan) add tens of millions of new patents each year. It is nearly impossible to efficiently process such enormous amounts of semistructured data (patent documents usually contain partially structured and partially textual data). Patent analysis with semiautomated software tools is one way to ease the processing of these very large databases.

#### A Representative Example of Patent Analysis

Eastman Kodak employs more than 5,000 scientists, engineers, and technicians around the world. During the twentieth century, these knowledge workers and their predecessors claimed nearly 20,000 patents, putting the company among the top 10 patent holders in the world. Being in the business of constant change, the company knows that success (or mere survival) depends on its ability to apply more than a century's worth of knowledge about imaging science and technology to new uses and to secure those new uses with patents.

Appreciating the value of patents, Kodak not only generates new patents but also analyzes those created by others. Using dedicated analysts and state-of-the-art software tools (including specialized text mining tools from Clearforest Corp.), Kodak continuously digs deep into various data sources (patent databases, new release archives, and product announcements) in order to develop a holistic view of the competitive landscape. Proper analysis of patents can bring companies like Kodak a wide range of benefits.

- (1) • It enables competitive intelligence. Knowing what competitors are doing can help a company to develop countermeasures.
- (2) • It can help the company make critical business decisions, such as what new products, product lines, and/or technologies to get into or what mergers and acquisitions to pursue.
- (3) • It can aid in identifying and recruiting the best and brightest new talent, those whose names appear on the patents that are critical to the company's success.
- It can help the company to identify the unauthorized use of its patents, enabling it to take action to protect its assets.
- It can identify complementary inventions to build symbiotic partnerships or to facilitate mergers and/or acquisitions.
- It prevents competitors from creating similar products and it can help protect the company from patent infringement lawsuits.

Using patent analysis as a rich source of knowledge and a strategic weapon (both defensive as well as offensive), Kodak not only survives but excels in its market segment defined by innovation and constant change.

1. Explain why do we consider patent analysis as one of the text mining applications?

because its extract valuable knowledge from patent database.

(exclusive rights) = Textual data

2. Highlight/Underline: What is the main contribution of using patent analysis in Kodak?
3. Highlight/Underline: List two benefits of using patent analysis at Kodak.

Good luck

Wes Danner

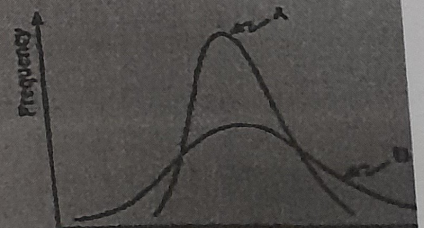


Q1: Choose the most suitable answer for each of the following statements then fill out the table [8 points, CLO 2]  
with your answers

1. At PSUT, adding a student to a closed section is a type of 7  
 A) strategic decision.  
 B) structured decision.  
 C) unstructured decision.  
 D) managerial control decision.
2. .... refer to the requirement that the data variables are given at a level of detail for the intended use.  
 A) data source reliability  
 B) data accessibility  
 C) data granularity  
 D) data currency
3. The number of tourists visited Petra during 2021 is a type of  
 A) numeric data  
 B) ordinal data  
 C) interval data  
 D) nominal data
4. Your major in the school (BIT, BA, Accounting, eMarketing) is a type of  
 A) numeric data  
 B) ordinal data  
 C) interval data  
 D) nominal data
5. Creating a new attribute for the age out of the date of birth attribute is typically a part of which data preprocessing step?  
 A) Data Consolidation  
 B) Data Cleaning  
 C) Data Transformation  
 D) Data Reduction
6. Removing outlier and inconsistent data is typically a part of which data preprocessing step?  
 A) Data Consolidation  
 B) Data Cleaning  
 C) Data Transformation  
 D) Data Reduction
7. Given a set of customers' ages, if the mean=mode=median then the ages are:  
 A) normally distributed  
 B) the value of standard deviation equals to 1  
 C) there are many outliers in the data
8. Given the following datasets and the figure below, which statement is correct?  
 $X = \{2, 9, 13, 14, 20, 20, 24, 26, 32, 40\}$   
 $Y = \{10, 11, 14, 20, 20, 20, 22, 24, 28, 31\}$

Q	A
1	C
2	C
3	A
4	D
5	C
6	B
7	A
8	B

- A) Curve A represent dataset X and curve B represents dataset Y  
 B) Curve A represent dataset Y and curve B represents dataset X





Q2: Answer each of the following questions

[6 points, CLO 1]

- Do you consider Amazon data as an example of big data? Explain your answer. [2 points]

Yes Amazon is an example of BigData, this is because bigdata includes data that cannot be stored in a single storage and it comes in many different forms

- In DSS, what is the relationship between data management subsystem and model management subsystem. How does this relation end up with a recommended decision? [3 points]

The Data management subsystem and model management subsystem work together. Data management subsystem contains relevant data for the situation. Data is given to the model management subsystem where it analyzes data and draws a model that represents it, if there is model management subsystem there is DSS. The analyzed data is then given to the user, the user communicates and commands DSS through user interface. When the analyzed data reach the user the ~~data~~ recommended decision is made

- What is the role of transactional information systems in the high-level architecture of BI. [1 point]

BI is not transaction processing. Transactional processing system constantly involved in handling updates to operational database



1. List two problems that are encountered in the old predictive models used at Humana?  
A delay in the submission & processing of claim data, this delay is reflected in the time it takes to identify high risk members, this issue is relevant when new members join as they don't have a claim history

2. What is new in NMPPM model which makes it better than the old predictive models used at Humana?

NMPPM stands for new member predictive model that works in identifying at risk individuals as soon as they sign up with Humana rather than waiting for sufficient claim history to become available for compiling clinical profiles and predicting future health risk.

3. What is the result of implementing the new model (NMPPM)? (2 results are enough to answer this question)

- 31,000 new members enrolled in clinical programs compared to 4,000 in the same period a year earlier
- Increased volume of clinical program enrollments





Q1: Choose the most suitable answer for each of the following statements then fill out the table with your answers

1) ..... is often used to explore the relationship between two or three variables (in 2D or 3D visuals).

- A) line chart
- B) bar chart
- C) pie chart
- ☒ D) scatter plot

[5 points, CLO 2]

Q	A
1	D
2	C
3	D
4	A
5	B

2) The combination of visualization and predictive analytics is called:

- A) visualization.
- B) predictive analytics.
- ☒ C) visual analytics.
- D) advanced histograms.

3) According to kdnuggets.com, which data mining process/methodology is the most comprehensive?

- A) SEMMA
- B) proprietary organizational methodologies
- C) KDD Process
- ☒ D) CRISP-DM

4) In the data mining process, data preparation is also known as \_\_\_\_\_.

- ☒ A) data preprocessing
- B) data harvesting
- C) knowledge discovery
- D) pattern discovery

5) Because of its successful application to retail business problems, association rule mining is commonly called

- A) sequence analysis
- ☒ B) market basket analysis
- C) link analysis

Q2: Answer each of the following questions

[6 points, CLO 3]

4 1/2

- Explain the statement: "The first 3 steps in the data mining process consume around 85% of the total project time".

The business understanding, Data understanding and data preparation consume 85% of the project time due to ~~the large amount of time~~ to the fact that real-world data are generally incomplete, they lack attribute values, lacking certain attributes of interest or containing only aggregate data. They are also noisy (contain errors and outliers) and inconsistent. Therefore the first three steps require the most effort and time to complete.



- List and briefly describe the three main types of business reports.

- 1) Metric management reports: Help manage business performance through metrics. It is a summary of KPIs, can be external (SLAs for external, EPR for internal)
- 2) Dashboard type reports: Graphical representation of KPIs on a single screen
- 3) Balanced scorecard - type reports: Used to show financial customer, business process, learning and growth indicators

- Give one example of using NLP that is NOT discussed in class.

One example of using NLP is Speech, it converts spoken words (voice) to machine readable input.

### Q3: Data mining

3 [4 points, CLO 3]

Choose the most suitable data mining tool (classification, clustering, regression, time series analysis, association, outlier analysis) for each of the following problems.

- Given the number of students who applied to join the university for the last 15 academic years, you are asked to build a predictive model to estimate the number of students for the next academic year.  
...regression...
- There is a plan to start a new student club for the business technology school. You are asked to study the student's profiles and find the common interests to define the new club's aim. ...Clustering...
- PSUT is planning to study the factors affecting the students' performance in their exams. So that, they can work on these factors to improve the academic files of its students. ....association....
- The admission and registration deanship is looking to build a predictive model using students' historical data that can be used to tell whether a new applicant (student) fits into the BIT program or not. ...Classification...