Data Mining Final Module Assessment

**Please 1) answer the question and 2) provide rationale for your response. In 1-3 sentences, explain why you chose that answer.**

* \_\_\_\_\_\_\_\_\_ is a process for the automatic extraction of non-obvious, hidden knowledge from large volumes of data.
* Artificial Intelligence
* Data Mining
* Cognitive Computing
* Database Extraction

**RATIONALE:**

Data Mining also known as Knowledge Discovery in Databases, it refers to the nontrivial extraction of implicit, previously unknown, and potentially useful information from data stored in databases.

* Data mining is not an iterative process.
* True
* False

**RATIONALE:**

Data Mining is an iterative process where the mining process can be refined, and new data can be integrated to get more efficient results.

* Place the data mining steps in the correct order:
* 2: Collect the data
* 4: Interpret the model/draw conclusions
* 5: Preprocessing the data
* 3: Estimate the model
* 1: State the problem/form a hypothesis

1: State the problem/form a hypothesis

2: Collect the data

3: Estimate the model

4: Interpret the model/draw conclusions

5: Preprocessing the data

**RATIONALE:**

Project Goal Setting:

For anything to succeed, it has to have a plan.

Data Gathering & Preparation:

For every good kind of data, there is a mountain of bad data.

Data Modeling:

With the use of mathematical models and various data visualization tools, there are meaningful patterns discovered in the data.

Data Analysis:

After the modeled data is analyzed, it is then extracted, transformed, and visualized. Data analysis helps bring together useful information to give insights.

Deployment:

In the last stage of Data Mining, relevant partners test the hypothesis.

* Which of the follow is considered a “threat” to data mining?
* Intrusion of privacy
* Large projected market growth
* High quality data
* Low cost of data movement

**RATIONALE:**

Threat can be anything that can take advantage of a vulnerability to breach security and negatively alter, erase, harm object or objects of interest.

* Is Data mining a brute-force crunching of large amounts of data.
* True
* False

**RATIONALE:**

The process of mining consists of finding an input to a cryptographic hash function which hashes below or equal to a fixed target value. It is brute force because at every iteration the content to be hashed is slightly changed in the hope to find a valid hash.

* Raw data has which of the following problems?
* Missing data
* Redundant
* Differing formats
* Misrecorded data
* All of the above

**RATIONALE:**

Raw data can often be out-of-date, denormalized, or poorly structured. There is no built-in capacity for consistency, version control, and collaboration.

* Which of the following are data preprocessing techniques?
* Normalization
* Outlier Detection & Removal
* Sampling
* Data Cleansing
* All of the above

**RATIONALE:**

The data preprocessing phase is crucial for determining the correct input data for the machine learning algorithms. As we saw previously, without applying the proper techniques, you can have a worse model result. For example, the k-nearest neighbors algorithm is affected by noisy and redundant data, is sensitive to different scales, and doesn’t handle a high number of attributes well. If you use this algorithm, you must clean the data, avoid high dimensionality and normalize the attributes to the same scale.

* Which of the following is not a common mechanism for missing data?
* Random occurrences
* Specific cause (e.g., patient dropping out of a study due to adverse reaction)
* Database breaches
* Structural deficiencies

**RATIONALE:**

There are three main types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR).

* Missing data is rare in real world data sets.
* True
* False

**RATIONALE:**

Missing data, or missing values, occur when you don't have data stored for certain variables or participants. Data can go missing due to incomplete data entry, equipment malfunctions, lost files, and many other reasons

* One method for automatic replacement of missing data is
* Replace all missing values with a global constant
* Ignore missing data
* Replace the missing value with the value closest to it in the data set
* Replace the missing value with its feature mode

**RATIONALE:**

One of the most widely used imputation methods in such a case is the last observation carried forward (LOCF). This method replaces every missing value with the last observed value from the same subject.

* Is de-identified data secure?
* Yes
* No
* Not Always

**RATIONALE:**

Health Information is not subject to the HIPAA Privacy Rule if it is de-Identified in accordance with the HIPAA Privacy Rule. No authorization from an Individual is required to use or disclose Health Information that is de-Identified.

* The \_\_\_\_\_\_\_\_\_\_\_\_\_ is defined as the number of data records within the neighborhood (mass) divided by the area of the neighborhood (volume).
* Mean of neighborhood
* Density of neighborhood
* Outlier of neighborhood
* Radius of neighborhood

**RATIONALE:**

The radius-based approach to locating neighbors is appropriate for those datasets where it is desirable for the contribution of neighbors to be proportional to the density of examples in the feature space.

* DIANA is an example of what type of clustering algorithm?
* Density
* K-means
* Hierarchical
* Partitioned

**RATIONALE:**

DIANA is also known as DIvisie ANAlysis clustering algorithm. It is the top-down approach form of hierarchical clustering where all data points are initially assigned a single cluster.

* This type of clustering analysis is efficient, but sensitive to noisy data and does not support nominal data.
* Density
* K-means
* Hierarchical
* Partitioned

**RATIONALE:**

It requires to specify the number of clusters (k) in advance.

It can not handle noisy data and outliers.

It is not suitable to identify clusters with non-convex shapes.

* Which of the following is a deviation-based algorithm for outlier analysis?
* Index-based
* Nested-loop b
* Cell-based
* OLAP data cube

**RATIONALE:**

This method integrates both outlier detection concept in data mining and ideas from OLAP field. An outlier score function is defined over OLAP cells to measure the extremeness level of the cell, and when the outlier score is significant enough, we say the records contained in the cell are associated to each other.

* The \_\_\_\_\_\_\_\_ set is used to fit the models. The \_\_\_\_\_\_\_\_ set is used to estimate prediction error for model selection.
* Validation, Training
* Learning, Validation
* Test, Training
* Training, Validation

**RATIONALE:**

The training set is used to fit the models; the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the generalization error of the final chosen model

* What are the two primary goals of data mining?
* Prediction & description
* Prediction & conclusion
* Prediction & data discovery
* Prediction & interpretation

**RATIONALE:**

The two "high-level" primary goals of data mining, in practice, are prediction and description. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Description focuses on finding human-interpretable patterns describing the data.

* Raw data is ready for data mining.
* True
* False

**RATIONALE:**

Raw Data needs to be cleaned before mining

* Cluster analysis falls under what learning algorithm?
* Supervised Learning
* Reinforcement Learning
* Unsupervised Learning
* Semi-supervised Learning

**RATIONALE:**

Clustering is an unsupervised machine learning method of identifying and grouping similar data points in larger datasets without concern for the specific outcome

* Which of the following is not considered a cluster analysis approach?
* Partitioning algorithms
* kNN algorithms
* Hierarchy algorithms
* Density algorithms

**RATIONALE:**

the K - nearest neighbor method is not a clustering method

* Which of the following is not a basic operation in the data-reduction process?
* Delete a column
* Delete a row
* Reduce the number of values in a column
* Delete a table

**RATIONALE:**

reducing the values isnt part of the reduction process

* Data mining was first introduced under the name in what year?
* 2001
* 1996
* 1990
* 1973

**RATIONALE:**

In the 1990s, the phrase "data mining" was invented. Data mining emerged from the convergence of three scientific disciplines: artificial intelligence, machine learning, and statistics.

* There is not enough data available to take full advantage of what data mining can offer.
* True
* False

**RATIONALE:**

there are endless amount of data that are siting around not being used or mined.

* Data mining is a very database intensive task.
* True
* False

**RATIONALE:**

Data Mining refers to a set of methods applicable to large and complex databases to eliminate the randomness and discover the hidden pattern. Data mining methods are almost always computationally intensive

* \_\_\_\_\_\_\_\_ is the average error that results from using a machine learning method to predict the response on a new observation.
* Training error
* Test error
* Standard error
* Linear Regression error

**RATIONALE:**

Test error is the average (or expected) error that results from using a statistical learning method to predict a response on a new observation, one not used in training the model. - The average is over all test points and all corresponding responses.