SIGNATURE

Name: XIAN HUI B. CHENG ELECTIVE 2 BSIT 3.5

1. compare and contrast data mining w/ traditional statistical analysis. What are the key differences & similarities

|DEFINITION|

Traditional statistical analysis

↳ is hypothesis-driven and deals with statistical data that is structured.

↳ Goal is to provide summary & insights by examining historical data (sales, reports, customer ratings, market trends

↳ Employs techniques like regression analysis, hypothesis testing, and probability distributions

Data Mining

↳ Overall process of identifying patterns & getting useful insights from big data sets

↳ often used on large unstructured data sets. instead of starting w/ hypothesis, it employs machine learning algorithms

|SIMILARITIES|

↳ Both methods aim to uncover relationships within data to inform decision making. They both share the same goal, using techniques to analyze & interpret information to define trends

|DIFFERENCES|

	TRADITIONAL STATISTICAL ANALYSIS	DATA MINING
Approach	hypothesis-driven (deductive)	exploratory (inductive)
Dataset	structured, small dataset	large, unstructured dataset
scale	small - medium dataset	large & big data applications
Automation	mostly manual	highly automated
hypothesis	reavired before analysis	not required, find patterns automatic
flexibility	low (strict assumptions)	high (adaptable to various data types)
compitational power	lower, reavires statistical formulas	higher, requires computing power

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

|Problem|

Tuberculosis is one of the leading cause of morbidity in th Philippines. According to DOH & World Health Organization (WHO), the Philippines is among the top countries with the highest burden of TB cases. Early detection & timely intervention are used to control spread of disease but our problem is ~~[illegible]~~ under reporting and delayed diagnosis persist

|Solution|

Data mining has been applied in the PH healthcare systed, particularly in predictive TB diagnosis, to improve early detection rates & optimize healthcare resources. Given the large volume of patient records in public hospitals & centers, data mining was utilized

|Data used|

Large volume of patien records was used & patient data w/ the ff where collected:

→ demographics (age, gender, location, socio economic background)

→ medical history (previous clinical illness, TB history)

→ clinical symptoms (fever, cough, etc)

→ diagnostic test results

→ lifestyle factors

|Techniques Applied|

→ Decision Trees where a classification algorithm used to analyze patient records & symptoms to determine whether a person is high risk in TB

→ Artificial Neural Networks was used in X-Ray images to detect TB related abnormalities, reducing the need for radiologist in remote areas

→ Natural Language Processing (NLP) → extracts patient information from unstructured clinical notes in electronic medical records

|Example| DOH's TB innovations & health system strengthening project

DATE ____________

Benefits

AI-driven TB screening programs have made diagnosis accessible to more Filipinos. It also helps in efficient resource allocation since by identifying high risked individuals, DOH can allocat testing resources to areas w/ highest needs plus reduce diagnostic errors & automate the process

3. Explain the importance of data preparation in data mining, ~~particularly concerning privacy & security~~ process. What are some common data preprocessing techniques & why are they necessary

→ **IMPORTANCE**

↳ data mining revolves data & raw data is often incomplete, not clean, or inconsistent. Without preparation, data mining models can be inaccurate.

TECHNIQUES

Data Cleaning →	Data Integration →	Data Transformation →	Data Reduction
remove duplicate or incorrect entries	merging data from multiple sources	normalization and standardation	dimensionality of data, if dataset is too large then information must be simplified

4. Discuss the ethical considerations surrounding data mining. process. ~~what are some~~ particularly concerning privacy & security. What measures can be taken to mitigate these concerns.

↳ According to GDPR, ethical concerns in data mining revolve aroud privacy, security, and the potential misuse of personal data. Since many ~~soco~~ organizations collect data of the users, some of them without explicit consent, this raises concerns about how it's stored & used.

↳ For example, the COMELEC experienced a data breach, hacking & exposed 55 million Filipino voters information. All the sensitive data like name, address, passport numbers, fingerprint data were ~~leaked~~ and was publicly accessed online which puts Filipino's at risk of identity theft & fraud

MITIGATION → All the measures in mitigating this revolves around

DATE

implementing strong data protection:

① Compliance w/ Data Privacy Laws → Data Privacy Act 2012 ensures responsible data collection.

② Data Encryption & Anonymization → to protect data, we need to encrypt them especially sensitive data, by doing this it ensures protection when analyzing large dataset.

③ User Consent & Transparency → Before collecting the data, companies should ask for concent & communicate how it will be used.

④ Accountability & Ethical AI Development → AI & ML models should be designed to avoid biases & discrimination in decision-making

5. How can biases in data affect thei results in data mining models? what steps can be taken to identify & mitigate bias in data mining?

↳ Bias in data mining distorts outcomes which leads to incorrect conclusions. Bias comes from unbalanced datasets or selection bias.

↳ Example in hiring algorithms, if training data consists of past hiring decisions that favored a particular demographic, the hiring model may continue to exhibit discriminatory behavior.

MITIGATION

↳ organizations must ensure data diversity and regularly audit models

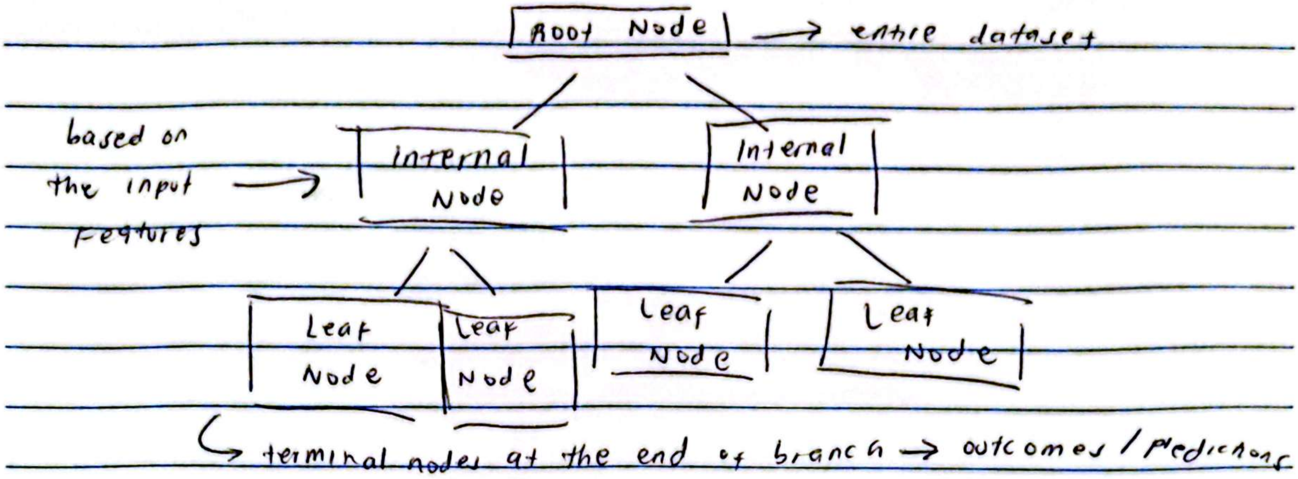
↳ techniques like re-sampling and re-weighting balances the data, what it does is creating new samples based on one observe sample

↳ use explainable AI (XAI) techniques to improve transparency in decision-making. This is a field of AI re-search that helps humans understan how AI systems make decision

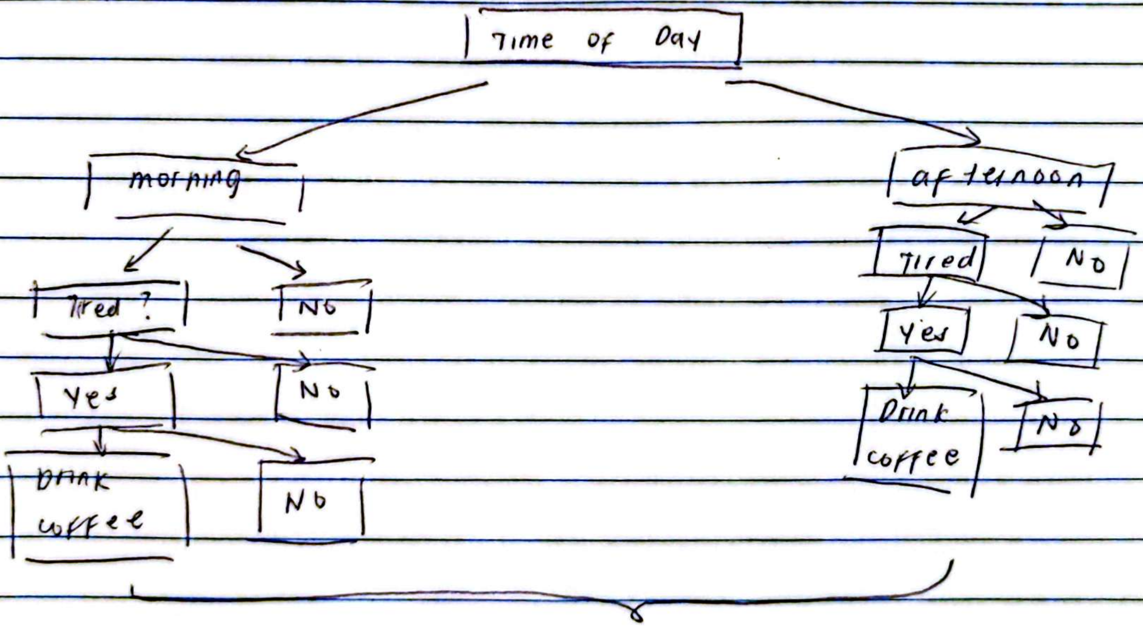
6. Choose one data mining algorithm. Briefly explain how it works, what types of problem its suited for, and its strengths & weakness

↳ Decision Tree — is a popular data mining algorithm used for classification & regression tasks.

↳ it solves problems through hierarchical tree structure with one top call root node & branches out to different possible outcomes.


Root Node
→ entire dataset
based on the input features →
Internal Node
Internal Node
Leaf Node
Leaf Node
Leaf Node
Leaf Node
↳ terminal nodes at the end of branch → outcomes / predictions

example: Deciding if you're tired the tree suggest if you need to drink coffee



Explanation

Tree checks time of day (morning / afternoon)

↓

ask if you're tired

↓

if tired, it suggests drink coffee

↓

if it's afternoon it asks again if you're tired

↓

'Final conclusion is made based on the answer