DATA MINING ASSIGMNET 1

SECTION 1

NAME :- DURESA UDESSA

ID :- RU2526/13

Question 1: List and explain the typical operations used in data warehouse?

A data warehouse is a single, centralized data storage area that collects data from multiple sources within an organization. It is optimized to support the BI operations such as data analysis, reporting, and decision making. Data warehouses store all data that is historical and are designed for analysis and query rather than the operations of transaction processing.

In the case of data warehouses, different areas of action are handled, which are mainly oriented to managing, transforming and analyzing of the data carefully. These revolutions guarantee that databases warehouse is correct, has good performance and useful for decision-making.

- **Extract, Transform, Load (ETL)**
  - Extract: Fetching from various data sources such as databases, flat files, and APIs constitute this procedure.
  - Transform: Data transform means that data cleaning, verification, and remodeling of the extorted data for the analysis is performed.
  - Load: The data is then transformed before being loaded into the data warehouse, this allows the data to be stored and processed further.
- **Data Integration**: - Data integration is used to consolidate existing data sources typically from several systems into one central data store which in most of the cases is the data warehouse and the data is ready to be utilized in a unified format.
- **Data Cleaning: -** Data discrepancy, like duplicate or erroneous data, technological incompatibility, or coding issues are the causes of data inaccuracy. One should spot all these types of inaccuracy to achieve trustworthiness and credibility.

- **Data Aggregation: -** Amongst the data aggregation techniques, a few stand out, including combining multiple data reports into composite reports like sums and averages with a fairly high response speed of queries and analysis.
- **Data Mining: -** Data mining technology is made possible by utilizing commonly applied statistical, mathematical and machine learning models to uncover the underlying patterns, relations and other data which are valuable to have with all the data included in the data warehouse.

- **Data Querying: -** It is not only that the customers can revisit their data warehouse through SQL or any other query language but that they can also be able to have an authority over the data and be able to make sure that the exact information that is needed for analysis, reporting, and making informed decisions is obtained.
- **Data Reporting and Visualization: -** Building reports, dashboards and adding visualizations for non-technical audiences is necessary to make the data real by analyzing the outcome.
- **Data Backup and Recovery: -** The key feature is multiple copies of the data in the data warehousing central storage unit. Data recovery procedures act not only to carry out the data in the event of equipment breakdowns, software errors or disasters, but also to minimize holes and restore them.
- **Data Archiving: -** Storage of data in the manner of archiving, when cold data is transferred into offline storage for reading, even though it might be necessary for the future investigation of data or any processing, the data is well organized and easily accessible.
- **Data Security: -** Set of security data including access controls, encryption, and monitoring assures (the) safekeeping(s) of data warehouse(s)/ systems from cases like unauthorized access and data vulnerability, breach, among other cybersecurity problems.

Question 2: What is the distinction between a data warehouse and a database management system?

A data warehouse and a database management system (DBMS) serve distinct purposes and have different characteristics.

Feature	Data warehouse	Database management system
Purpose	Central repository for analytical and reporting data	Manages and organizes structured data for transactional applications
Architecture	Dimensional model (e.g., star schema, snowflake schema)	Various data models (relational, hierarchical, NoSQL)
Data integration	Optimized for ETL processes	Supports CRUD operations (Create, Read, Update, Delete)

Functionality	Emphasizes read-heavy operations for analysis	Supports transactional operations, ACID complianceelete)
Data model	Denormalized, optimized for analysis	Relational, hierarchical, document-based, depending on application
Usage	Business intelligence, reporting, decision support	Transactional applications, real-time data processing
Security	Includes data mining, OLAP, advanced analytics features	Emphasizes data integrity, concurrency control, security

Question 2: What are the challenges we faced in data mining, and what countermeasures we took to overcome those challenges?

Data mining pose problem such as information quality issue; high dimensionality, scaling and privacy and security risk. Bad quality data usually contain many omissions or inconsistencies that can mislead the direction of analysis results. To overcome this, affected firms use data cleaning and verification methods in improving data quality.

Multidimensional data complexity may lead to a loss of information due to redundancies. To tackle this obstacle, The PCA or projection methods along with the dimensionality reduction method are used to simplify data and maximize the efficiency of data mining algorithms.

Scalability is yet another issue which ususally comes hand in hand with large amounts of data. Firms effectively use distributed computing and parallel processing, through open source tools like Apache Hadoop or Spark, to solve big data computation problems and address the data mining tasks in a timely manner.

Privacy and Security are of utmost importance in data mining, especially in the context of messages that contain privacy related data. For the protection of data, one of the major precautions taken by the firms includes data anonymization, encryption as well as access control in order to protect their data from unauthorized access or potential data leaks.

Interpretability and complexity of data mining algorithms could pose another related issue at stake. To fix the opaqueness, the organizations try out various approaches with technical algorithms. At the same time, they actively apply the clear machine learning models, such as decision trees, to

deconstruct the data mining results. Moreover, model-agnostic methods, such as LIME and SHAP, view complex models effectively for an intending explained.

As a result, organizations make data cleaning, feature selection, distributed processing, safety measures, and interpretable modeling techniques a mix to conquer the challenges data mining comes with as this guarantees the reliability of the outcomes, scalability e and interpretability.

Question 4: List and explain the quality of metrics to measure the performance of any classifier algorithm.

Performance metrics are essential tools for evaluating the effectiveness and reliability of classifier algorithms in machine learning. Here are some of the key quality metrics used to measure the performance of classifier algorithms:

- **Accuracy:** Accuracy measures the proportion of correctly predicted instances out of the total instances in the dataset.
    - Formula: Accuracy =Number of Correct Predictions/Total Number of Predictions
    - Usage: Accuracy is a widely used metric for balanced datasets but may be misleading for imbalanced datasets where the majority class dominates.

- **Precision:** Precision measures the proportion of true positive predictions out of all positive predictions made by the classifier.
    - Formula: Precision =True Positives / (True Positives + False Positives)
    - Usage: Precision is particularly useful when the cost of false positives is high, such as in medical diagnoses.
- **Recall (Sensitivity):** Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions out of all actual positive instances in the dataset.
    - Formula: Recall = True Positives True Positives + False Negatives
    - Usage: Recall is important when the cost of false negatives is high, such as in fraud detection.
- **F1 Score**: F1 score is the harmonic mean of precision and recall, providing a balance between these two metrics.
    - Formula: F1 Score= 2 × Precision × Recall

- Usage: The F1 score is particularly useful for imbalanced datasets where accuracy can be misleading.

- **Specificity**: Specificity measures the proportion of true negative predictions out of all actual negative instances in the dataset.
  - Formula: Specificity=True Negatives / (True Negatives + False Positives)
  - Usage: Specificity is useful in scenarios where correctly identifying negative instances is crucial, such as in disease screening.
- **ROC Curve and AUC**: The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a classifier across various threshold values, and the Area Under the Curve (AUC) quantifies the overall performance of the classifier.
  - Usage: ROC curves and AUC are useful for evaluating and comparing the performance of classifiers, especially when the class distribution is imbalanced.
- **Confusion Matrix**: A confusion matrix is a table that summarizes the performance of a classifier by listing true positive, true negative, false positive, and false negative predictions.
  - Usage: Confusion matrices provide a detailed overview of the classifier's performance and are often used to compute other metrics such as accuracy, precision, recall, and specificity.
- **Balanced Accuracy**: Balanced accuracy calculates the average accuracy of each class, taking into account the imbalance in the dataset.
  - Formula: (Balanced Accuracy=Sensitivity + Specificity)/2
  - Usage: Balanced accuracy provides a more reliable measure of classifier performance for imbalanced datasets compared to regular accuracy.

Question 5 Assume you have the data set shown below. The data contains 10 instances, V1-V11 features, and a binary class (0 and 1)

Instance	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	Category
1	1	2	3	1	0	1	3	4	5	10	10	1
2	0	1	2	4	6	3	2	5	5	9	9	
3	1	2	4	3	6	5	7	9	8	10	10	0
4	10	3	5	9		8	4	7	6	4	4	0
5	2	1	10	8	9	1000	1	3	7	8	8	0
6	2	2	-100	5	7	8	10			1	1	1
7	9	9	9	9	9	9	9	9	9	9	9	0
8	2	4	6	3		10	1	200	1	7	7	1
9	6	6	6	6	0	0	0	0	0	0	0	1
10	5	5	5	5	5	5	5	5	5	5	5	1

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
data = {
    'V1': [1, 2, 1, 0, 10, 10, 0, 3, 10, 200],
    'V2': [1, 0, 4, 2, 1, 7, 5, 1, 1, 1],
    'V3': [2, 1, 6, 4, 1, 9, 2, 3, 7, 7],
    'V4': [3, 2, 3, 3, 7, 9, 4, 7, 8, 1],
    'V5': [1, 4, 3, 6, 9, 9, 5, 8, 10, 7],
    'V6': [0, 6, 6, 5, 9, 9, 5, 8, 1, 7],
    'V7': [1, 3, 5, 7, 9, 9, 5, 0, 2, 1],
    'V8': [3, 2, 7, 9, 9, 9, 6, 6, 2, 0],
    'V9': [4, 5, 9, 8, 9, 9, 0, 2, 1, 0],
    'V10': [5, 5, 8, 10, 9, 9, 0, -100, 10, 1],
    'V11': [10, 10, 9, 10, 8, 9, 1, 5, 10, 7],
    'Category': [1, 0, 1, 1, 0, 0, 0, 1, 1, 0]
}
df = pd.DataFrame(data)

X = df.drop('Category', axis=1)
y = df['Category']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)

X_test_scaled = scaler.transform(X_test)

knn = KNeighborsClassifier(n_neighbors=5)

knn.fit(X_train_scaled, y_train)

y_pred = knn.predict(X_test_scaled)

accuracy = accuracy_score(y_test, y_pred)

print("Accuracy:", accuracy)
```

Question 6. Distinguish between the following points.

I. Classification and clustering II. Training and testing dataset III. Structured, semi-supervised and unstructured data IV. Data warehouse and DBMS conceptual modeling V. Feature selection and feature extraction VI. Data mining and machine learning VII. Fact table and data dictionary

I. Classification and clustering

Classification and clustering are both techniques used in machine learning, but they serve different purposes. Classification is a supervised learning technique where the model is trained on labeled data to predict the class label of new instances. The model is given a set of features and a corresponding class label, and it learns to associate these features with the class label. On the other hand, clustering is an unsupervised learning technique where the model is given a set of features without any corresponding class labels. The model then groups similar instances together based on their features, creating clusters of data points.

II. Training and testing dataset

Training and testing datasets are used in machine learning to evaluate the performance of a model. The training dataset is used to train the model by providing it with a set of features and corresponding class labels. The model learns to associate these features with the class labels and creates a model that can predict the class label of new instances. The testing dataset is used to evaluate the performance of the model by providing it with a new set of features without any class labels. The model then predicts the class labels for these new instances, and the predictions are compared to the actual class labels to evaluate the performance of the model.

III. Structured, semi-supervised, and unstructured data

Structured data is data that has a well-defined structure, such as a table with rows and columns. Semi-supervised data is data that has some labeled instances and some unlabeled instances. Unstructured data is data that does not have a well-defined structure, such as text, images, or videos.

IV. Data warehouse and DBMS conceptual modeling

Data warehouses and database management systems (DBMS) are both used to store and manage data, but they serve different purposes. A data warehouse is a large, centralized repository of data that is used for analytical purposes. It is designed to handle large volumes of data and provide fast query performance. A DBMS is a software system that is used to manage and organize data in a database. It provides a set of tools and interfaces for creating, modifying, and querying the data in the database.

Conceptual modeling is the process of creating a conceptual model of the data, which is a high-level representation of the data that describes its structure and relationships. In the context of data warehouses and DBMS, conceptual modeling is used to design the data model for the system.

V. Feature selection and feature extraction

Feature selection and feature extraction are both techniques used to reduce the dimensionality of the data. Feature selection is the process of selecting a subset of the original features that are most relevant for the task at hand. This is done by analyzing the correlation between the features and the class labels and selecting the features that have the highest correlation. Feature extraction is the process of creating new features from the original features by applying mathematical

transformations to them. This is done to capture the underlying patterns in the data that are not apparent in the original features.

VI. Data mining and machine learning

Data mining and machine learning are both techniques used to extract insights from data, but they serve different purposes. Data mining is the process of discovering patterns and trends in large datasets. It involves using statistical and machine learning techniques to analyze the data and identify patterns that are not apparent through simple analysis. Machine learning is the process of training a model on a dataset to make predictions or decisions without being explicitly programmed. It involves using algorithms to learn from the data and create a model that can be used to make predictions or decisions on new data.

VII. Fact table and data dictionary

A fact table is a table in a data warehouse that contains the measures or facts of the data. It is used to store the quantitative data that is used to answer business questions. A data dictionary is a repository of metadata that describes the data in a database or data warehouse. It contains information about the structure of the data, such as the names of the tables and columns, the data types, and the relationships between the tables.