Pointers/Keywords:

1. Data mining concepts
   a. is the process of discovering interesting patterns from massive amounts of data.
2. What is Prediction?
   a. we need to predict the missing data for a new observation, depending on the previous data.
3. Applications using Regression, classification, clustering, association rule

ASSOCIATION RULE LEARNING

Summary: Identifies meaningful relationships between variables in large datasets. Used in market basket analysis to discover patterns of co-occurrence, such as identifying products frequently bought together for cross-selling strategies.

Application: Analyzing market basket transaction data to understand customer purchase behavior and find associations and correlations among various items purchased.

CLUSTERING

Summary: Groups similar data points into clusters without prior knowledge of their structure. Used in marketing segmentation, anomaly detection, and image processing.

Application: Dividing sets of items into homogeneous groups or clusters, such as clustering insurance policyholders based on risk factors or clustering documents based on their content for efficient retrieval.

CLASSIFICATION

Summary: Identifies patterns in data and assigns them to predefined classes or categories based on features. Commonly used in fraud detection, customer segmentation, and sentiment analysis.

Application: Categorizing data into predefined classes, such as classifying policyholders into high, medium, or low risk categories for insurance claim prediction or classifying documents into different topics or sentiment categories.

REGRESSION

Summary: Establishes relationships between dependent and independent variables to predict values of the dependent variable. Used in demand forecasting, price optimization, and trend analysis.

Application: Predicting numerical values, such as predicting insurance claim costs or predicting the severity of claims based on various factors.

Text Mining:

Application: Analyzing unstructured text data to extract useful patterns and concepts, such as using text mining techniques to preprocess and analyze accident descriptions to predict the likelihood of attorney involvement or claim severity in insurance processes.

4. Data mining Attributes
    a. Nominal Data:
    Summary: Represents categories with no inherent order, such as gender or occupation. Used in classification and clustering tasks.
    b. Ordinal Data:
    Summary: Represents categories with an inherent order, like education level or social status. Used in ranking and classification tasks.
    c. Binary Data:
    Summary: Consists of only two possible values, like yes/no or true/false. Used in classification and association rule mining tasks.
    d. Interval Data:
    Summary: Represents quantitative data with equal intervals between values but no absolute zero point, like temperature or time. Used in clustering and prediction tasks.
    e. Ratio Data:
    Summary: Similar to interval data but with an absolute zero point, allowing for meaningful ratios. Examples include height, weight, and income. Used in prediction and association rule mining tasks.

5. Outliers
    a. **Outliers** are data points that significantly deviate from the rest of the dataset, often indicating unusual behavior or measurement errors. They are not necessarily noise but can be indicative of different data generation processes.

    Types of Outliers:
    b. Global Outliers:
        i. are individual data points that deviate significantly from the overall distribution of the dataset. They can arise from errors in data collection or represent truly unusual events.
        ii. Detection: Techniques include statistical methods like z-score, machine learning algorithms like isolation forest, and data visualization.
        iii. Handling: Options include removal, correction, or using robust methods to mitigate their impact on analysis and modeling.
    c. Collective Outliers:
        i. are groups of data points that together exhibit unusual behavior compared to the overall dataset. While individual points may not be outliers, their collective behavior is significant.
        ii. Detection: Techniques involve clustering algorithms, density-based methods, and subspace-based approaches.
        iii. Handling: Requires further analysis of group behavior, identification of contributing factors, or considering contextual information.
    d. Contextual Outliers:
        i. deviate significantly from expected behavior within specific contexts or subgroups. They may not be outliers in the entire dataset but show unusual behavior within certain contexts.

ii. Detection: Involves contextual anomaly detection, clustering, and context-aware machine learning approaches, considering contextual information like time or location.

iii. Handling: Considers contextual normalization, transformation, or using context-specific models to properly address their impact.

iv. Proper understanding of the context and domain knowledge is crucial for accurate detection and interpretation of outliers, as they can vary based on specific contexts or subgroups.

6. R Syntax
7. Data Preprocessing
    a. Data Cleaning
        i. This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.
    b. Data Transformation
        i. This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.
    c. Data Reduction
        i. This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.
    d. Data Integration
        i. This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.
    e. Data smoothing
        i. Smoothing is a technique where you apply an algorithm in order to remove noise from your dataset when trying to identify a trend. Noise can have a bad effect on your data and by eliminating or reducing it you can extract better insights or identify patterns that you wouldn’t see otherwise.
    f. Generalization
        i. Data generalization refers to the process of transforming low-level attributes into high-level ones by using the concept of hierarchy. Data

generalization is applied to categorical data where they have a finite but large number of distinct values.

g. Data Aggregation
   i. Data aggregation is possibly one of the most popular techniques in data transformation. When you're applying data aggregation to your raw data you are essentially storing and presenting data in a summary format.

h. Normalization
   i. This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.

i. Segmentation
   i. is the process of taking the data you hold and dividing it up and grouping similar data together based on the chosen parameters so that you can use it more efficiently within marketing and operations. Examples of Data Segmentation could be: Gender. Customers vs.

Normalization – refers to data transformation methods where data are scaled in a specified ranges for faster data extraction.

Formula:

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

Exercise: For the following group of data: 2000, 4000, 5000, 6500, 8000, 9560, 10000 normalize them with min = 1 and max = 10. Complete the table below.

Solution:

Values	Normalized Value (v')
2000	1
4000	3.25
5000	4.375
6500	6.0625
8000	7.75
9560	9.505
10000	10

Given:

$\min_A$ = 2000, the minimum value
$\max_A$ = 10000 the maximum value
new_ $\min_A$ = 1
new_ $\max_A$ = 10

V’ = (2000 – 2000) / (10000 – 2000) * (10-1) + 1

= 0 * 9 + 1

= **1**

V’ = (4000 – 2000) / (10000 – 20000) * (10 – 1) + 1

= 0.25 * 9 + 1

= **3.25**

1. What would be the output of the following code?

   ```
   > x <- 11:14
   > y <- 16:19
   > z <- x + y
   ```

   Output: 27 29 31 33

```
y <- c(10,NA,20,30,NA)
x <- is.na(y)
```

Output: False True False False True

```
x <- c(15:18, NA,45:47, NA, NA, 30:32, NA)
print(x)
```

Output: 15 16 17 18 NA 45 46 47 NA NA 30 31 32

```
x <- seq(12,18)
print(x)
```

Output: 12 13 14 15 16 17 18

```
x <- seq(12,18)
y <- mean(x)
print(y)
```

Output: 15

Sol: 12+13+14+15+16+17+18=105/7= 15

```
x <- seq(12,18)
y <- sum(x)
print(y)
```

Output: 105

Sol: 12+13+14+15+16+17+18=105

Levels of Measurement | Nominal, Ordinal, Interval and Ratio

https://www.scribbr.com/statistics/levels-of-measurement/#:~:text=Nominal%3A%20the%20data%20can%20only%20be%20categorized.,and%20has%20a%20natural%20zero.

Few examples as below for the Nominal variable:

- Red, Yellow, Pink, Blue
- Singapore, Japan, USA, India, Korea
- Cow, Dog, Cat, Snake
-

Example of Ordinal variables:

- High, Medium, Low
- “Strongly agree,” Agree, Neutral, Disagree, and “Strongly Disagree.”
- Excellent, Okay, Bad

Exercise: Determine which of the four levels of measurement (nominal, ordinal, interval, ratio) is the most appropriate.

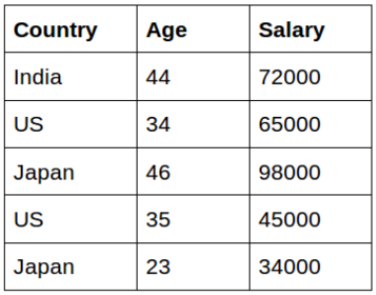
Data	Type
Economic Status (Low, Medium, High)	**ordinal**
Number of voters	**ratio**
Bus Speed	**ratio**
Religion	**nominal**
Weight	**ratio**
Salary	**ratio**

Encoding Categorical Data

It is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the different models.

Types of Encoding

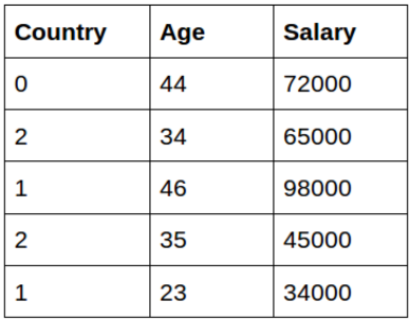
Sample Dataset



Country	Age	Salary
India	44	72000
US	34	65000
Japan	46	98000
US	35	45000
Japan	23	34000

1. Label Encoding

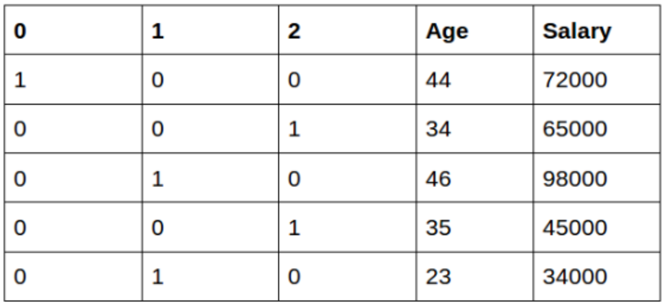
   It is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.



Country	Age	Salary
0	44	72000
2	34	65000
1	46	98000
2	35	45000
1	23	34000

2. One-Hot Encoding

   It is another popular technique for treating categorical variables. It simply creates additional features based on the number of unique values in the categorical feature. Every unique value in the category will be added as a feature. One-Hot Encoding is the process of creating dummy variables. Each category is mapped with a binary variable containing either 0 or 1. Here, 0 represents the absence, and 1 represents the presence of that category.



0	1	2	Age	Salary
1	0	0	44	72000
0	0	1	34	65000
0	1	0	46	98000
0	0	1	35	45000
0	1	0	23	34000

Exercises:

Perform Label and One-Hot Encoding to the given dataset below:

Shoe	Item Sold	Sales
Nike	144	?
Adidas	127	148000
Under Armour	130	154000
Nike	138	161000
Adidas	140	179000
Under Armour	135	158000
Nike	?	152000
Under Armour	148	?
Adidas	137	167000

Label Encoding:

Shoe	Item Sold	Sales	Nike_Shoe	Adidas_Shoe	UA_Shoe
Nike	144	?	**1**	**0**	**0**
Adidas	127	148000	0	1	0
UA	130	154000	0	0	1
Nike	138	161000	1	0	0
Adidas	140	179000	0	1	0
UA	135	158000	0	0	1
Nike	?	152000	1	0	0
UA	148	?	0	0	1
Adidas	137	167000	0	1	0

One-Hot Encoding

ShoeLabelEncoding	Item Sold	Sales

100	144	?
010	127	148000
001	130	154000
100	138	161000
010	140	179000
001	135	158000
100	?	152000
001	148	?
010	137	167000

Handling Missing Values
Visit the site to answer the tables below
https://www.youtube.com/watch?v=2qHbeFXyqi8

Exercises

1. **General Mean (Generalize Imputation)**

Shoe	Item Sold	Sales
Nike	144	159857
Adidas	127	148000
Under Armour	130	154000
Nike	138	161000
Adidas	140	179000
Under Armour	135	158000
Nike	137	152000
Under Armour	148	159857
Adidas	137	167000

2. **Attribute Mean (Similar Case Imputation)**

Shoe	Item Sold	Sales
Nike	144	156500
Adidas	127	148000
Under Armour	130	154000
Nike	138	161000
Adidas	140	179000
Under Armour	135	158000
Nike	141	152000
Under Armour	148	156000
Adidas	137	167000

Sol 1. General Mean (Generalize Imputation):

Total = 144 + 127 + 130 + 138 + 140 + 135 + 148 + 137
= 1099

Nike Item Sold
= 1099 / 8
= 137

Total = 148000 + 154000 + 161000 + 179000 + 158000 + 167000
= 1119000

Nike Sales = 1119000/7
= **159857**

2. Attribute Mean (Similar Case Imputation)

Total Item Sold = 144 + 138
= 282
= 282 / 2
= 1**41**

Nike Sales = 161000 + 152000
= 313000
= **156500**

Under Armour Sales = 154000 + 158000
= 312000
= **156000**