NAME: VEDANTH.V. BALIGA

SEC: 5th 'H'

USN: ENG20DS0044

COURSE: Data Warehouse And KNOWLEDGE Mining.

COURSE CODE:

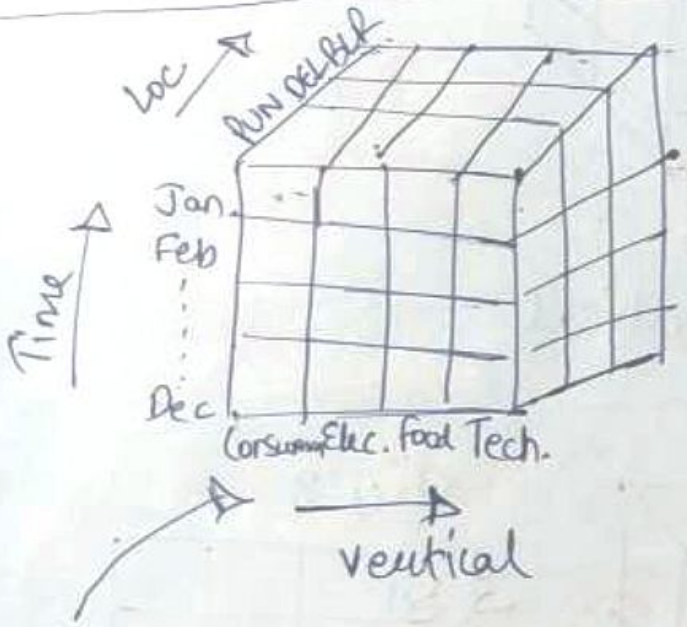
TOPIC: ASSIGNMENT-I

Module 1

Choose a realtime domain & apply all OLAP operations on it.

A OLAP Stands for Online Analytical Processing.

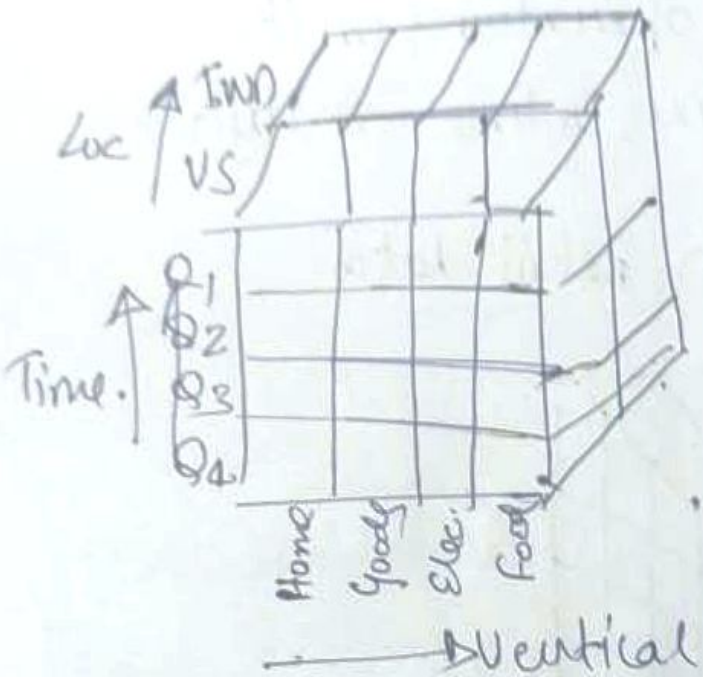
OLAP operations on Retail data



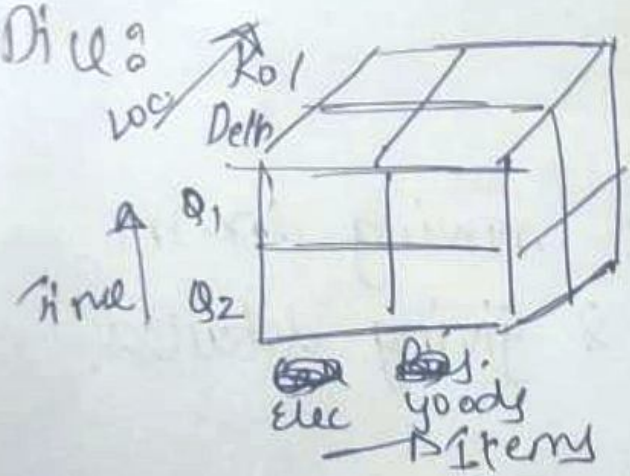
Drill Down Operation: moving down a concept hierarchy & giving detailed summary.

Roll Up: Climb up concept hierarchy & reduce the dimensions.

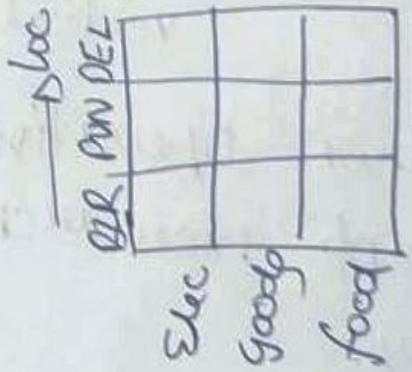
5. Design multidimensial multilevel pattern mining for Flipkart.



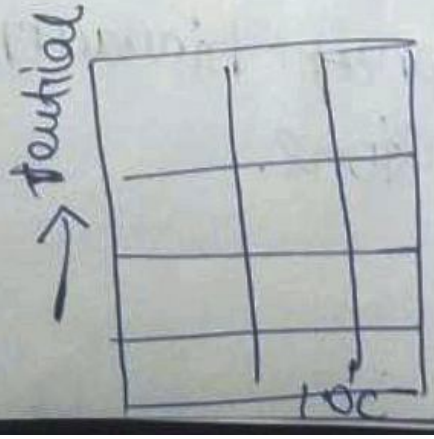
Dice:



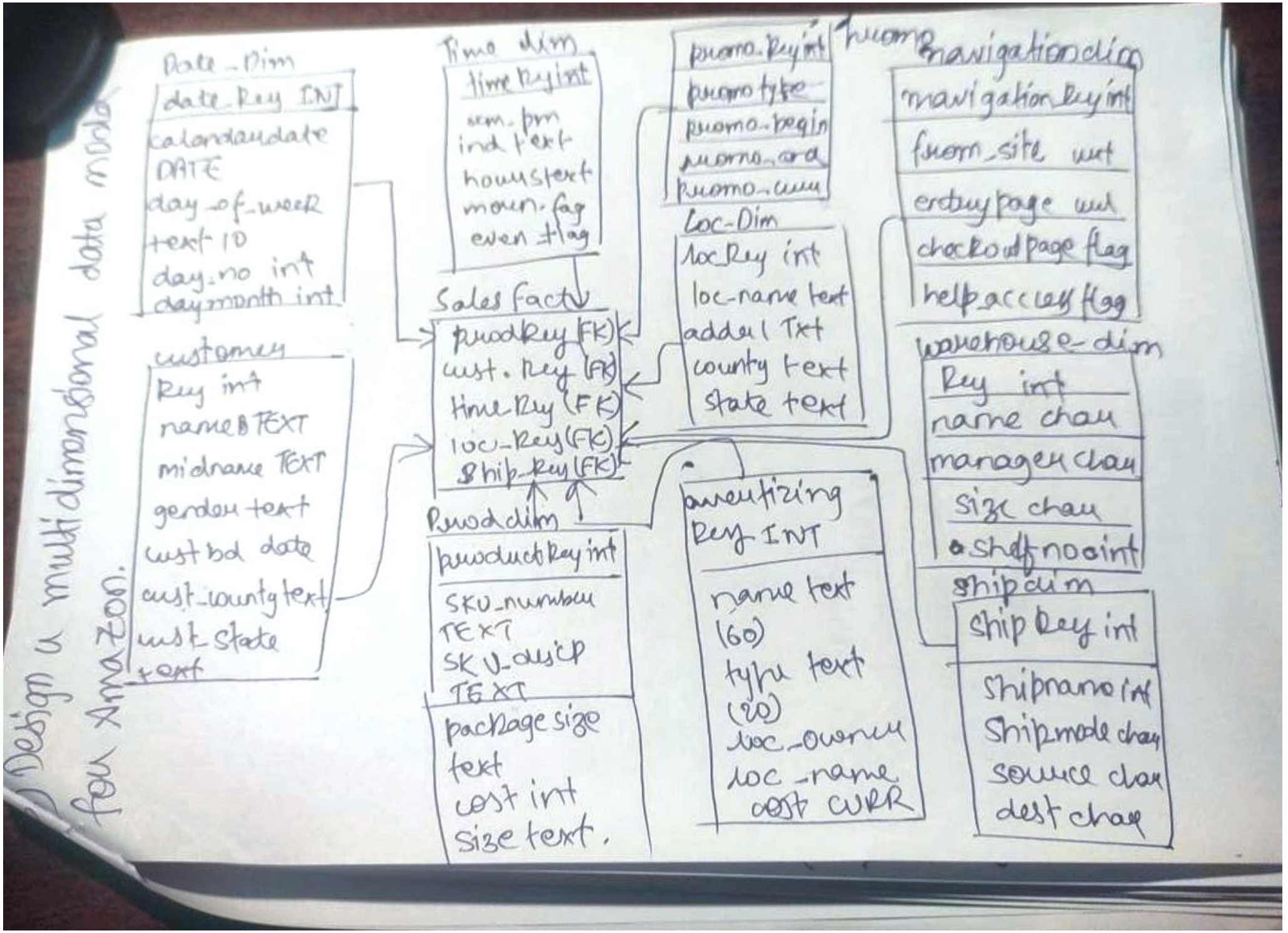
Slice:



Pivot:



Design a multidimensional data model for Amazon.



Design data warehouse for parallel processing.

Parallel execution: break down a task into subtasks & each task performed by many processors.

When to execute these queries.

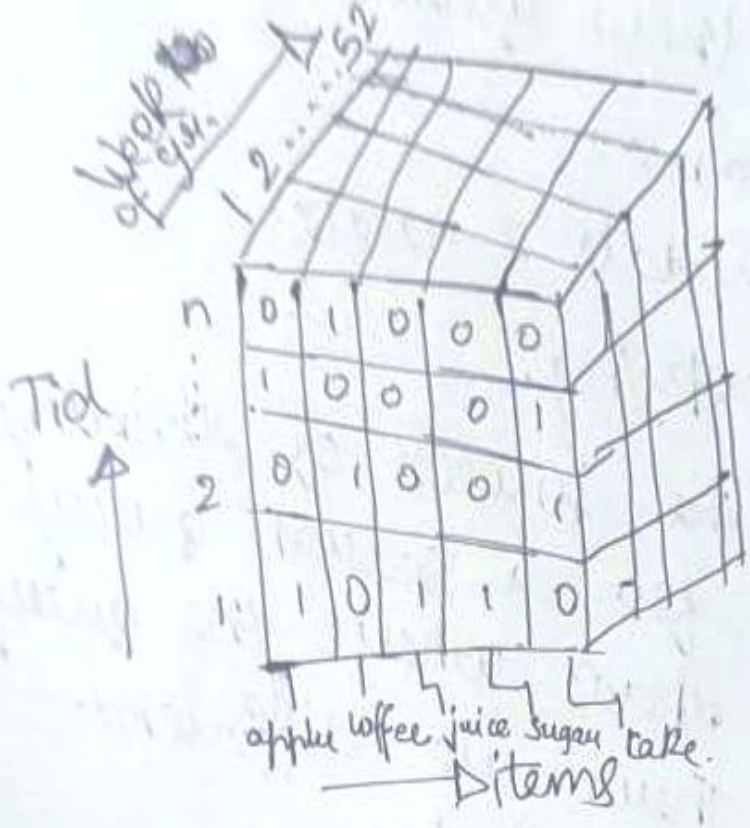
- Large table scans
- Bulk insert, update & deletes
- aggregating & copying.

Partitioning is the process of dividing a large table for a search query to a smaller space where the query may be found. There are different types of queries partitioning:

- Hash Partitioning
- List Partitioning.
- Composite Partitioning.
- Index Partitioning.

the folder "SQL_Partitioning" in the zipped code file. The model is same as in Q3) module file. The.

) Design Data Cube for market basket analysis.



→ The x-axis represents items, y axis represents t-id & z axis represents week of yr.

→ we represent week of yr & not per day to save space & the time is not an imp. factor here.

→ For every transaction we have a 1,0 attribute. 1 represents item.

purchased & 0 represented not purchase. This technique is called binary value indexing to save space.

In SQL the cube is built using CUBE operator. more about CUBE operator is in the SQL folder of attached zip file.

CUBE OPERATORS

Retail Sales db

Syntax

```
Select
    c1, c2, AGGR(c3)
From
    table
Group by cube (c1, c2);
```

Table inventory table.

warehouse	product	model	quantity
SF	IPhone	6S	50
SF	IPhone	7S	10
SF	IPhone	X	200
SF	Samsung	Galaxy	200
SF	Samsung	Note	100
SJ	iPhone	6S	100
SJ	iPhone	7	50
SJ	iPhone	X	150
SJ	Samsung	8S	200
SJ	Samsung	8	150

TASK 1: Return total inventory of all inventory in warehouses.

```
SELECT
    warehouse,
    sum(quantity)
FROM Inventory
Group by CUBE (warehouse)
ORDER BY
    warehouse.
```

Warehouse	QUANTITY
SF	560
SJ	650
(null)	1210.

Task 2: Group quantity by sum of the diff. brands.

```
SELECT
    warehouse,
    product,
    Sum(quantity)
FROM
    inventory
Group by
    cube (warehouse, product)
ORDER by
    warehouse,
    product;
```

(2) or A.

warehouse	Product	Sum(qua
SF	Samsung	300
SF	IPhone	260
SF	(null)	560
SJ	Samsung	350
SJ	IPhone	300
SJ	Null	650
null	samsung	650
null	iPhone	560
nubl	null	1210.

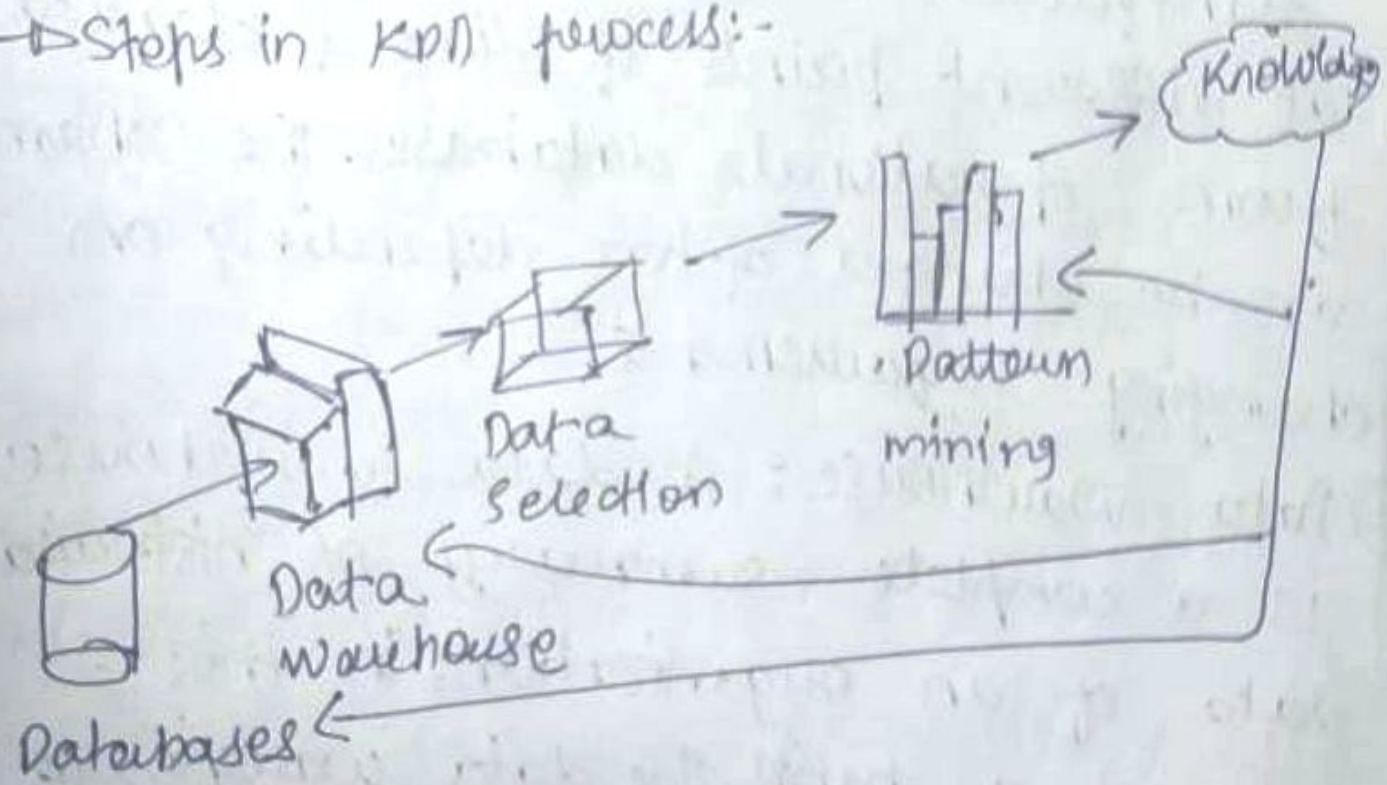
Module 2

1) Design a KDD model for credit card fraud detection.

Q A What is KDD model?

KDD stands for Knowledge discovery in databases & it involves mining patterns from transactions, scientific & sensor data, pictures, text etc.

→ Steps in KDD process:-



KDD Process in Credit Card Fraud De

Q2 or A

① Databases: A database usually stores data of the current activity & the data that powers the application. In the case of credit card fraud detection, a bank stores data of all its transactions including transaction id, amount & time. This is mostly transferred to data warehouses at different points of time & cleaned from operational databases. The schema can be changed often depending on changing requirements.

② Data Warehouse: A data warehouse is a complete summary of historical data of an organization, in this case of a bank. The data warehouse can have different kinds of data features.

Some of them are cc_num (customer number), card merchant, category of transaction_data & time, city, state, zipcode & population. This data may or may not have missing values, so we need to preprocess it first.

Data Preprocessing: Since our data set is very imbalanced, we have to fix the balance first. On observing distributions of transaction_amt & time, we can see that they are very skewed.
Some methods to solve this imbalanced dataset problem is subsampling, random undersampling & SMOTE.

Correlation Analysis: Correlation analysis helps us understand whether there is any relation between the various features of the dataset.

Boxplots & Correlation Plots help to detect underlying distribution of data

Anomaly Detection

IQR: Any data beyond 75th percentile & below 25th percentile is an anomaly.
We should be careful about the threshold we set here to remove outliers.

Feature Selection

Feature selection can be done on credit card fraud data using multiple algorithms. Some of them are PCA (Principal Component Analysis) where we reduce the data from high-dimensions to lower dimensions preserving as much info as possible.
t-sne preserves the neighbourhood embeddings based on the thresholds of distance measured from the

~~wrong~~ points.

5. Pattern mining: Pattern mining is the process of visualizing the different features of the data set.
Some of the visualizations we can plot are:

1] Category of transaction v/s count
2] gender of clients v/s count:
3] States / geographical area with most frauds.
4] No of credit card frauds by metro city.
5] No of credit card frauds by job of transaction.
6) How much skewness in the amt v/s density of amount.
7) We can use dob column to extract age & then devise age-groups v/s fraud amt.

e) Using date field, we can extract & devise month name v/s count of frauds plot.

Prediction using mined Patterns:

There are many classifiers that can be used:

1) Logistic Regression: Since its a two class classification, we can expect Logistic regression to perform well. Other models like SVM & Naive Bayes can also be tested & parameters can be tuned.

Explain the data preprocessing methods for Info Retrival Applications.

Soln

① We can use info method to give statistical description of the data.

② missing data can be checked using missingno matrix

③ Correlation heatmap allows us to measure nullity correlations

④ Dendogram show the hierarchical nullity relationship between columns. It uses a hierarchical clustering algorithm.

⑤ A simple numerical survey of data can show how many types of data counts we have in the dataset.

⑥ We can delete data to remove whole attributes or only a sample of attributes on data.

③ Evaluate the statistical descpn for stock market analysis with data visua

A ① Extract Tesla stock market data using Beautiful Soup.

② The data contains the date feature & Revenue column.

③ We can plot the date v/s prices graph

④ Plotting the data of years v/s revenue would give historical data of Tesla Stock.

⑤ Historical revenue of GameStop would be very uneven due to its downfall in 2020.

Module 3

④ Find the frequent itemsets & strong association rules for the db using Apriori Algorithm. sup count = 2, conf = 50%

Tid	items
T_1	I_1, I_2, I_5
T_2	I_2, I_4
T_3	I_2, I_3
T_4	I_1, I_2, I_4
T_5	I_1, I_3
T_6	I_2, I_3
T_7	I_1, I_3
T_8	I_1, I_2, I_3, I_5
T_9	I_1, I_2, I_3

A 1-itemset.

	sup
I_1	6/9
I_2	7/9
I_3	6/9
I_4	2/9
I_5	2/9

1 frequent item set has all Items.

2-itemset

Item	Support.
$I_1\ I_2$	4/9
$I_1\ I_3$	4/9
(X) $I_1\ I_4$	1/9
$I_1\ I_5$	2/9
$I_2\ I_3$	4/9
$I_2\ I_4$	2/9
$I_2\ I_5$	2/9
(X) $I_3\ I_4$	0
(X) $I_3\ I_5$	1/9
(X) $I_4\ I_5$	0

2 freq itemset

Item	Sup
$I_1 I_2$	4/9
$I_1 I_3$	4/9
I_1, I_5	2/9
$I_2 I_3$	4/9
$I_2 I_4$	2/9
$I_2 I_5$	2/9.

3 freq itemset

Item	Sup.
$I_1 I_2 I_3$	2/9
$I_1 I_2 I_5$	2/9
$I_1 I_2 I_4$	1/9 (X)
$I_2 I_3 I_4$	0
$I_2 I_3 I_5$	1/9 (X)

3 freq Set

Item	Sup
$I_1 I_2 I_3$	2/9
$I_1 I_2 I_5$	2/9.

nd valid association rules.

$I_1 \rightarrow I_2 I_3$

$$\text{Support} = \frac{\sigma(I_1 I_2 I_3)}{|T|} = \frac{2}{9}$$

$$\text{confidence} = \frac{\sigma(I_1 I_2 I_3)}{\sigma(I_1)} = \frac{2/9}{6/9} \times 100$$

$$= \frac{2}{6} \times 100$$

$$= 33.33\% < 50\%$$

$I_2 \rightarrow I_1 I_3$

$$\text{conf} = \frac{\sigma(I_1 I_2 I_3)}{\sigma(I_2)} = \frac{2/9}{7/9} \times 100$$

$$= 28.57\% < 50\%$$

$I_3 \rightarrow I_1 I_2$

$$\text{conf} = \frac{\sigma(I_1 I_2 I_3)}{\sigma(I_3)} = \frac{2/9}{6/9} \times 100$$

$$= 33.33\% < 50\%$$

$I_1 I_2 \rightarrow I_3$ ✓

$$\text{conf} = \frac{\sigma(I_1 I_2 I_3)}{\sigma(I_1 I_2)} \frac{2/9}{4/9} \times 100 = 50\% \quad (= 50\%)$$

$I_1 I_3 \rightarrow I_2$ ✓ conf $= \frac{\sigma(I_1 I_2 I_3)}{\sigma(I_1 I_3)}$

$= \frac{2/9}{4/9} \times 100 = 50\%$ 50%.

$I_2 I_3 \rightarrow I_1$ ✓ conf $= \frac{\sigma(I_1 I_2 I_3)}{\sigma(I_2 I_3)} = \frac{2/9}{4/9} \times 100$

$= 50\% <= 50\%$

$I_1 \rightarrow I_2 I_5$

Support $= \frac{2}{9} \times 100 = 22.22\% > 20\%$

conf $= \frac{\sigma(I_1 I_2 I_5)}{\sigma(I_1)} \times 100$

$= \frac{2}{9} \times \frac{9}{6} \times 100 = 33.33\% <$ 50%.

$I_2 \rightarrow I_1 I_5$

conf $= \frac{\sigma(I_1 I_2 I_5)}{\sigma(I_2)} \times 100$

$= \frac{2}{9} \times \frac{9}{7} \times 100 = 28.57\% < 50\%$

$I_5 \rightarrow I_1 I_2$

(✓) $\text{conf} = \frac{\sigma(I_1 I_2 I_5)}{\sigma(I_5)} = \frac{2}{9} \times \frac{9}{2} \times 100$

$= 100\% \geq 50\%$

$I_1 I_2 \rightarrow I_5$

(✓) $\text{conf} = \frac{\sigma(I_1 I_2 I_5)}{\sigma(I_1 I_2)} = \frac{2}{9} \times \frac{9}{4} \times 100$

$= 50\% \geq 50\%$

$I_1 I_5 \rightarrow I_2$

(✓) $\text{conf} = \frac{\sigma(I_1 I_2 I_5)}{\sigma(I_1 I_5)} = \frac{2}{9} \times \frac{9}{2} \times 100$

$= 100\% \geq 50\%$

Rules $I_1 I_3 \rightarrow I_2$

$I_2 I_3 \rightarrow I_1$

$I_5 \rightarrow I_1 I_2$

$I_1 I_2 \rightarrow I_5$

$I_1 I_5 \rightarrow I_2$.

① Apply Apriori Algorithm for Market Basket Analysis.

Soln • Dataset used has 20,000 entries over 9000 transactions

• Coffee & bread are most frequent items

• Afternoon is peak hours of sales.

• most sales are on Fri, Sat & Sun.

• most productive month is march

• The support for Olptajones & Bread is highest.

• Bread → Pastry, Cake → Coffee, Coffee → Cake, Cake → Tea are

• Seeing connection graph, coffee has most connections: 8 followed by cake with 2.

• The rules are refined further to being Tea → Bread & Coffee → Bread as frequent patterns.

Design FP Tree for breast cancer detection.

Soln The FP Tree growth algorithm is used mainly to decrease the number of features in the dataset to feed in to the neural network. After dim. reduction, the features are fed to NN.

Some terms: confidence $(X \rightarrow Y) = \frac{\text{Sup}(X \cup Y)}{\text{support}(X)}$

Database:

#Records	#Malignant	#benign
699	241	458

Summary of db characteristics:

Code	Desp	mean	Std
I	Thickness	...	...
II	Uniformity of cell size	...	...
⋮	⋮	⋮	⋮
IX	mitoses	...	...

③ Apply vertical data ~~representation~~ mining domain.

Now our task is to bring down the number of features to be fed to the ANN.

1st method: Suppose we have association rule $X \rightarrow Y$ where X can take up multi itemsets & Y also. if the set items in X pass the minSup & minConf levels, we can remove Y.

Eg $\{I, III, VIII, IX\} \rightarrow \{II\}$

Here I, III, VIII, IX have values > minSup & minConf, so II can be removed.

(II) Max Miner method

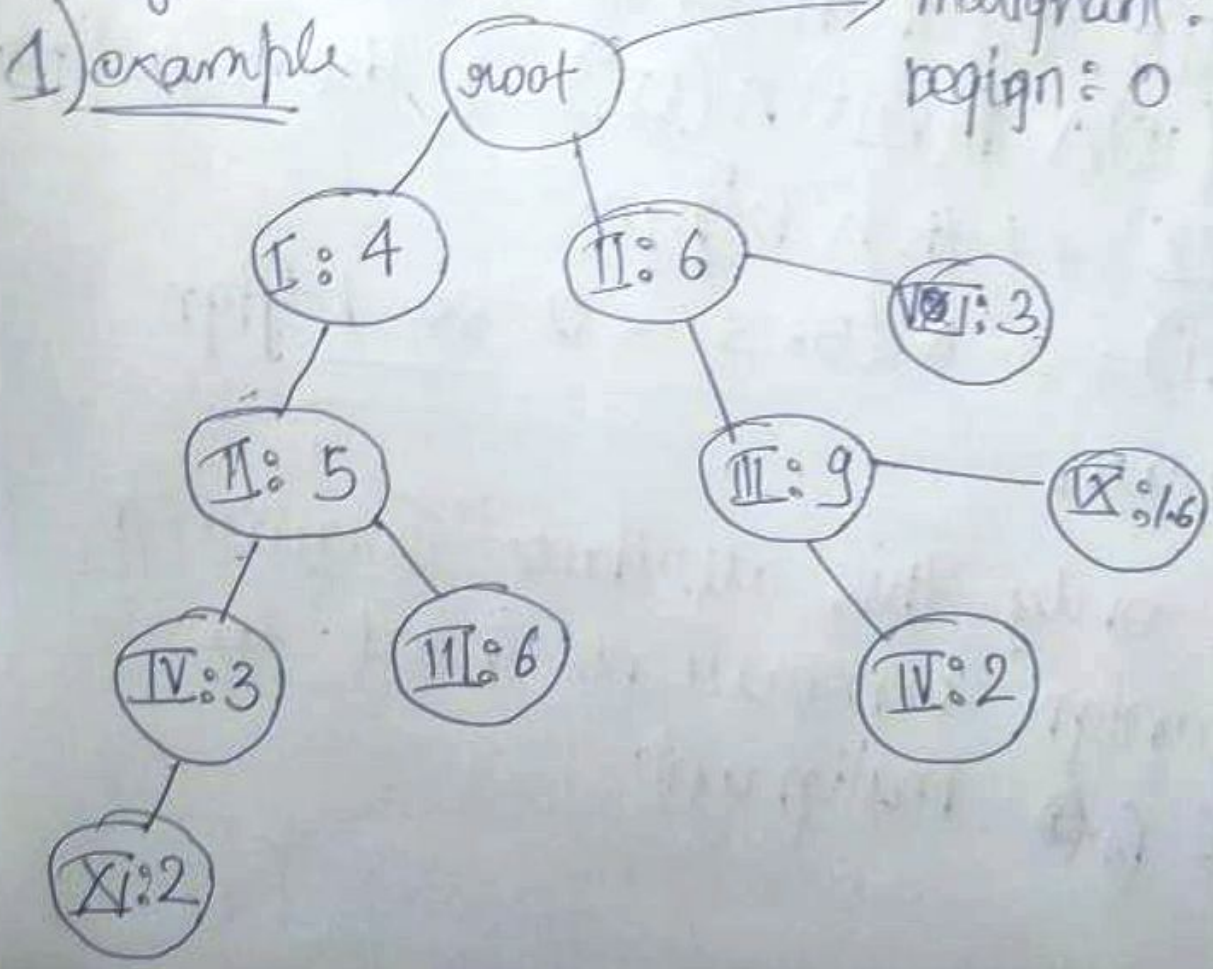
Here we choose only supersets of features that do not have subsets that are frequent. This will help reduce the # of features.

Class Begign (Value = 1):
Input: {II, VIII, IX}
Class malignant (Value = 1):
Input: {VI}

This implies classes II, VIII IX enough for malignant class detection & VI enough for begign class detection.
This can be fed to ANN.

FP growth tree

1) example



∴ To solve above example, first lets find the means of each of the features that are important

	Mean	
II	3.1	⟶ Uniformity of Cell
VIII	2.3	⟶ Normal Nucleoli
IX	1.6	⟶ mitoses.
VI	3.5	⟶ Bare Nuclei

Now these means act as minSup. we back track from leaf to root: If the value > minSup assign 1 else 0. Then and these values.

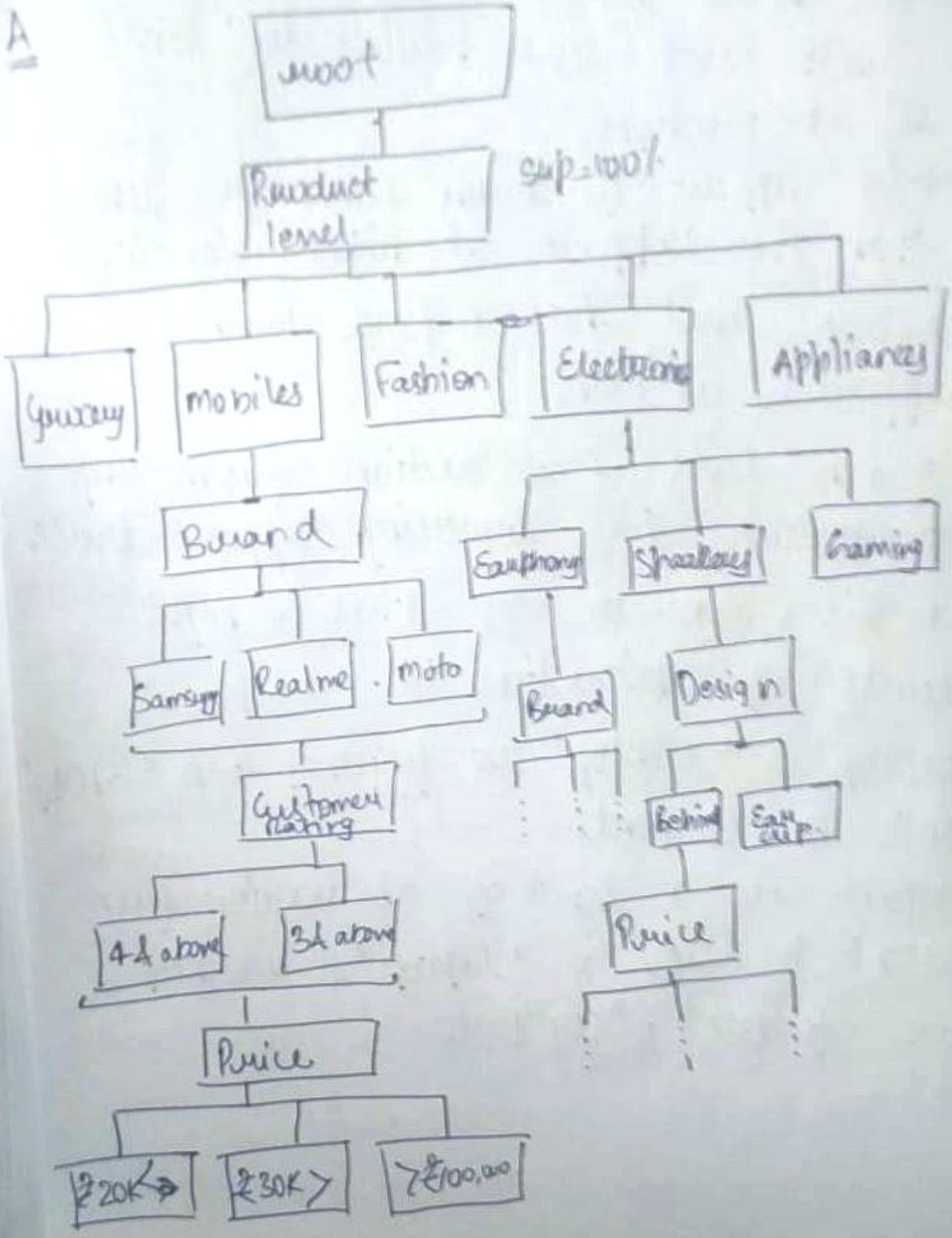
If (II) ∧ (VIII) ∧ (IX) = 1 then maligrant & VI = 1 then begign.

⟶ (VI) = 3 < 3.5 = 0 so begign tumour.

Since only this attribute required for benign tumour no need to check for maligrant.

3) Design multilevel multidimensional pattern mining for FlipKart.

A



The previous page shows a pattern mining format for Flipkart.

→ I have built a multilevel model with each level having diff level of abstraction.

→ i.e support @ lower levels is less than the support at higher levels.

→ Each level has a fixed threshold support & confidence

→ Each level of abstraction from top to bottom has lowering support levels.

→ This means to say that @ higher levels of abstraction if "computer", "laptop" & "desktop" are frequent then "laptop" will be frequent.

→ But at a level of abstraction from parent to leaf, if "laptop" & "desktop" are frequent, "computer" is not..

Module 4

1. Design DT for play tennis example.

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Overcast	Hot	High	False	Y
Rainy	Mild	High	False	Y
Rainy	Cool	Normal	False	Y
Rainy	Cool	Normal	True	N
Overcast	Cool	Normal	True	Y
Sunny	Mild	High	False	N
Sunny	Cool	Normal	False	Y
Rainy	Mild	Normal	False	Y
~~Overcast~~ Sunny	Mild	Normal	True	Y
Overcast	Mild	High	True	Y
Overcast	Hot	Normal	False	Y
Rainy.	Mild	High.	True	N.

Entropy (D) $= -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14}$

$= 0.4097 + 0.530$

$= 0.940.$

E(Temp)

E(hot) = (2+, 2-) = 1

$E_{mild} = (4+, 2-) = -\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6}$

$= 0.389 + 0.528$

$= 0.917.$

$E_{cool} = (3+, 1-) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}$

$= 0.311 + 0.5$

$= 0.811$

$E_{not} = (2+, 2-) = 1$

$Gain(Temp) = 0.940 - \frac{4}{14} \times 1 - \frac{6}{14} \times 0.917 - \frac{4}{14} \times 0.811$

$= \boxed{0.029}$

E(humidity).

$E_{High}(3+, 4-) = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}$

$= 0.5283 + 0.461$

$= 0.984$

$E(6+, 1-) = -\frac{6}{7}\log\frac{6}{7} - \frac{1}{7}\log\frac{1}{7}$

$= 0.190 + 0.401 = 0.591$

$E_{normal}(6+, 1-)$

$= -\frac{6}{7}\log\frac{6}{7} - \frac{1}{7}\log\frac{1}{7}$

$= 0.401 + 0.190$

$= 0.591$

$E_{high}(3+, 4-)$

$= -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}$

$= 0.523 + 0.461$

$= 0.984$

Gain (humidity) $= 0.940 - \frac{7}{14} \times 0.591 - \frac{7}{14} \times 0.984$

$= 0.940 - 0.2955 - 0.49$

$= \boxed{0.1745}$

$E(wind)$

$E_{True}(3+, 3-) = 1$

$E_{False}(6+, 2-) = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}$

$= 0.311 + 0.5 = 0.811$

Gain (wind) $= 0.940 - \frac{6}{14} \times 1 - \frac{8}{14} \times 0.811$

$= 0.940 - 0.428 - 0.463$

$= \boxed{0.049}$

Outlook (Information Gain)

$E(\text{Sunny}) (2+, 3-)$

$= -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}$

$= 0.233$

$E(\text{Outlook}) = 0.940$

$= E(\text{Overcast}) (4+, 0-)$

$= 0$

$E(\text{Rainy}) (3+, 2-)$

$= -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}$

$= 0.442 + 0.528 = 0.970$

$\text{Gain}(\text{Outlook}) = 0.940 - \frac{5}{14} \times 0.233 - \frac{5}{14} \times 0.970$

$= \boxed{0.510}$

Root Node = 0.029 (T) | 0.049 (W) | 0.510 (O) | 0.1785 (H)

Outlook is root node

(2) Design Bayesian classifier for buying PC

PTO.

age	income	student	credit	com
<=30	high	N	fair	N
<=30	high	N	exc	N
31...40	high	N	fair	Y
>40	medium	N	fair	Y
>40	low	Y	fair	Y
>40	low	Y	excl	N
31...40	low	Y	excl	Y yes.
<=30	medium	N	fair	N
<=30	low	Y	fair	Y
>40	medium	Y	fair	Y
<=30	medium	Y	excl	Y
31...40	medium	N	excl	Y
31...40	high	Y	fair	Y
>40	medium	N	excl	N

$P(Y) =$

age	Y	N
<=30	2/9	3/5
31...40	4/9	0
>40	3/9	2/5

student	Y	N
Y	6/9	1/5
N	3/9	4/5

income	Y	N
high	2/9	2/5
medium	4/9	2/5
low	3/9	1/5

credit	Y	N
fair	6/9	2/5
excl	3/9	3/5

Apply SVM for breast cancer domain.

Preprocessing: ① Check for null values.

EDA: Check for value counts of benign & malignant tumours here. Box Plots & Violin Plots allow us to see what distributions the data follows.

Classification Models: Split the data into train & test sets.

KNN gave an accuracy of 92%. If there are normalization problems, we can scale the data.

We can further use SVM to classify the type of cancer & tweak with multiple kernels. SVM gave accuracy of 98%.

Logistic Regression gave an accuracy of 0.92.

So SVM can be the best model out of all of them.

④ Apply outliers detection in t-Dom indu

① Tukey's IQR method.

① Find Q1 ② find Q3 ③ IQR = $Q_3 - Q_1$

④ Define the normal data range with lower limit Q1 - 1.5 IQR & upper limit as $Q_3 + 1.5$ IQR.

Any point beyond this range is outliers.

② Standard deviation

68% of data lies within 1 std deviation of mean.

95% of data lies in 2 std dev of mean.

99.7% of data lies in 3 std dev of mean.

As the std dev increases the outliers increases.

Z score method: We rescale & center the data & look for pts which are too far from zero.

modified Z score value is robust to outliers

Module 5

Analyse the sentiments & polarize into +ve & -ve sentiments

① I have used nltk to classify tweets.
② check the value-counts of +ve, -ve & 0.
③ Remove null value tweets
④ Remove URLS from tweets
⑤ Tokenize text, Remove emojis, newline characters & punctuation signs.
⑥ Lowercase all text.

Model Building

Some models I built were single LSTM with 0.74 accuracy.

Bidirectional LSTM with 0.704 accuracy

One D CNN with 0.59 accuracy

Hyperture parameters to increase accuracy.

(2) Analyse sarcasm using clustering approach on Twitter dataset.

A (1) Drop article_link attribute
(2) Check value_counts of sarcastic & non sarcastic tweets
(3) Plot word cloud of tweets.
(4) Split data into train & testset.
(5) Use tfidf to vectorize the text.
(6) Use logistic regression → 0.76 f1 score.
Use Naive bayes → 0.784 f1 score
(7) Voting classifiers → 0.73 f1 score.
(8) Linear SVC gave the best results
(9) RNN gave accuracy of 77%.