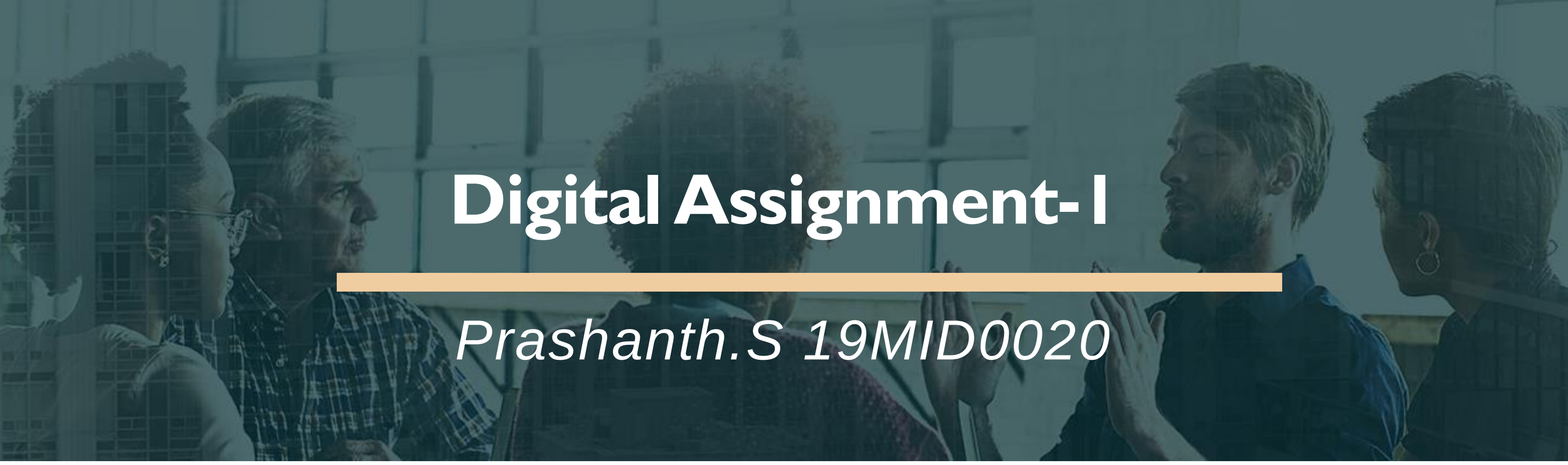
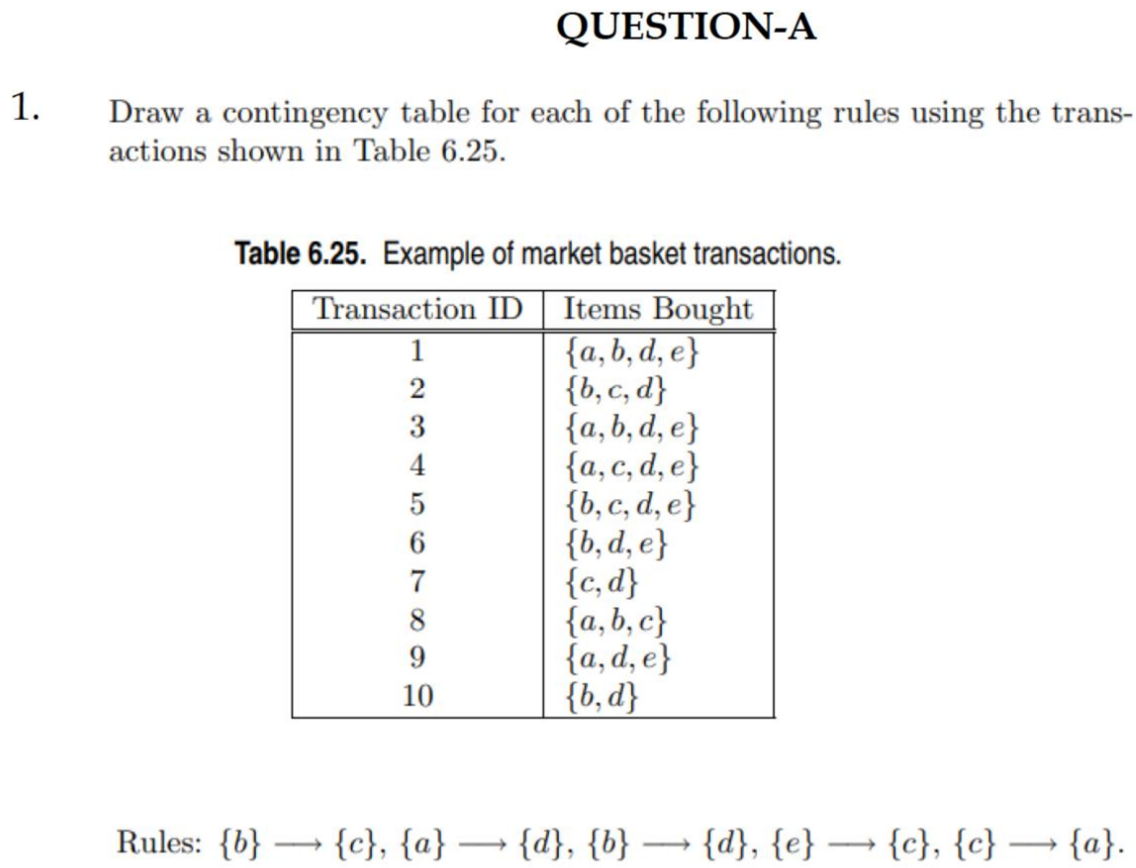


# Digital Assignment-1

*Prashanth.S 19MID0020*

## Question-1



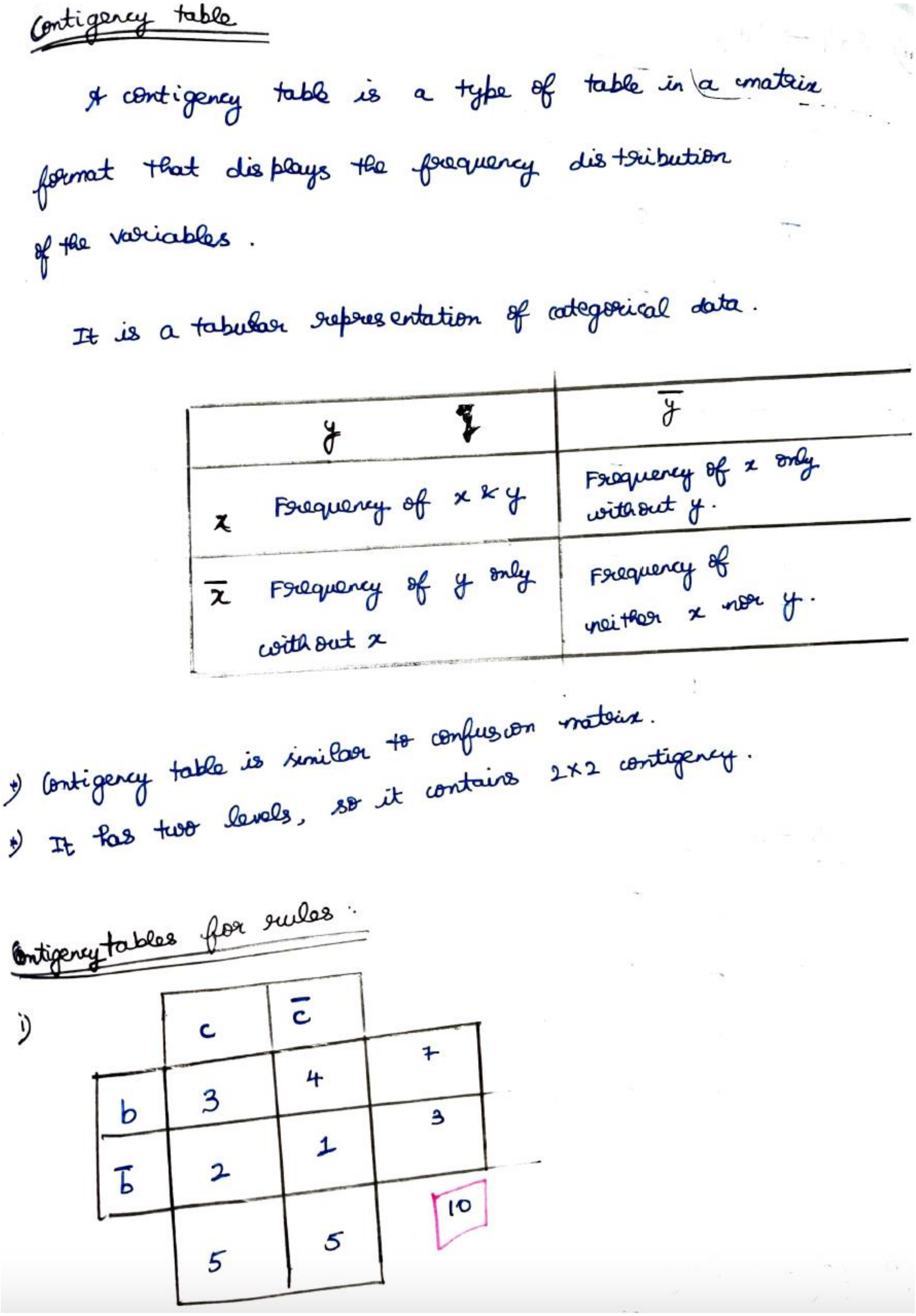
### QUESTION-A

1. Draw a contingency table for each of the following rules using the transactions shown in Table 6.25.

**Table 6.25.** Example of market basket transactions.

| Transaction ID | Items Bought |
|---|---|
| 1 | $\{a, b, d, e\}$ |
| 2 | $\{b, c, d\}$ |
| 3 | $\{a, b, d, e\}$ |
| 4 | $\{a, c, d, e\}$ |
| 5 | $\{b, c, d, e\}$ |
| 6 | $\{b, d, e\}$ |
| 7 | $\{c, d\}$ |
| 8 | $\{a, b, c\}$ |
| 9 | $\{a, d, e\}$ |
| 10 | $\{b, d\}$ |

Rules: $\{b\} \longrightarrow \{c\}$, $\{a\} \longrightarrow \{d\}$, $\{b\} \longrightarrow \{d\}$, $\{e\} \longrightarrow \{c\}$, $\{c\} \longrightarrow \{a\}$.



## Contigency table

A contigency table is a type of table in a matrix format that displays the frequency distribution of the variables.

It is a tabular representation of categorical data.

|  | $y$ | $\bar{y}$ |
|---|---|---|
| $x$ | Frequency of $x$ & $y$ | Frequency of $x$ only without $y$. |
| $\bar{x}$ | Frequency of $y$ only without $x$ | Frequency of neither $x$ nor $y$. |

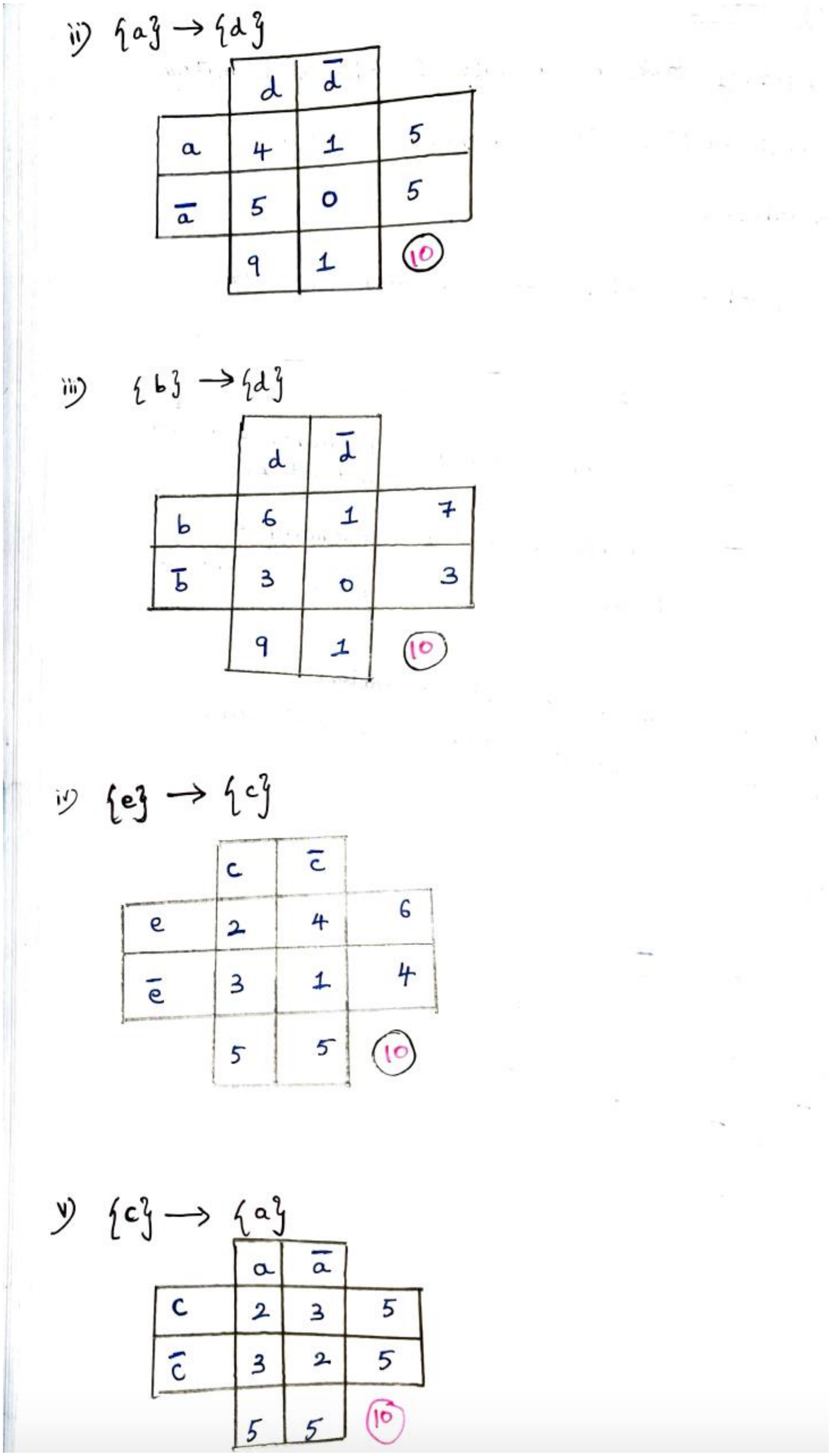
*) Contigency table is similar to confusion matrix.

*) It has two levels, so it contains 2×2 contigency.

## Contigency tables for rules :

i)

|  | c | $\bar{c}$ |  |
|---|---|---|---|
| b | 3 | 4 | 7 |
| $\bar{b}$ | 2 | 1 | 3 |
|  | 5 | 5 | 10 |



ii) {a} → {d}

|  | $d$ | $\bar{d}$ |  |
|---|---|---|---|
| $a$ | 4 | 1 | 5 |
| $\bar{a}$ | 5 | 0 | 5 |
|  | 9 | 1 | 10 |

iii) {b} → {d}

|  | $d$ | $\bar{d}$ |  |
|---|---|---|---|
| $b$ | 6 | 1 | 7 |
| $\bar{b}$ | 3 | 0 | 3 |
|  | 9 | 1 | 10 |

iv) {e} → {c}

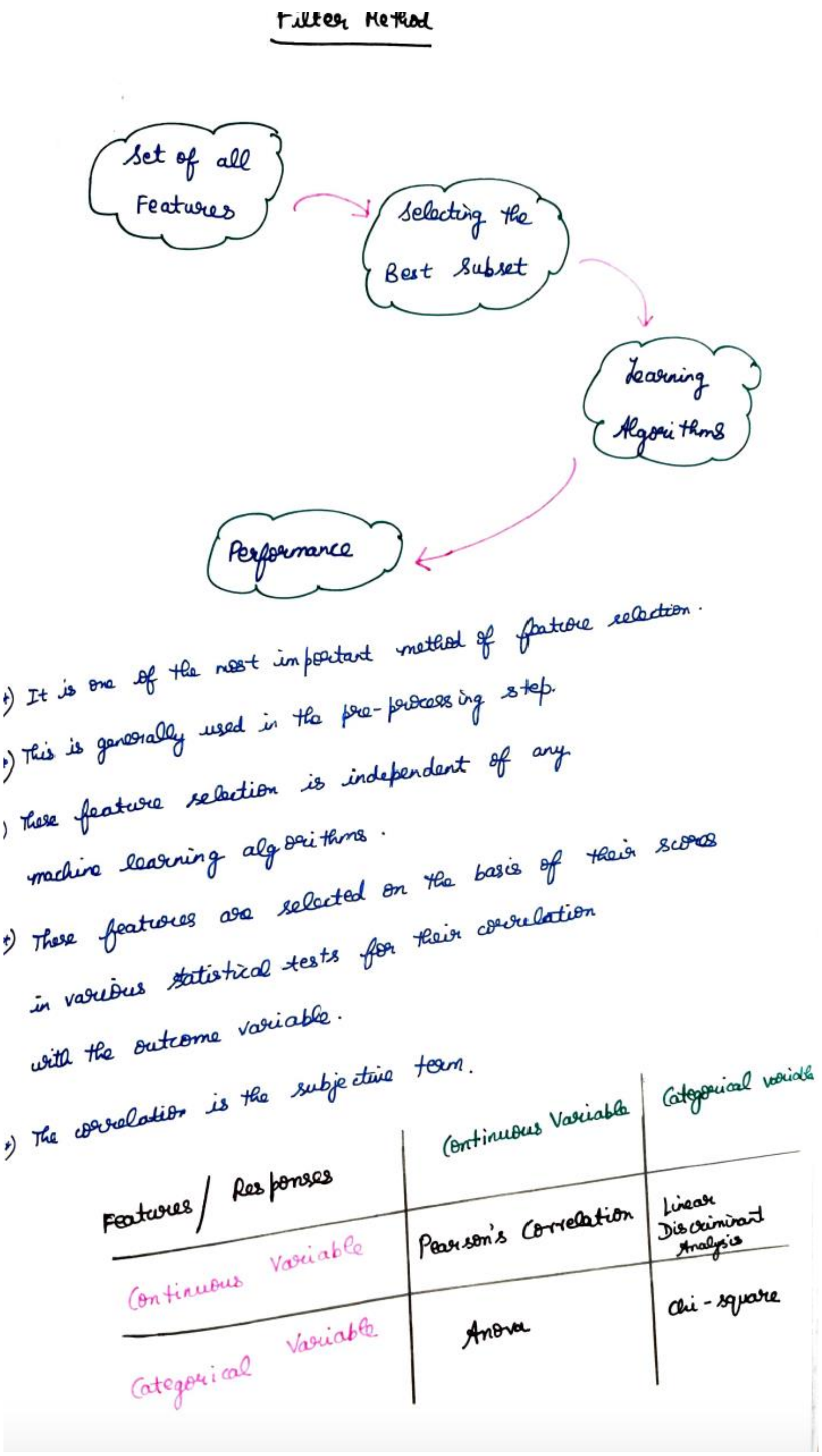
|  | $c$ | $\bar{c}$ |  |
|---|---|---|---|
| $e$ | 2 | 4 | 6 |
| $\bar{e}$ | 3 | 1 | 4 |
|  | 5 | 5 | 10 |

v) {c} → {a}

|  | $a$ | $\bar{a}$ |  |
|---|---|---|---|
| $c$ | 2 | 3 | 5 |
| $\bar{c}$ | 3 | 2 | 5 |
|  | 5 | 5 | 10 |

# Question-2

2. Narrate the procedure with sample dataset about the preference of filter approach and wrapper approach in feature/variable selection of data pre-processing of datamining.



Filter Method

Set of all Features
Selecting the Best Subset
Learning Algorithms
Performance

*) It is one of the most important method of feature selection.

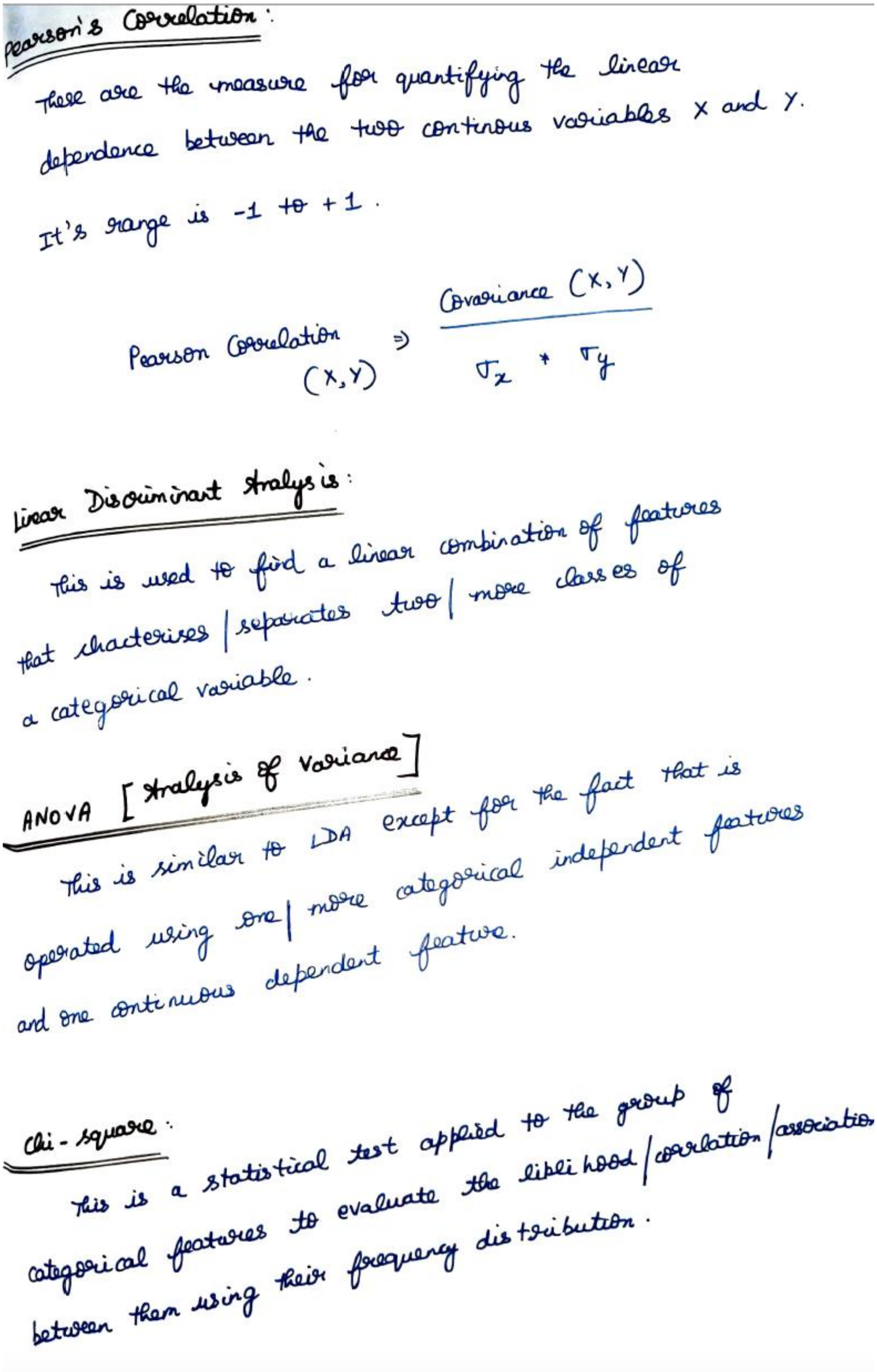
*) This is generally used in the pre-processing step.

) These feature selection is independent of any machine learning algorithms.

*) These features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable.

*) The correlation is the subjective term.

| Features/ Responses | Continuous Variable | Categorical variable |
|---|---|---|
| Continuous Variable | Pearson's Correlation | Linear Discriminant Analysis |
| Categorical Variable | Anova | Chi-square |



## Pearson's Correlation :

These are the measure for quantifying the linear dependence between the two continous variables X and Y.

It's range is -1 to +1.

$$\text{Pearson Correlation } (X,Y) \Rightarrow \frac{\text{Covariance } (X,Y)}{\sigma_x * \sigma_y}$$

## Linear Discriminant Analysis :

This is used to find a linear combination of features that chacterises|separates two|more classes of a categorical variable.

## ANOVA [Analysis of Variance]

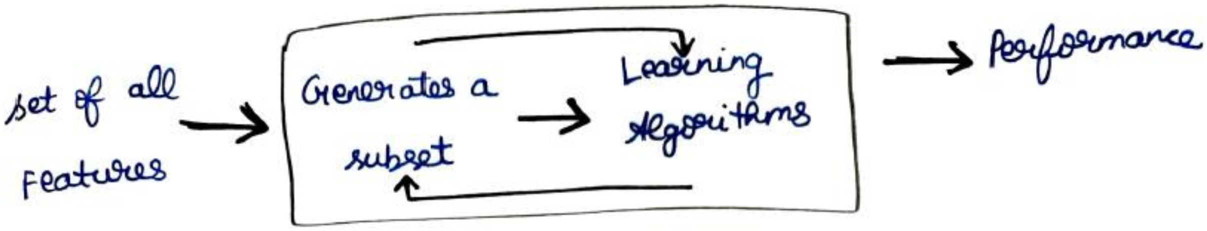
This is similar to LDA except for the fact that is operated using one|more categorical independent features and one continuous dependent feature.

## Chi-square :

This is a statistical test applied to the group of categorical features to evaluate the liklihood|correlation|association between them using their frequency distribution.

# Wrapper Method

## Selecting the Best Subset



*) In wrapper method, we try to use a subset of features and train a model using them.

*) Based on the inferences that we draw from the previous model, we decide to add/remove features from your sub-set.

*) The problem is essentially reduced to a search problem.

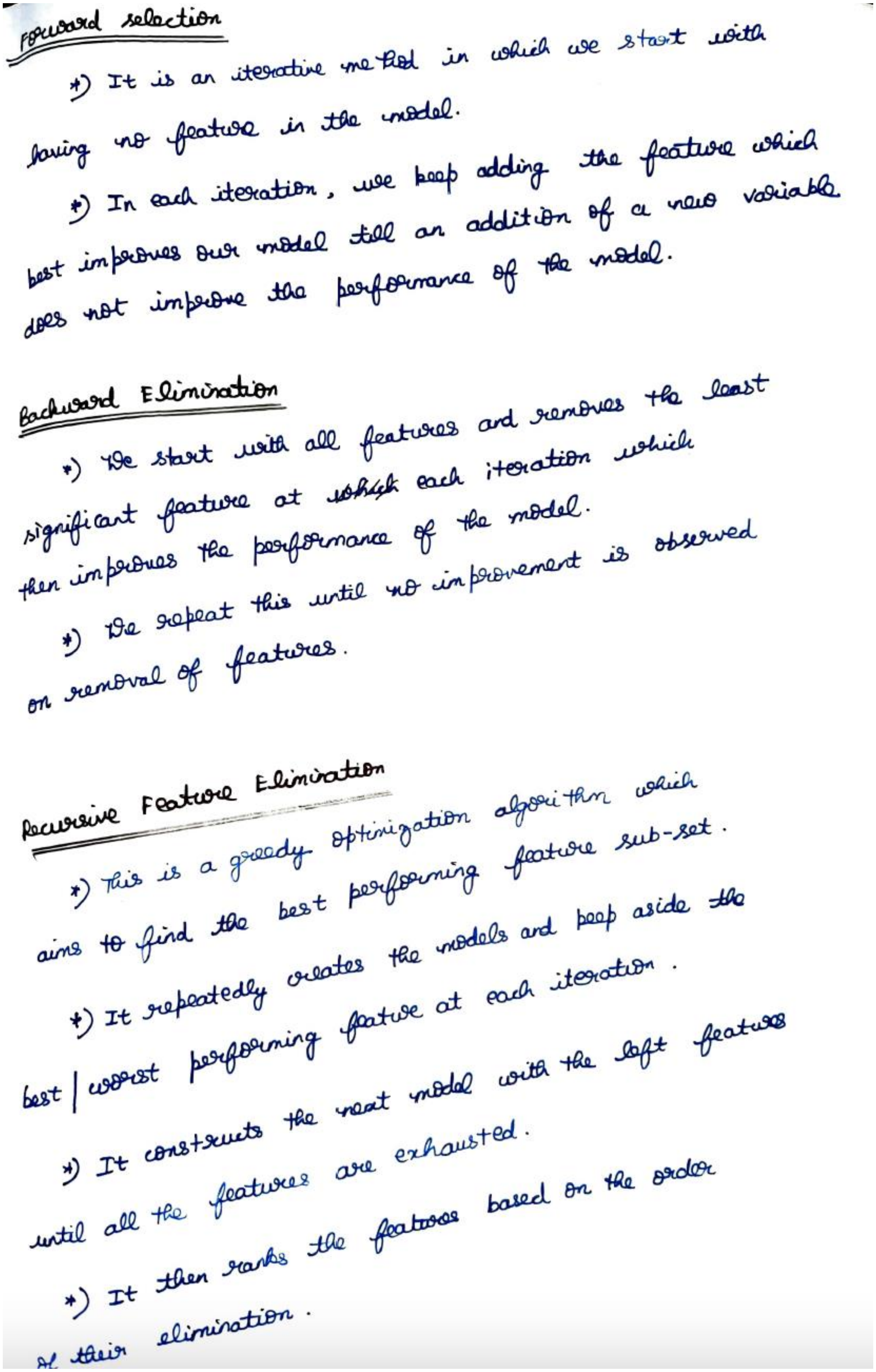
*) These methods are usually computationally very expensive.

Common examples of wrapper method

Forward Feature Selection

Backward feature elimination

recursive feature elimination



## Forward selection

*) It is an iterative method in which we start with having no feature in the model.

*) In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

## Backward Elimination

*) We start with all features and removes the least significant feature at ~~which~~ each iteration which then improves the performance of the model.

*) We repeat this until no improvement is observed on removal of features.

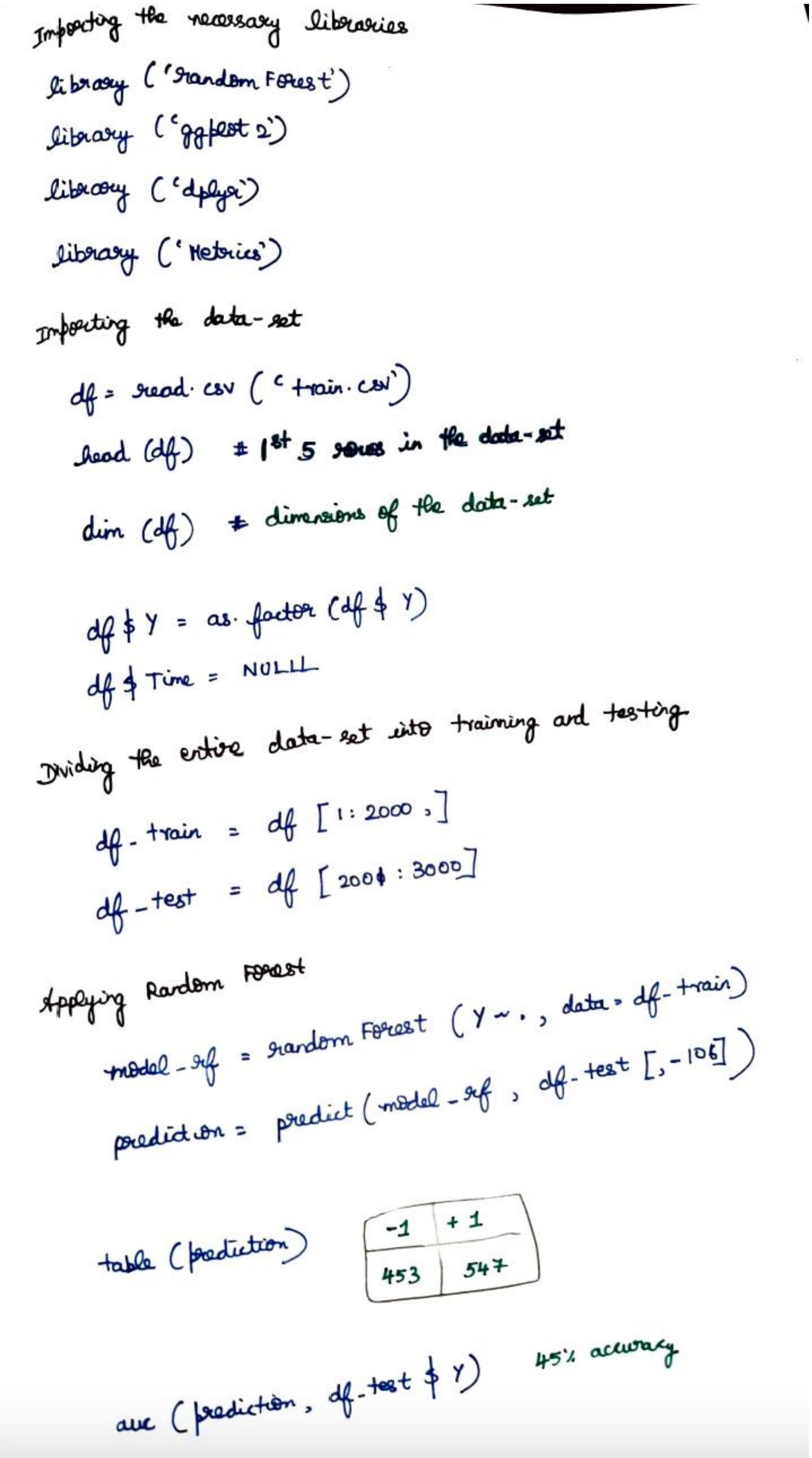
## Recursive Feature Elimination

*) This is a greedy optimization algorithm which aims to find the best performing feature sub-set.

*) It repeatedly creates the models and keep aside the best | worst performing feature at each iteration.

*) It constructs the next model with the left features until all the features are exhausted.

*) It then ranks the features based on the order of their elimination.



Importing the necessary libraries

```
library ('randomForest')
library ('ggplot 2')
library ('dplyr')
library ('Metrics')
```

Importing the data-set

```
df = read.csv ('train.csv')
head (df)   # 1st 5 rows in the data-set
dim (df)   # dimensions of the data-set

df $ Y = as.factor (df $ Y)
df $ Time = NULL
```

Dividing the entire data-set into training and testing

```
df-train = df [1:2000,]
df_test = df [2001:3000]
```

Applying Random Forest

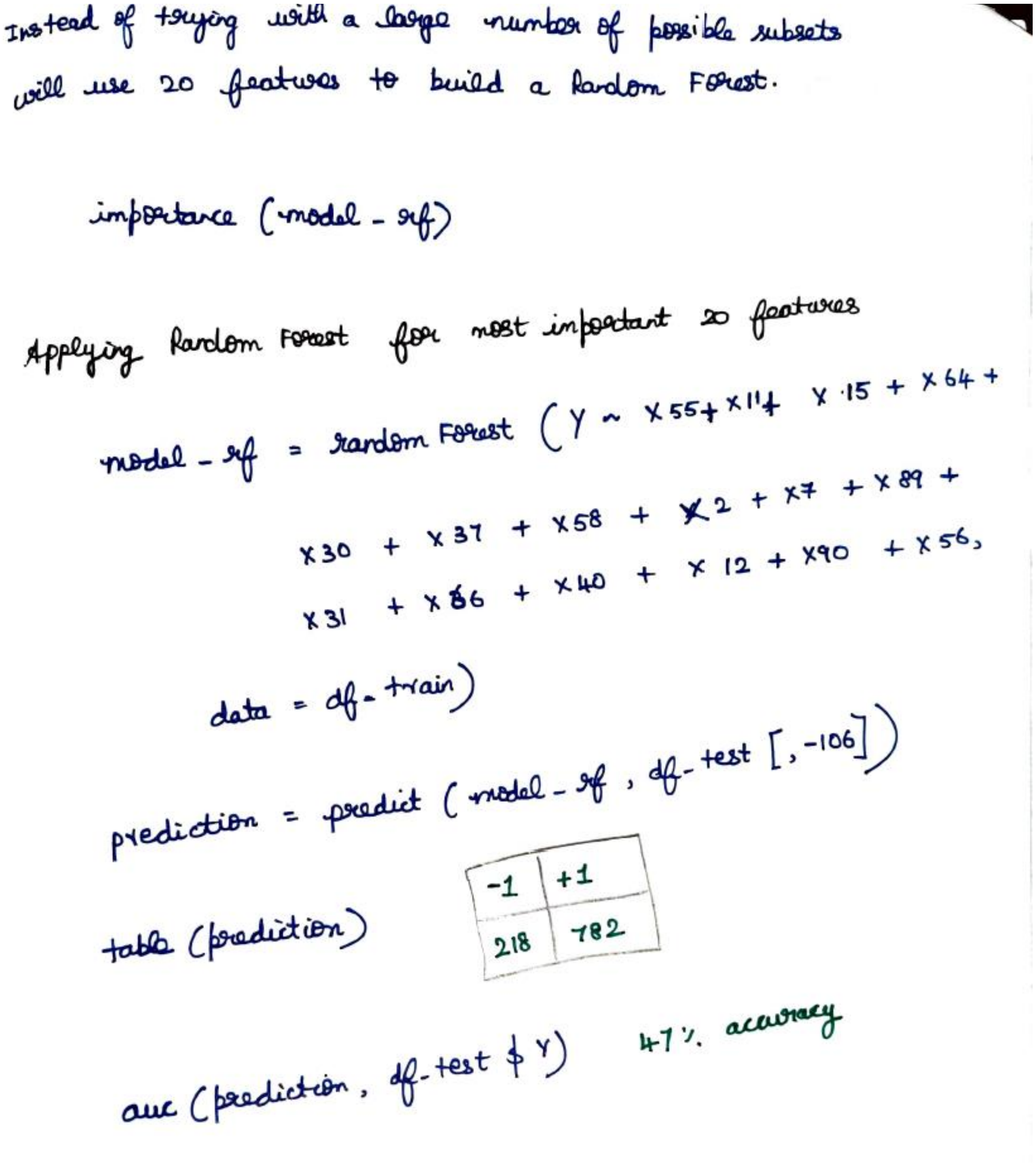
```
model_rf = randomForest (Y~., data = df-train)
prediction = predict (model_rf, df-test [,-106])

table (prediction)
```

| -1 | + 1 |
|---|---|
| 453 | 547 |

```
auc (prediction, df-test $ Y)
```

45% accuracy



Instead of trying with a large number of possible subsets will use 20 features to build a Random Forest.

importance (model - rf)

Applying Random Forest for most important 20 features

model - rf = random Forest (Y ~ X 55 + X11 + X ·15 + X 64 +
X 30 + X 37 + X58 + X2 + X7 + X 89 +
X 31 + X 86 + X40 + X 12 + X90 + X 56,
data = df - train)

prediction = predict (model - rf , df - test [, -106])

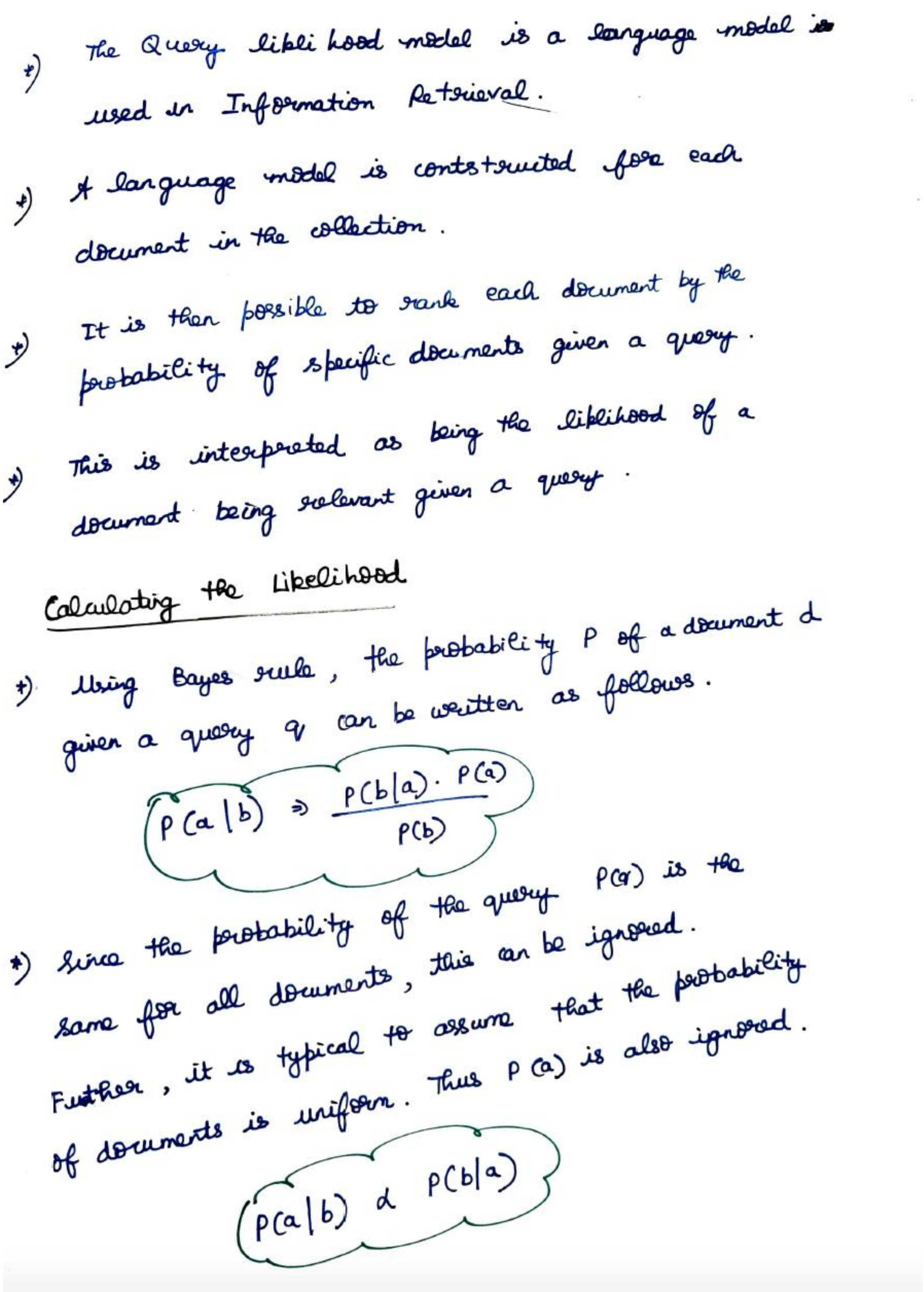
table (prediction)

| -1 | +1 |
|---|---|
| 218 | 782 |

auc (prediction , df - test $ Y) 47 % accuracy

# Question-3

3. Elaborate with your example about the usage of the Text Mining for Query Likelihood Estimation



*) The Query likli hood model is a language model is used in Information Retrieval.

*) A language model is contstructed for each document in the collection.

*) It is then possible to rank each document by the probability of specific documents given a query.

*) This is interpreted as being the liklihood of a document being relevant given a query.

Calculating the Likelihood

*) Using Bayes rule, the probability P of a document d given a query q can be written as follows.

$$P(a|b) \Rightarrow \frac{P(b|a) \cdot P(a)}{P(b)}$$

*) Since the probability of the query P(q) is the same for all documents, this can be ignored. Futher, it is typical to assume that the probability of documents is uniform. Thus P(a) is also ignored.

$$P(a|b) \alpha P(b|a)$$

*) Documents are than ranked by the probability that a query is observed on a random sample from the document model .

*) the multinomial unigram language model is commonly used to achieve this .

*) we have

$$P(b \mid M_a) = K_b \prod_{t \in V} P(t \mid M_a)$$

where . the multi-nomial coefficient is

$$K_q \Rightarrow \frac{L_q!}{t \cdot f_{t1,q}!\ t \cdot f_{t2,q}! \ldots\ldots t \cdot f_{tN,q}!}$$

for query q

$$L_q \Rightarrow \sum_{1 \leq i \leq N} t f_{ti,q}$$ is the length of the query q.

given the term frequency tf in the query vocabulary N.

*) In practice, the multinomial coefficient is usually removed from the calculation. The reason is that it is a constant for a given bag of words. The language model Ma should be the true language model calculated from the distribution of words underlying each retrieved document. In practice this language model is un-known so it is usually approximated by considering each term (unigram) from the retrieved document together with its probability appearance.

*) This calculation is repeated for all other documents to create a ranking of all documents in document collection.