Answer to Q1:

a) Insufficient training data -

Ans.**1. Linear Discriminant Analysis**

Because Logistic regression can become unstable when there are few examples from which to estimate the parameters.

b) Class imbalance -

Ans. **3. Either of the two.** We need to use additional sampling in case of class imbalance to improve the accuracy of model

c) Different co-variance matrices for the classes with Gaussian distribution.

Ans. **2. Logistic Regression** works well when co-variance matrices differ for the classes, because LDA expects Homogeneity of variance /covariance among classes.

d) Uniform distribution instead of Gaussian distribution.

Ans. **2. Logistic Regression** because LDA for simplification assumes that data is gaussian.

Answer to Q2:

Initial Entropy:

$H(x) = -(((4/10)*\log2(4/10))+((6/10)*\log2_2(6/10)))$

$= 0.971$

Information Gain (Weather):

Entropy(Weather) =
(4/10)*(-((3/4)*log2(3/4)+(1/4)*log2(1/4))) (Weather=Fine)
+ (4/10)*(-((3/4)*log2(3/4)+(1/4)*log2(1/4))) (Weather=Rain)
+ (2/10)*0 (Weather=Cloudy)
= 0.646
Information Gain(Weather) = Initial Entropy - Entropy(Weather) =0.971 - 0.646= 0.325

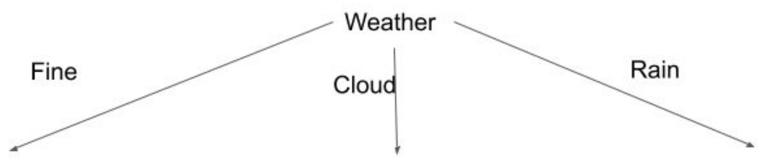
Information Gain (Humidity):
Entropy(Humidity) =
(5/10) * (-((⅖)*log2(⅖)+(⅗)*log2(⅗))   (Humidity = High)
+ (5/10) * (-((⅘)*log2(⅘)+(⅕)*log2(⅕)) (Humidity = Medium)
= 0.843

Information Gain (Humidity) = Initial Entropy - Entropy(Humidity)
= 0.971 - 0.843 = 0.128

Information Gain (Wind):
Entropy(Wind) =
(7/10) * (-((5/7)*log2(5/7)+(2/7)*log2(2/7))   (Wind = None)
+ (3/10) * (-((⅓)*log2(⅓   )+(⅔)*log2(⅔)) (Wind = Breezy)
= 0.88

Information Gain (Wind) = Initial Entropy - Entropy(Wind)
= 0.971 - 0.88 = 0.091

Since, Information gain is high if we split data set on attribute weather, we will pick Weather as an attribute to split initially.



RID	Humidity	Wind	Play Golf
1	high	none	no
3	high	none	no
4	medium	none	yes
5	high	breezy	no

RID	Humidity	Wind	Play Golf
2	medium	breezy	yes
6	high	none	yes

RID	Humidity	Wind	Play Golf
7	high	none	yes
8	medium	none	yes
9	medium	breezy	no
10	medium	none	yes

Now new computing information gain for each individual sub data sets

1. Sub Data Set Weather=Fine

Initial Entropy:

$H(x) = -(((1/4)*\log2(1/4))+((3/4)*\log2(3/4)))$

$= 0.811$

Information Gain (Humidity):

Entropy(Humidity) =

$(3/4) * (-((3/3)*\log2(3/3)+(0/3)*\log2(0/3)))$ (Humidity = High)

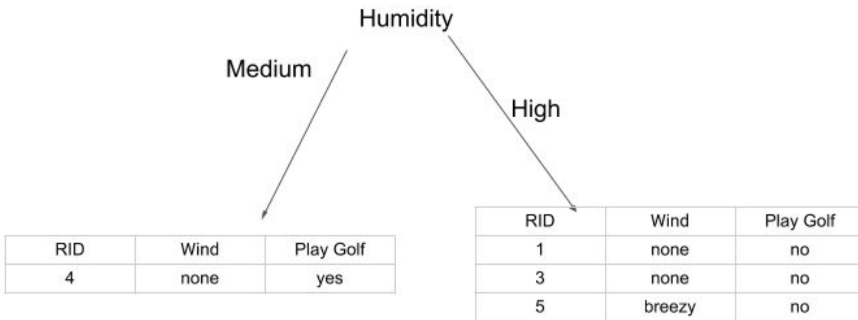
$+ (1/4) * (-((1/1)*\log2(1/1)+(0/1)*\log2(0/1)))$ (Humidity = Medium)

$= 0$

Information Gain (Humidity) = Initial Entropy - Entropy(Humidity)

$= 0.811 - 0 = 0.811$

Since here information gain is maximum here for this split, we can skip other attributes and split on Humidity



RID	Wind	Play Golf
4	none	yes

RID	Wind	Play Golf
1	none	no
3	none	no
5	breezy	no

2. Sub Data Set Weather=Cloud

We can stop examining further attributes for this split since information gain is maximum here.

3. Sub Data Set Weather=Rain

Initial Entropy:

H(x) = -(((¼)*log2(1/4))+((¾)*log2(¾)))
= 0.811

Information Gain (Humidity):

Entropy(Humidity) =
(3/4) * (-((2/3)*log2(2/3)+(1/3)*log2(1/3)) (Humidity = Medium)
+ (1/4) * (-((1/1)*log2(1/1)+(0/1)*log2(0/1)) (Humidity = High)
= 0.68875

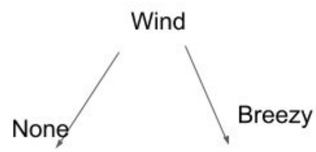
Information Gain (Humidity) = Initial Entropy - Entropy(Humidity)
= 0.811 - 0.68875 = 0.12225

Information Gain (Wind):

Entropy(Wind) =
(3/4) * (-((3/3)*log2(3/3)+(0/3)*log2(0/3)) (Wind = none)
+ (1/4) * (-((1/1)*log2(1/1)+(0/1)*log2(0/1)) (Wind = breeze)
= 0

Information Gain (Wind) = Initial Entropy - Entropy(Wind)
= 0.811 - 0 = 0.811

So for this split, if we split of Wind we get maximum gain.



RID	Humidity	Play Golf
7	high	yes
8	medium	yes
10	medium	yes

RID	Humidity	Wind	Play Golf
9	medium	breezy	no

Final Decision Tree will look like this.