**Time Limit: 3 Hours**

**For Question 2 and 3, if workings of the solutions are not shown, no marks will be given. If you copy and paste directly, it will be considered as plagiarism and that paper shall not be marked. You are encouraged to give your personal opinion and analysis.**

# Question 1:

1. Which of the following statements are correct? (Select all that are correct)
2. Information gain is measured by the reduction in entropy by splitting the data on the basis of a single attribute
3. Decision trees are used for classification but not regression
4. A Gini value of 0 implies a pure node
5. What is the distributional assumption underlying Linear Discriminant Analysis? (Select all that are correct)
6. Classes follow the lognormal distribution
7. Classes follow the bionomial distribution
8. Classes follow a normal distribution
9. Which of the following is a metric to use a split into a decision tree into additional nodes? (Select all that are correct)
10. Median
11. Entropy
12. Gini impurity
13. Standard Deviation

# Question 2:

Import the dataset attached (Invistico\_Airline.csv) in python notebook (Jupyter Notebook/Google Colab) and develop an interesting investigative theme for a data driven story which **should** include the following –

* No. of rows and columns
* Null Analysis
* Duplicated items
* EDA of total count of Satisfied vs Unsatisfied responses
* Correlation amongst variables (this include Customer Satisfaction column)
* EDA of total count of Satisfied vs Unsatisfied responses based on –

1. Gender
2. Customer Type
3. Age
4. Type of Travel
5. Class
6. Flight Distance
7. Seat Comfort

Submit the notebook file (.ipynb) along with a html file of your answer.

**Note:** Your analysis should not be just python codes and output but also 2-3 sentences describing your analysis in markdown.

# Question 3:

Import the dataset zoo.csv in python notebook (Jupyter Notebook/Google Colab) which contains 100 observations on different animals, with a number of features that describe each animal. The data has a column for species as well as a **class\_type** which is a numerical representation of the species. (the **class\_type** goes from 1 to 7, but the lables may require adjustments).

You need to build an XGBoost based classification model to predict the **class\_type** (which corresponds to species) using other features in the dataset. Train-Test split will be 80-20 ratio and **random\_state** is to be set to 1 (**random\_state = 1).**

Find the overall accuracy rate of the model.