Question 1

Correct answer - Minimum Bounding Rectangles (MBRs) are used in the Filter and Refine step of query processing.

-1 : Partially wrong answer
-2 : Completely incorrect answer

Question 2

Correct answer - To minimize or eliminate any variances or biases between the individual learners in the ensemble
-1 : Partially wrong answer
-2 : Completely incorrect answer

Question 3
ARFF file format
Example:

```
@RELATION house

@ATTRIBUTE houseSize NUMERIC
@ATTRIBUTE lotSize NUMERIC
@ATTRIBUTE bedrooms NUMERIC
@ATTRIBUTE granite NUMERIC
@ATTRIBUTE bathroom NUMERIC
@ATTRIBUTE sellingPrice NUMERIC

@DATA
3529,9191,6,0,0,205000
3247,10061,5,1,1,224900
4032,10150,5,0,1,197900
2397,14156,4,1,0,189900
2200,9600,4,0,1,195000
3536,19994,6,1,1,325000
2983,9365,5,0,1,230000
```

ARFF- 1mark

Partially correct example- 0.5

Fully correct example -1

Description of the format - 0.25

Question 4:

Neural nets – to determine the output value of a neuron; logistic regression – to

classify the incoming (unknown) data point into one of two classes, depending on

the sigmoid function's value being <0.5 or >0.5.

Question 5:

- List of techniques: https://en.wikipedia.org/wiki/Binary_classification
    - Decision trees
    - Random forests
    - Bayesian networks
    - Support vector machines
    - Neural networks
    - Logistic regression
    - kNN
- Do not accept clustering techniques
- For each technique, giving name gets 1 point, giving explanation gets 1 point
- Hieu: 166 - 220
- Haoyu: 221 - 270
- Yingjun: 271 - 320
- Duc: 321- 370

Question 8(a)
YARN (Yet Another Resource Negotiator) capabilities which MR v1 does not have:

graph processing, iterative modeling
compatible with v.1.0, ie. can run MapReduce jobs
offers better scalability
better cluster utilization
create (near) real-time applications.

Marks deduction:
-1 for 1 wrong capability
-2 for 2 wrong

Question 8(b)
Bulk Synchronous Parallel (BSP) model alternative to map reduce

Marks deduction:
-1 for wrong answer

Question 9 (2points)
Expected Answer:
  9a. Data Science based analysis - 1 point
  9b. Data Mining, Machine Learning, Regression, Classification - 1 point
Rubrics :
  9a. Must mention **Data Science based analysis/ Data Science Life Cycle** to get full 1 point

   Partial points (0.5) if Data Science/ Data analysis/Data Lifecycle/ Data based analysis is mentioned.
  9b. Must mention all three: **Data Mining, Machine Learning, Regression** and one other Method to get full 1 point.
   Partial points(0.5) for any two methods mentioned.

Question 10
Expected Answer :
10a. R programming language - 1 point
10b. Neural Network trains on the square root of 50 random numbers and predicts the sqrt of the 10 numbers in the test data.

Or

a neural network is taught how to predict square roots

Any rephrasing of the above is acceptable. Explanation is awarded 1 point

The question clearly states **to be specific** with the answer. If you have only mentioned about neural network or the hidden layers basically rephrasing all the comments given in the question - **Only 0.5/1 marks are awarded**

Question 11

1st card and last card needs to be flipped. (A and 7)
So minimum card to be flipped to know answer is 2.
No partial marking. Student needs to mention the cards and the number of cards to be flipped