Birla Institute of Technology and Science, Pilani
Second Semester 2021-22
CS F415 Data Mining
**Mid-Sem Test**

**Date: 11.03.2022** **Max. Time: 90 minutes** **Maximum Marks: 30**

Note: **Answers to all questions must be analytical, precise and complete.**

1. Write short answers for the following:
    a. Compare and contrast the Oblique decision tree and constructive induction approach for building the decision trees.
    b. Explain the meaning of penalty term in pessimistic estimation of generalization. What is the meaning when its value is 0.5?
    c. When do you day that decision tree is not giving a statistically significant decision? Give a relevant example and suggest a technique to handle it. What is the standard name of this problem?
    d. When do you say an attribute is redundant? How do you identify them? Mention some of the techniques to handle them.
    e. Describe two distance normalization and standardization methods.
    f. Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

    [2*6=12 marks]
2. Explain how the multiple comparison method is related to model overfitting with an example. [3 marks]
3. Assume you find out that two attributes have a correlation of 0.02; what does this tell you about the relationship of the two attributes? Answer the same question assuming the correlation is -0.98! [2 marks]
4. The probability that a person owns a sports car given that they subscribe to at least one automotive magazine is 40%. We also know that 3% of the adult population subscribes to at least one automotive magazine. Finally, the probability of a person owning a sports car given that they don't subscribe to at least one automotive magazine is 30%. Use this information together with Bayes theorem to compute the probability that a person subscribes to at least one automotive magazine given that they own a sports car. [2 marks]
5. Compute the Information Gain and Gini-gain for a 3-way split for a 3-class classification problem; the class-distribution before the split is (10, 5, 5) and after the split the class distribution is (0,0,5), (9, 2,0) and (1,3,0). Discuss the results. [4 marks]
6. A mobile phone company notices that 1% of its customers leave them each month for another service provider. The average cost of gaining a new customer is $12. They decide that it would be better to send a $10 coupon to customers who are likely to leave soon, in the hope of retaining them. You have been retained as an independent consultant to help them solve this problem.
   Describe the steps you would take, including at least answers to the following questions:

- what attributes would you like them to collect for you?
- over what time periods would you like the data collected?
- what kind of data mining techniques(s) would you apply (and in what order)?
- why are these the right techniques?
- what results do you expect to see?
- how would you validate your results?

[7 marks]