Warm-up on k-NN, Decision Trees, Random Forests

1. Why is the k-NN algorithm called a ‘lazy-learner’?

2. Which of the following distance metrics can not be used for k-NN?
   A) Manhattan
   B) Minkowski
   C) Tanimoto
   D) Jaccard
   E) Mahalanobis
   F) All of the above

3. What are the assumptions/pre-requisites of the k-NN algorithm?

4. Which of the following distance measure do we use in case of categorical variables in k-NN?

   (i) Hamming Distance
   (ii) Euclidean Distance
   (iii) Manhattan Distance

   A) (i)
   B) (ii)
   C) (iii)
   D) (i) and (ii)
   E) (ii) and (iii)
   F) (i), (ii), and (iii)

5. What would be the relation between the time taken by 1-NN,2-NN,3-NN:

   A) 1-NN >2-NN >3-NN
   B) 1-NN < 2-NN < 3-NN
   C) 1-NN = 2-NN = 3-NN (approx.)
   D) None of these

6. Discuss the differences and similarities between the following in groups:

(i) GINI Impurity
(ii) Entropy
(iii) Misclassification Error

7. What are the limitations of information gain?

8. How do you fit decision trees in the presence of missing values?

9. What is the difference between re-substitution error and generalization error? Discuss in groups.

10. What is the difference between rule-based classifiers and instance-based classifiers?

11. Write a pseudo-code for the random forest algorithm.