Show your working in all calculations.

5. This question is related to **Anomaly/Outlier Detection** and **Data Properties and Preprocessing**

   a) Briefly describe TWO possible applications of outlier detection: explain what the application is, what an outlier would indicate in that context and why it would be important to detect.

   b) What is the best distance (or similarity) measure for each of the following applications?

      i) measure the dissimilarity between two dogs based on 7 numeric attributes.

      ii) compare similar diseases with a set of medical tests that show results as either positive or negative.

      iii) find similar documents in a plagiarism checking system.

   c) For the following two vectors, $p = [1, 1, 0, 0, 0, 0, 1, 0, 0, 0]$ and $q = [0, 1, 0, 0, 0, 0, 1, 0, 1, 0]$, compute the following similarities:

      - Simple Matching Similarity
      - Jaccard Similarity
      - Cosine Similarity

   d) Explain the difference between **ordinal** and **continuous** attributes. Give TWO examples of each type of attribute.

END OF EXAMINATION PAPER