1 Bayesian Belief Network

1.1 Procedure:

1. Read the CSV file and clean the data by getting only the columns which contain the data required to build the Bayesian Network.
   (a) Each node in the network has 4-5 questions and each question has a rating from 1 to 5
   (b) Take the average of the ratings for each node and round it to the nearest integer.
   (c) Return only the nodes and the average ratings for each student in form of a Pandas DataFrame.
2. Create the graph as mentioned in the paper, add it in the Bayesian Network.
3. Calculate the conditional probabilities for each of the nodes.
   (a) Network nodes can be divided into two categories:
      i. **Root Nodes** : Nodes which do not have a parent node.
      ii. **Child Nodes** : nodes which have atelast one parent node.
   (b) Consider each node corresponds to a random variable (X) where X takes values from 1 to 5 ($X \in [1, 5]$).
   (c) To calculate the probability for the root nodes:
      i. $\Pr_X(x = k)$ is defined as the probability that the variable X takes the value k.
      ii. The numerator is calculated as the number of rows having value or X as k, i.e., number of rows with average rating equal to k for the given node.
      iii. The denominator would contain the total number of nodes.
      iv.
      $$\Pr_X(x = k) = \frac{\text{rows with rating = k}}{\text{total number of rows}}$$
   (d) To calculate the probability for the child nodes:
      i. $\Pr_{X|Y}(x = k|y = l)$ is defined as the probability that X takes the value k given that Y has the value l. In our case, as we can have multiple parents to the given node, Y can denote a tuple of all the parent nodes and l denotes a tuple of all its values.
      ii. The numerator is defined as the number of rows where X has value k and Y has value l, i.e., the number of rows which have the rating of the node as k and rating of the parent nodes as the tuple l
      iii. The denominator is defined as the number of rows where value of Y is l, i.e., the number of rows with rating of all the nodes mentioned in Y as l.
      iv.
      $$\Pr_{X|Y}(x = k|y = l) = \frac{\text{rows with rating of node and parent nodes = k and l respectively}}{\text{total number of rows with ratings of parent nodes = l}}$$
4. Adding the Bayesian check:
   (a) As we can have atuple which does not occur in any of the rows, then the probability (as mentioned in **??**, **??**) would be undefined. We mention a uniform probability in this case i.e. $\frac{1}{5}$.