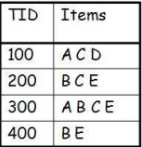
**CSC 3251 – DATA ANALYTICS**

**PRACTICE PROBLEMS**

1. Apply Apriori algorithm on the given transactional dataset and generate frequent itemsets and also determine the association rules, given that the min\_sup = 2 and minimum confidence threshold is 50%.



1. Find a linear regression equation for the following data.

|  |  |
| --- | --- |
| **Year** | **Life expectancy** |
| 1982 | 74.5 |
| 1987 | 75 |
| 1992 | 75.7 |
| 2010 | 78.7 |

Find the estimated life expectancy for an individual born in 2020.

1. Cluster the given set of points using **K-Means algorithm** (Euclidean distance measure) and **complete link agglomerative clustering** algorithm.

P1 (1,2)

P2 (1.5,1.5)

P3 (2,3)

P4 (3,4)

P5 (3.5,4.5)

P6 (6, 5)

1. Explain about Hadoop MapReduce and HDFS architecture with a neat diagram.
2. Explain about the given components of Hadoop ecosystem - Pig, Hive.