CS F415: DATA MINING
Assignment-4

Team Members :-

1. **Ayush Pandey [2015B3A70517H]**
2. **Abhinav kumar [2015B5A70674H]**
3. **Shivam Bhagat [2015B5A70460H]**
4. **Rohan Jain [2015B4A70676H]**

DATASET DESCRIPTION

The data-set used is NURSERY data-set (link) mentioned in assignment description. After loading the data into memory we made a 85%-20% train and validation split. Both the algorithm just uses the data from the training set (actively learning in Naive Bayes and passively in k-Nearest Neighbours). Apart from this no extra pre-processing were done on the dataset.

NAIVE BAYES ALGORITHM

1. **Accuracy on Training Set** : **90.33%**
2. **Accuracy on Validation Set** : **89.81%**
3. **Training Confusion Matrix:**

Predictions --------->

Labels | Pred	not_recom	recomm	very_recom	priority	spec_priority
not_recom	3693	0	0	0	0
recomm	0	0	2	0	0
very_recom	0	0	11	267	0
priority	0	0	0	3248	359
spec_priority	0	0	0	437	3000

4. **Validation Confusion Matrix**:

Predictions --------->

Labels | Pred	not_recom	recomm	very_recom	priority	spec_priority
not_recom	628	0	0	0	0
recomm	0	0	0	0	0
very_recom	0	0	2	48	0
priority	0	0	0	592	67
spec_priority	0	0	0	83	524

K-Nearest Neighbor ALGORITHM

5. Accuracy on Validation Set : 91.82%

6. Validation Confusion Matrix:

Predictions --------->

Labels | Pred	not_recom	recomm	very_recom	priority	spec_priority
not_recom	896	0	0	0	0
recomm	0	0	0	0	0
very_recom	0	1	60	7	0
priority	0	823	22	0	0
spec_priority	0	0	0	190	693