Name - Sparsh Palkhiwala
ASU ID - 1228656470
Subject - Data Mining (Hw-3)

TASK - 1

Q1: Run K-means clustering with Euclidean, Cosine and Jarcard similarity. Specify K= the number of categorical values of y (the number of classifications). Compare the SSEs of Euclidean-K-means, Cosine-K-means, Jarcard-K-means. Which method is better? (10 points)
Ans :
Euclidean K-means

- SSE: 25517481494.00
- Accuracy: 13.39%

Cosine K-means

- SSE: 697.81,
- Accuracy: 5.59%

Jaccard K-means

- SSE: 3653.15,
- Accuracy: 7.95%.

We see that Cosine Distance function has the lowest sum of squared errors. Cosine K-means outperforms Euclidean and Jaccard K-means in terms of SSE

Q2: Compare the accuracies of Euclidean-K-means Cosine-K-means, Jarcard-K-means. First, label each cluster using the majority vote label of the data points in that cluster. Later, compute the predictive accuracy of Euclidean-K-means, Cosine-K-means, Jarcard-K-means. Which metric is better? (10 points)
Ans :

- Euclidean K-means with Majority Vote - Accuracy: 59.72%
- Cosine K-means with Majority Vote - Accuracy: 57.37%
- Jaccard K-means with Majority Vote - Accuracy: 60.12%

The choice of distance metric can significantly impact the performance of clustering algorithms. The effectiveness of a distance metric depends on the characteristics of the data and the underlying structure of the clusters.

Q3: Set up the same stop criteria: “when there is no change in centroid position OR when the SSE value increases in the next iteration OR when the maximum preset value (e.g., 500, you can set the preset value by yourself) of iteration is complete”, for Euclidean-K-means, Cosine-Kmeans, Jarcard-K-means. Which method requires more iterations and times to converge? (10 points)
Ans:

Euclidean K-means

- Iterations: 48,
- SSE: 25392039606.17,
- Time to Converge: 57.5660 seconds

Cosine K-means

- Iterations: 59,
- SSE: 682.22,
- Time to Converge: 124.8682 seconds

Jaccard K-means

- Iterations: 69,
- SSE: 3660.35,
- Time to Converge: 182.2073 seconds

We see that Euclidean Kmeans take the least amount iterations and the time to coverge, Euclidean distance is sensitive to the scale of features. The high SSE may indicate that the algorithm struggles to converge when dealing with data that has varying feature scales.

Q4: Compare the SSEs of Euclidean-K-means Cosine-K-means, Jarcard-K-means with respect to the following three terminating conditions: (10 points)

- when there is no change in centroid position
- when the SSE value increases in the next iteration
- when the maximum preset value (e.g., 100) of iteration is complete

Ans:

- Euclidean K-means - SSE: 25373275133.23
- Cosine K-means - SSE: 684.57
- Jaccard K-means - SSE: 3708.24

Q5: What are your summary observations or takeaways based on your algorithmic analysis? (5 points)

1. Cosine K-means SSE:Cosine K-means has the lowest SSE among Euclidean and Jaccard. Tighter, more coherent clusters.
2. Majority Vote Accuracy: Jaccard K-means achieves highest accuracy in labeling clusters based on majority vote.The use of majority voting improves accuracy in all cases.
3. Convergence Speed: Euclidean K-means converges in fewer iterations and less time. Sensitivity to feature scales noted.
4. Sensitivity to Feature Scales: Euclidean K-means exhibits high SSE, indicating sensitivity to varying feature scales.
5. Terminating Conditions SSE: SSEs under different terminating conditions:
    - Euclidean: 25373275133.23
    - Cosine: 684.57
    - Jaccard: 3708.24

Emphasizes impact of distance metric on clustering behavior.
Euclidean K-means converges faster than Cosine and Jaccard K-means with stop criteria.

TASK - 2

1. Read data from “ratings small.csv” with line format: 'userID movieID rating timestamp'.
   a.

	userId	movieId	rating	timestamp
0	1	31	2.5	1260759144
1	1	1029	3.0	1260759179
2	1	1061	3.0	1260759182
3	1	1129	2.0	1260759185
4	1	1172	4.0	1260759205

2. MAE and RMSE are two famous metrics for evaluating the performances of a recommender system. The definition of MAE can be found via: https://en.wikipedia.org/wiki/Mean_absolute_error. The definition of RMSE can be found via: https://en.wikipedia.org/wiki/Root-mean-square_deviation.
3. Compute the average MAE and RMSE of the Probabilistic Matrix Factorization (PMF), User based Collaborative Filtering, Item based Collaborative Filtering, under the 5-folds cross-validation (10 points)

a.

```
Evaluating MAE, RMSE of algorithm NMF on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
MAE (testset)     0.7276  0.7268  0.7271  0.7309  0.7280  0.7281  0.0015
RMSE (testset)    0.9457  0.9470  0.9456  0.9530  0.9494  0.9481  0.0028
Fit time          2.42    2.41    2.40    2.48    2.45    2.43    0.03
Test time         0.20    0.11    0.19    0.10    0.19    0.16    0.04
Computing the msd similarity matrix...
Done computing similarity matrix.
Computing the msd similarity matrix...
Done computing similarity matrix.
Computing the msd similarity matrix...
Done computing similarity matrix.
Computing the msd similarity matrix...
Done computing similarity matrix.
Computing the msd similarity matrix...
Done computing similarity matrix.
Evaluating MAE, RMSE of algorithm KNNBasic on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
MAE (testset)     0.7504  0.7434  0.7438  0.7463  0.7444  0.7457  0.0026
RMSE (testset)    0.9750  0.9666  0.9681  0.9710  0.9708  0.9703  0.0029
Fit time          0.14    0.18    0.16    0.15    0.15    0.16    0.01
Test time         1.28    1.37    1.26    1.38    1.24    1.30    0.06
Computing the msd similarity matrix...
Done computing similarity matrix.
Computing the msd similarity matrix...
Done computing similarity matrix.
Computing the msd similarity matrix...
Done computing similarity matrix.
Computing the msd similarity matrix...
Done computing similarity matrix.
Computing the msd similarity matrix...
Done computing similarity matrix.
Evaluating MAE, RMSE of algorithm KNNBasic on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
MAE (testset)     0.7236  0.7215  0.7134  0.7219  0.7264  0.7214  0.0043
RMSE (testset)    0.9414  0.9363  0.9254  0.9356  0.9387  0.9355  0.0054
Fit time          2.61    2.66    2.60    2.61    2.62    2.62    0.02
Test time         6.25    6.23    6.26    6.27    6.48    6.30    0.09
```

4. Compare the average (mean) performances of User-based collaborative filtering, item-based collaborative filtering, PMF with respect to RMSE and MAE. Which ML model is the best in the movie rating data? (10 points)

```
Test time          6.25    6.23    6.26    6.27    6.48    6.30    0.09
Average MAE and RMSE for SVD: 0.690066815755929 0.896744104934777
Average MAE and RMSE for PMF: 0.7280792262814966 0.9481393840129846
Average MAE and RMSE for User-based CF: 0.7456667983333296 0.9703143355288111
Average MAE and RMSE for Item-based CF: 0.7213709145201285 0.935487848781376
```

Here we see the result for

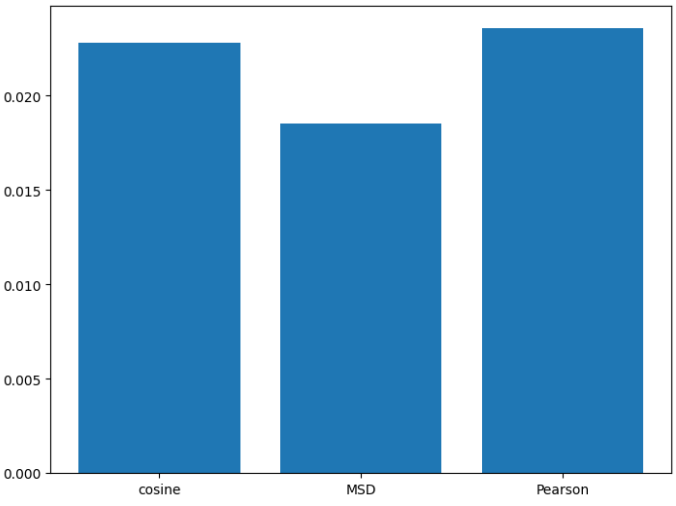
- Singular Value Decompostion - MAE : 0.69 and RMSE : 0.89
- Probabilistic Matrix Factorization - MAE : 0.72 and RMSE : 0.94
- User-Based Collabrative Function - MAE : 0.74 and RMSE : 0.96
- Item-Base Collabrative Function - MAE : 0.72 and RMSE : 0.93

a.

b. We see that SVD has the lowet Mean Average Error and Root Mean Squared Error, thus making it the best option for movie recommendation

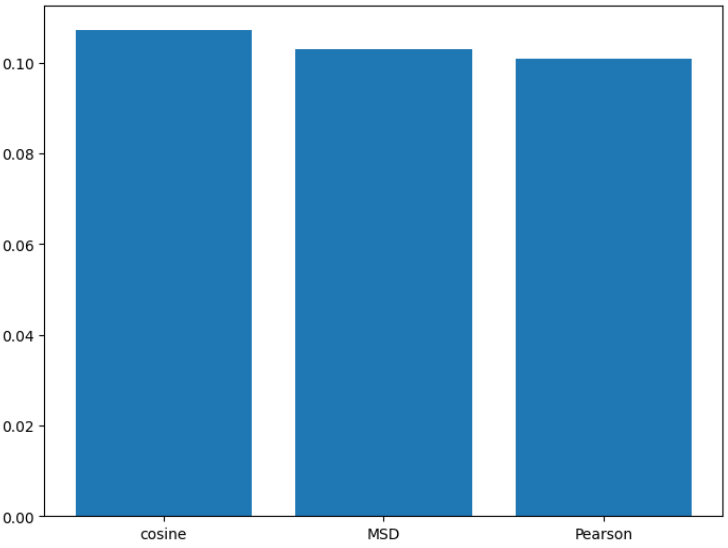
5. Examine how the cosine, MSD (Mean Squared Difference), and Pearson similarities impact the performances of User based Collaborative Filtering and Item based Collaborative Filtering. Plot your results. Is the impact of the three metrics on User based Collaborative Filtering consistent with the impact of the three metrics on Item based Collaborative Filtering? (10 points)
   a. User Based - RMSE

i.



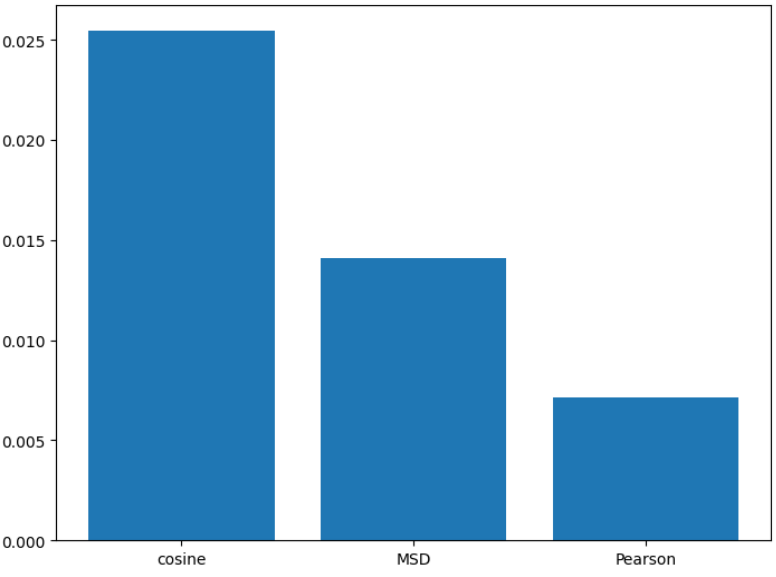
   b. User Based - MAE

i.



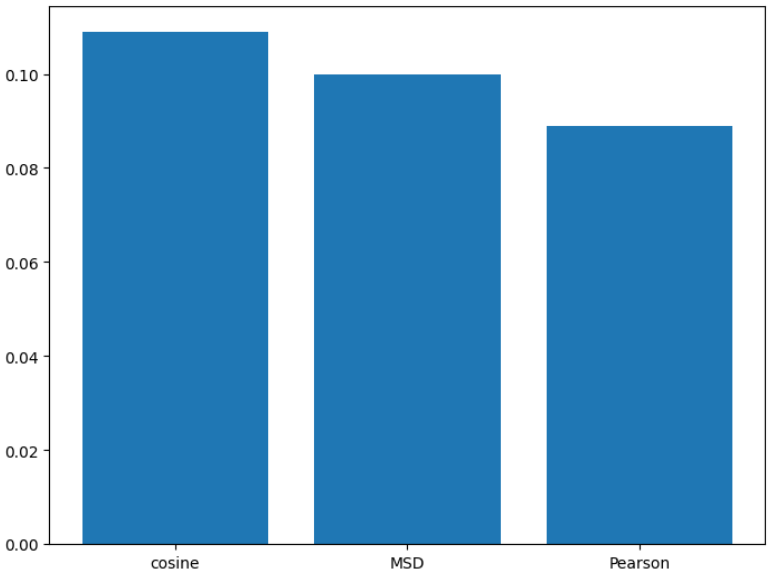
c. Item Based - RMSE

i.



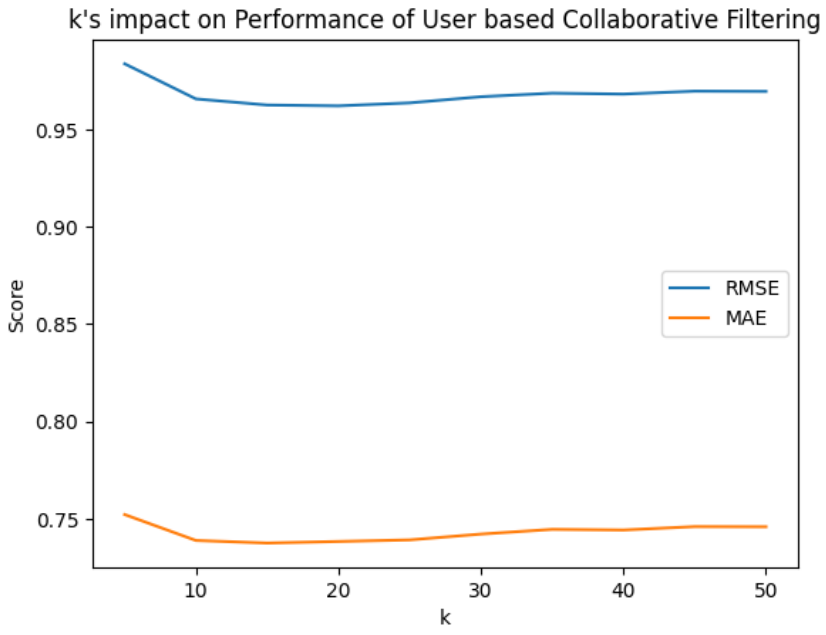
d. Item Based - MAE

i.



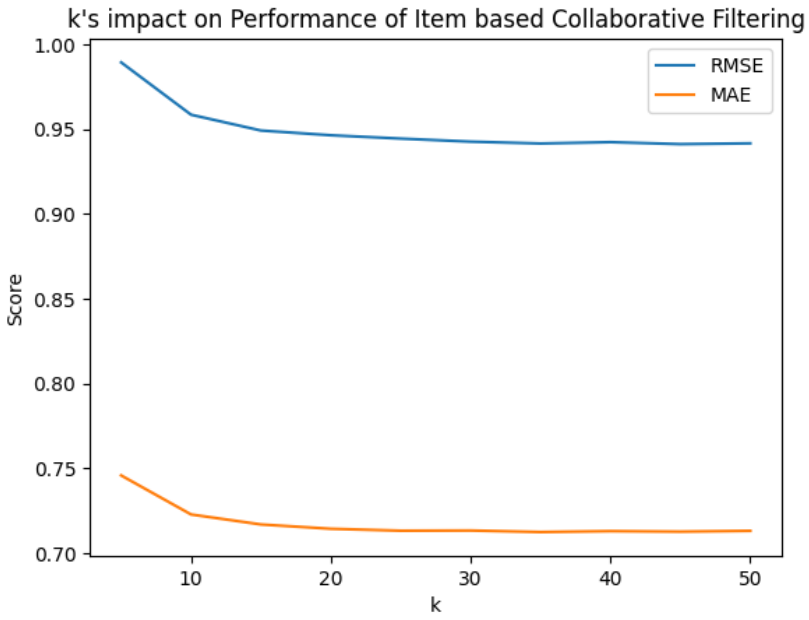
6. Examine how the number of neighbors impacts the performances of User based Collaborative Filtering and Item based Collaborative Filtering? Plot your results. (10 points)
    a. User Based

i.



    b. Item Based

i.



7. Identify the best number of neighbor (denoted by K) for User/Item based collaborative filtering in terms of RMSE. Is the best K of User based collaborative filtering the same with the best K of Item based collaborative filtering? (10 points)
    a. We see that RMSE and MAE for both UBCF and IBCF, start to reduce as the value of k increases
    b. We see that K's impact on UBCF start to plateau around 20, so the best k = 20
    c. We see that K's impact on IBCF starts to plateau around 30, the best k = 30