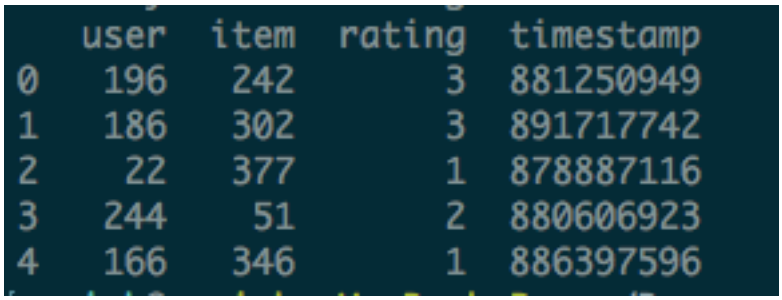
REPORT
Manish Jaisinghani

1. Objective: Demonstrate functioning of recommender systems with different Accuracy matrices such as RMSE (Root Mean Square Deviation), MAE (Mean absolute error) and FCP (Fraction of Concordant Pairs) and using different similarity matrices such as MSD (Mean Squared Difference), Cosine and Pearson.

2. Data: The data has already been provided in the form of a text file ("restaurant_ratings.txt") which has data columns as below:



	user	item	rating	timestamp
0	196	242	3	881250949
1	186	302	3	891717742
2	22	377	1	878887116
3	244	51	2	880606923
4	166	346	1	886397596

Column 1 represents user ID, Column 2 represents Item ID, Column 3 represents rating given by the user and Column 4 represents the timestamp when the rating was given.

3. Data Processing: The data processing is done with respect to reading data from a text file and assigning column names to the same, the code lines that help to perform this task are as below:

```
file_path = os.path.expanduser('restaurant_ratings.txt')
reader = Reader(line_format='user item rating timestamp', sep='\t')
data = Dataset.load_from_file(file_path, reader=reader)
```

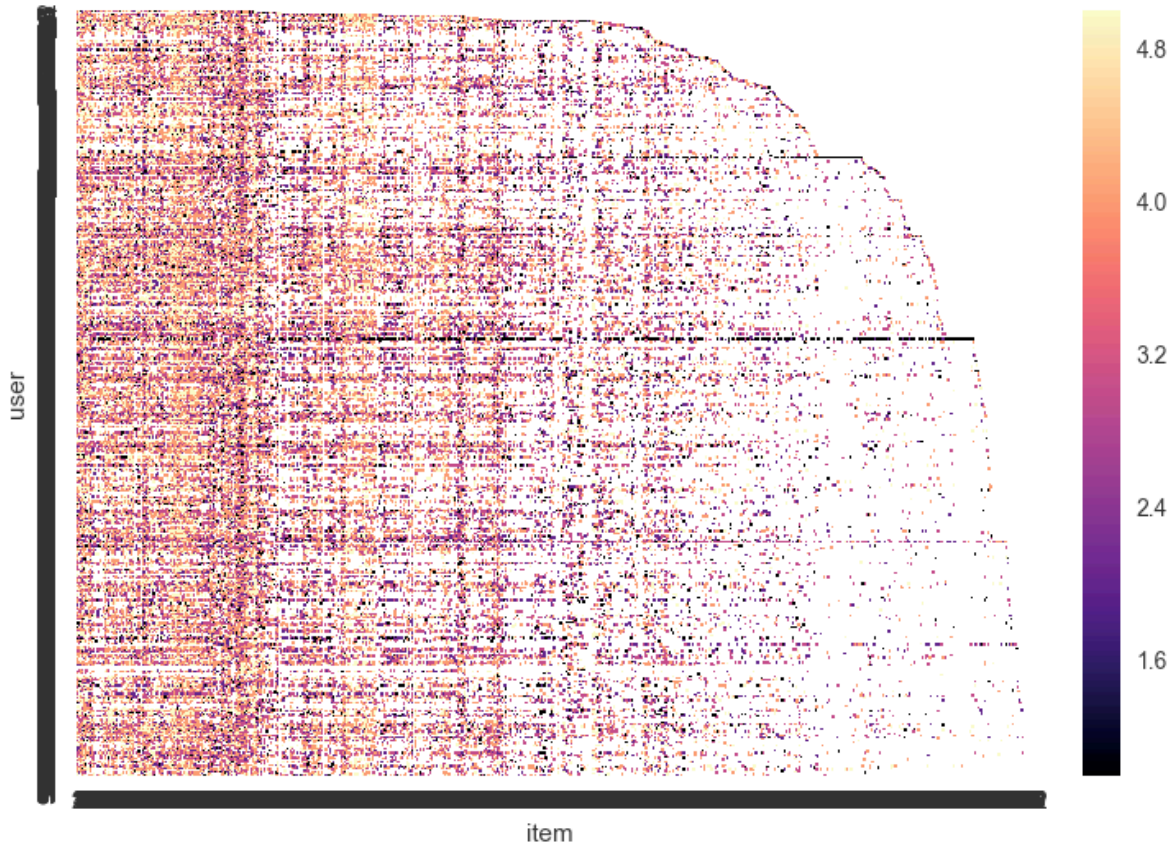
The data seems pre-Processed and mimics real life data for reading submission from users. This was confirmed by below analysis:

```
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 4 columns):
user         100000 non-null int64
item         100000 non-null int64
rating       100000 non-null int64
timestamp    100000 non-null int64
dtypes: int64(4)
memory usage: 3.1 MB
None
```

item	1	2	3	4	5	6	7	8	9	10	...	1673
user											...	
1	5.0	3.0	4.0	3.0	3.0	5.0	4.0	1.0	5.0	3.0	...	NaN
2	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0	...	NaN
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
5	4.0	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
6	4.0	NaN	NaN	NaN	NaN	NaN	2.0	4.0	4.0	NaN	...	NaN
7	NaN	NaN	NaN	5.0	NaN	NaN	5.0	5.0	5.0	4.0	...	NaN
8	NaN	NaN	NaN	NaN	NaN	NaN	3.0	NaN	NaN	NaN	...	NaN
9	NaN	NaN	NaN	NaN	NaN	5.0	4.0	NaN	NaN	NaN	...	NaN
10	4.0	NaN	NaN	4.0	NaN	NaN	4.0	NaN	4.0	NaN	...	NaN
11	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.0	5.0	NaN	...	NaN
12	NaN	NaN	NaN	5.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
13	3.0	3.0	NaN	5.0	1.0	NaN	2.0	4.0	3.0	NaN	...	NaN
14	NaN	NaN	NaN	NaN	NaN	NaN	5.0	NaN	4.0	NaN	...	NaN
15	1.0	NaN	NaN	NaN	NaN	NaN	1.0	NaN	4.0	NaN	...	NaN

The above snippet is the output for the pivot table, where Columns predict “Item ID”, the rows predict “User ID” and the corresponding values in the matrix represent the user ratings given to particular items.

We can observe a lot of null values and this is pretty normal as it will be very difficult for every user to try all the items and submit a rating for them, most of the users will try 5 to 10 items and submit a rating for around 4 to 7 from them. Hence, the data depicts the real world situation. Below is the heat map representation of the data:



Above heat map is a bit clumsy and the reason behind that is the data set is very large as well as contains a lot of null values as confirmed through heat map representation as well.

4. Module specifications:

The algorithms have been implemented using python 3 and below listed libraries have been used:

- **Pandas** – Deals with csv, text data, acts as a supporting block to the "Surprise Library" and helps in manipulation.
- **Numpy** – Deals with data in the form of arrays and assists in data manipulation
- **Matplotlib** – Helps to visualize the data.
- **Seaborn** – Another library which helps in data visualization, the only difference here is seaborn is vast than matplotlib.
- **Surprise** – Library that helps to implement recommendation systems. The library contains many predefined functions to implement algorithms such as SVD (Singular Value Decomposition), PMF (Probabilistic Matrix Factorization) and NMF (Non-Negative Matrix Factorization) and KNN (K Nearest Neighbors) using different similarity matrices such as MSD, Cosine and Pearson.

5. Terminology & Explanation:

- *Mean Squared Difference Similarity:* The similarity metric falls under predictive accuracy measures collaborative filtering recommender systems and is given as below:

$$|\bar{E}^2| = \frac{\sum_{(c,s)\in W}(u^p(c,s) - u(c,s))^2}{|W|}$$

- *Cosine Similarity*: Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.

$$\mathbf{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

- *Pearson Similarity:* Pearson similarity is a measure of the linear correlation between two variables *X* and *Y*.

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

  Where cov is the covariance and sigma of x and y are the standard deviations

- *RMSE (Root Mean Square Deviation):* The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(x_{1,t} - x_{2,t})^2}{n}}.$$

- *MAE (Mean absolute error):* the mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|.$$

- *FCP (Fraction of Concordant Pairs)):* a **concordant pair** is a pair of observations, each on two variables, $\{\mathbf{X}_1,\mathbf{Y}_1\}$ and $\{\mathbf{X}_2,\mathbf{Y}_2\}$, having the property that

$$\operatorname{sgn}(X_2 - X_1) = \operatorname{sgn}(Y_2 - Y_1)$$

Where,

$$\operatorname{sgn} x = \begin{cases} -1 & : \quad x < 0 \\ 0 & : \quad x = 0 \\ 1 & : \quad x > 0 \end{cases}$$

6. Implementation details: Below are the details on how the algorithms have been written and pre-requisites for successfully running the same:-

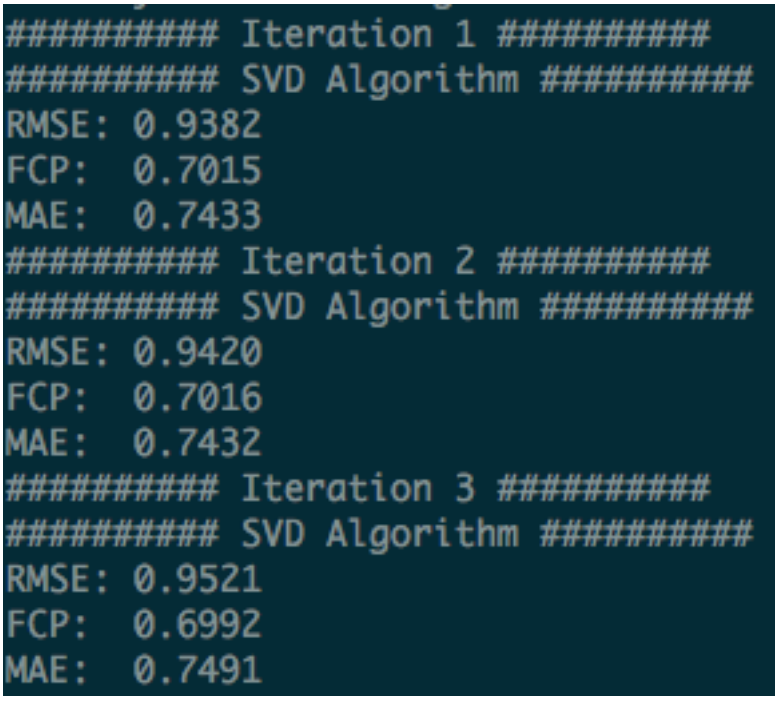
- **Rest_recommend.py** - The file contains the implementation of SVD, NMF, PMF, User based collaborative filtering and item based collaborative filtering.
- **Rest_Surprise_13.py** – The file contains the implementation of User based and Item based collaborative filtering with different matrices such as the MSD, cosine and pearson similarity.
- **Rest_Surprise_14.py** – The file contains the implementation of User based and Item based collaborative filtering with varying K value (K ranges from 1 to 20)
- **graphs.py** – The file contains the implementation of data visualization for analysis purposes.

- **FCP_Measure.csv** – This file contains the data with respect to FCP accuracy measure for all the algorithms, collected by running the algorithms for 3 folds of train and test data.
- **MAE_Measure.cvs** – This file contains the data with respect to MAE accuracy measure for all the algorithms, collected by running the algorithms for 3 folds of train and test data.

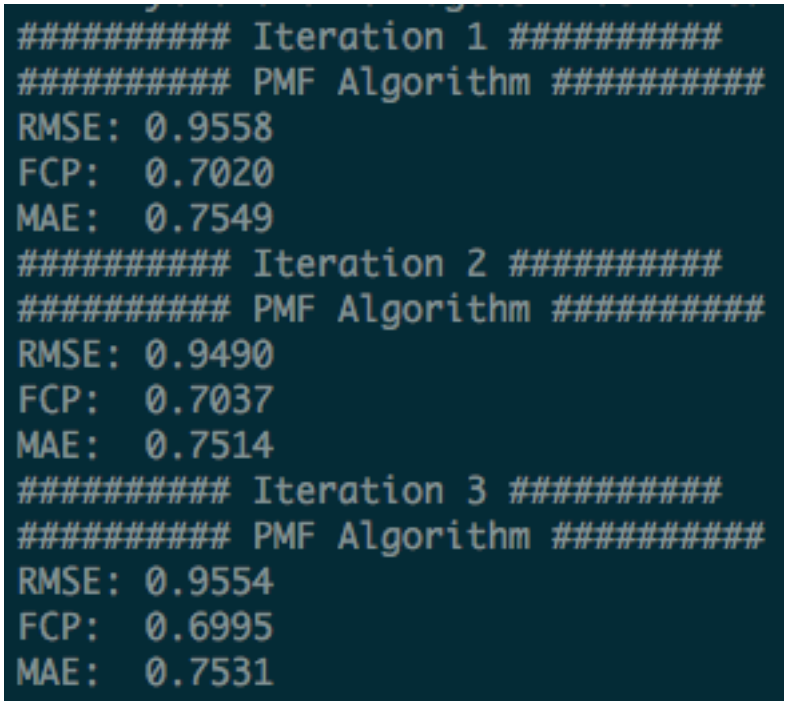
- **RMSE_Measure.cvs** – This file contains the data with respect to RMSE accuracy measure for all the algorithms, collected by running the algorithms for 3 folds of train and test data.
- **UBCF_IBCF_FCP.cvs** – This file contains the data with respect to FCP accuracy measure for User based Collaborative filtering and Item Based Collaborative filtering.
- **UBCF_IBCF_MAE.cvs** – This file contains the data with respect to MAE accuracy measure for User based Collaborative filtering and Item Based Collaborative filtering.
- **UBCF_IBCF_RMSE.cvs** – This file contains the data with respect to RMSE accuracy measure for User based Collaborative filtering and Item Based Collaborative filtering.
- **restaurant_ratings.txt** – This file contains all the data used for performing the analysis. The file consists of four columns of data, user ID, item ID, rating given by user and the timestamp.
- **Constraints** – The system should be installed with all python modules listed in "Module Specifications" section.
- The python script should be executed using python 3.0.

7. Data Analysis:

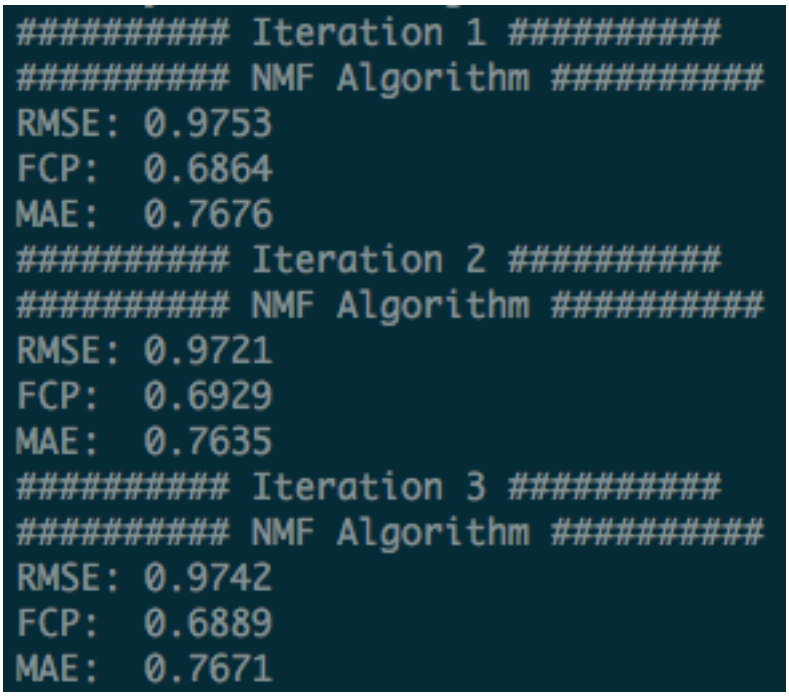
a. Compute the MAE and RMSE of the SVD (Singular Value Decomposition) algorithm on 3 folds of data:



b. Compute the MAE and RMSE of the PMF (Probabilistic Matrix Factorization) algorithm on 3 folds of data:



c. Compute the MAE and RMSE of the NMF (Non-Negative Matrix Factorization) algorithm on 3 folds of data:



d. Compute the MAE and RMSE of the UBCF (User Based Collaborative Filtering) algorithm on 3 folds of data:

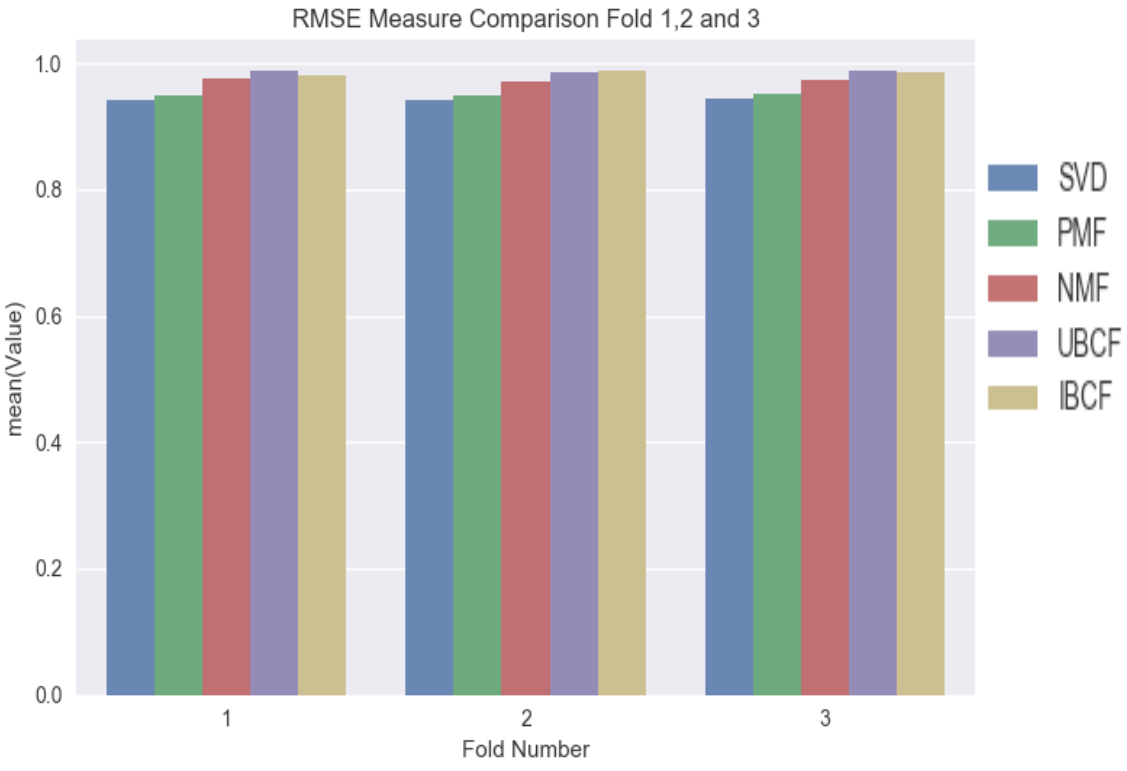
```
########## Iteration 1 ##########
########## User based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9829
FCP:  0.7085
MAE:  0.7772
########## Iteration 2 ##########
########## User based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9923
FCP:  0.7134
MAE:  0.7832
########## Iteration 3 ##########
########## User based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9883
FCP:  0.7165
MAE:  0.7818
```

e. Compute the MAE and RMSE of the IBCF (Item Based Collaborative Filtering) algorithm on 3 folds of data:

```
########## Iteration 1 ##########
########## Item based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9891
FCP:  0.6662
MAE:  0.7814
########## Iteration 2 ##########
########## Item based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9796
FCP:  0.6675
MAE:  0.7764
########## Iteration 3 ##########
########## Item based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9849
FCP:  0.6621
MAE:  0.7807
```

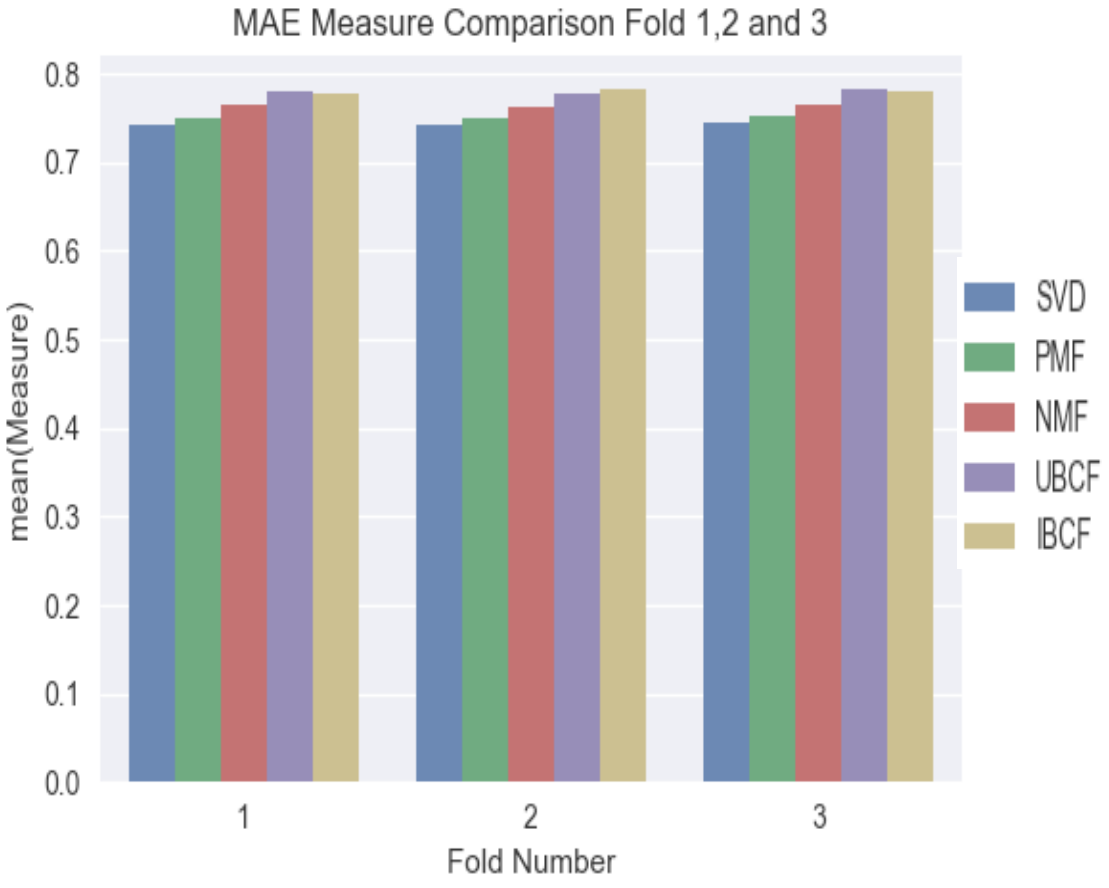
f. Compare the Performances of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF on **fold-1** with respect to RMSE and MAE

```
########## Iteration 1 ##########
########## SVD Algorithm ##########
RMSE: 0.9409
FCP:  0.6993
MAE:  0.7423
########## PMF Algorithm ##########
RMSE: 0.9503
FCP:  0.6988
MAE:  0.7510
########## NMF Algorithm ##########
RMSE: 0.9750
FCP:  0.6849
MAE:  0.7649
########## User based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9880
FCP:  0.7126
MAE:  0.7804
########## Item based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9814
FCP:  0.6657
MAE:  0.7782
```



g. Compare the Performances of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF on **fold-2** with respect to RMSE and MAE

```
########## Iteration 2 ##########
########## SVD Algorithm ##########
RMSE: 0.9427
FCP:  0.7005
MAE:  0.7426
########## PMF Algorithm ##########
RMSE: 0.9500
FCP:  0.7008
MAE:  0.7502
########## NMF Algorithm ##########
RMSE: 0.9708
FCP:  0.6909
MAE:  0.7632
########## User based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9849
FCP:  0.7116
MAE:  0.7774
########## Item based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9880
FCP:  0.6652
MAE:  0.7819
```

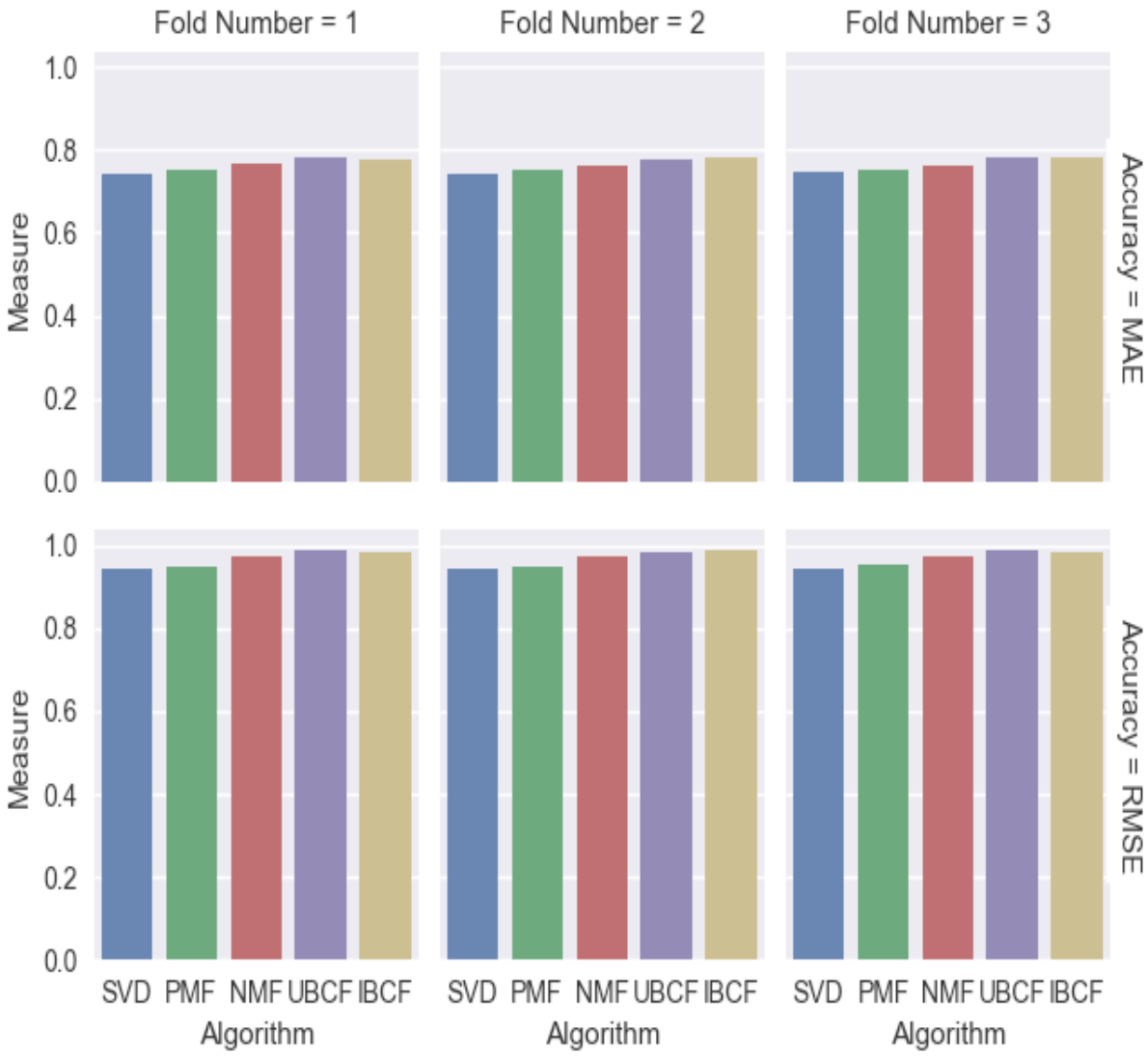


h. Compare the Performances of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF on **fold-3** with respect to RMSE and MAE.

```
######### Iteration 3 #########
######### SVD Algorithm #########
RMSE: 0.9440
FCP:  0.7043
MAE:  0.7452
######### PMF Algorithm #########
RMSE: 0.9522
FCP:  0.7043
MAE:  0.7531
######### NMF Algorithm #########
RMSE: 0.9739
FCP:  0.6914
MAE:  0.7644
######### User based Collaborative Filtering Algorithm #########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9873
FCP:  0.7128
MAE:  0.7815
######### Item based Collaborative Filtering Algorithm #########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9851
FCP:  0.6659
MAE:  0.7813
```

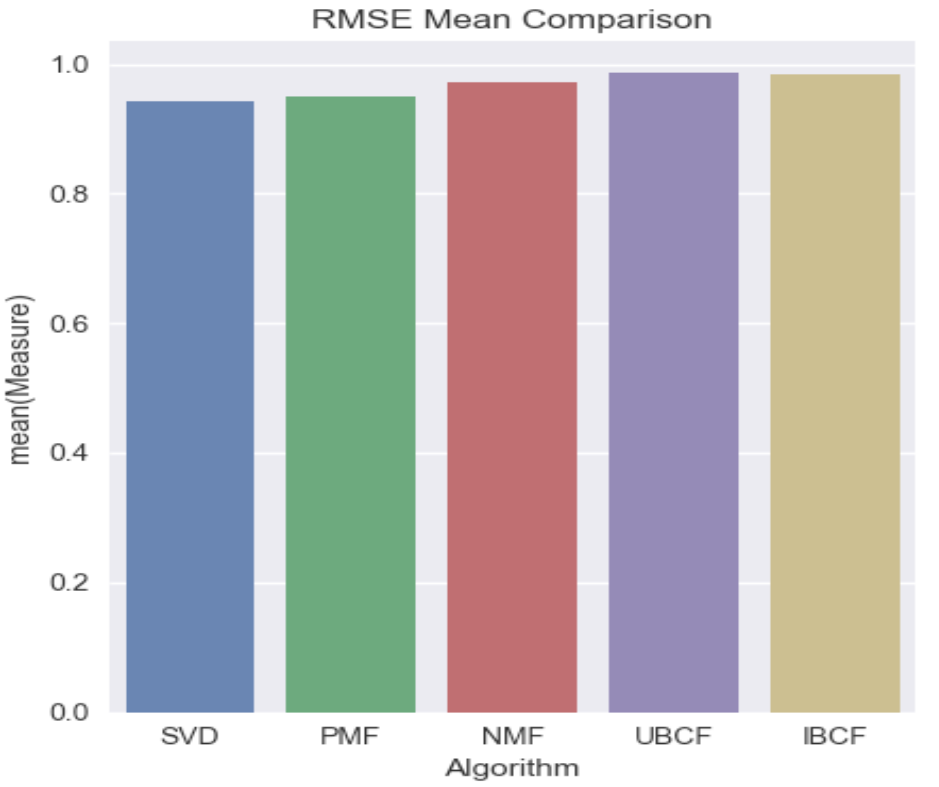
Comparison of all the algorithms with respect to all the accuracy matrices over 3 folds of data is shown below:

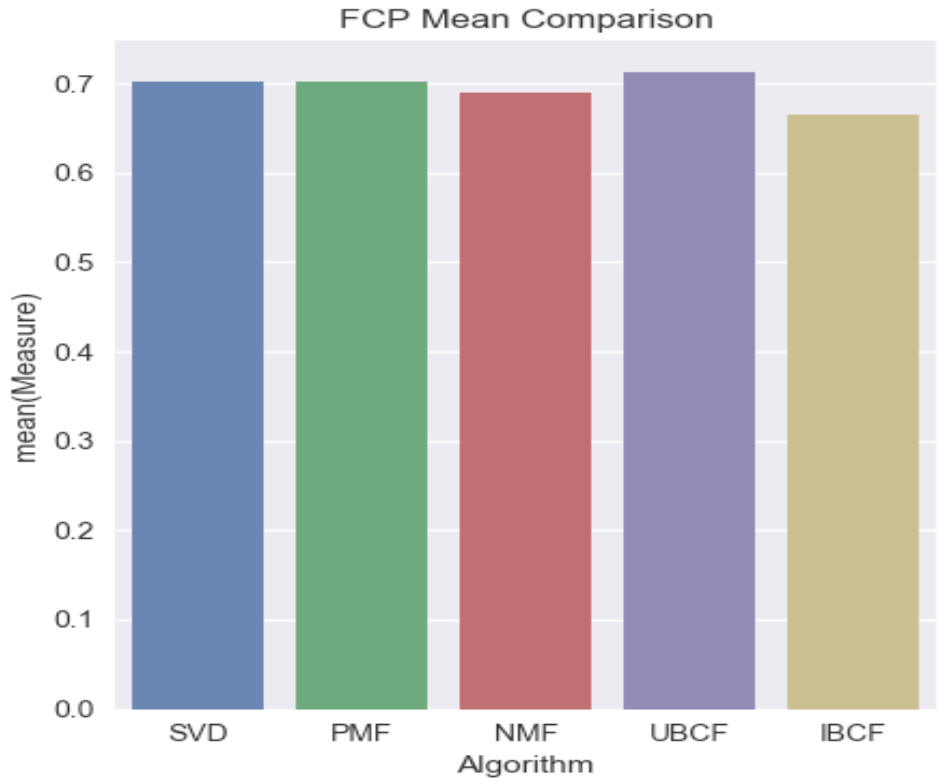
- UBCF stands for User Based Collaborative Filtering
- IBCF stands for Item Based Collaborative Filtering



NOTE – All data has been compared on the same folds, the data is result of single execution and visualized as per the requirement of problem scenario.

i. Compare the **average (mean)** Performances of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF with respect to RMSE and MAE.





j. Examine how the cosine, MSD (Mean Squared Difference), and Pearson similarities impact the performances of User based Collaborative Filtering and Item based Collaborative Filtering

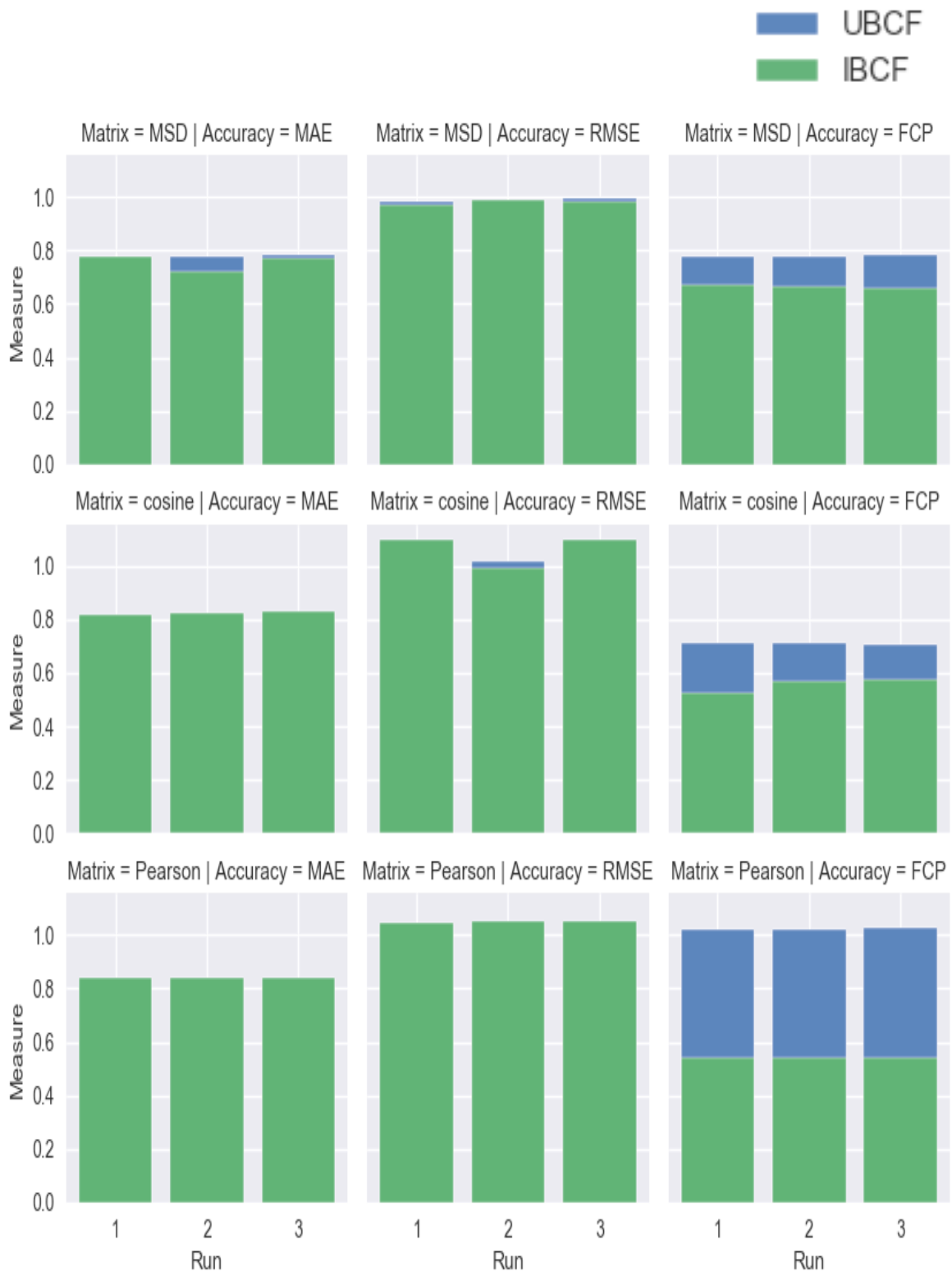
```
########## Iteration 1 ##########
########## MSD User based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9817
FCP:  0.7143
MAE:  0.7784
########## MSD Item based Collaborative Filtering Algorithm ##########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9798
FCP:  0.6701
MAE:  0.7763
########## Cosine User based Collaborative Filtering Algorithm ##########
Computing the cosine similarity matrix...
Done computing similarity matrix.
RMSE: 1.0139
FCP:  0.7119
MAE:  0.8050
########## Cosine Item based Collaborative Filtering Algorithm ##########
Computing the cosine similarity matrix...
Done computing similarity matrix.
RMSE: 1.0330
FCP:  0.5722
MAE:  0.8203
########## Pearson User based Collaborative Filtering Algorithm ##########
Computing the pearson similarity matrix...
Done computing similarity matrix.
RMSE: 1.0106
FCP:  0.7152
MAE:  0.8046
########## Pearson Item based Collaborative Filtering Algorithm ##########
Computing the pearson similarity matrix...
Done computing similarity matrix.
RMSE: 1.0451
FCP:  0.5433
MAE:  0.8372
```

```
######### Iteration 2 #########
######### MSD User based Collaborative Filtering Algorithm #########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9853
FCP:  0.7154
MAE:  0.7788
######### MSD Item based Collaborative Filtering Algorithm #########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9877
FCP:  0.6624
MAE:  0.7816
######### Cosine User based Collaborative Filtering Algorithm #########
Computing the cosine similarity matrix...
Done computing similarity matrix.
RMSE: 1.0181
FCP:  0.7120
MAE:  0.8057
######### Cosine Item based Collaborative Filtering Algorithm #########
Computing the cosine similarity matrix...
Done computing similarity matrix.
RMSE: 1.0375
FCP:  0.5684
MAE:  0.8231
######### Pearson User based Collaborative Filtering Algorithm #########
Computing the pearson similarity matrix...
Done computing similarity matrix.
RMSE: 1.0181
FCP:  0.7146
MAE:  0.8081
######### Pearson Item based Collaborative Filtering Algorithm #########
Computing the pearson similarity matrix...
Done computing similarity matrix.
RMSE: 1.0500
FCP:  0.5428
MAE:  0.8396
```

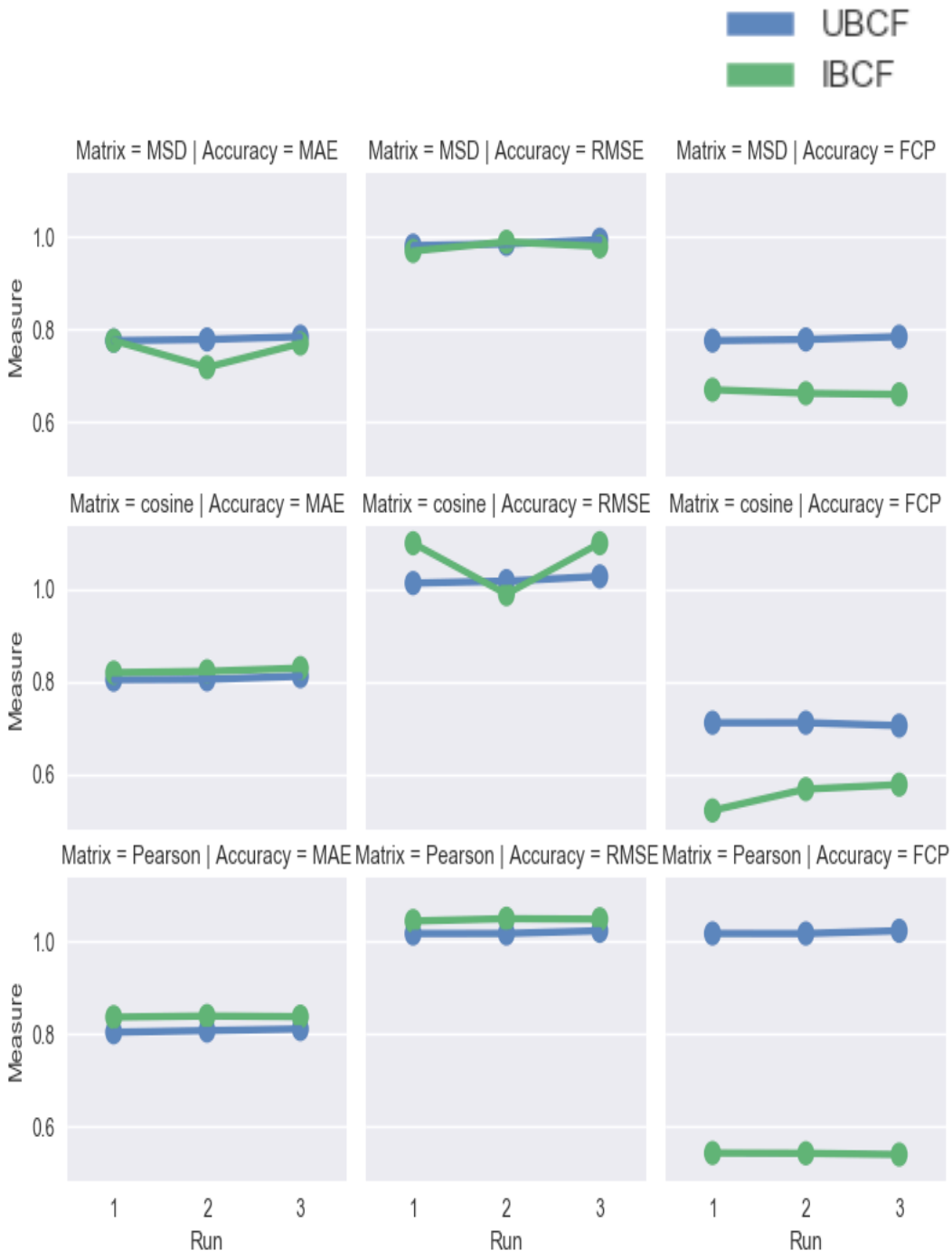
```
######### Iteration 3 #########
######### MSD User based Collaborative Filtering Algorithm #########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9944
FCP:  0.7093
MAE:  0.7848
######### MSD Item based Collaborative Filtering Algorithm #########
Computing the msd similarity matrix...
Done computing similarity matrix.
RMSE: 0.9881
FCP:  0.6658
MAE:  0.7816
######### Cosine User based Collaborative Filtering Algorithm #########
Computing the cosine similarity matrix...
Done computing similarity matrix.
RMSE: 1.0283
FCP:  0.7057
MAE:  0.8126
######### Cosine Item based Collaborative Filtering Algorithm #########
Computing the cosine similarity matrix...
Done computing similarity matrix.
RMSE: 1.0377
FCP:  0.5780
MAE:  0.8237
######### Pearson User based Collaborative Filtering Algorithm #########
Computing the pearson similarity matrix...
Done computing similarity matrix.
RMSE: 1.0239
FCP:  0.7093
MAE:  0.8116
######### Pearson Item based Collaborative Filtering Algorithm #########
Computing the pearson similarity matrix...
Done computing similarity matrix.
RMSE: 1.0490
FCP:  0.5401
MAE:  0.8382
```

Data Comparison Visualization:

- UBCF stands for User Based Collaborative Filtering
- IBCF stands for Item based Collaborative Filtering



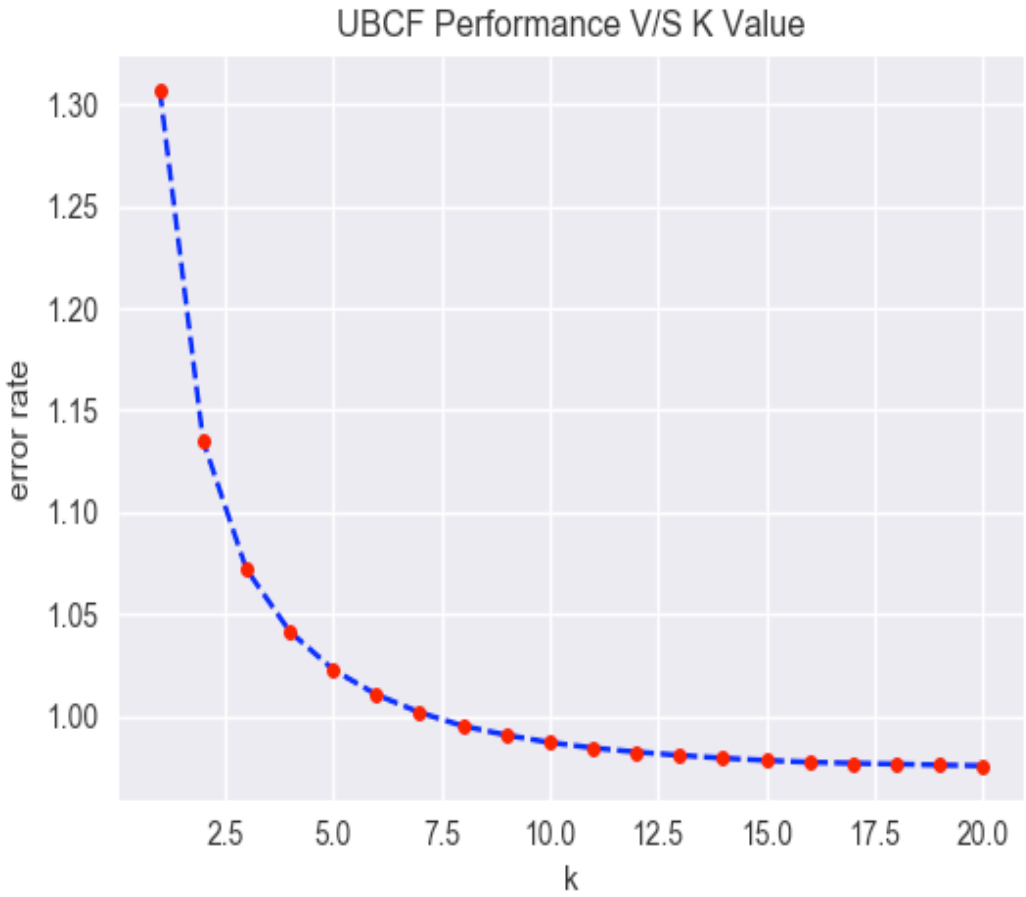
An another view of graphical representation



k. Examine how the number of neighbors impacts the performances of User based Collaborative Filtering or Item based Collaborative Filtering

The output of all the values of k has been attached in a text file by name K_Output_UBCF.txt

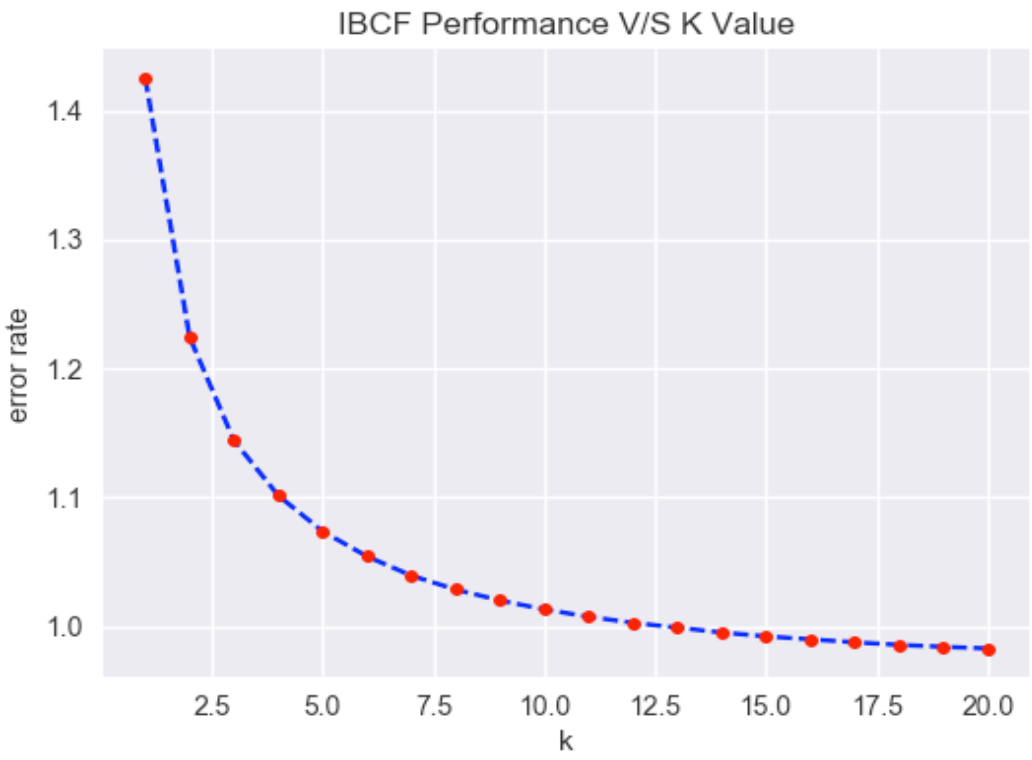
Below is the graphical representation of change in K-value V/S Performance of User based Collaborative Filtering



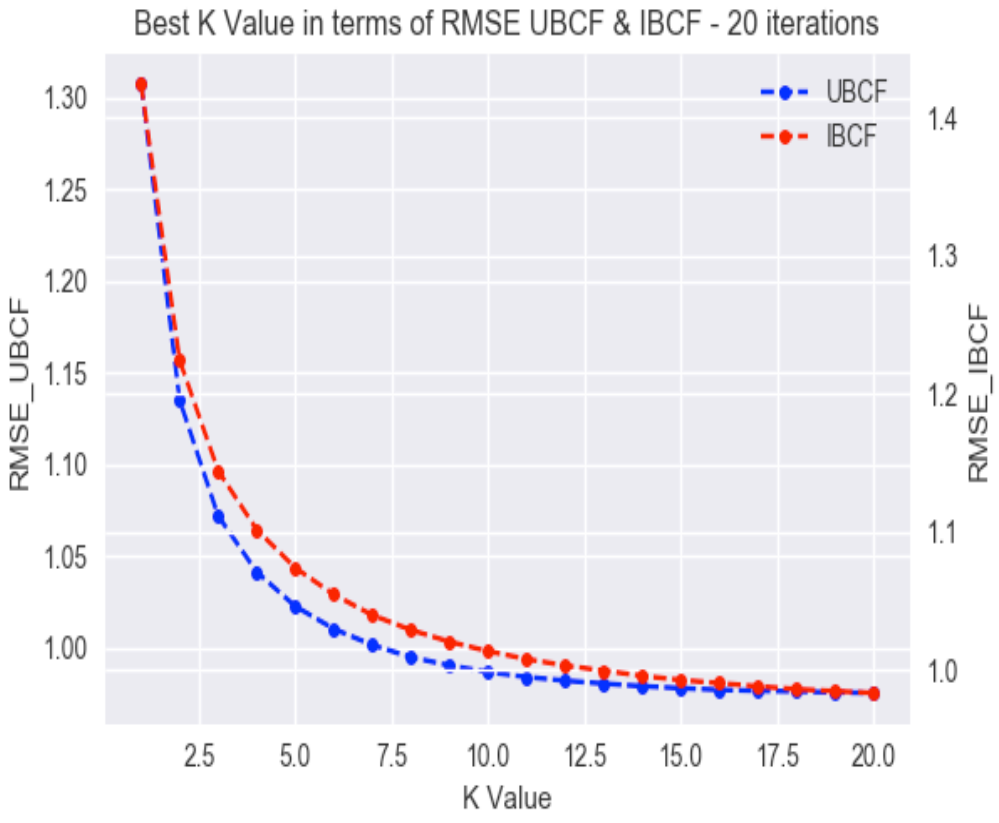
l. Examine how the number of neighbors impacts the performances of Item based Collaborative Filtering or Item based Collaborative Filtering

The output of all the values of k has been attached in a text file by name K_Output_IBCF.txt

Below is the graphical representation of change in K-value V/S Performance of Item based Collaborative Filtering



Comparing K-Values for both User based Collaborative Filtering and User Based Collaborative Filtering we get below:



8.Conclusion:

1. Comparison of performance of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF on **fold-1** with respect to RMSE and MAE resulted in below ranking (Highest to lowest):

RMSE	**MAE**
1. UBCF	1. UBCF
2. UBCF	2. IBCF
3. NMF	3. NMF
4. PMF	4. PMF
5. SVD	5. SVD

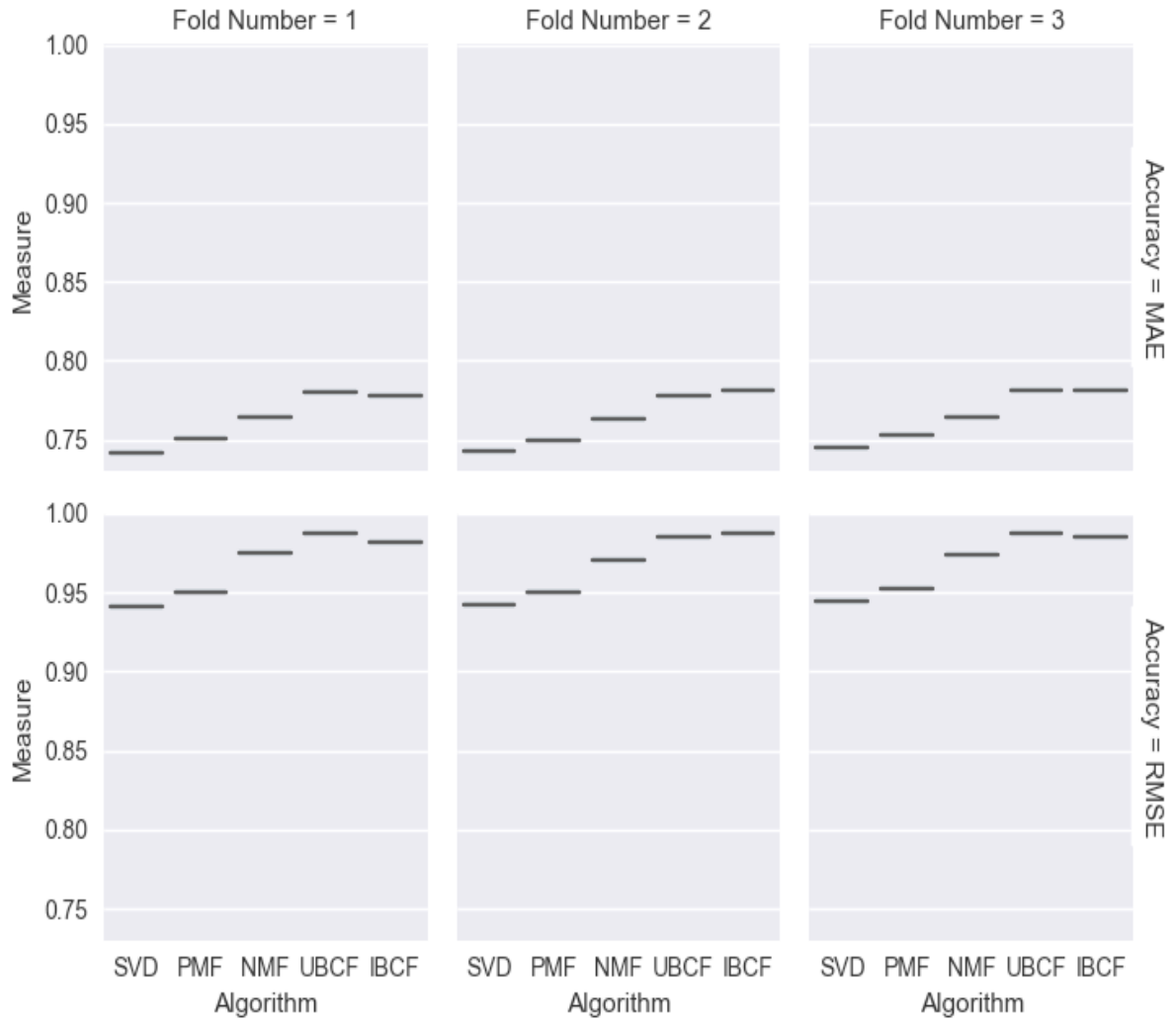
2. Comparison of performance of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF on **fold-2** with respect to RMSE and MAE resulted in below ranking (Highest to lowest):

RMSE	**MAE**
1. IBCF	1. IBCF
2. UBCF	2. UBCF
3. NMF	3. NMF
4. PMF	4. PMF
5. SVD	5. SVD

3. Comparison of performance of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF on **fold-3** with respect to RMSE and MAE resulted in below ranking (Highest to lowest):

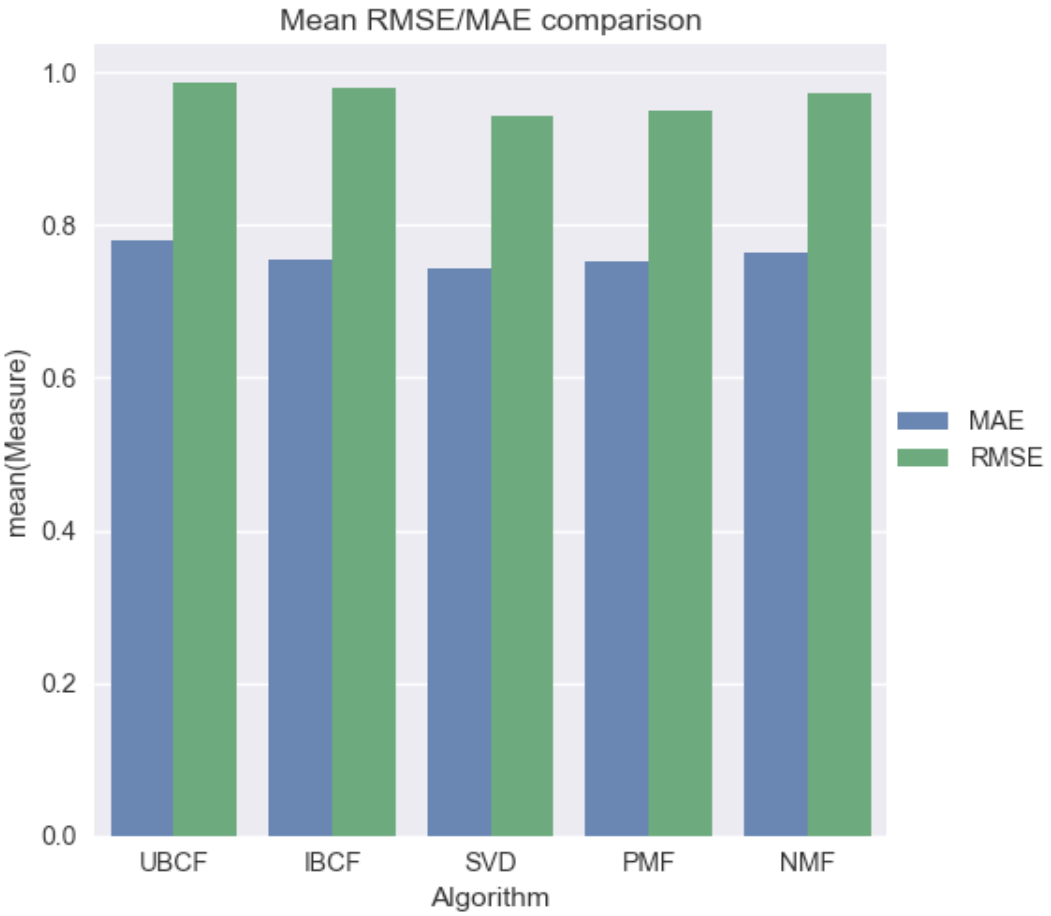
RMSE	**MAE**
1. IBC, UBCF	1. UBCF
2. NMF	2. IBC
3. PMF	3. NMF
4. SVD	4. PMF
	5. SVD

Below is the graphical representation depicting the results:



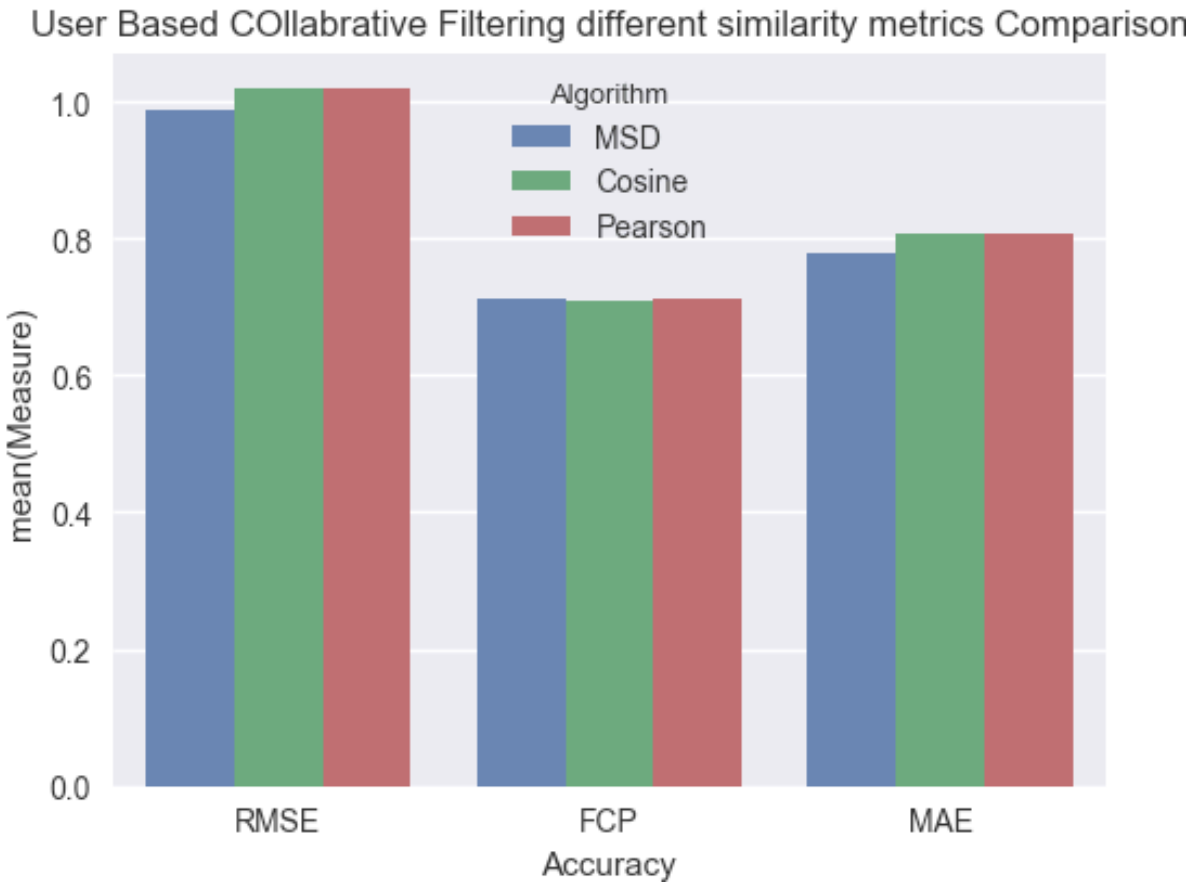
4. Comparison of performance of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF on **Average (Mean)** with respect to RMSE and MAE resulted in below ranking (Highest to lowest):

RMSE	**MAE**
1. UBCF	1. UBCF
2. IBC	2. IBC
3. NMF	3. NMF
4. PMF	4. PMF
5. SVD	5. SVD



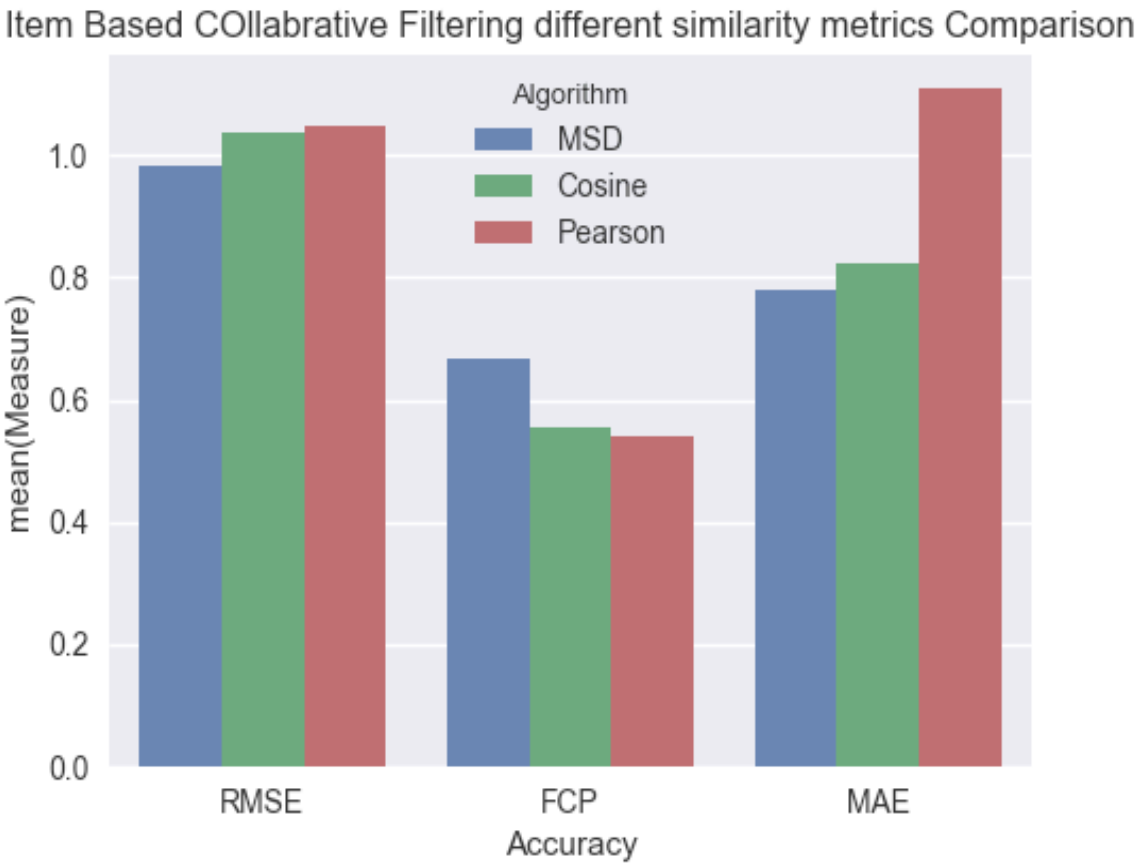
5. Examining the impact of cosine, MSD (Mean Squared Difference), and Pearson similarities on performances of User based Collaborative Filtering we get below ranking results:

RMSE	**FCP**	**MAE**
1. Pearson, Cosine	1. MSD, Pearson	1. Cosine, Pearson
2. MSD	2. Cosine	2. MSD

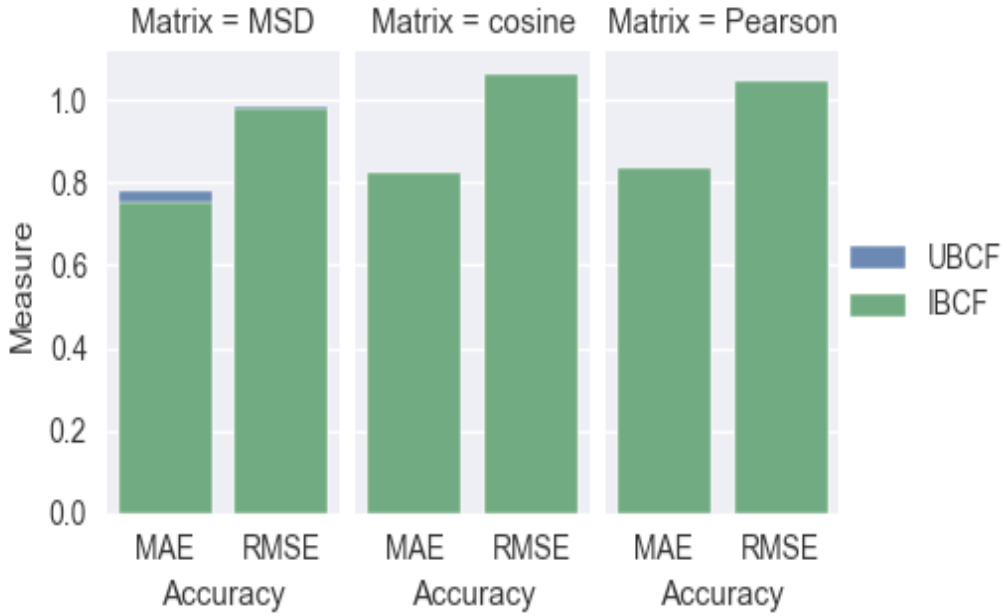


6. Examining the impact of cosine, MSD (Mean Squared Difference), and Pearson similarities on performances of Item based Collaborative Filtering we get below ranking results:

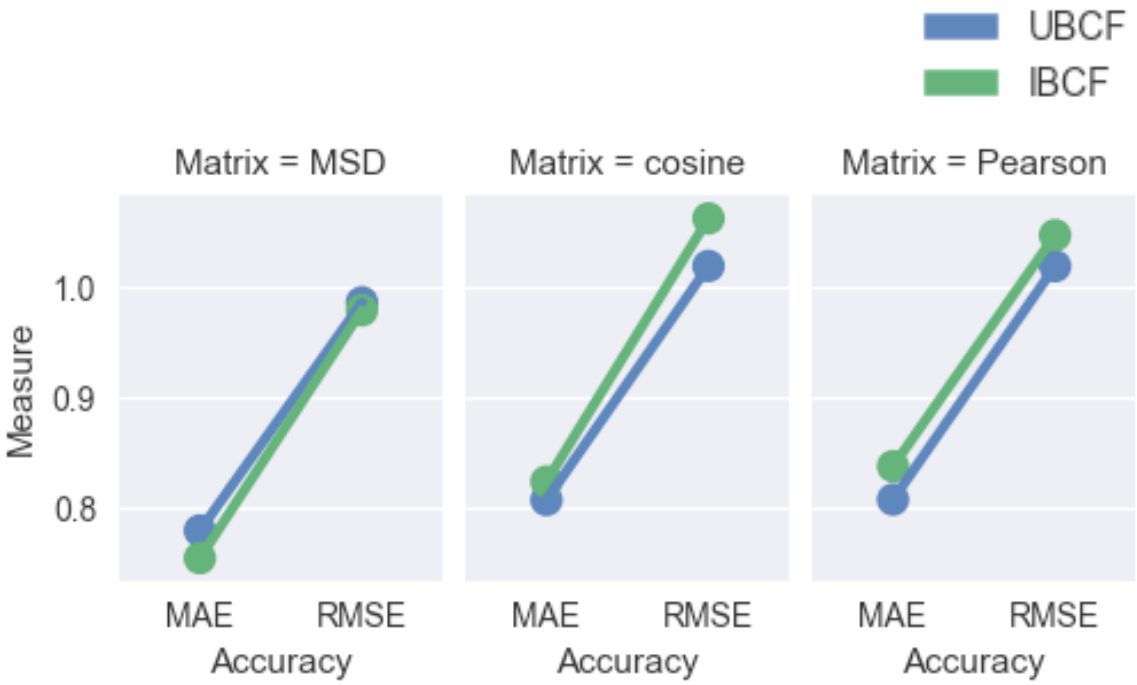
RMSE	**FCP**	**MAE**
1. Pearson	1. MSD	1. Pearson
2. Cosine	2. Cosine	2. Cosine
3. MSD	3. Pearson	3. MSD



The impact of the three metrics on User based Collaborative Filtering is consistent with the impact of the three metrics on Item based Collaborative Filtering, the above representation confirms the same point

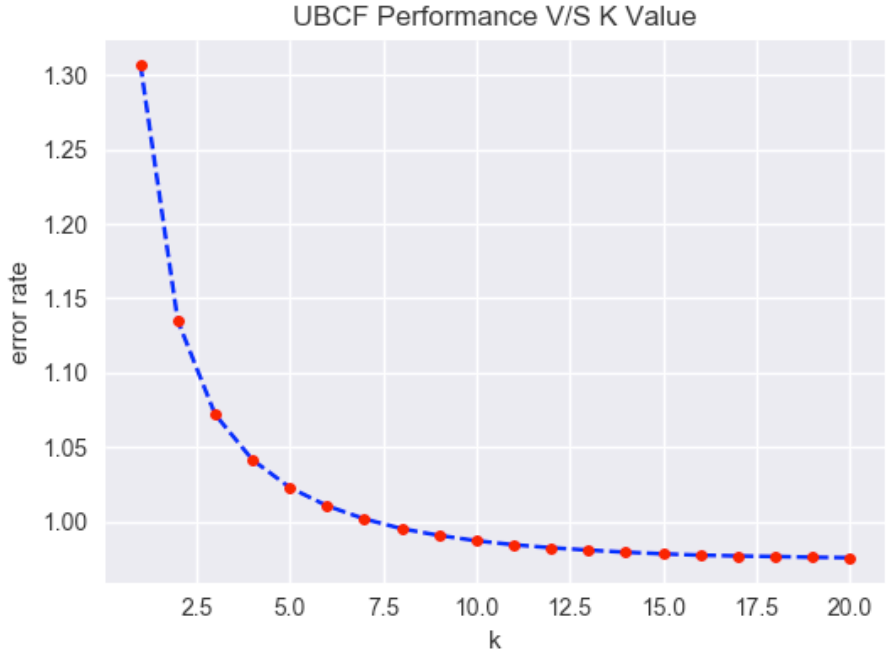


Clearly for the above figure we cannot see the difference in hue which says that results for UBCF and IBCF are similarly taking the impacts of similarity matrix change. An another view for the same:

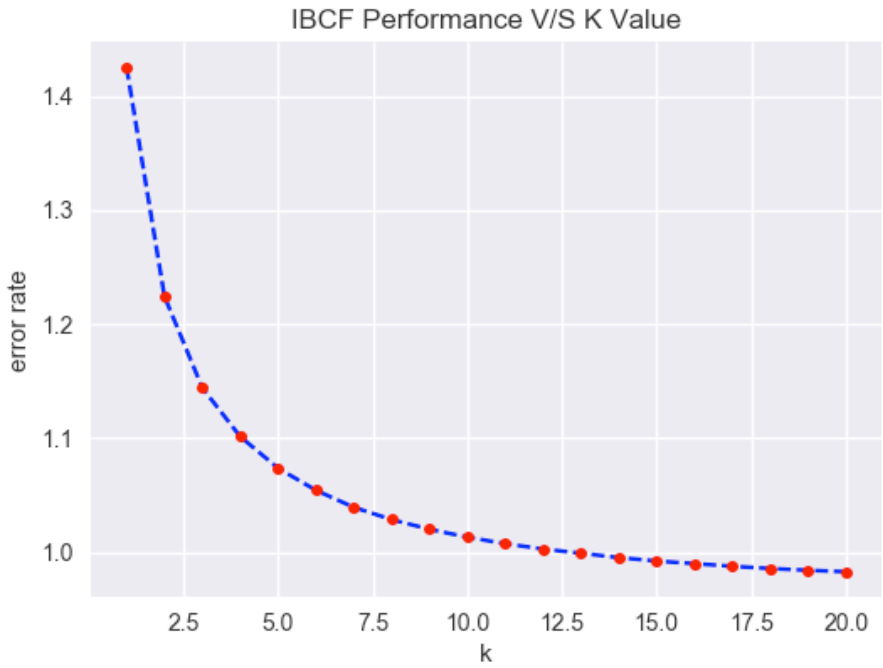


7. Examination of number of neighbors over the performances of User based Collaborative Filtering or Item based Collaborative Filtering

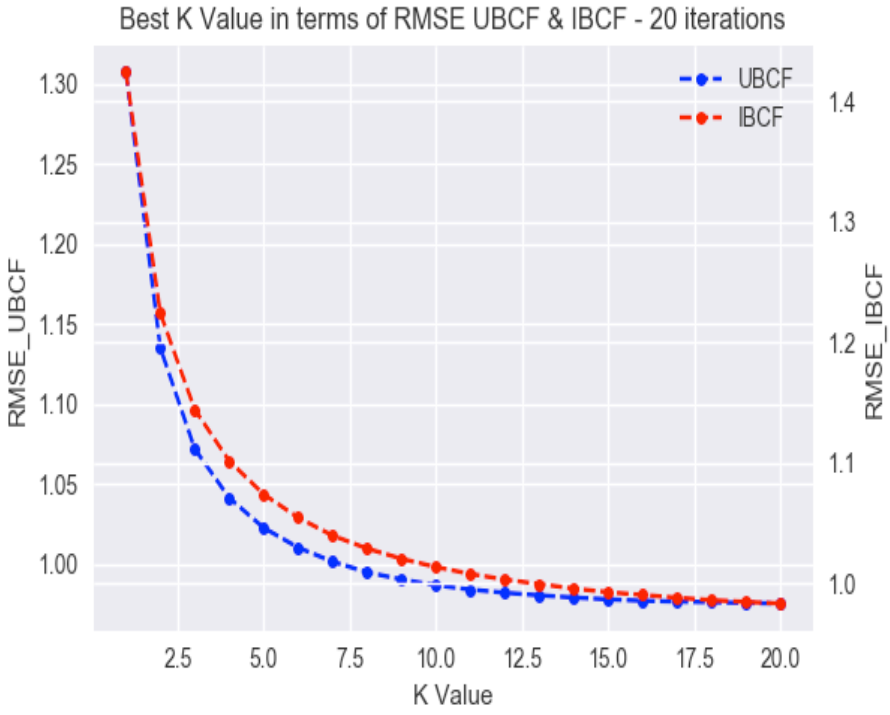
   a. Impact of K value on User Based Collaborative filtering:



b. Impact of K value on Item based Collaborative filtering:



c. Best K value for UBCF and IBCF



As we can observe from the above figure the best K value can be any value from k=12 to k=20 and the best case is similar for User based collaborative filtering and Item based Collaborative filtering.