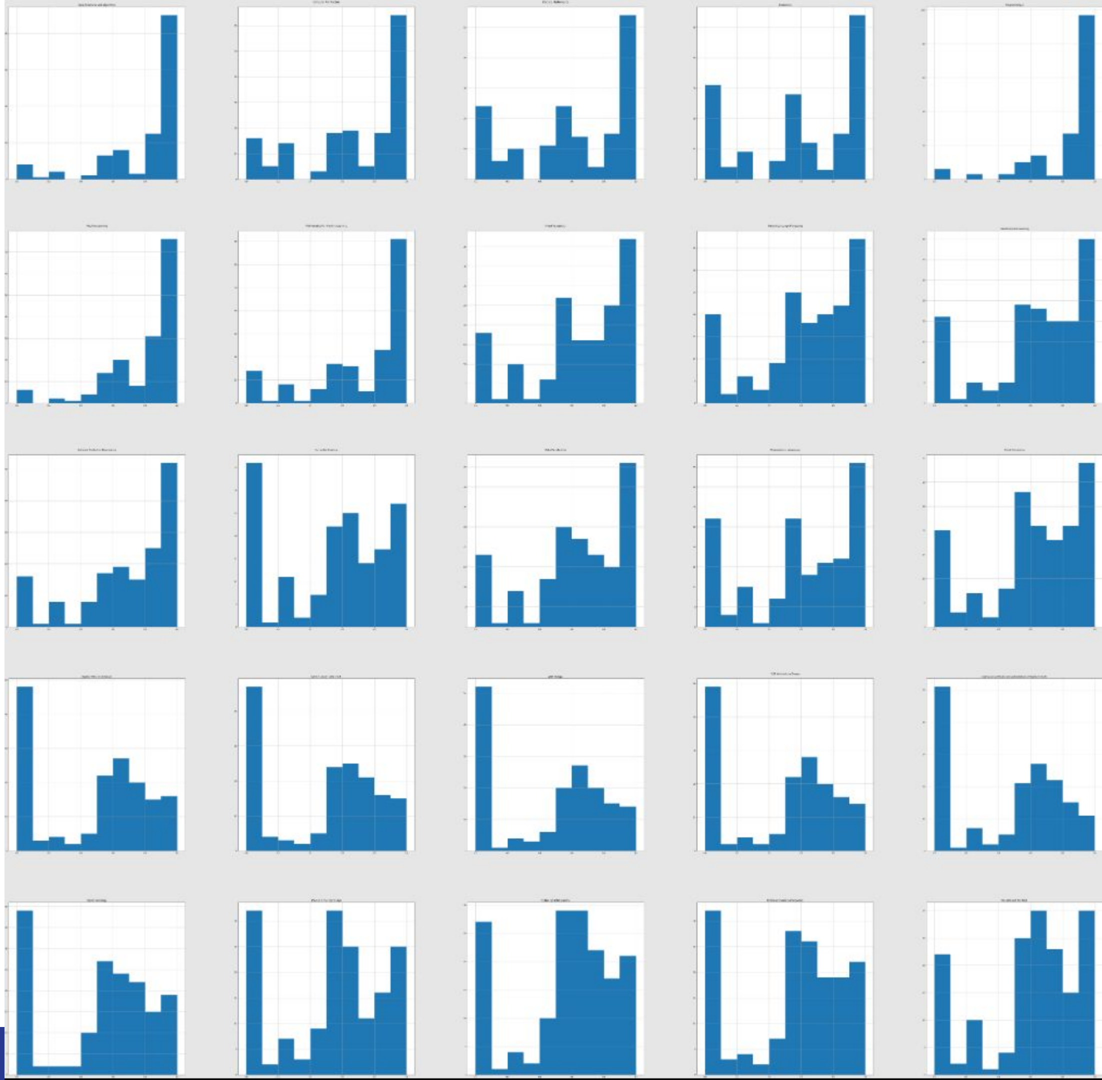
Recommendation Systems

Midterm Project

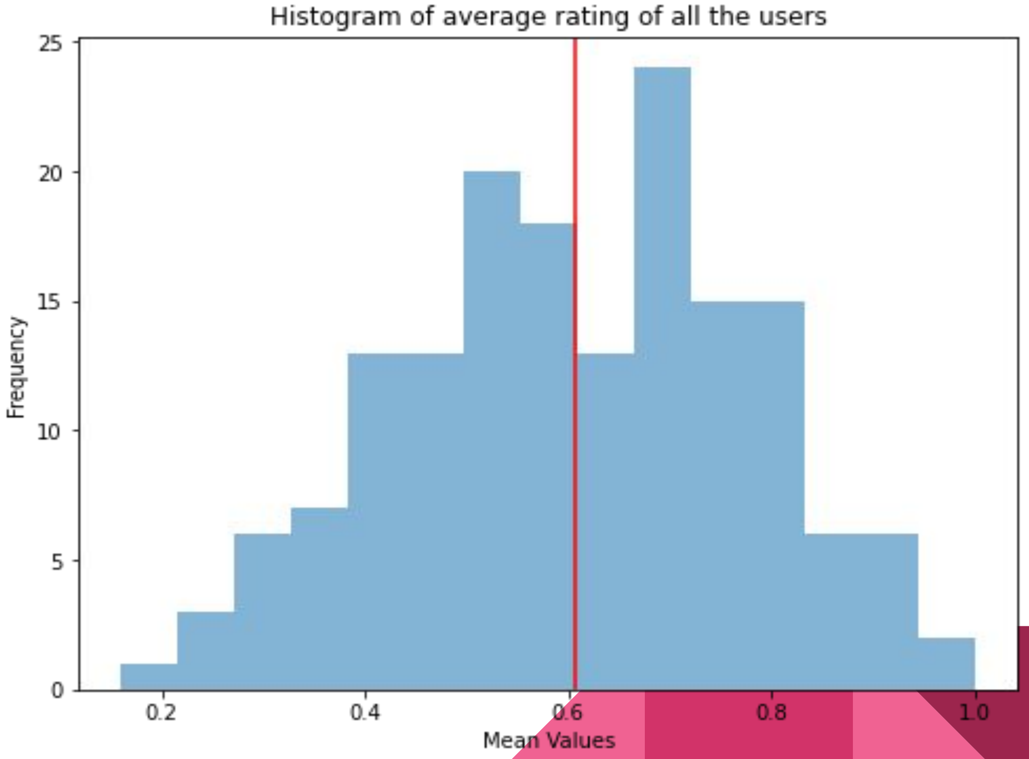
Preprocessing

1. Histogram of all the courses shows that student are biased towards some courses.
2. Replacing the NaN with the user average is more beneficial because mean value for every user is different so it won’t make sense to replace the NaN with that course average.

There is no ECE core courses in the 5 core courses which makes us difficult to differentiate between an ECE student and CSE student. We can also see that mostly all the ECE elective courses have a biased towards not taking the course(i.e 0 rating) because CSE student have given 0 rating instead of leaving that course.



This graph shows that average rating of given by a user is somewhat gaussian distribution. We have replaced the NaN with the average rating given by that user. Still every user have different range of rating between 0 to 1. We used MinMaxScaler row wise. Theoretically it should increase the accuracy but practically it is decreasing the overall accuracy.

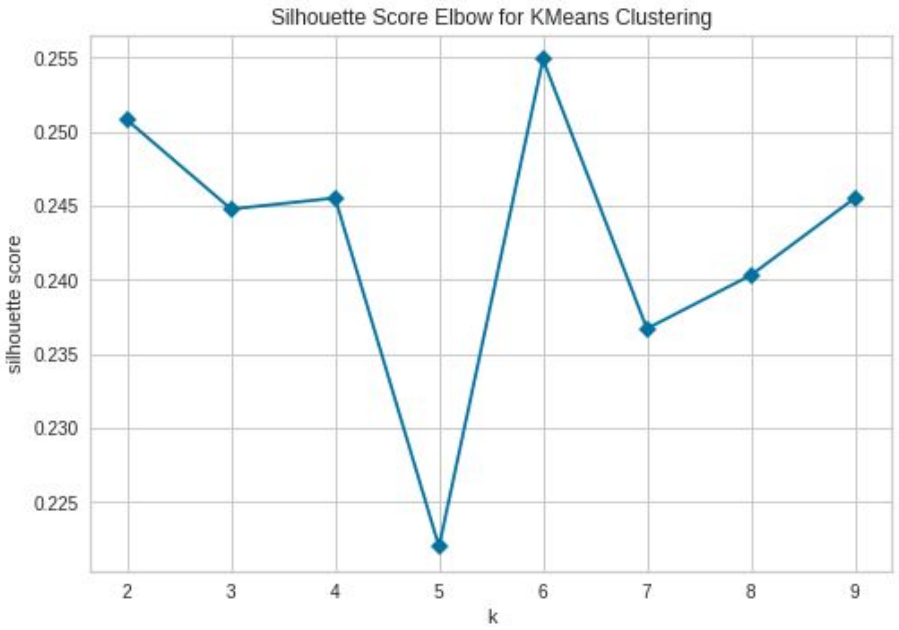


Our Approach

- The dataset had 5 core courses and 20 elective courses.
- For new users who completed the core courses, we recommended the best elective course based on the preferences of users in the same cluster.
- We handled missing data by taking the average of the rows.
- We used k-means clustering to group users based on similar preferences.
- We applied SVD to reduce the dimensionality of the dataset and identify important features.

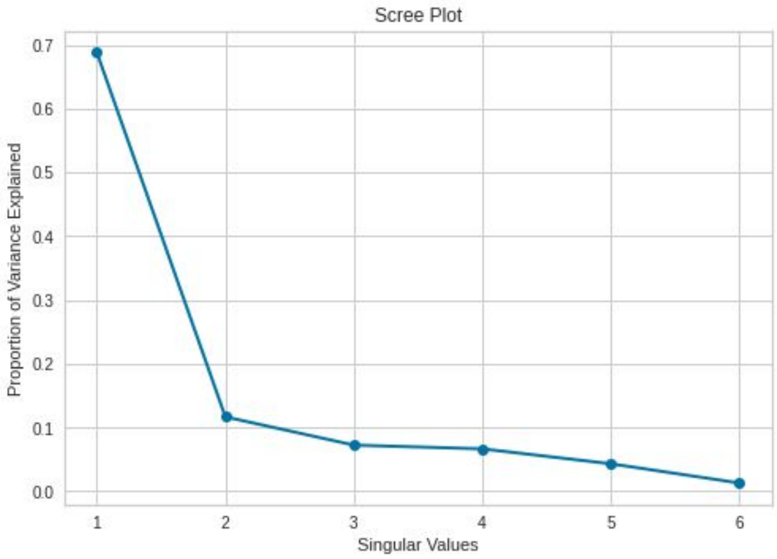
Choosing optimal K

- We used elbow method to find optimal K
- Here we can observe we get a peak for k = 6



Choosing optimal number of singular values

- We have used scree plot to get optimal number of singular values
- Here we can observe an elbow at 2 so it will be our optimal choice



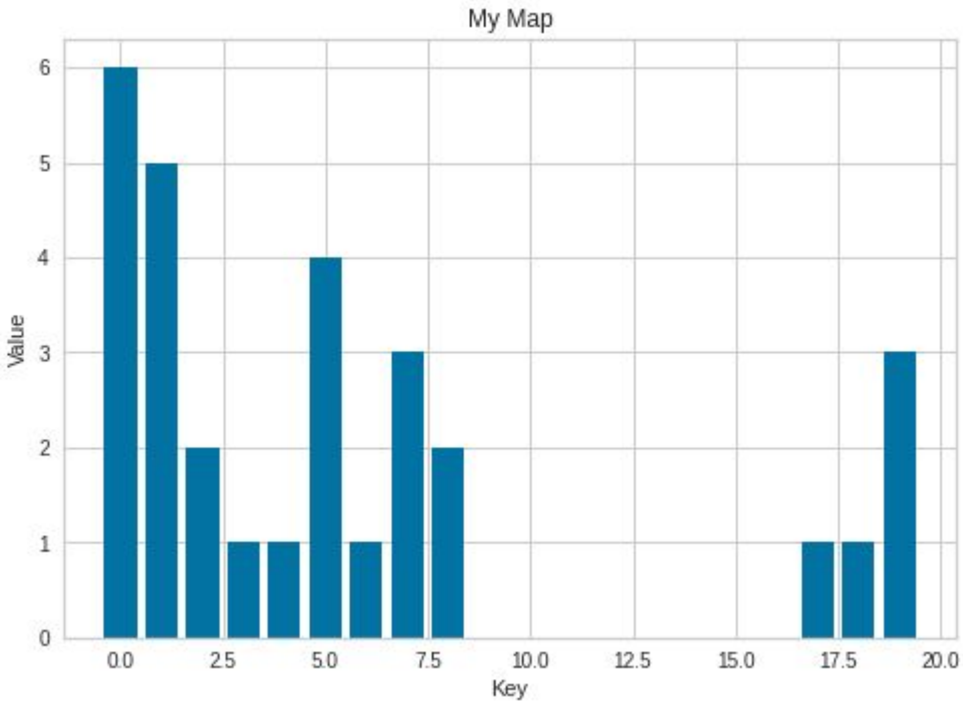
Observation for different singular values

- K = 6 (fixed)

Singular Values	Training Acc	Testing Acc
6 (Without SVD)	47.03	47.05
5	45.10	48.23
4	48.55	51.76
3	44.41	47.05
2	46.89	49.41
1	47.99	51.76

- A course will be recommended by many cluster if th course is likable by many people. Which is similar tc suggest a comedy movie to a new person.

```
{
'Machine Learning\n': 6,
 'Mathematics For Machine Learning\n': 5,
 'Data Visualization\n': 3,
 'Software Production Engineering\n': 4,
 'The Web and the Mind ': 3,
 'Reinforcement Learning': 1,
 'Techno-economics of networks ': 1,
 'Technology Ethics and AI ': 1,
 'Programming Languages\n': 2,
 'Visual Recognition\n': 2,
 'Natural Language Processing\n': 1,
 'Computer Graphics\n': 1
}
```



Novelty

- Normally, for a new data point(user), we check which cluster this new data point belongs to and then use the ratings for electives of that cluster as the prediction for this new user.
- But restricting the new user to one cluster gives ordinary results. Instead, we can take the contributions for multiple clusters for a new user.
- So we decided that we can predict a weighted sum of the ratings of clusters, with the weights being inversely proportional to the distance of the new data points to each cluster.
- With this, we observed a jump in testing accuracy from 50.5% to 76%

Team members

- Pushkar Pawar - IMT2020015
- Rudransh Dixit - IMT2020056
- Teja Janaki Ram - IMT2020100
- Manan Patel - IMT2020121
- Darpan Singh - IMT2020133