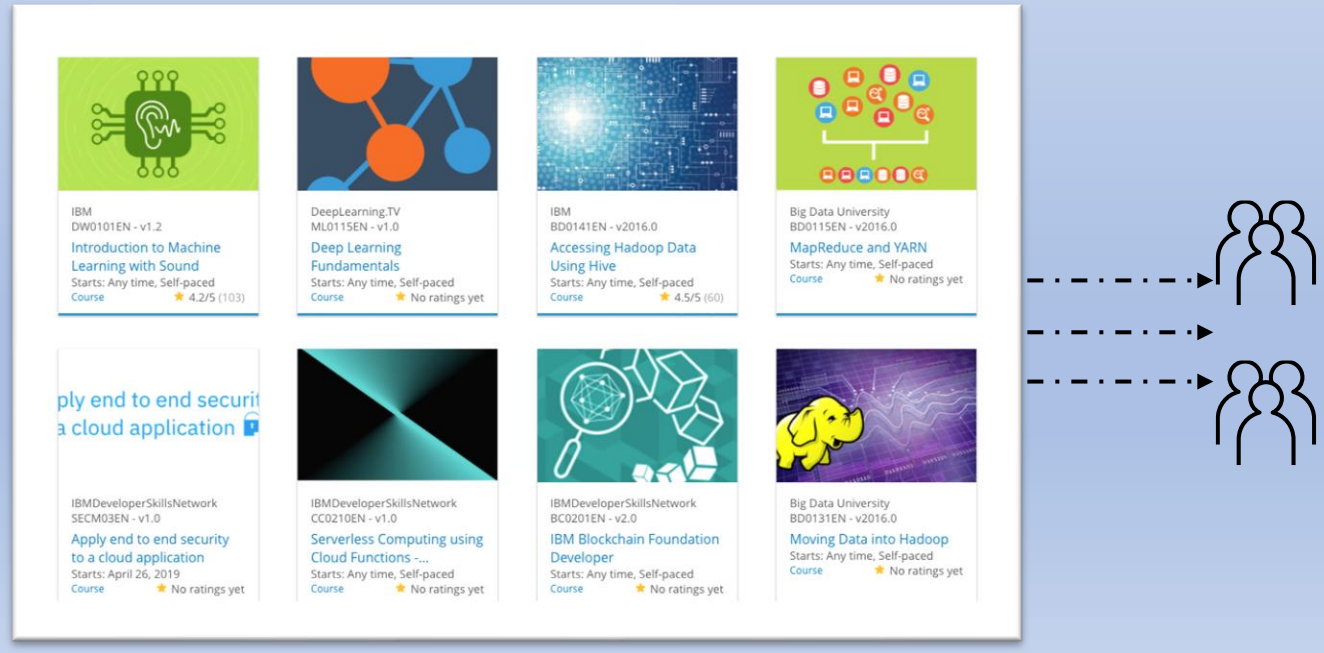
Build a Personalized Online Course Recommender System with Machine Learning

Taiwo Matthew Adekoya

2nd-April-2023



Outline

- Introduction and Background
- Exploratory Data Analysis
- Content-based Recommender System using Unsupervised Learning
- Collaborative-filtering based Recommender System using Supervised learning
- Conclusion
- Appendix

Introduction

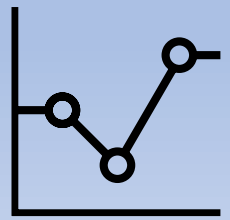
- Project background and context

  There are different courses that can be studied online. However, once a course has been taken, it is important to recommend courses that the student might be interested in. For this to be done, a good recommender system is needed.

- Problem states and hypotheses

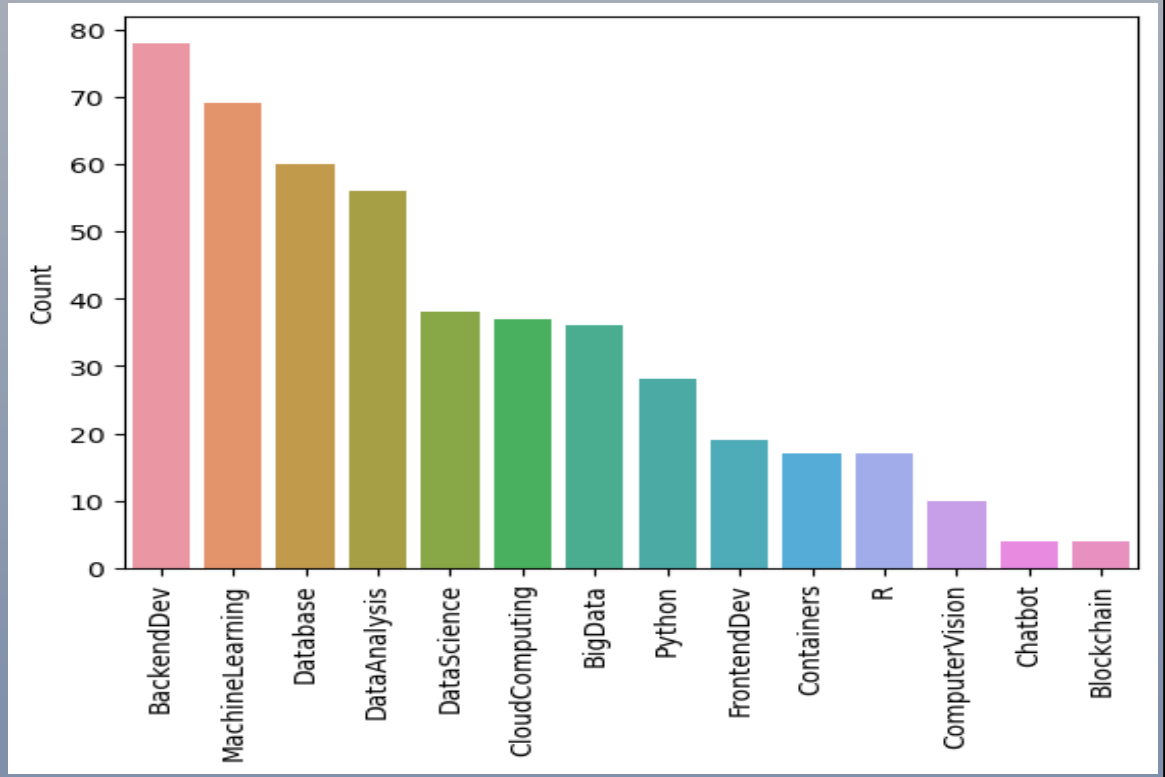
  After taking an online course, what course should be recommended to a user?

Exploratory Data Analysis



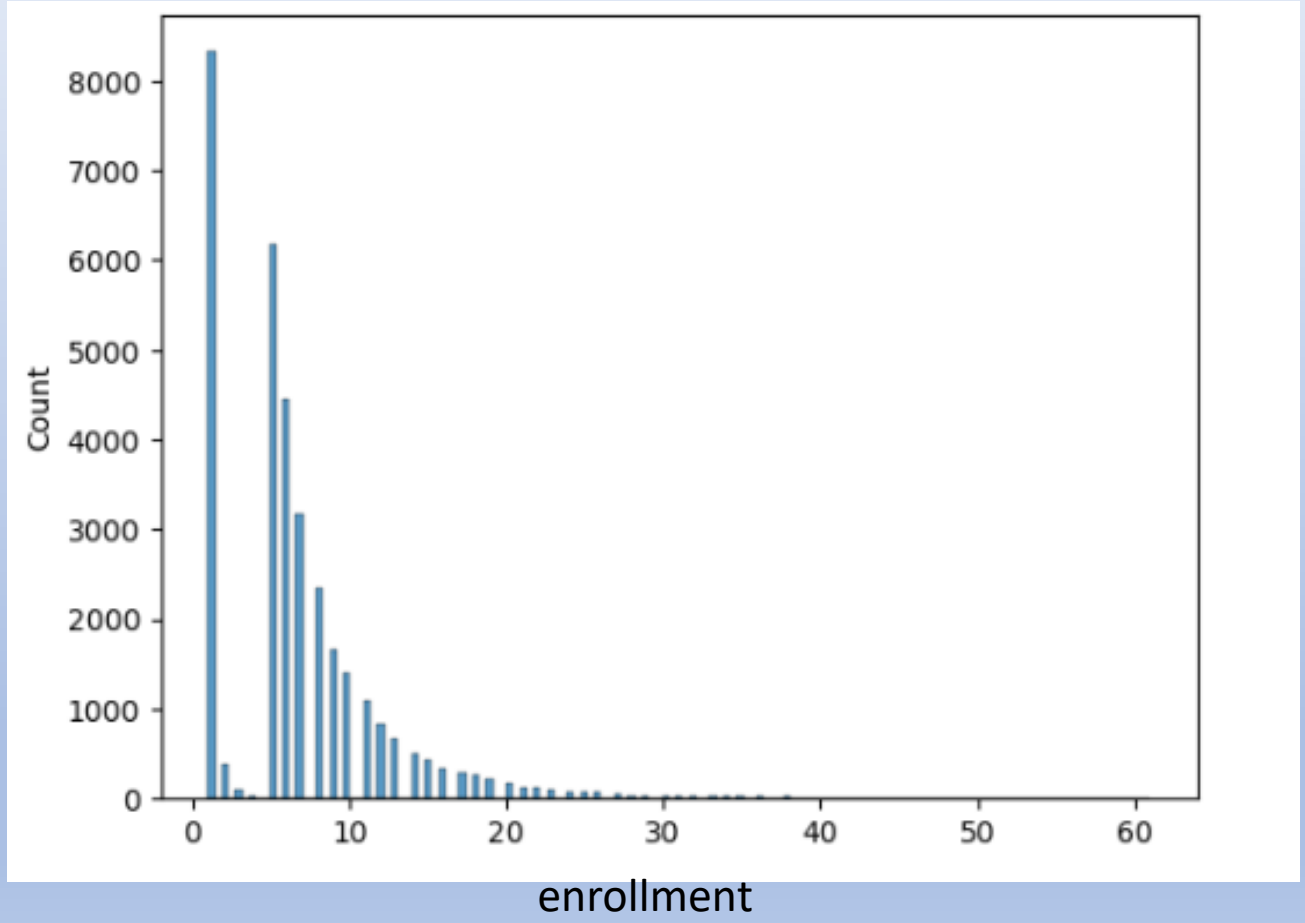
Course counts per genre

- The number of courses in each genre is represented.
- As seen, BackendDev has the highest number of courses, this is followed by machine learning while the least is Blockchain.



Course enrollment distribution

- The distribution is right tailed
- A lot of users enrolled and rated one item
- Generally, ratings per user decreased as enrolment increased.



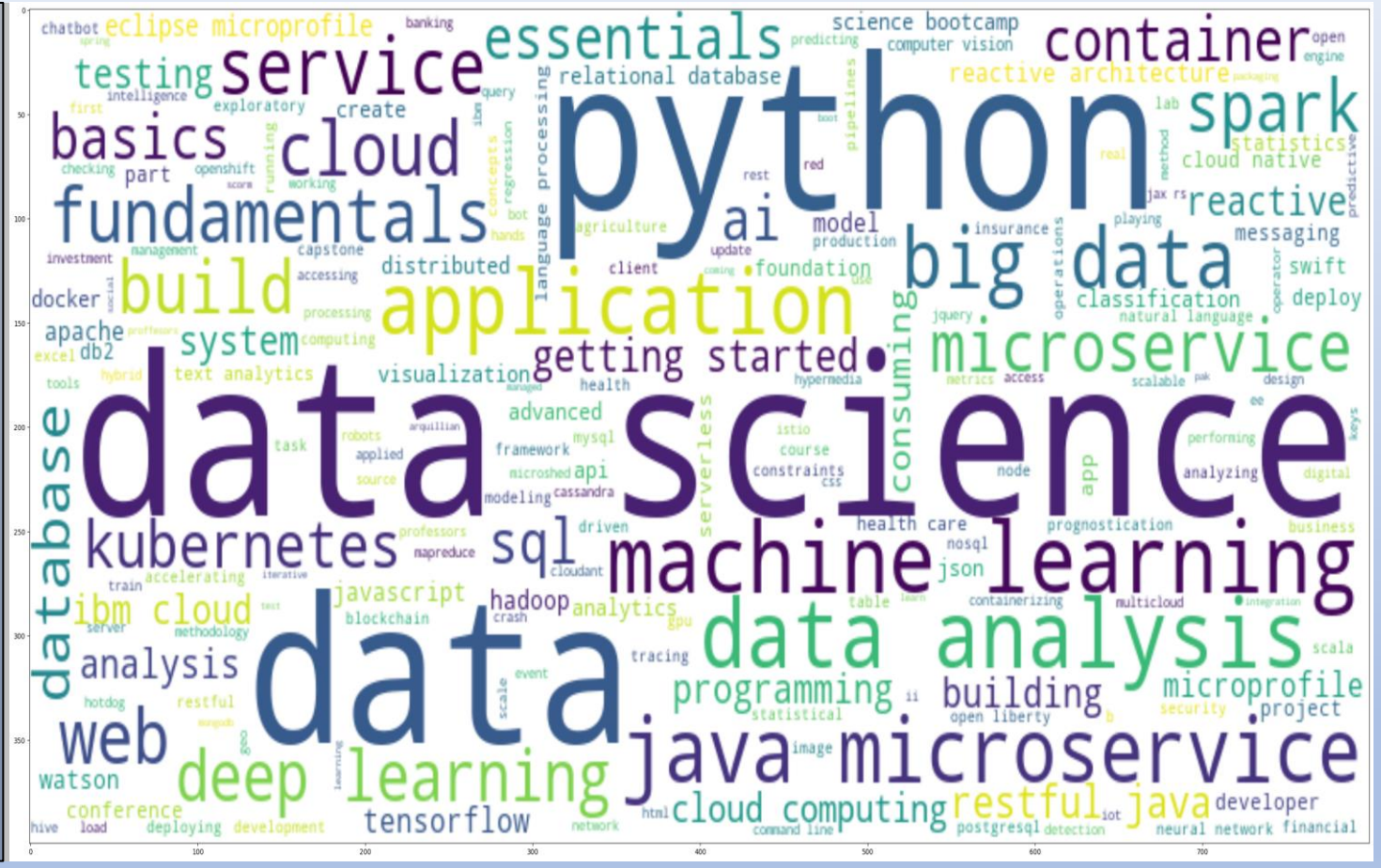
20 most popular courses

- The most popular courses are related to data science, data analysis and machine learning
- From the top 20 courses, almost half is related to data science and visualization.

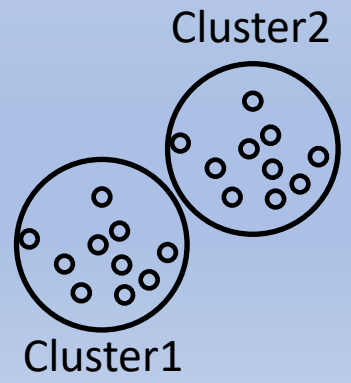
	TITLE	Ratings
0	python for data science	14936
1	introduction to data science	14477
2	big data 101	13291
3	hadoop 101	10599
4	data analysis with python	8303
5	data science methodology	7719
6	machine learning with python	7644
7	spark fundamentals i	7551
8	data science hands on with open source tools	7199
9	blockchain essentials	6719
10	data visualization with python	6709
11	deep learning 101	6323
12	build your own chatbot	5512
13	r for data science	5237
14	statistics 101	5015
15	introduction to cloud	4983
16	docker essentials a developer introduction	4480
17	sql and relational databases 101	3697
18	mapreduce and yarn	3670
19	data privacy fundamentals	3624

Word cloud of course titles

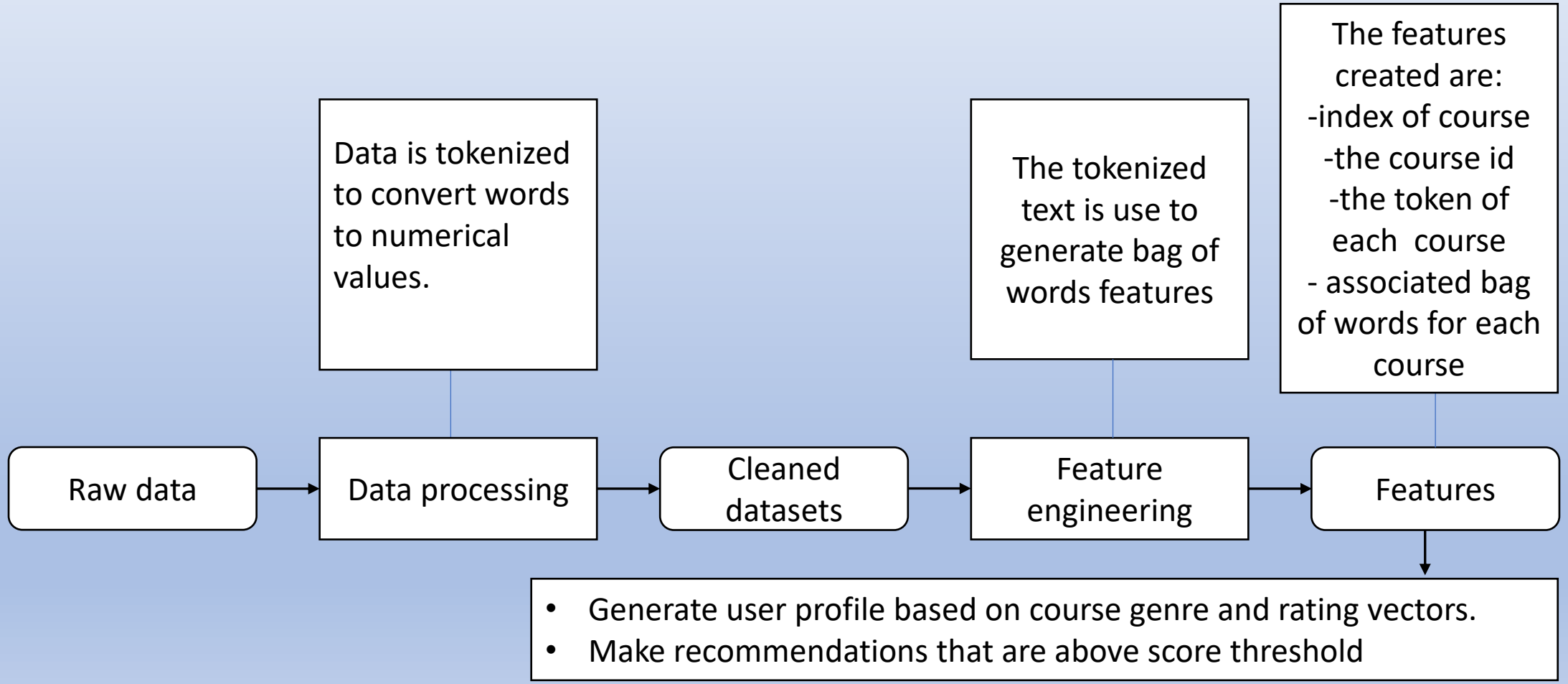
- The word cloud shows that there are many popular IT related keywords such as python, data science, machine learning, big data etc.
- Thus, most courses in the data are focused on IT skills.



Content-based Recommender System using Unsupervised Learning



Flowchart of content-based recommender system using user profile and course genres



Evaluation results of user profile-based recommender system

New/unseen courses have been recommended per user

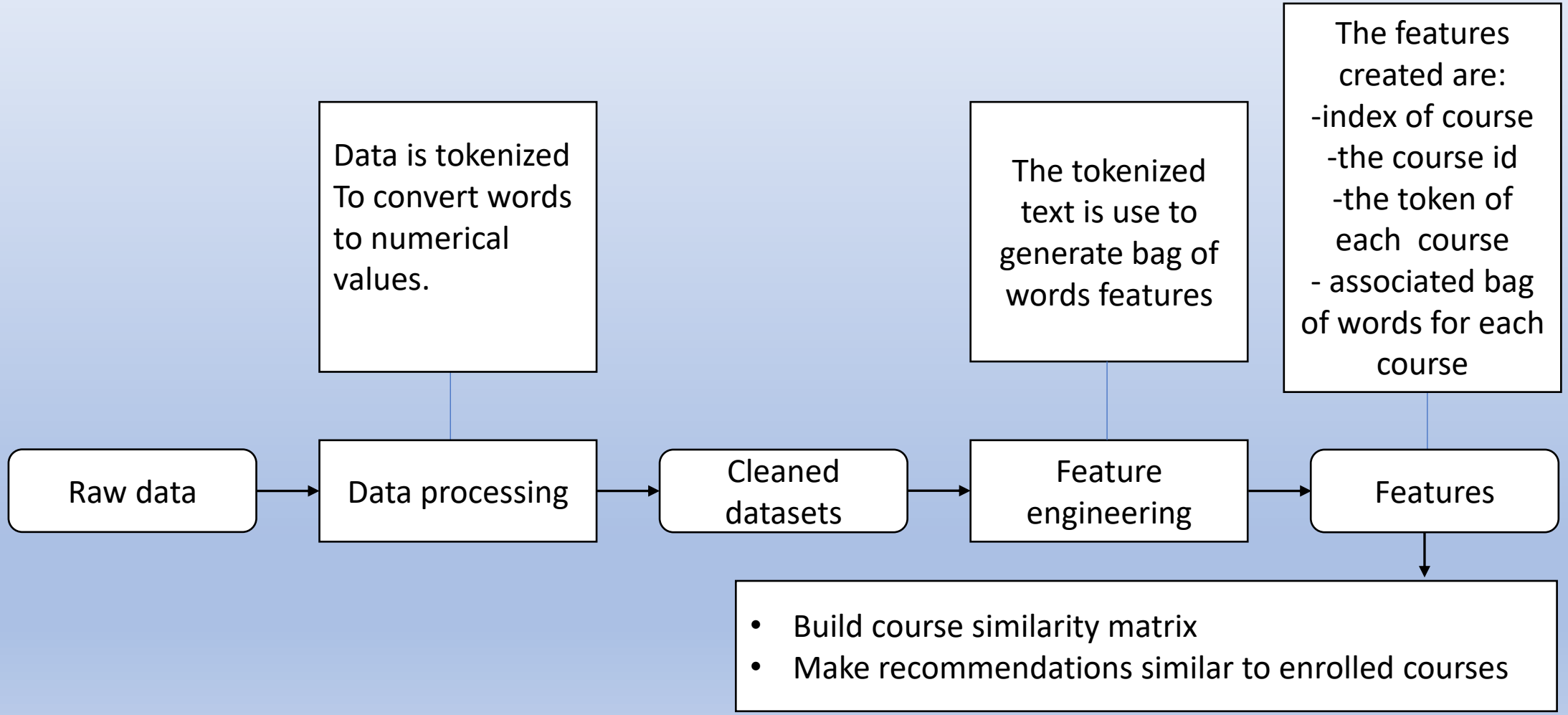
Score - Threshold	Average Courses Recommended per user
10	62
20	28
24	24
26	21

The higher the threshold, the lower the courses recommended

The top-10 commonly recommended courses (with score threshold set at 10) are:

COURSE_ID	
TA0106EN	608
GPXX0IBEN	548
excourse22	547
excourse21	547
ML0122EN	544
excourse06	533
excourse04	533
GPXX0TY1EN	533
excourse31	524
excourse73	516

Flowchart of content-based recommender system using course similarity



Evaluation results of course similarity based recommender system

Average number of /unseen courses recommended per user is 11

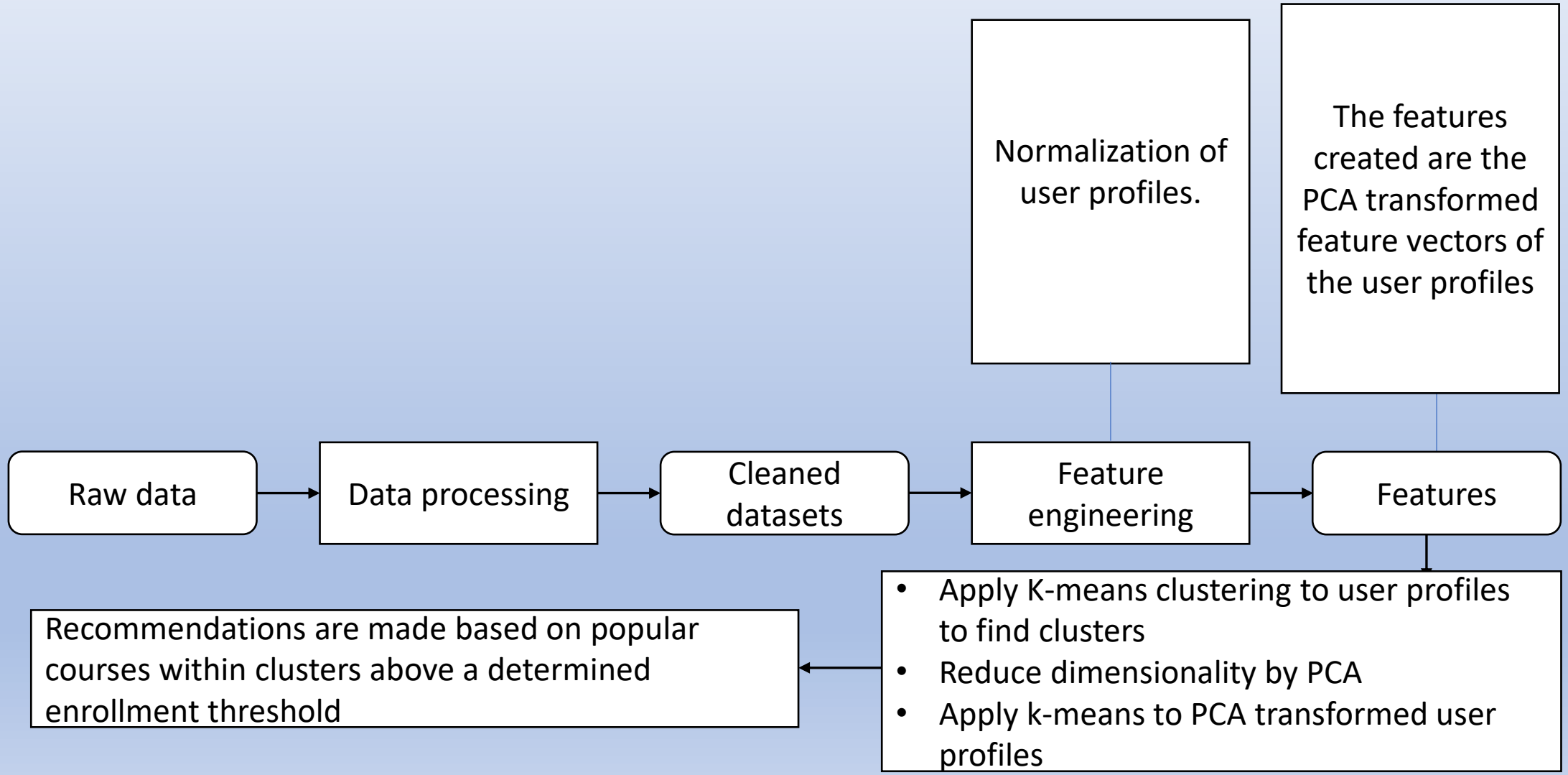
```
res_dict = {}
users, courses, sim_scores = generate_recommendations_for_all()
res_dict['USER'] = users
res_dict['COURSE_ID'] = courses
res_dict['SCORE'] = sim_scores
res_df = pd.DataFrame(res_dict, columns=['USER', 'COURSE_ID', 'SCORE'])

user = res_df.groupby('USER')['COURSE_ID'].size().sort_values(ascending=False)
user
```

The top 10 most frequently recommended courses are

COURSE_ID	
excourse62	579
excourse22	579
DS0110EN	562
excourse65	555
excourse63	555
excourse72	551
excourse68	550
excourse74	539
excourse67	539
BD0145EN	506

Flowchart of clustering-based recommender system



Evaluation results of clustering-based recommender system

Average new courses recommended per user :

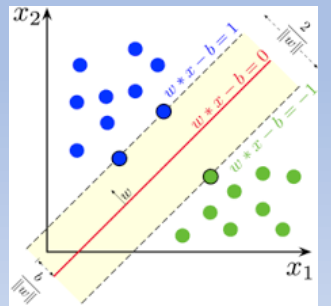
Enrollment threshold	Average Courses Recommended per user
30	20
40	13
59	11

The higher the threshold, the lower the courses recommended

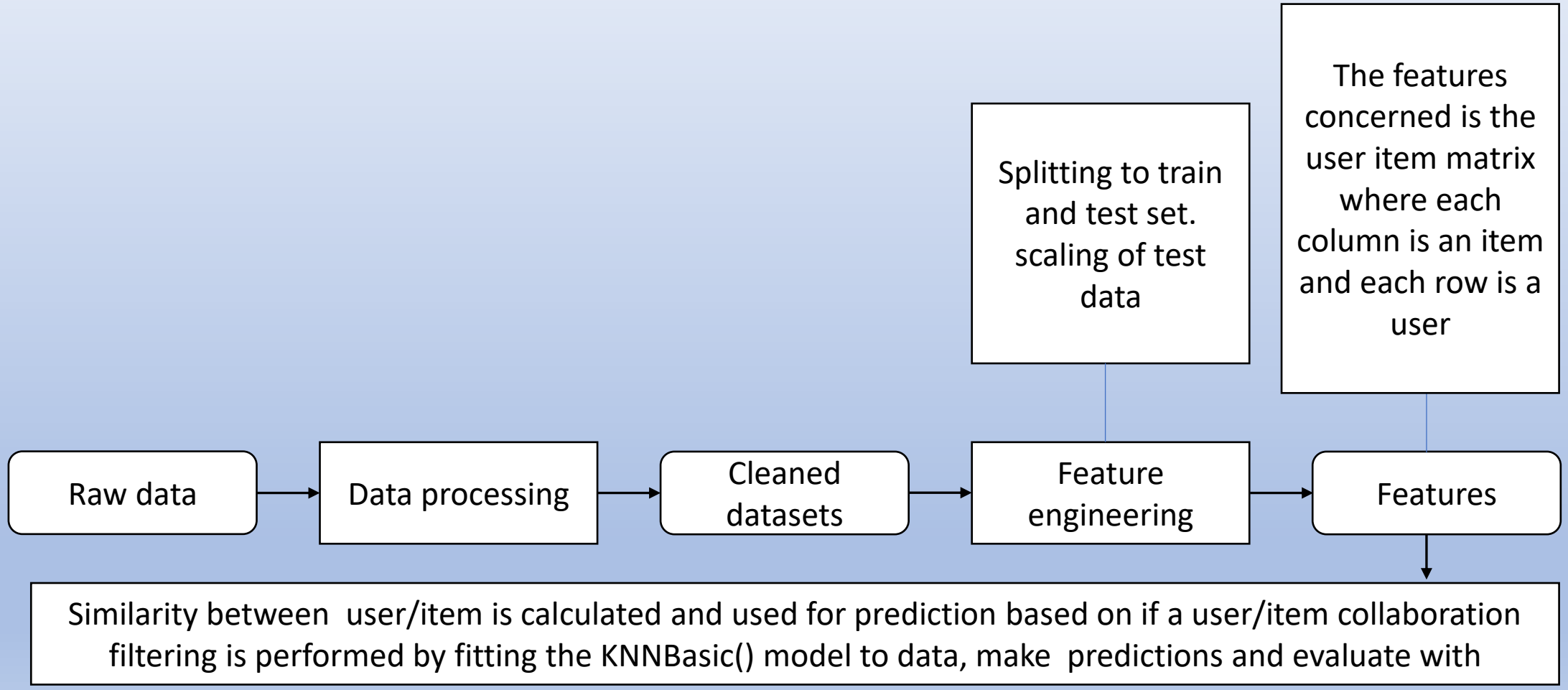
Top-10 commonly recommended courses when enrollment threshold is set at 30:

COURSES	
PY0101EN	2524
DS0101EN	2304
BD0101EN	2013
BD0111EN	1535
BC0101EN	979
CB0103EN	644
ML0115EN	589
DA0101EN	560
DS0105EN	551
DS0103EN	519

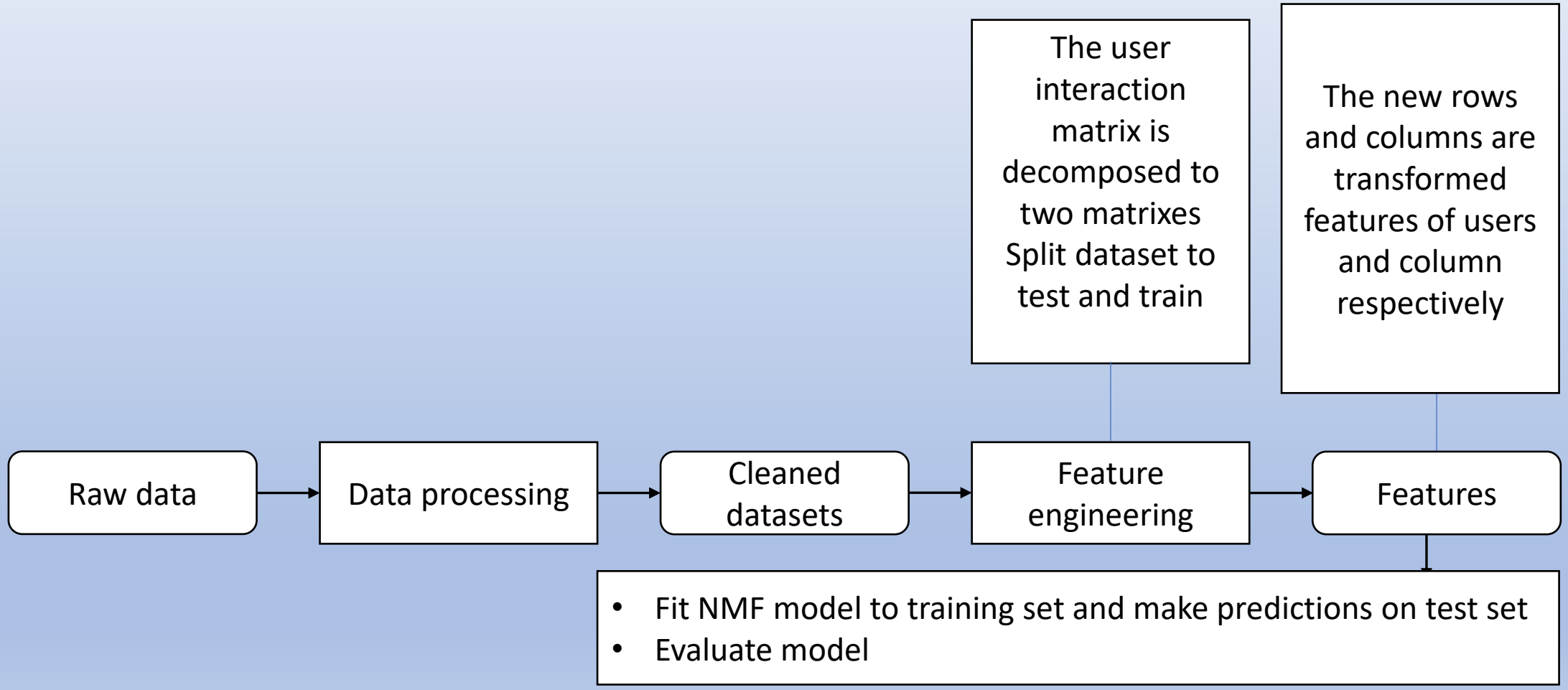
Collaborative-filtering Recommender System using Supervised Learning



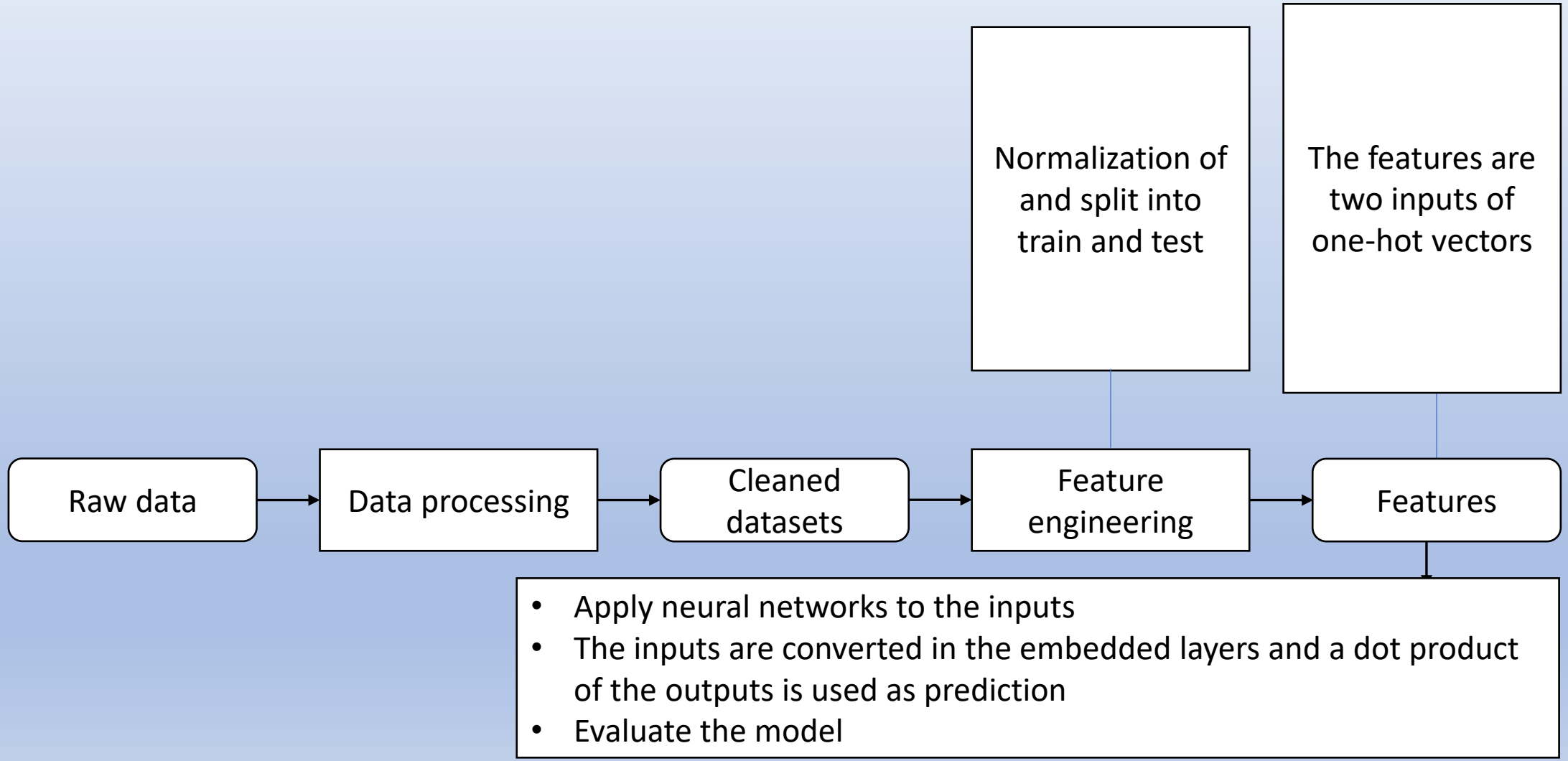
Flowchart of KNN based recommender system



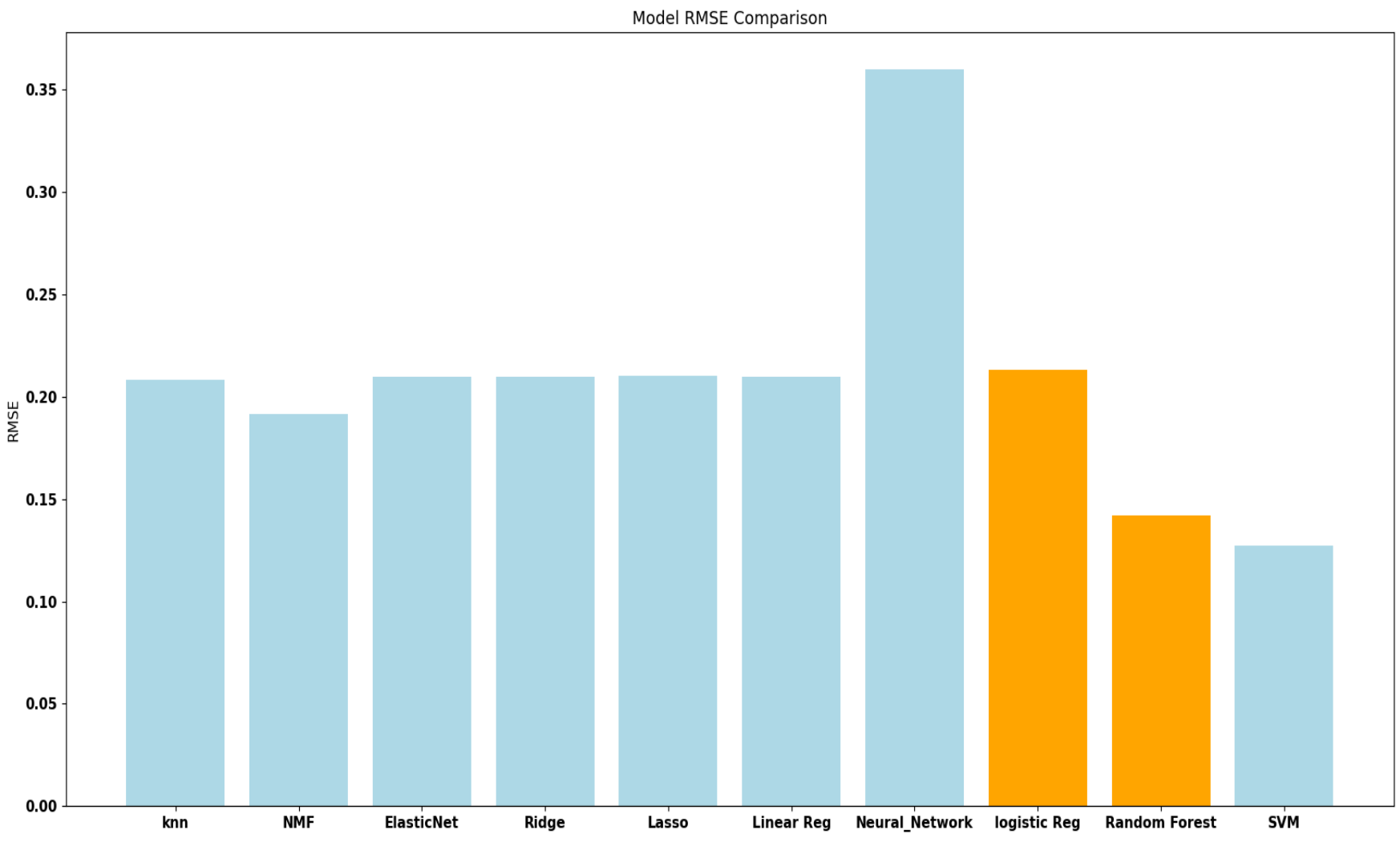
Flowchart of NMF based recommender system



Flowchart of Neural Network Embedding based recommender system



Compare the performance of collaborative-filtering models



The model with the lowest root mean square error is the Support Vector Machine (SVM) model

Conclusions

- Most popular courses are those related to data science
- The higher the threshold used in the recommender system , the lower the recommendations
- Among the models used for collaborative filtering recommender system, the Support Vector Model had the least error.

Appendix

- https://drive.google.com/drive/folders/1OMiydEyUTsb70jibm9lEXkvrhUzi_XP7?usp=share_link