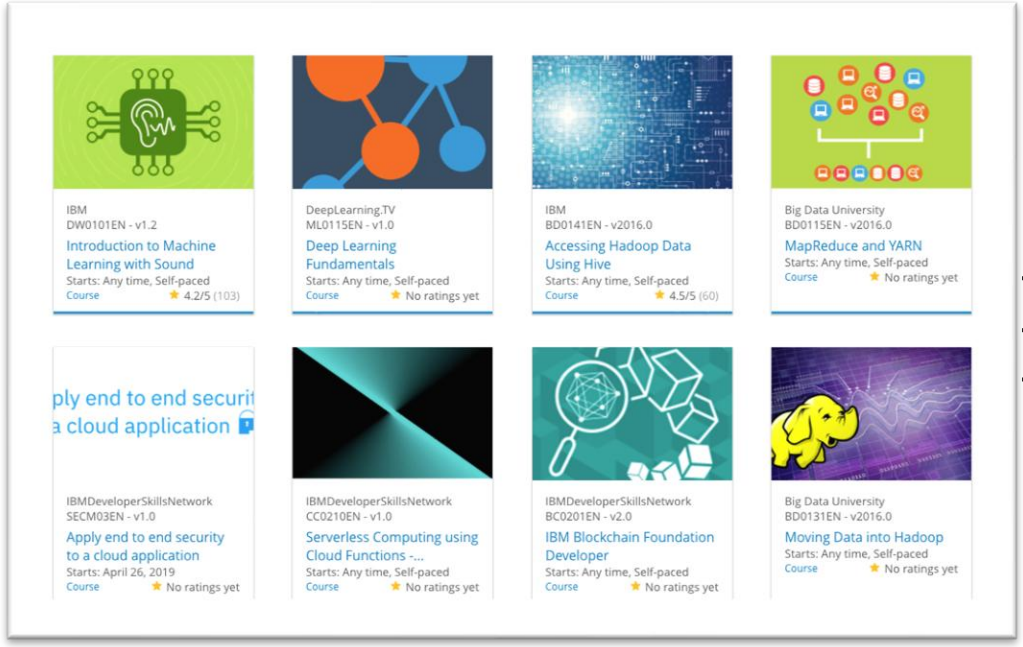
Build a Personalized Online Course Recommender System with Machine Learning

Abhishek Taneja
10th Jan 2025



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

Background and Context

- The ever-increasing availability of online courses creates a challenge for learners: how to find the courses that best match their needs and interests.
- A recommender system can bridge this gap by helping users discover relevant learning opportunities.
- Mention the specific data sources

Problem Statement

- There is a vast selection of courses, and it can be overwhelming to find the right ones.
- A generic course list will not satisfy specific learner interests.

Hypotheses

- We hypothesize that by leveraging various machine learning techniques, we can build a personalized recommender system that suggests courses with high relevance to individual users.
- We also hypothesize that different machine learning algorithms will have different performance with our data.

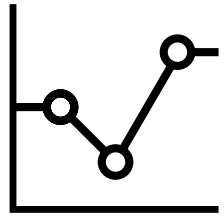
A course recommendation system will help in

- Finding better courses
- Finding courses that well suits each person's interests
- We aim to find the best courses to recommend to users based on their interests, their friend's interests, and the courses they are enrolled in.

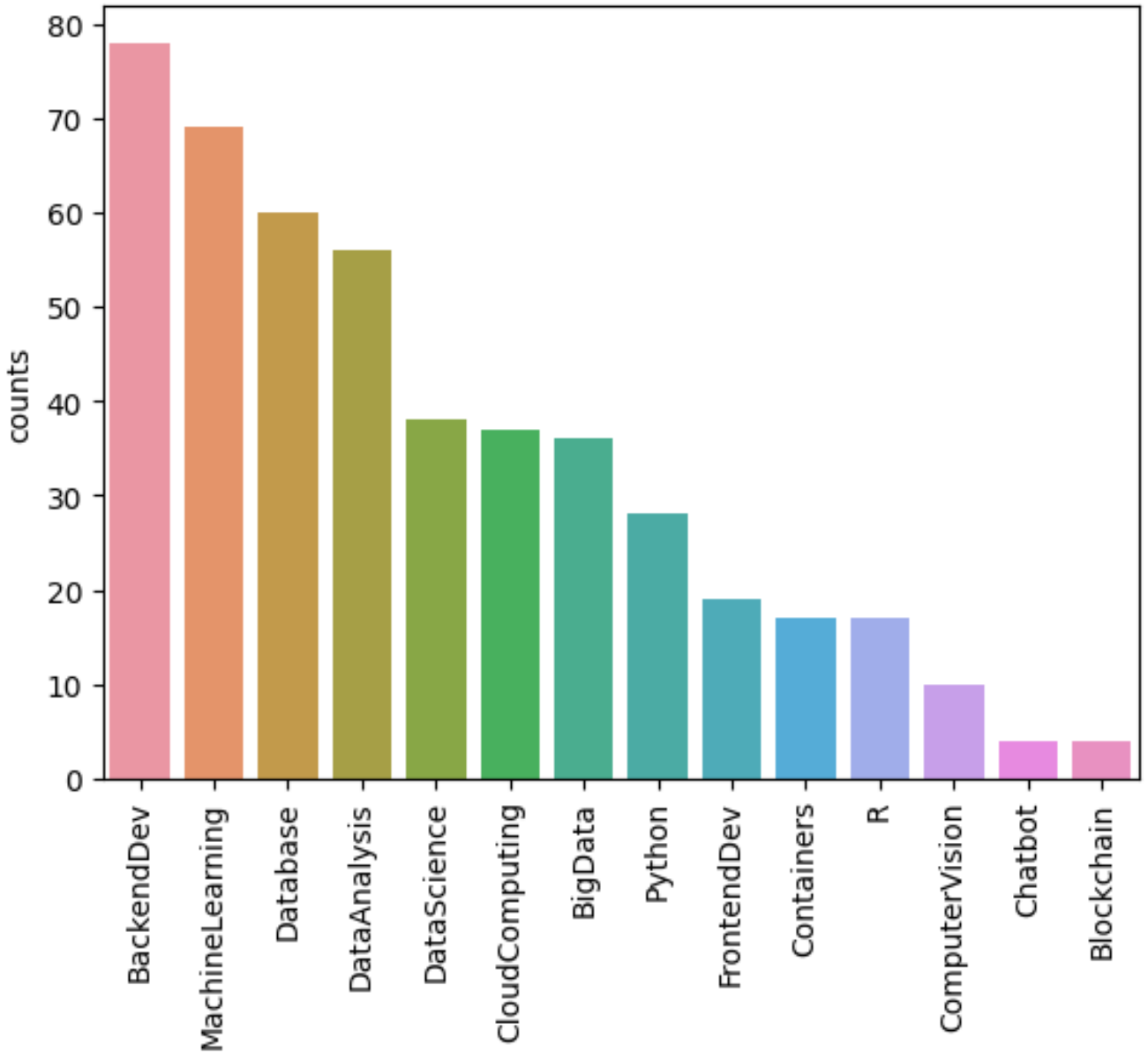
Obstacles

- We have many approaches
- Each approach has different assumptions

Exploratory Data Analysis

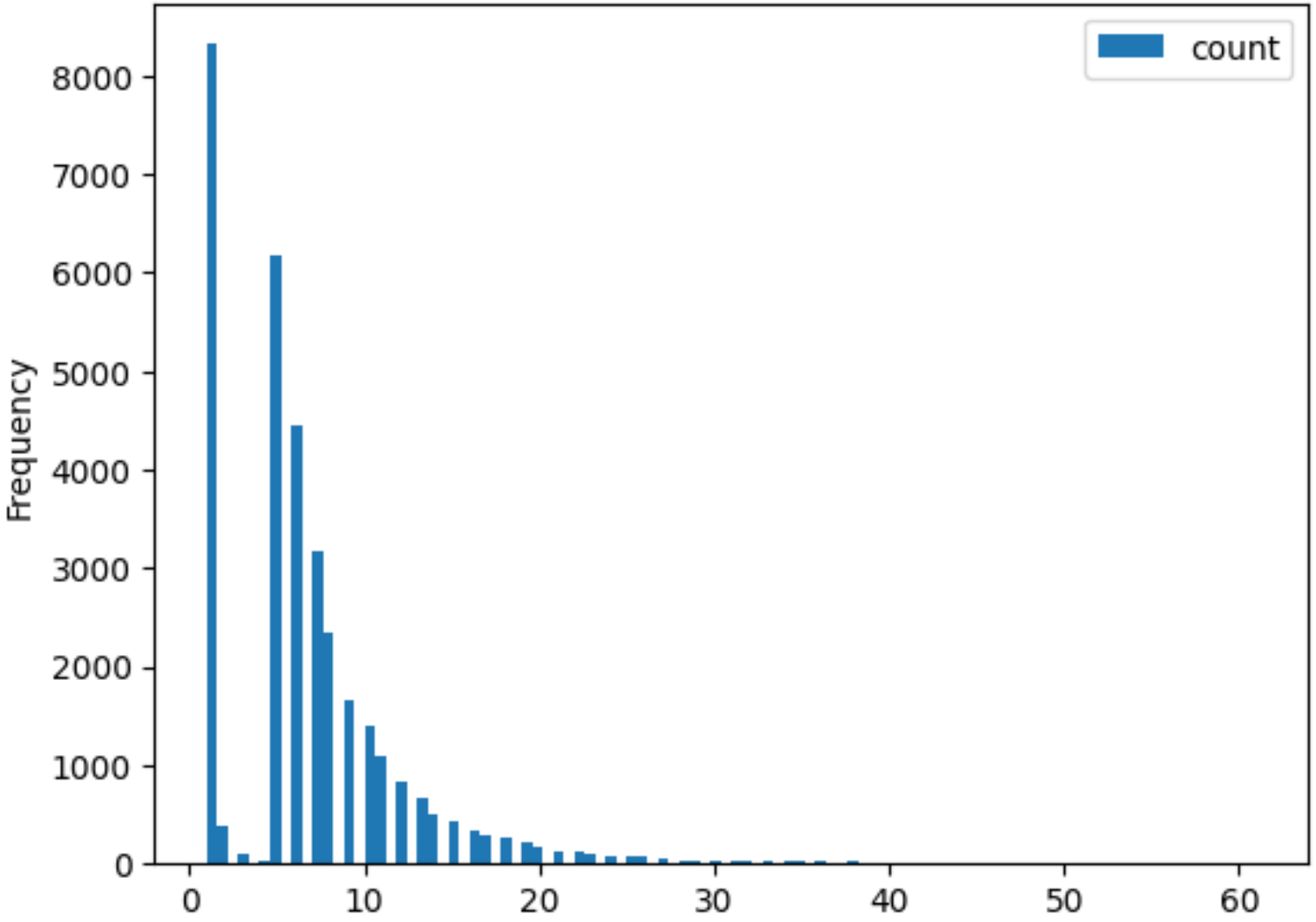


Course counts per genre



From the plot, BackendDev is the most commond coure genre, followed closely by MachineLearning, Database and DataAnalysis

Course enrollment distribution



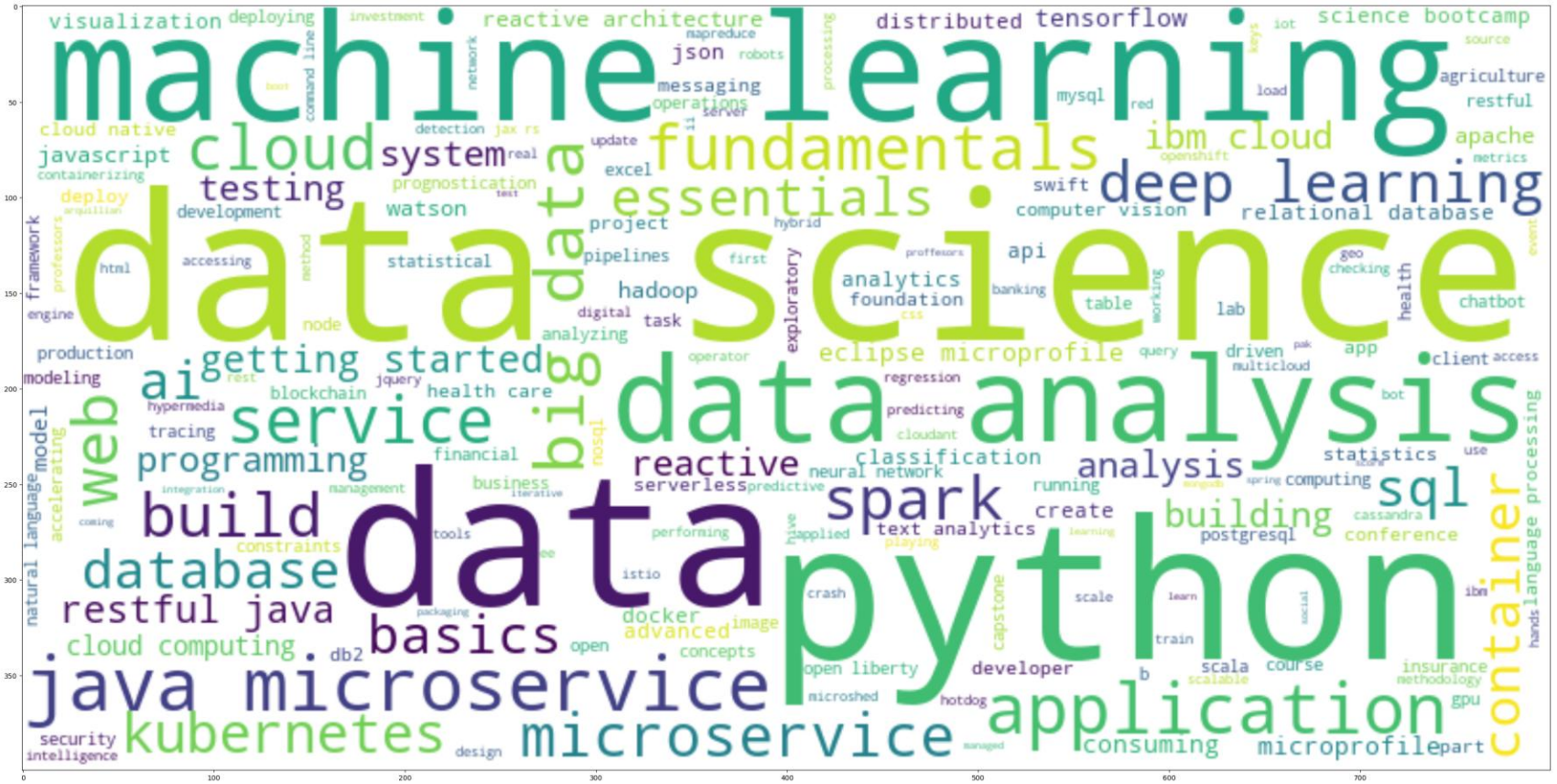
It could be seen that enrolling in just one course is the most common behavior among users. In addition, enrolling in just two or three courses is very rare. If a user enrolls in more than one course, the number of enrollments is likely to be 5 or greater.

20 most popular courses

Python for Data Science is the most popular course followed closely by introduction to data science. Intro to big data, Hadoop and data analysis are also in top positions. Data related courses are the most popular.

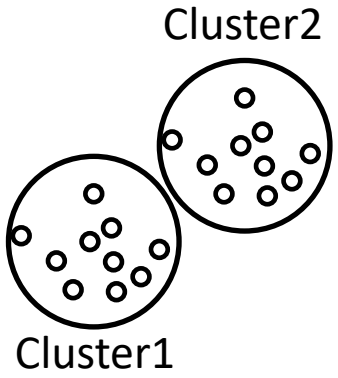
	TITLE	enrolls
0	python for data science	14936
1	introduction to data science	14477
2	big data 101	13291
3	hadoop 101	10599
4	data analysis with python	8303
5	data science methodology	7719
6	machine learning with python	7644
7	spark fundamentals i	7551
8	data science hands on with open source tools	7199
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

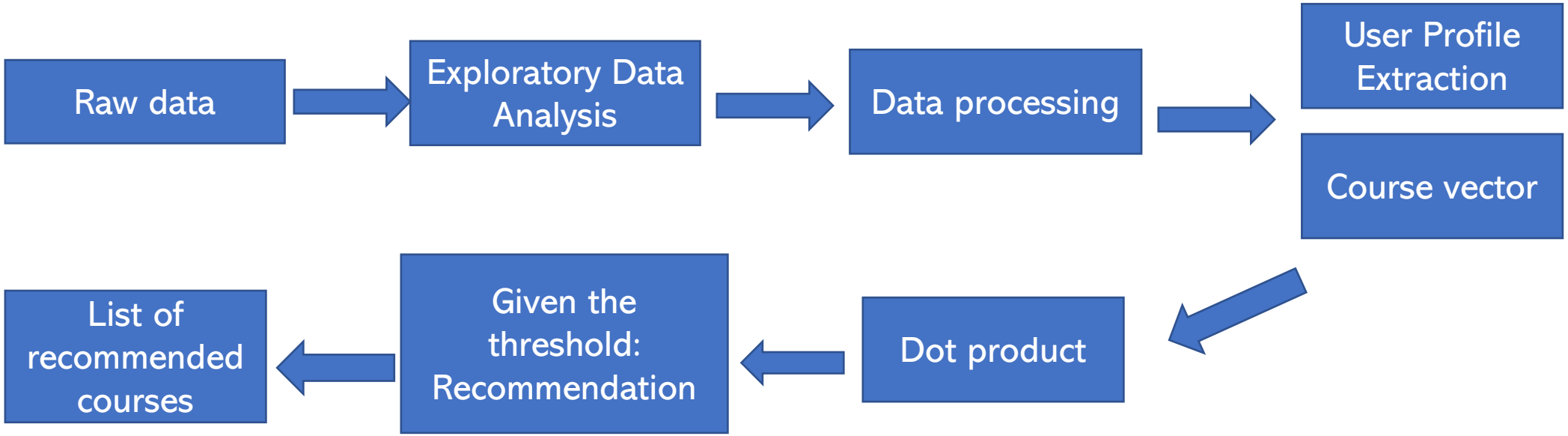


Data science, machine learning, data analysis and python are the most popular words in course titles. This gives an intuition about the nature of the courses in the dataset.

Content-based Recommender System using Unsupervised Learning



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



The input is raw data. After an EDA stage, the data has been processed and features were extracted, such as user profile and course vector. Then, dot product operation was made on new the course vectors using that user profile vector to get a recommendation score. Then, courses above the threshold were recommended.

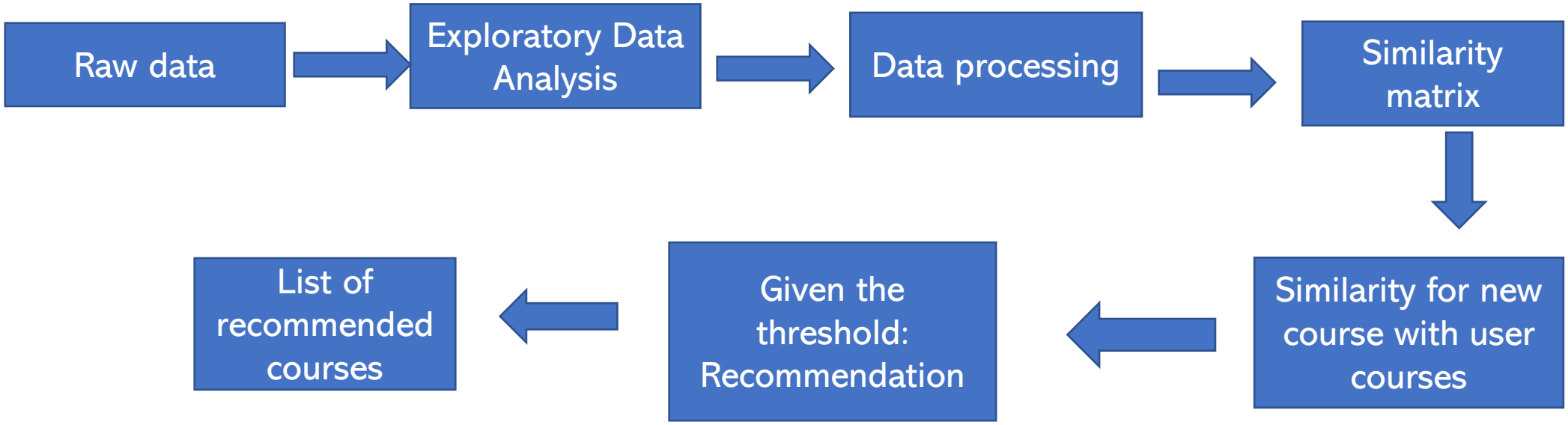
Evaluation results of user profile-based recommender system

On average, 19 new courses have been recommended to each user in the Data set

]:

	TITLE
0	analyzing big data in r using apache spark
1	getting started with the data apache spark ma...
2	spark fundamentals ii
3	spark overview for scala analytics
4	using the cql shell to execute keyspace operat...
5	\r\ndistributed computing with spark sql
6	cloud computing applications part 2 big data...
7	big data capstone project
8	foundations for big data analysis with sql
9	analyzing big data with sql

Flowchart of content-based recommender system using course similarity



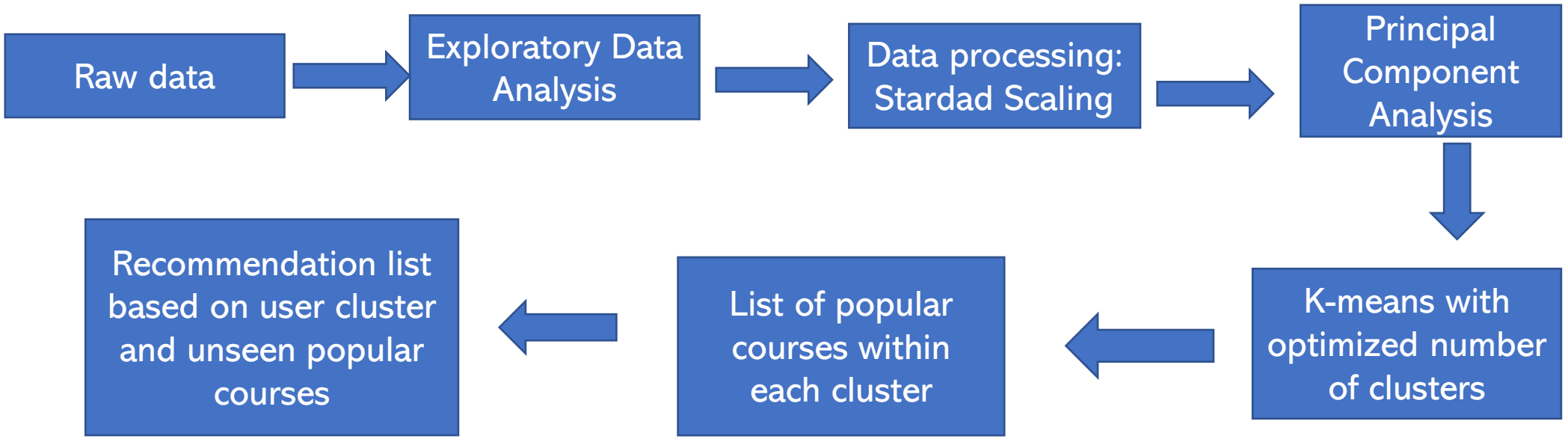
The input is raw data. After an EDA stage, the data has been processed and The similarity matrix for every course is obtained. Then, each new course for the user is compared with each courser of the user and a similarity score is extracted. If this score is greater than a threshold, it is recommended. This process is repeated for each user. At the end a list of recommended course is obtained.

Evaluation results of course similarity based recommender system

On average, 2 new courses have been recommended to each user in the Data set

TITLE
watson analytics for social media
text analytics 101
text analysis
data science with open data
introduction to data science in python
machine learning
machine learning for all
introduction to data science in python
a crash course in data science
data science fundamentals for data analysts

Flowchart of clustering-based recommender system



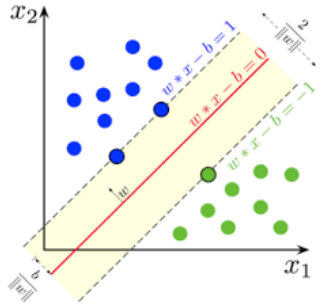
The input is raw data. After an EDA stage, the data has been standardized and normalized. Then, Principal Component analysis was performed for dimensionality reduction. With this new features, k-means algorithm was used for clustering. As a note, grid search was used to get the best hyperparameters for PCA and K-means. Finally, for each cluster a list of popular courses was made, and recommended courses were obtained from this list for each user. Unseen popular courses within the user cluster were the recommended ones.

Evaluation results of clustering-based recommender system

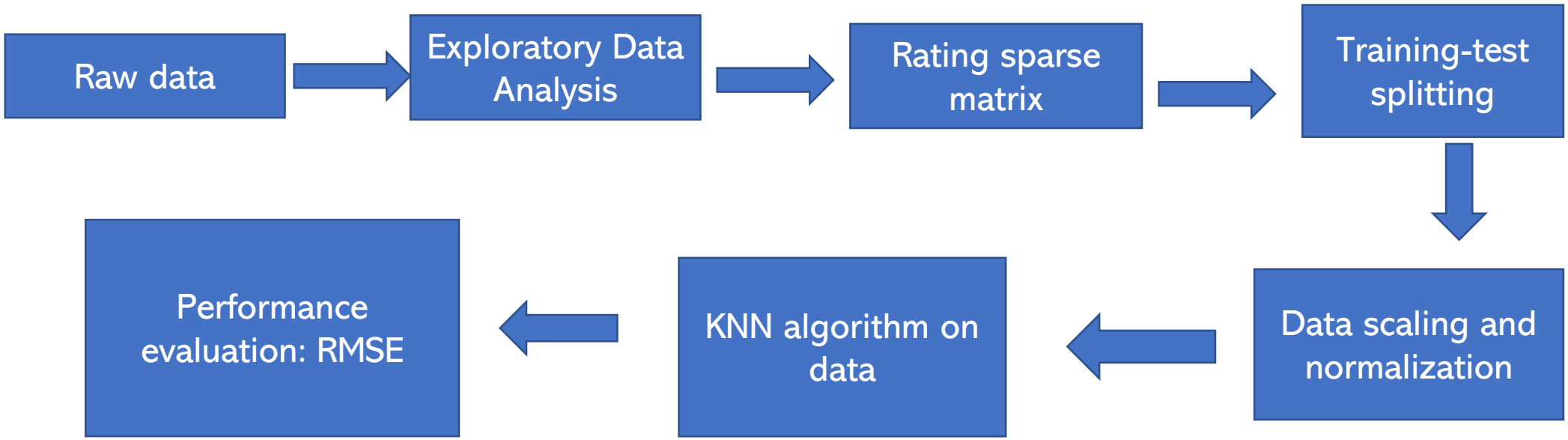
On average, 3 new courses have been recommended to each user in the Data set

	TITLE
0	data science hands on with open source tools
1	data science methodology
2	deep learning 101
3	big data 101
4	machine learning with python
5	introduction to data science
6	hadoop 101
7	python for data science
8	data visualization with python
9	r for data science

Collaborative-filtering Recommender System using Supervised Learning

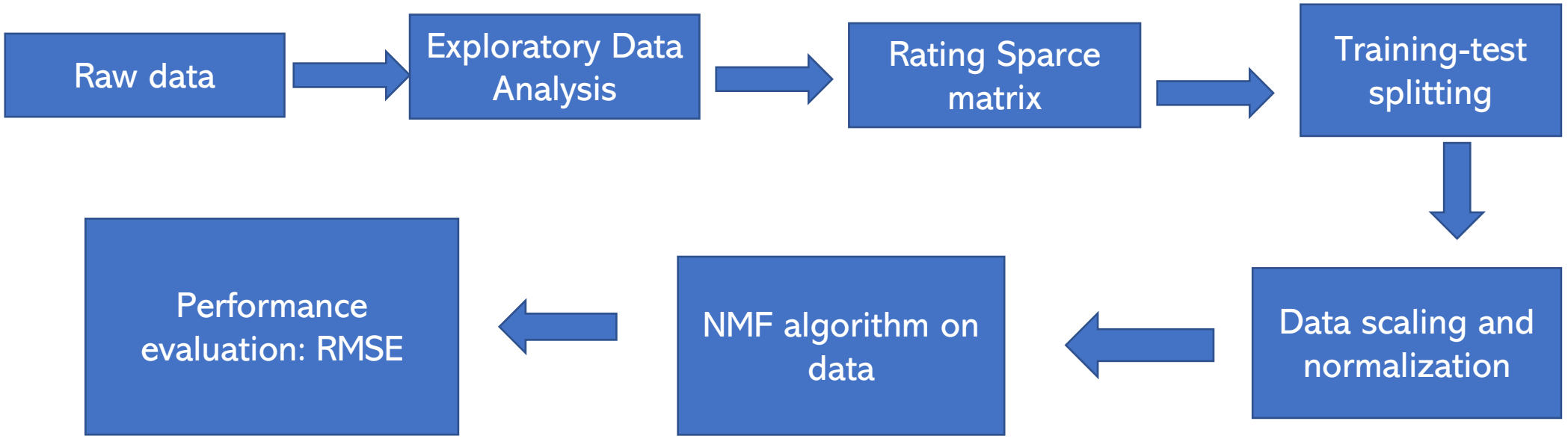


Flowchart of KNN based recommender system



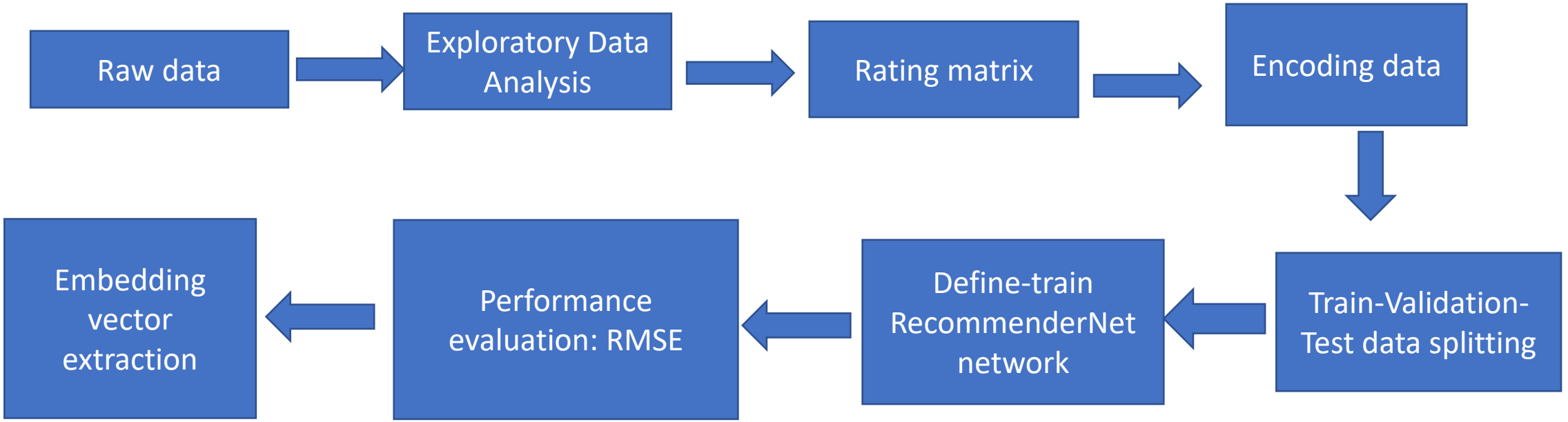
The input is raw data. After an EDA stage, the rating matrix in sparse form was built. This matrix was then used as the input data for KNN algorithm. Before the training process, the data was split in train test sets and standardized. The algorithm was fitted using the training set and then, performance was evaluated using the test set.

Flowchart of NMF based recommender system



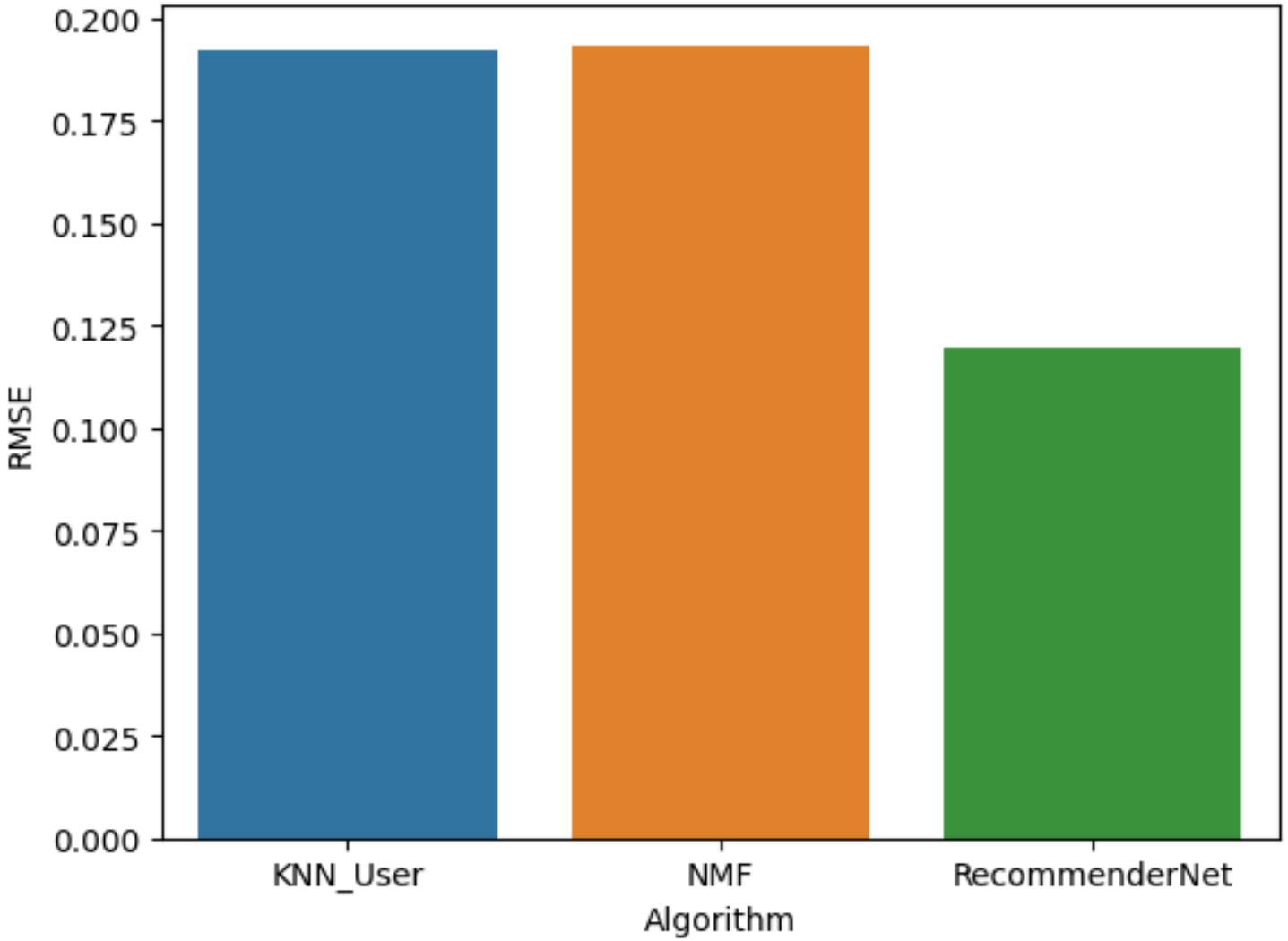
The input is raw data. After an EDA stage, the Rating matrix was built, this in sparse format. This matrix was then used as the input data for NMF algorithm from SciKit Learn. Before the training process, the data was split in train test sets and standardized. The algorithm was fitted using the training set and then, performance was evaluated using the test set.

Flowchart of Neural Network Embedding based recommender system



The input is raw data. After an EDA stage, the Rating matrix was built. Then, the data was encoded an split in train, validation, and test sets. This data was then used to train the previously defined RecommenderNet network. Once the model was built, performance was evaluated using the RMSE metric. Finally, the embedding vectors were extracted.

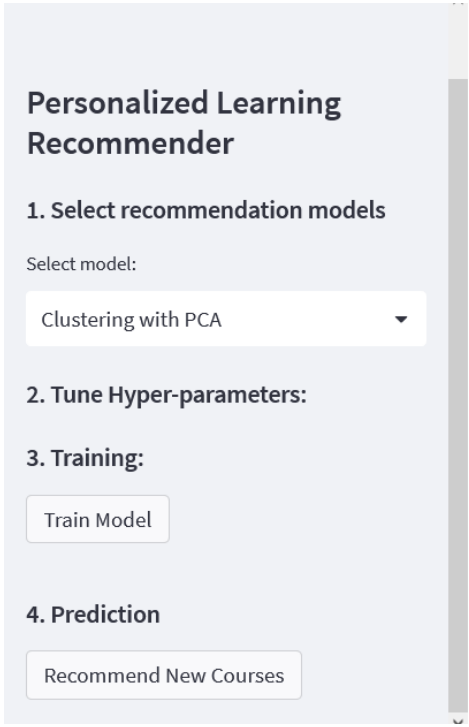
Compare the performance of collaborative-filtering models

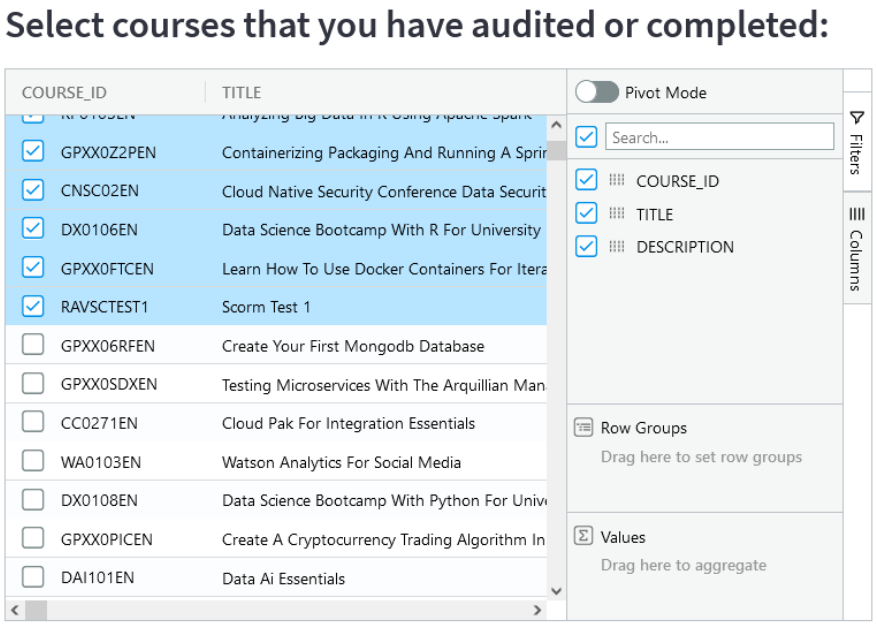


It is clear the RecommenderNet algorithm, which is a neural network, has the lowest rmse error. Therefore this is the best performing algorithm.

Optional: Build a course recommender system app with Streamlit

Streamlit app screenshot1





Streamlit app screenshot2

Your courses:

	COURSE_ID	TITLE
0	ML0201EN	Robots Are Coming Build Iot Apps With Watson Swift And Node Red
1	ML0122EN	Accelerating Deep Learning With Gpu
2	GPXX0ZG0EN	Consuming Restful Services Using The Reactive Jax Rs Client
3	RP0105EN	Analyzing Big Data In R Using Apache Spark
4	GPXX0Z2PEN	Containerizing Packaging And Running A Spring Boot Application
5	CNSC02EN	Cloud Native Security Conference Data Security
6	DX0106EN	Data Science Bootcamp With R For University Proffesors
7	GPXX0FTCEN	Learn How To Use Docker Containers For Iterative Development
8	RAVSCTEST1	Scorm Test 1

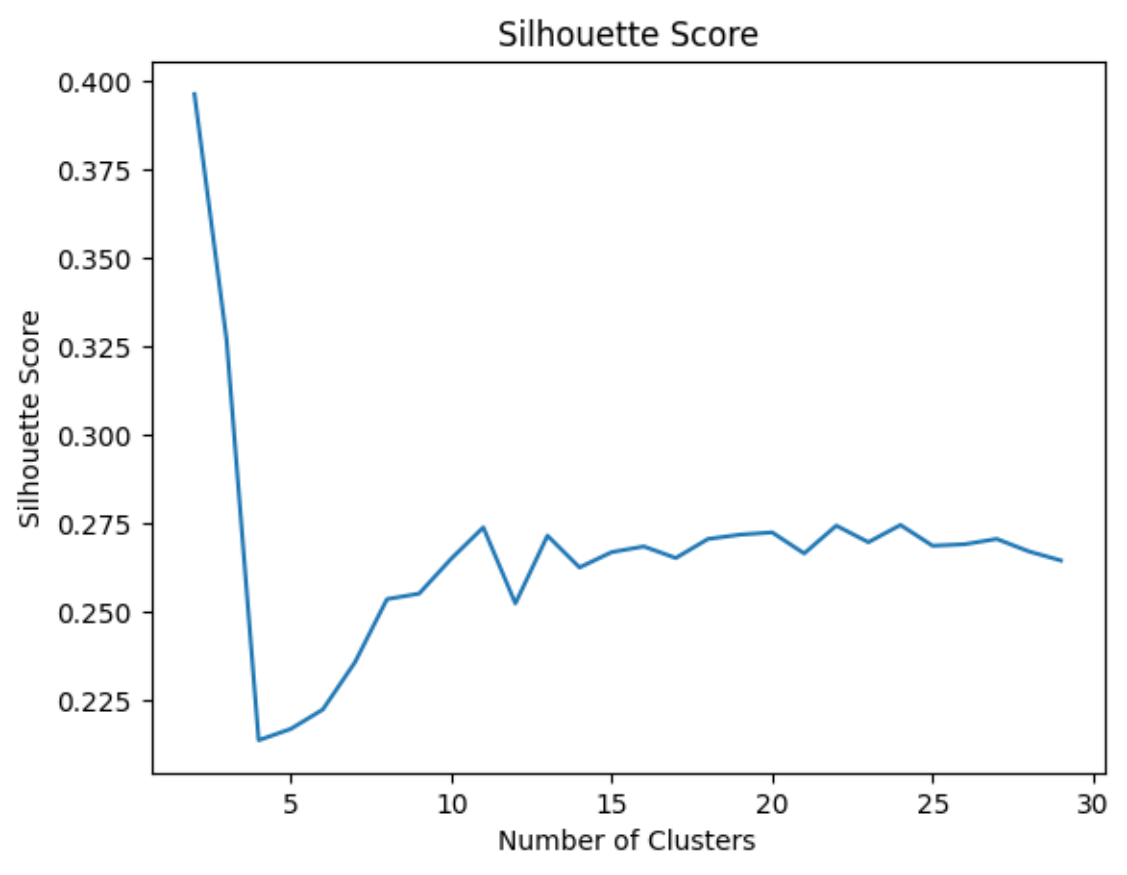
Recommendations generated!

Conclusions

- Most courses are data related ones. This explains why they are the popular ones.
- For content-based unsupervised algorithms, The top ten most recommended courses vary considerably from algorithm to algorithm. User-profile recommender system recommends more big data realated courses. Similarity recommender system suggests more text analytics courses, and the clustering algorithm recommends more data science courses.
- Evaluation of Unsupervised Algorithms, content-based, is difficult due to the lack of a performance metric. Selection of the right algorithm cold be made evaluating the recommendation outputs . Also, the number of recommended courses can be fine tuned by changing the hyperparameters for each algorithm.
- The best model for collaborative Filtering Recommender System is a neural network: RecommenderNet. The evaluation of supervised algorithms is simpler thanks to performance metrics being readily available.
- RecommenderNet is the algorithm that is going to be selected for the next stage of the project.
- Further work in fine tuning hyper-parameters is advised for future work.

Appendix

- To have a second metric to choose the right number of clusters using K-Means algorithm, the silhouette score was also evaluated:



Ignoring low and high values for number of clusters, where the score have a high value regardless of performance, the optimal value is around 11. This is for the case of using all features, without dimensionality reduction

Appendix

The link to the github repo for this project is at:

- https://github.com/taneja80/IBM-Machine-LearningAsset 3