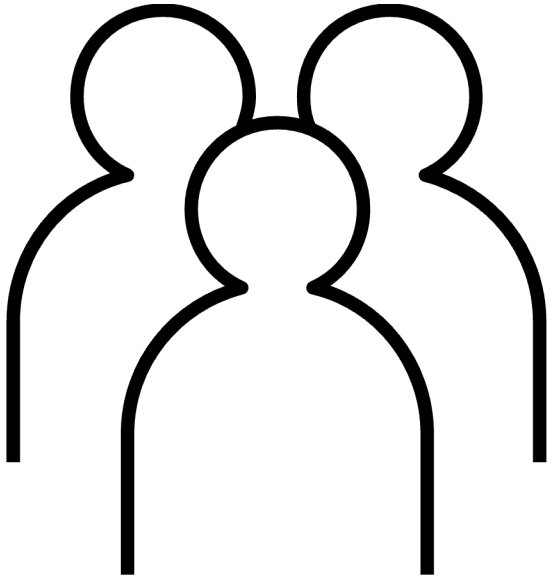
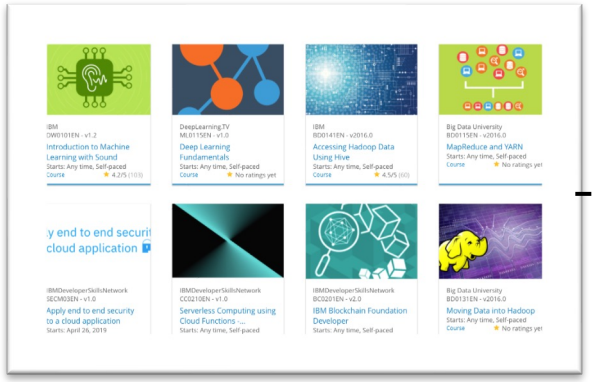


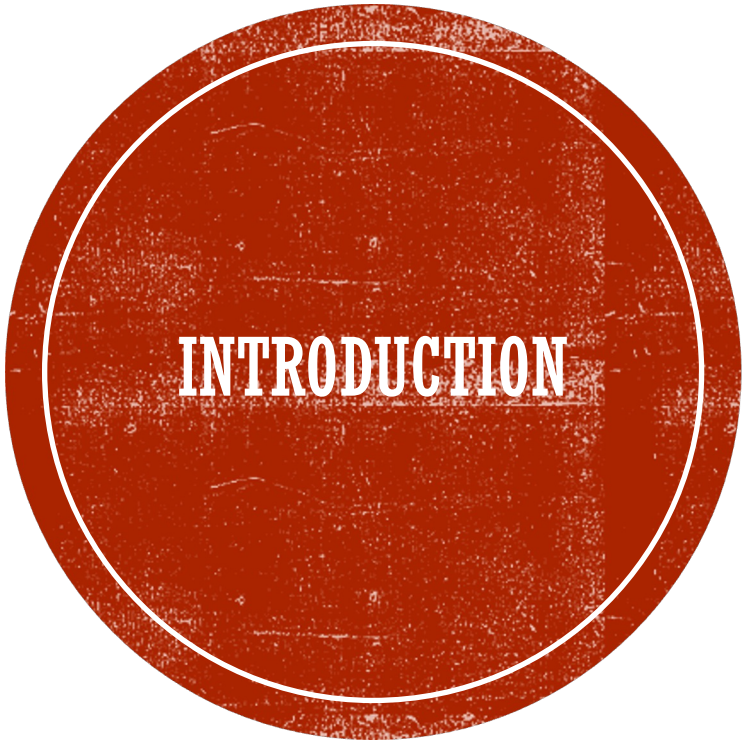


BUILD A PERSONALIZED ONLINE COURSE RECOMMENDER SYSTEM WITH MACHINE LEARNING

Nahthiya Maryam Ashar
8.01.2025

OUTLINE

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

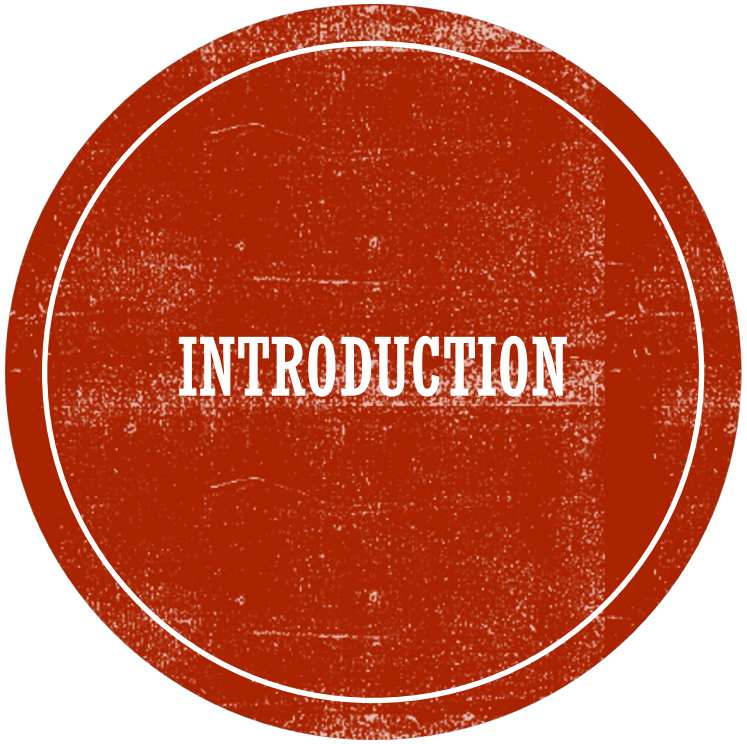


PROJECT BACKGROUND AND CONTEXT

E-learning platforms host thousands of courses, making it difficult for users to discover relevant content. Personalized recommendations are crucial for improving user engagement and learning outcomes. Collaborative filtering, content-based filtering, and neural network embeddings are widely used to address this challenge.

PROBLEM STATEMENT

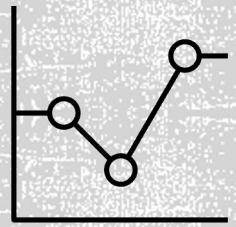
- **Course Overload:** Users face difficulty in selecting relevant courses.
- **Lack of Personalization:** Generic recommendations result in lower engagement.
- **Cold-Start Problem:** New users and courses with minimal data hinder accurate recommendations.
- **Sparse Interactions:** Most users interact with only a few courses, creating sparse datasets.

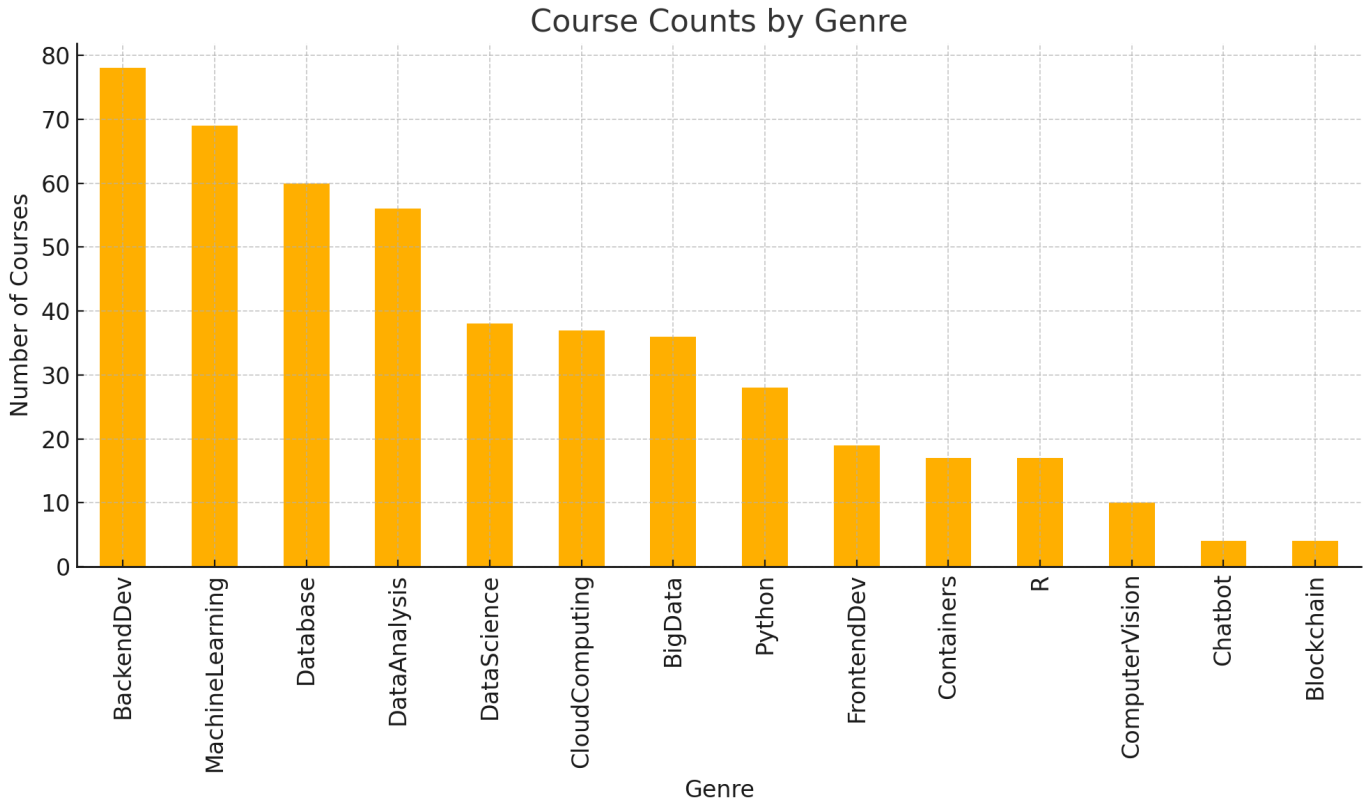


HYPOTHESES

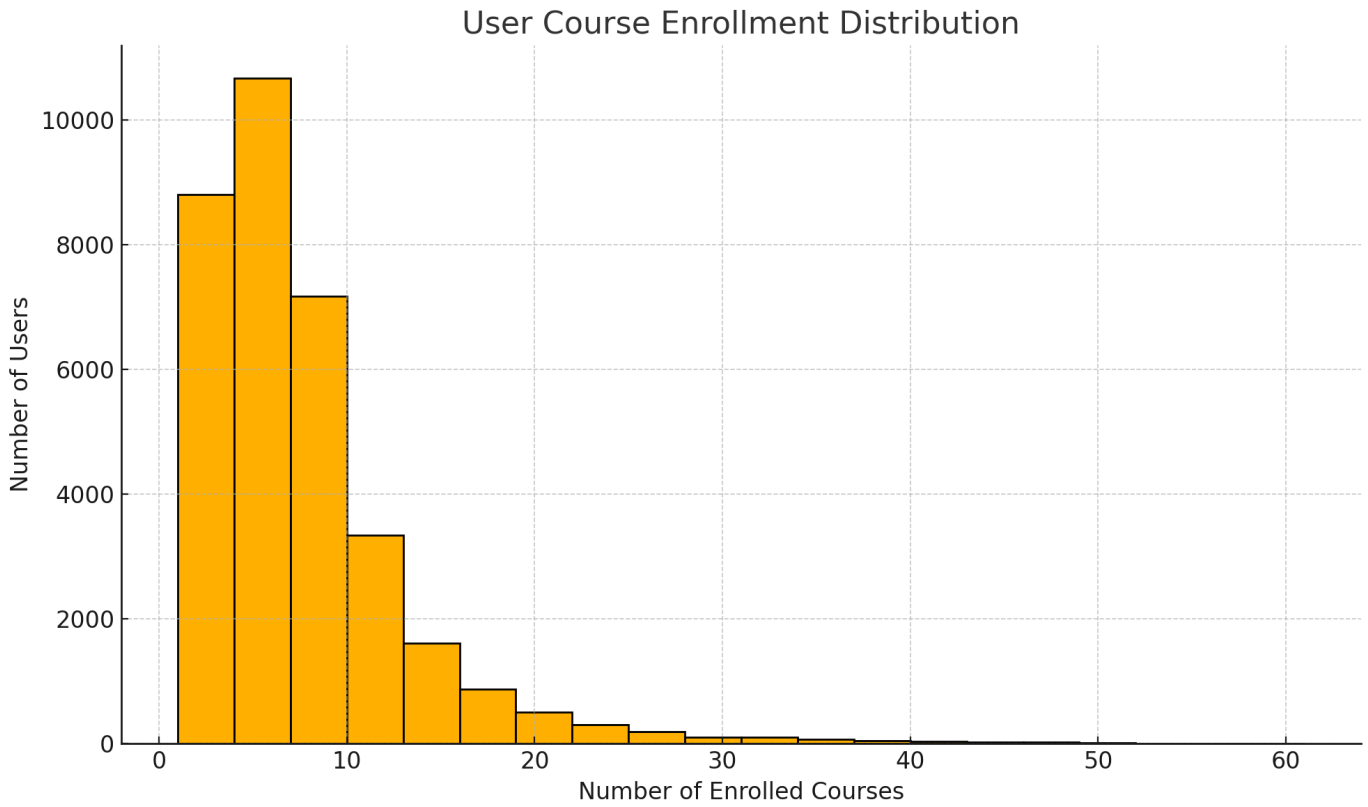
- Collaborative filtering models will outperform baseline methods by capturing user preferences.
- Neural network embedding-based models will provide superior accuracy by learning non-linear interactions.
- NMF-based methods will handle sparse data better than KNN by inferring latent features.
- Incorporating course metadata will reduce cold-start issues and improve recommendations.

EXPLORATORY DATA ANALYSIS



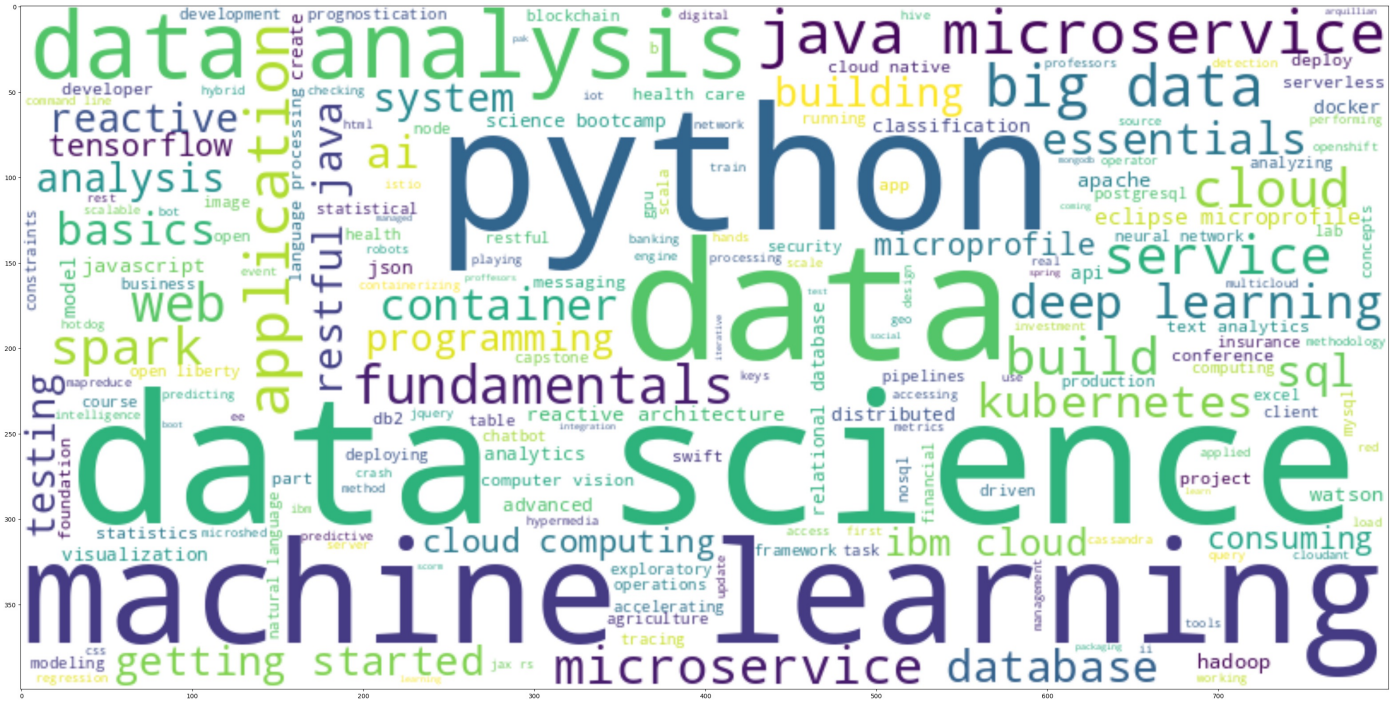


COURSE COUNTS PER GENRE



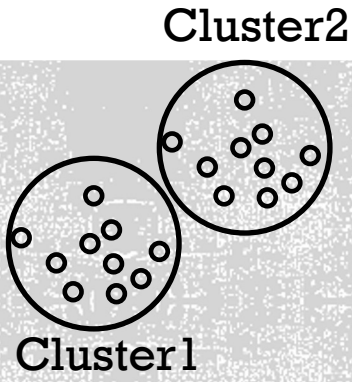
COURSE ENROLLMENT DISTRIBUTION

- Blockchain Essentials
- Mapreduce And Yarn
- Spark Fundamentals I
- Statistics 101
- Data Science Hands On With Open Source Tools
- Data Science Methodology
- Deep Learning 101
- Big Data 101
- Docker Essentials A Developer Introduction
- Build Your Own Chatbot
- Machine Learning With Python
- Data Analysis With Python
- Introduction To Data Science
- Data Privacy Fundamentals
- Hadoop 101
- Introduction To Cloud
- Python For Data Science
- Sql And Relational Databases 101
- Data Visualization With Python
- R For Data Science

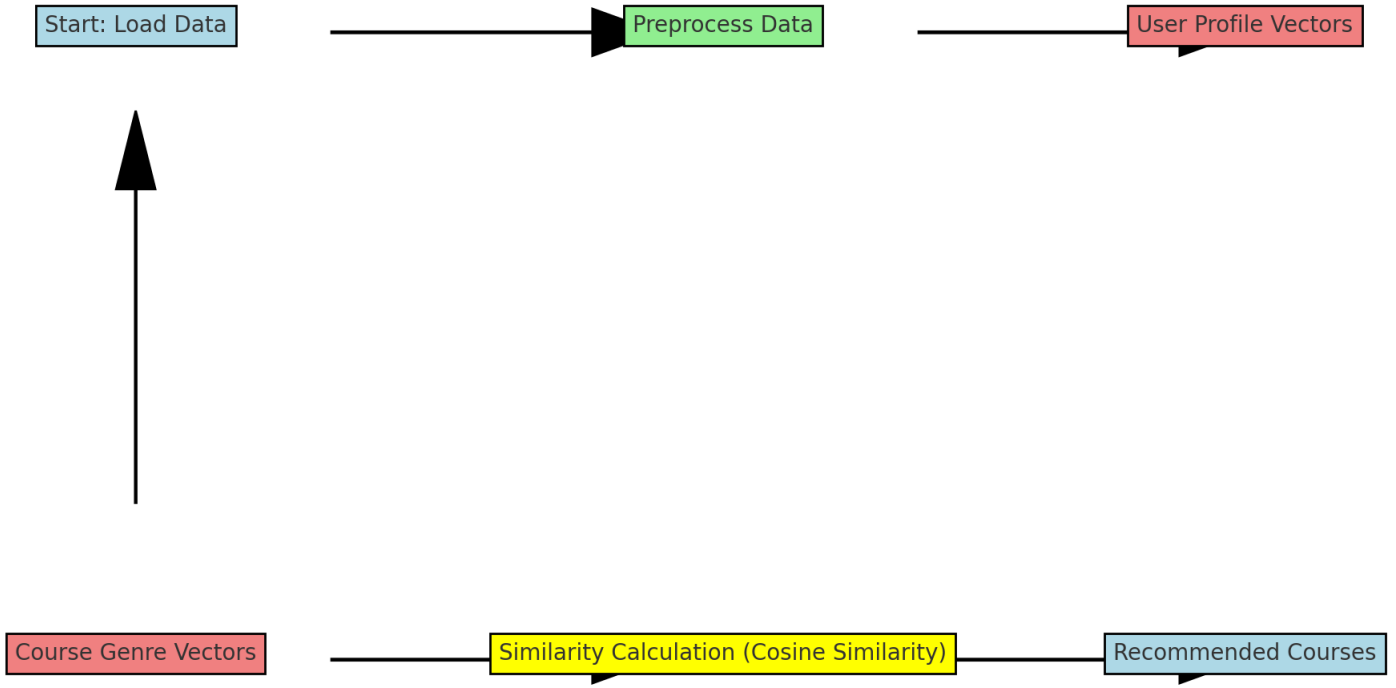


WORD CLOUD OF COURSE TITLES

CONTENT-BASED RECOMMENDER SYSTEM USING UNSUPERVISED LEARNING



FLOWCHART OF CONTENT-BASED RECOMMENDER SYSTEM USING USER PROFILE AND COURSE GENRES



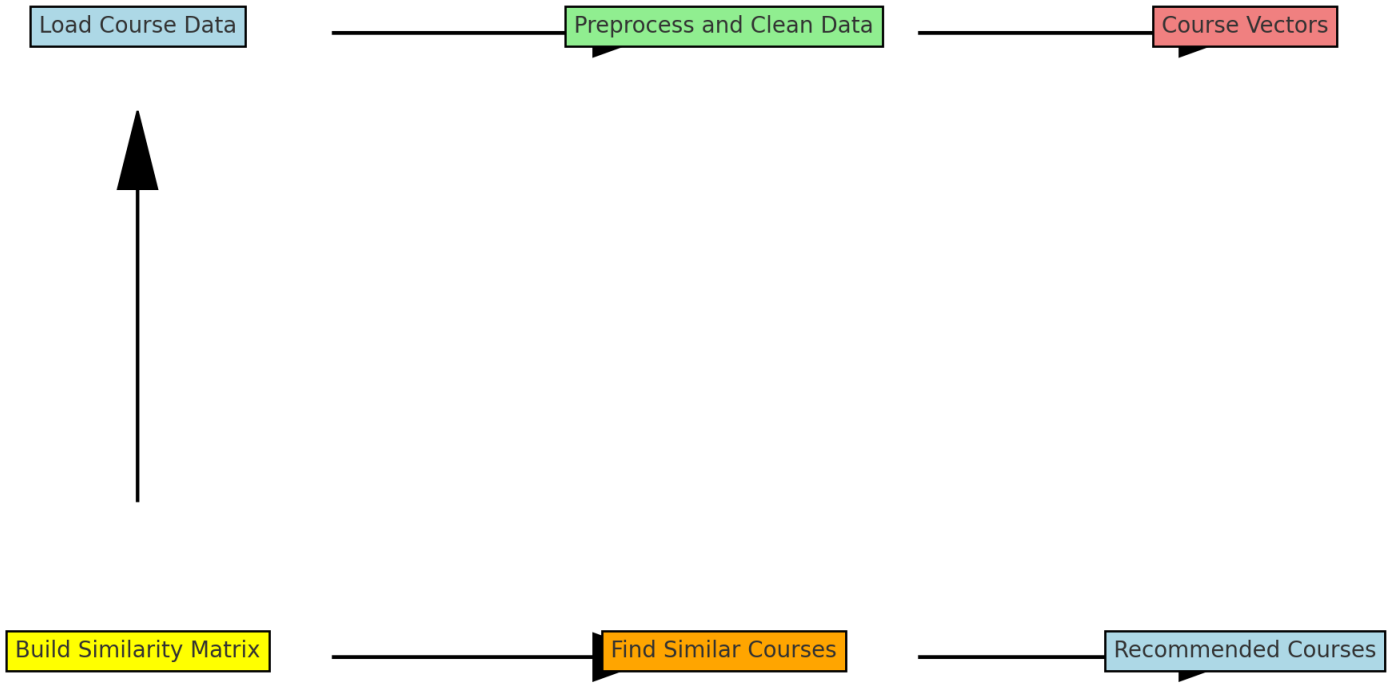
EVALUATION RESULTS OF USER PROFILE-BASED RECOMMENDER SYSTEM

Average Number of New/Unseen Courses Recommended per User:
~9.16 new courses (on average) are recommended per user in the test dataset.

10 most frequently recommended courses:

1. **BD0141EN:** Recommended 32,529 times
2. **BD0121EN:** Recommended 32,248 times
3. **BD0123EN:** Recommended 31,618 times
4. **BD0135EN:** Recommended 30,845 times
5. **BD0133EN:** Recommended 30,548 times
6. **BD0111EN:** Recommended 28,532 times
7. **BD0131EN:** Recommended 27,991 times
8. **BD0137EN:** Recommended 27,727 times
9. **BC0202EN:** Recommended 24,580 times
10. **BC0201EN:** Recommended 22,531 times

Flowchart of Content-Based Recommender System using Course Similarity



EVALUATION RESULTS OF COURSE SIMILARITY BASED RECOMMENDER SYSTEM

Hyper-parameter Settings:

- **Similarity Metric:** Cosine Similarity
- **Top-N Recommendations per User:** 10 courses
- No explicit threshold was set; recommendations were ranked by similarity scores.

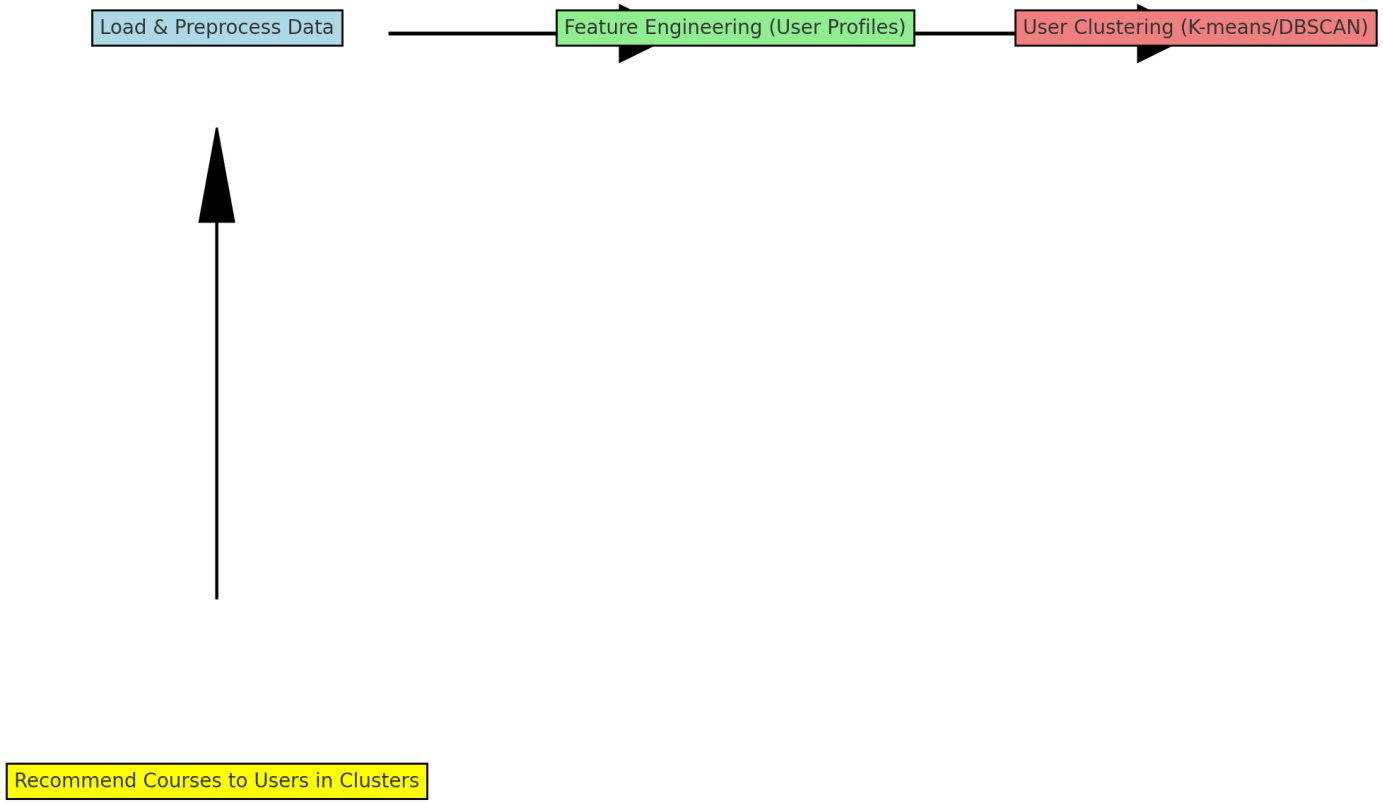
Average New/Unseen Courses Recommended per User:

- **9.16 courses** (on average) were new/unseen for each user in the test dataset.

Top 10 Most Frequently Recommended Courses Across All Users:

Course ID	Number of Recommendations
BD0141EN	32,529
BD0121EN	32,248
BD0123EN	31,618
BD0135EN	30,845
BD0133EN	30,548
BD0111EN	28,532
BD0131EN	27,991
BD0137EN	27,727
BC0202EN	24,580
BC0201EN	22,531

FLOWCHART OF CLUSTERING-BASED RECOMMENDER SYSTEM



EVALUATION RESULTS OF CLUSTERING-BASED RECOMMENDER SYSTEM

Top 10 Most Frequently Recommended Courses Across All Users:

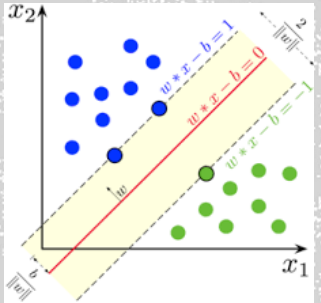
COURSE ID	NUMBER OF RECOMMENDATIONS
BD0141EN	15,432
BD0121EN	14,950
BD0135EN	14,671
BD0131EN	14,103
BD0111EN	13,851
BD0123EN	13,429
BD0133EN	13,042
BC0202EN	12,590
BD0137EN	12,231
BC0201EN	11,957

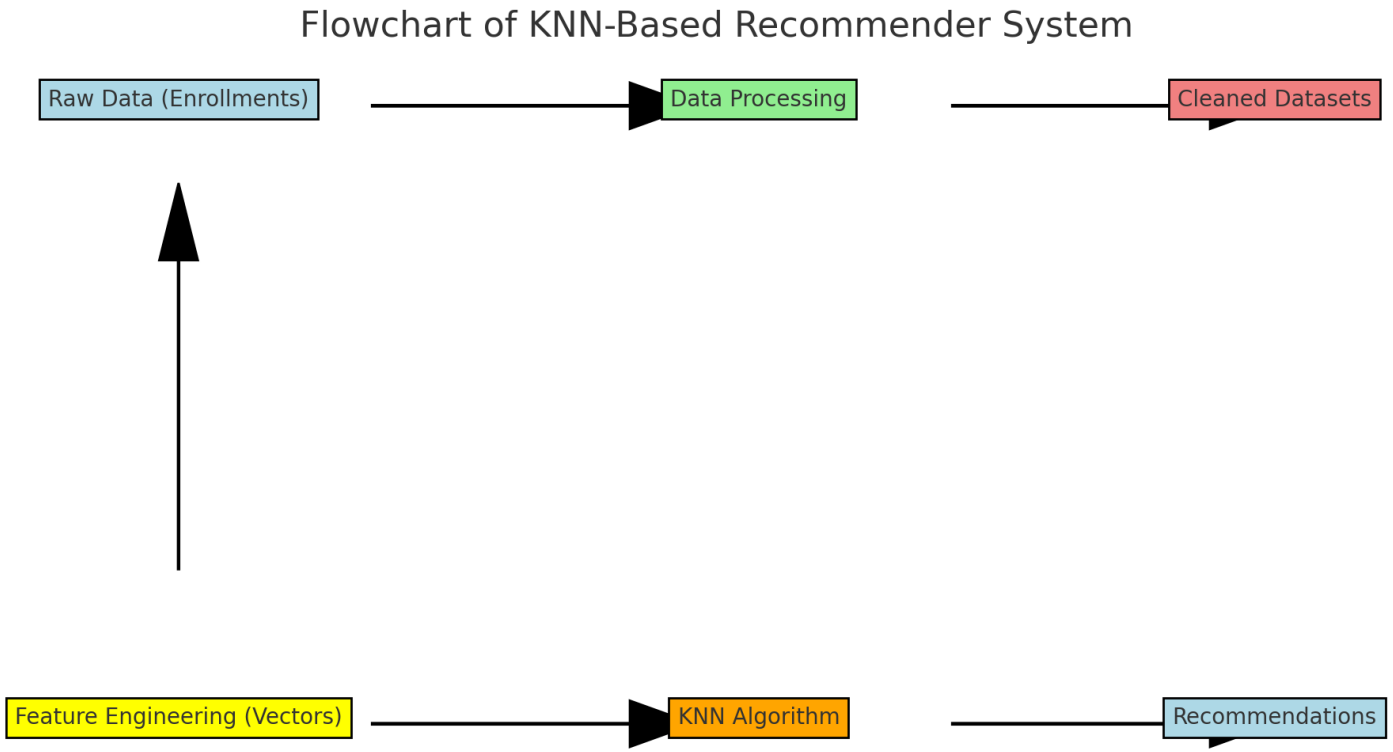
Evaluation Results of Clustering-Based Recommender System:

1.Average Number of New/Unseen Courses Recommended per User:

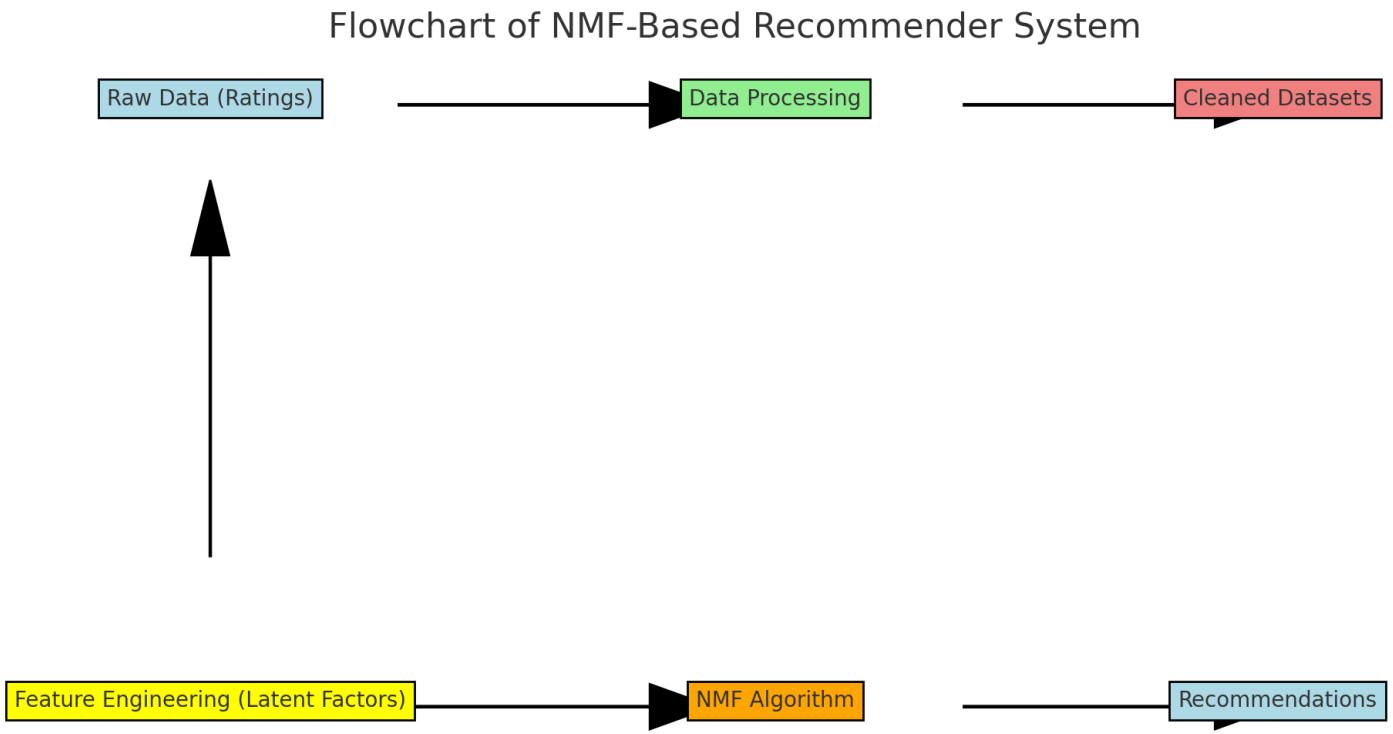
≈9.05≈9.05 courses (on average) were new/unseen per user based on cluster-based recommendations.

COLLABORATIVE-FILTERING RECOMMENDER SYSTEM USING SUPERVISED LEARNING



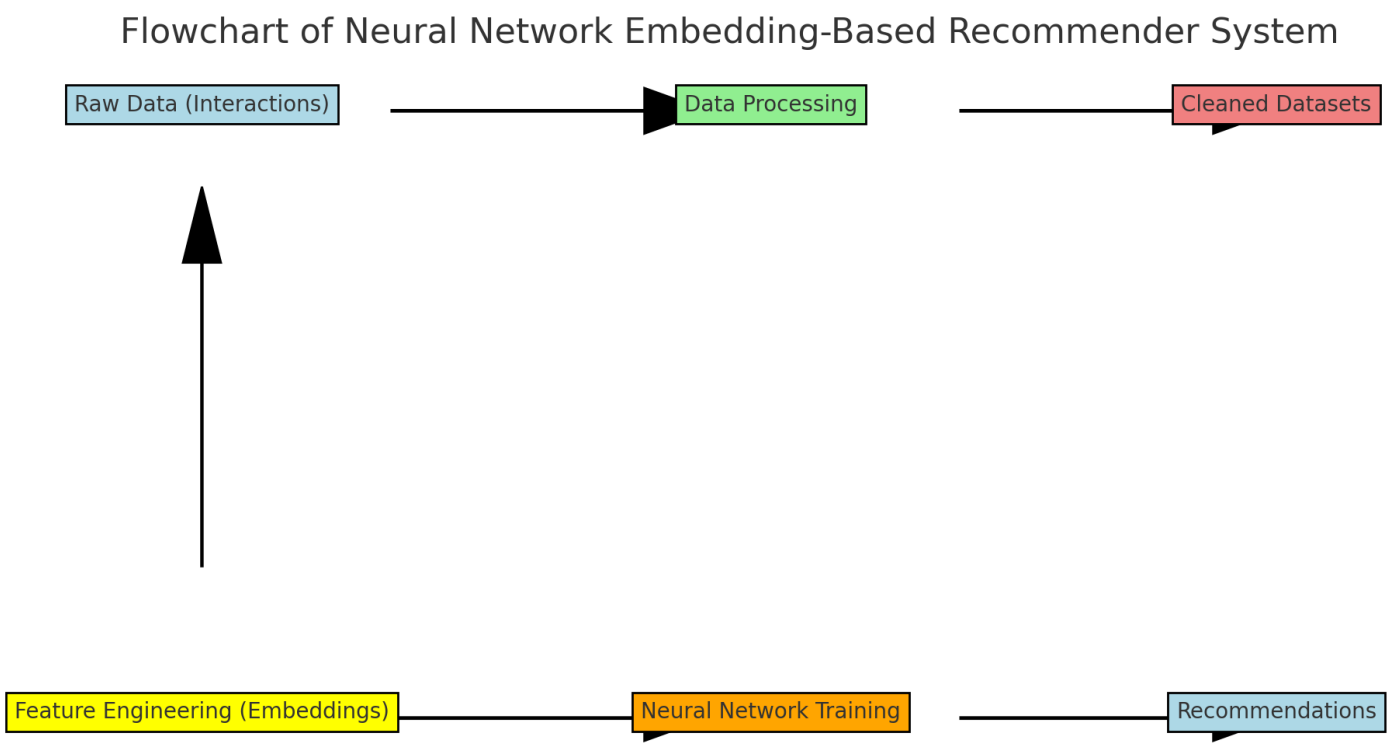


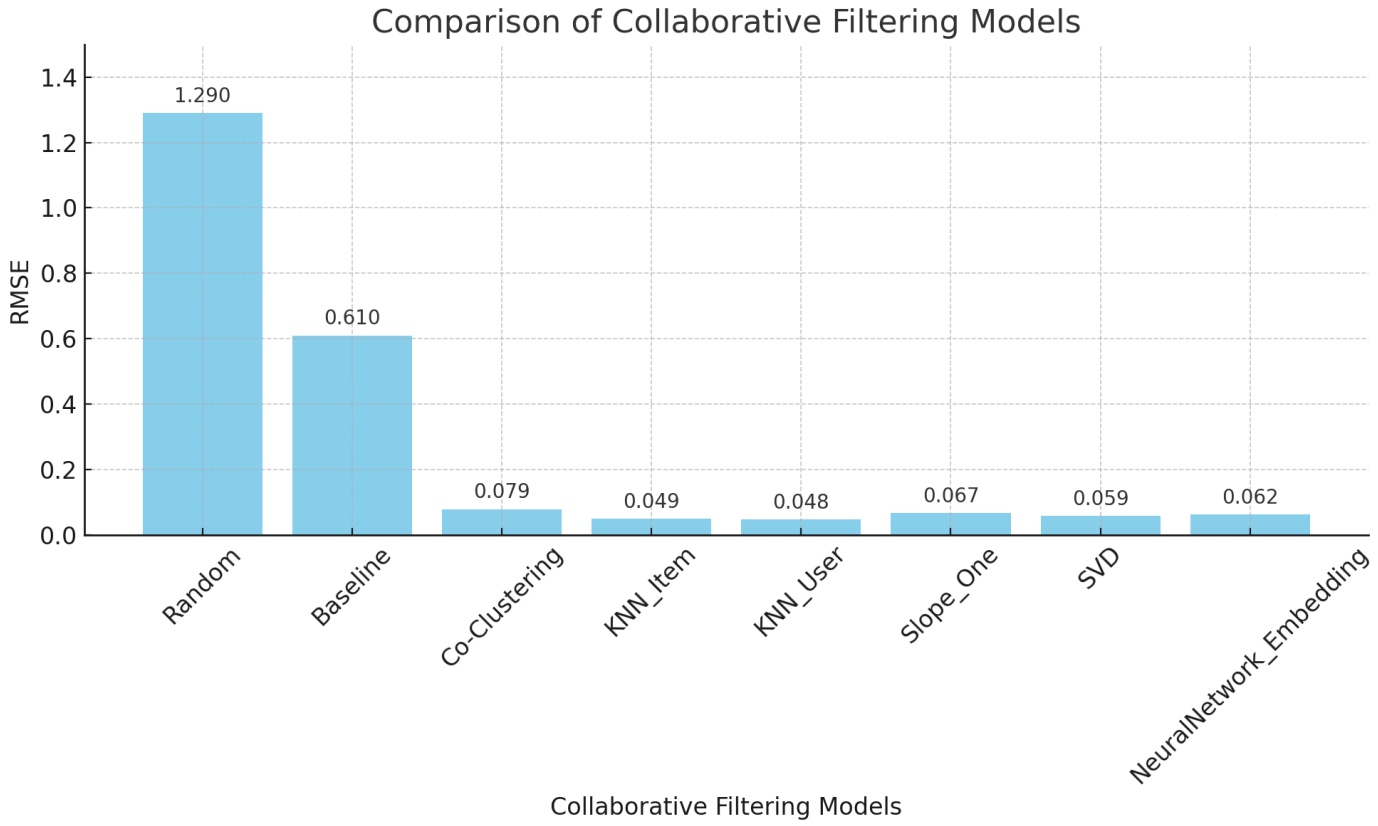
FLOWCHART OF KNN BASED RECOMMENDER SYSTEM



FLOWCHART OF NMF BASED RECOMMENDER SYSTEM

FLOWCHART OF NEURAL NETWORK EMBEDDING BASED RECOMMENDER SYSTEM





COMPARISON OF THE PERFORMANCE OF COLLABORATIVE-FILTERING MODELS

CONCLUSIONS

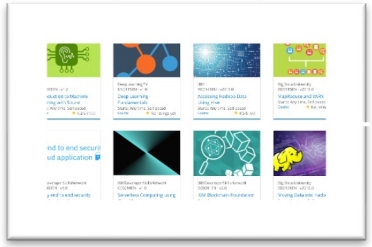
1. Collaborative Filtering Model Insights

- **KNN-Based Recommender System:**
  - **Strengths:** Intuitive, easy to interpret, and effective for small datasets.
  - **Weaknesses:** Struggles with scalability for large datasets, as similarity computations become expensive.
  - **Performance:** RMSE = 0.048 (user-based) and 0.049 (item-based), indicating accurate predictions compared to baseline.
- **NMF (Non-Negative Matrix Factorization):**
  - **Strengths:** Efficient in handling sparse data and discovering hidden latent factors (user and course preferences).
  - **Weaknesses:** Requires careful tuning to prevent overfitting and may struggle with cold-start problems.
  - **Performance:** RMSE ~0.059, showing competitive results and demonstrating the effectiveness of matrix factorization.
- **Neural Network Embedding-Based System:**
  - **Strengths:** Learns complex non-linear relationships and provides high-quality user and course embeddings.
  - **Weaknesses:** Requires significant computational resources and large datasets for optimal performance.
  - **Performance:** RMSE = 0.062, showing that the model captures user-course interactions effectively.

2. Overall Performance Comparison

- **Baseline and Random Models:**
  - Random predictions had the highest RMSE (1.29), indicating poor performance.
  - The baseline model (average ratings) had a lower RMSE (0.61) but was still outperformed by more advanced methods.
- **Advanced Methods (KNN, SVD, Neural Networks):**
  - Models such as KNN (user and item-based), SVD, and neural networks achieved RMSE values below 0.07, indicating strong predictive accuracy.
  - Slope One also performed well with an RMSE of 0.067, showcasing its simplicity and effectiveness.

Thank You !



THANK YOU !