University Degree Recommendation System

Islam Ehab Anwar

Faculty of Computer Science

University of Prince Edward Island, Cairo Campus

June 2023

A thesis submitted to University of Prince Edward Island, Cairo Campus in partial fulfillment of the requirements of the degree of Bachelor of Science

Acknowledgements

Words cannot express my appreciation and gratitude towards my professor, project adviser and Dean of the faculty of computer science Dr. Ahmed Elsheikh for his unwavering support and guidance. Additionally, this would not have been possible without my research assistants Heba Abdelkader and Ola Galal whom without their guidance and experience I would not have been able to achieve this. Furthermore, I would like to thank the university for providing the necessary data for this project.

I am also grateful to Dr. Amal Mohamed, professor at the faculty of computer science for her unconditional moral support and significant impact in inspiring me to pursue this field of study.

Lastly, I would be remiss in not mentioning my family, especially my parents and brother. Their belief and encouragement has kept my spirits and motivation high during this trying process. I would never have reached this stage without them.

Incomplete

Abstract

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Problem Statement

In recent years, the accelerating advancement of internet technology has left us with an abundant amount of information available. Various sources of data are now publicly accessible to users such as news content, e-commerce websites, as well as digital entertainment [1] [2]. A significant challenge with having an overload of data available is ensuring the quality of the information provided to users from their corresponding searches as well as results that match their interests and preferences. Preferences can vary depending on the background such as educational or entertainment purposes.

As a result, various recent researches have attempted to create tailor-made solutions to provide users with a series of recommendations catering to specific user preferences. However, there is a lack of research performed in the educational field pertaining to the choice of degree or major in which a student enrolls at a university. This provides suitable motivation to pursue this field of study in order to understand what factors affect the degree chosen by a prospective student.

A student degree recommendation system is a problem that has not been tackled by many recent publishers. With the vast diversity of university programs available to prospective students, many are left wondering what their real passion is or if a particular degree is the one for them. A recommendation algorithm is suitable to mitigate this issue, by analyzing student academic performance

during the latter end of high school as well as other factors such as familial status, personal status, and other daily lifestyle choices that could shape the overall performance of a student and impact the recommended degrees. Furthermore, we dive into the concept of interpretability of our proposed recommendation system and how that can build model trust and reliability by understanding why decisions were made in the local level of our data.

1.2 Thesis Outline

The rest of the thesis is as follows: Chapter 2 describes background research regarding terminologies and definitions, a literature review regarding supervised as well as deep learning techniques used with recommendation systems. Chapter 3 discusses the methodology to develop an effective student degree recommendation system. Chapter 4 offers insight into the results and discussion. Chapter 5 provides the conclusion and future work.

Chapter 2

Background

This background chapter provides the necessary background and related research to this project. This project applies a variety of approaches to create a viable recommendation system, the background section will be divided into five sections: Terminologies and Definitions, Content-Based Filtering, Collaborative Filtering Approaches, Knowledge-Based Approaches, and Hybrid Approaches.

2.1 Terminologies and Definitions

2.1.1 Recommendation Systems

Recommendation systems (RS) are a sophisticated machine learning algorithms that recommend a set of items or a collection of data to a user based on their specific preferences [1]. The concept of recommendation systems began in the 1990s to help people decide on what products to purchase [3]. Today, there is a plethora of recommender systems that are created and modified to cater to a vast range of fields such as e-commerce, digital entertainment, as well as the educational industry [1].

With the evolution of recommendation systems, they were subcategorized into various approaches to provide different kinds of recommendations based on different factors. One of the most popular recommendation approaches is collaborative filtering (CF), which works by repre-

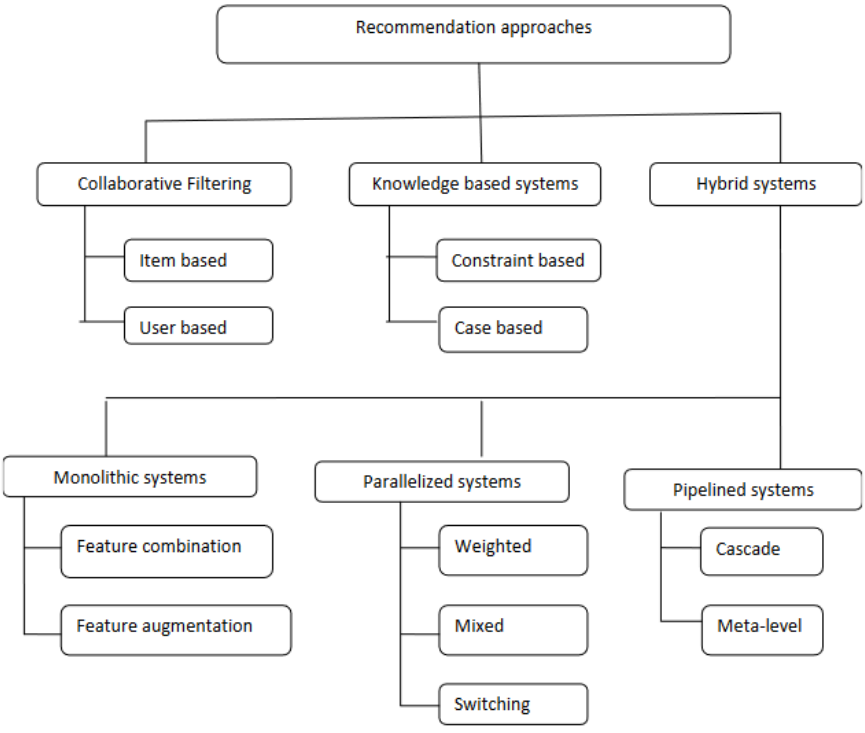

Figure 2.1: Classification of Recommendation Approaches

senting unique users and items with unique representation vectors. These representation vectors' interactions can be analyzed using various machine learning approaches such as neural networks (NN) [1]. However, a disadvantage of CF recommendation is suffering from data sparsity and the cold start issue, hence other techniques have been developed to tackle these problems [2].

Another common recommendation approach are knowledge-based systems. Knowledge-based recommendation systems are heterogeneous graphs, whose nodes represent the unique users, and the edges are the relations between the user and the corresponding items. Knowledge-based systems can be designed in a variety of ways, including embedding-based methods, path-based methods, as well as unified methods which combine the former techniques to build the latter [2].

2.1.2 Supervised Learning Techniques

Supervised learning is a subcategory within machine learning. Supervised learning is defined by labeled datasets that can be used to train algorithms to accurately classify data or predict out-

comes [4]. Training data is applied to a supervised algorithm to generate an effective classifier that can correctly classify data as well as form reasonable predictions. Many common algorithms are housed within supervised learning that operate on categorical as well as continuous data such as decision tree (DT) and random forest (RF).

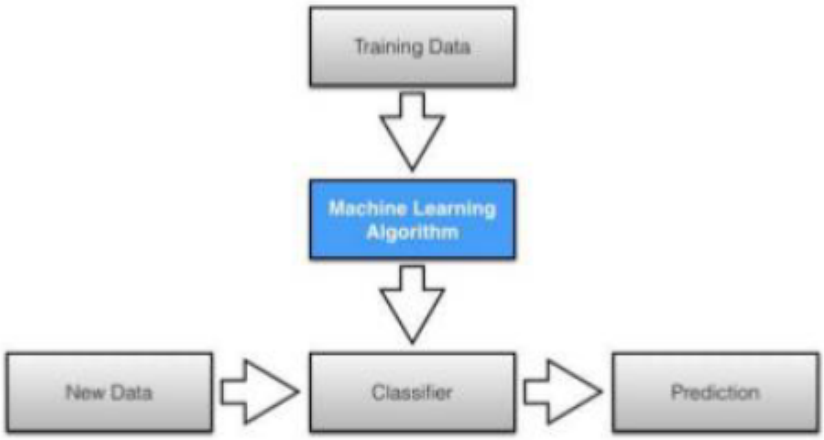

Figure 2.2: Supervised Learning Process

Decision trees are one of the simplest supervised learning algorithms due to their interpretability and simple execution compared to other algorithms [5]. The algorithm works by performing splits at each node in the tree according to a specific condition relating to a feature in the dataset. This is repeated until the tree cannot perform any more even splits and is providing little information.

Decision trees are very useful in classification problems to breakdown the labeled data through effective splits. However, they do have disadvantages such as the potential for overfitting, which is when the algorithm memorizes the data instead of generalization. Furthermore, decision trees struggle to handle large data since a single tree will result in many node splits and lead to overfitting [6].

Another method of supervised learning is random forest (RF). Random forest is an ensemble supervised learning algorithm that uses a collection of decision trees to perform its classification or regression. It's basic functionality relies on a method called bagging, in which subsets of features

of the training data are taken and used for training with the decision trees. The final classification or output is based on a majority vote between all the trees in the forest [7].

An advantage of random forests is that they are prone to overfitting due to their smoothing effect as we increase the number of decision trees in the forest. However, they are considered slow for training data and tend to generate bias when dealing with categorical variables [7].

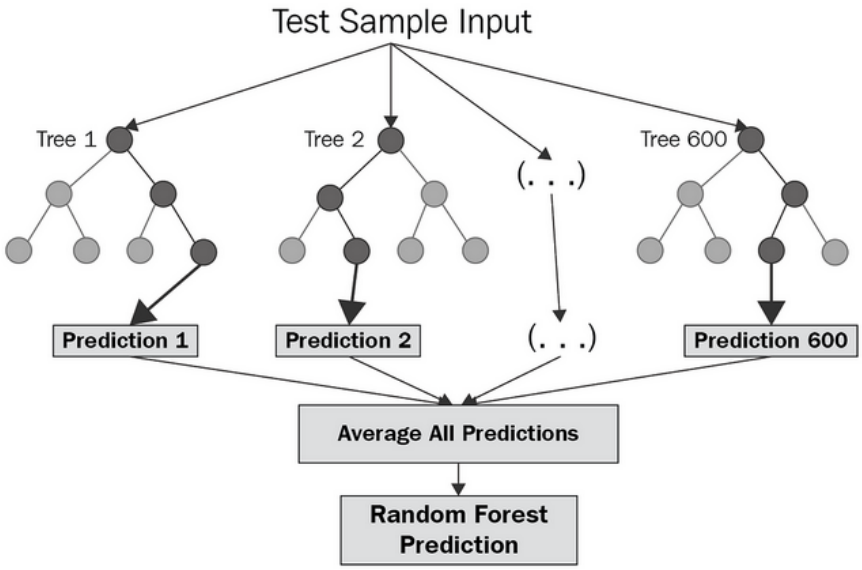

Figure 2.3: Random Forest Procedure

2.1.3 Deep Learning

Deep learning (DL) algorithms are another category of machine learning algorithms that aim to find multiple patterns within data using high level architecture. Deep learning has been used extensively in tasks that require high pattern recognition such as computer vision, natural language processing (NLP), as well as other multimedia-centered tasks [8].

Among the most simple and common neural networks is a multi-layer perceptron (MLP), which works by using back-propagation to update the weights of our neurons during each iteration. This is repeated until a convergence threshold is met or we reach the maximum number of iterations. Multi-layer perceptrons are used in many applications such as prediction and pattern classification [9].

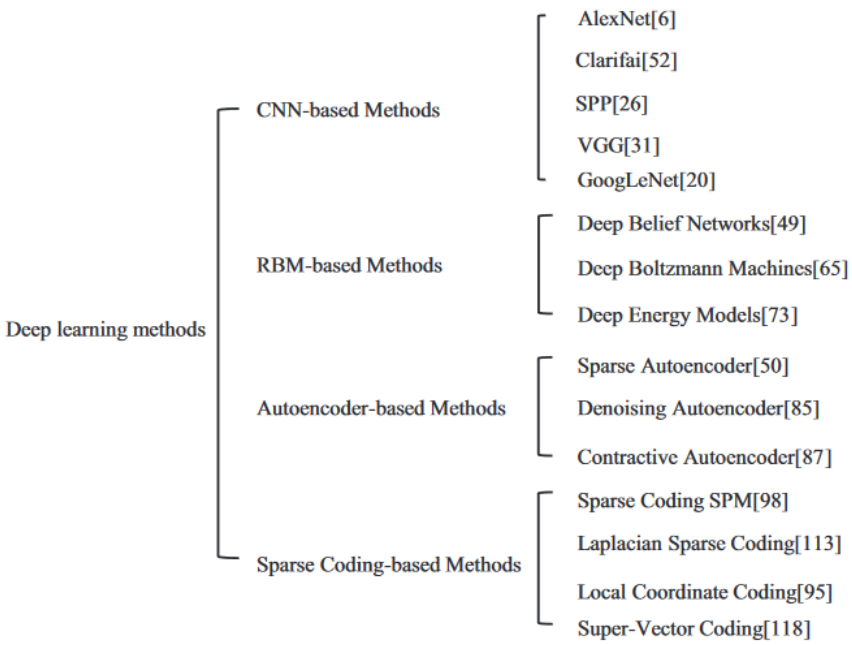

Figure 2.4: A categorization of the deep learning methods and their representative works.

However, one of the most impactful limitations of deep learning algorithms is their hunger for data. Small datasets will render deep neural networks ineffective as they require a large amount of data to continue the learning process. Another problem to consider is poor scalability, which is one of the many reasons supervised learning algorithms are chosen for most tasks that require a scalable model [10].

2.1.4 Model Interpretability

Recommendation systems provide personalized recommendations to each user. Therefore, interpreting why a certain output was given is a crucial aspect of recommendation systems. This is where model interpretability is introduced. Some algorithms that are used in recommendation systems such as random forest are classified as black box models that do not provide information regarding their internal functions and why a certain output was produced [11].

Interpretability is the degree to which a person can understand why a decision was made. It is among one of the most important parts in predictive modelling as understanding why a decision was made by a model implies model reliability and trust. Interpretability has a large scope that

pertains to algorithm transparency as well as global or local interpretations of our data [11]. Consequently, interpretability in a recommendation system is vitally important to invoke trust from the user that the predicted output was reliable and provided explanations as to why a certain decision was made by the system.

2.2 Content-Based Filtering

Content-based recommendation systems are a popular technique used to gather information about user preferences. They are based on past preferences of users and recommendations are suggested with similar items of similar characteristics [12]. Chen et al. [13] used correlation analysis in an experiment regarding the education field in order to group certain courses. This was performed by segmenting the data into three categories based on a rule-space model using a content-based approach to optimize the learning path for each individual.

Similarly, Shu et al. [14] utilised the historical data of students to create predictions regarding the provided learning materials using a content-based algorithm. The most common learning algorithms used in this domain are fuzzy-based as well as rule-based clustering that relies on a probabilistic methodology, and similarity between neighbors. However, a disadvantage of content-based recommendation is the overreliance on past data to create predictions. It cannot overcome the cold start problem of having no data in the initial stages as well as data sparsity [12].

2.3 Collaborative Filtering

Another favored technique in recommendation is collaborative filtering, which has been used frequently in recent systems as it mitigates the drawbacks of content-based filtering as we discussed earlier [12]. Liu [15] recommended a collaborative filtering approach that focused on the influence of e-learning group behavior to improve the accuracy of predictions even in presence of data sparsity. Moreover, collaborative filtering has been used with unsupervised learning.

El-Bishouty et al. [16] utilised a k-means algorithm to extract the learning path and objects of interest for each individual learner. Collaborative filtering does improve on content-based systems, but it still comes with its own set of difficulties. It can be difficult to create relations between attributes to their respective items, which can affect recommendation accuracy. It also suffers from cold-start and scalability issues [12].

2.4 Knowledge-Based Approaches

Knowledge-based approaches provide recommendations based on how certain item features meet user needs. H. Wang et al. [1] proposed a knowledge graph convolutional network. The purpose of this architecture was to capture relationships between several items of interest to a user through data mining techniques. Some of the featured techniques involved association rules to identify relations between attributes on a graph. This was performed by sampling data from respective neighbors per entity in a particular graph and fusing the information already gained with the bias in order to calculate an accurate representation of each entity's graph relations. Three datasets were used for testing including movie, book and music recommendation datasets. Results indicated that the proposed netwrok outperformed the baseline recommendation techniques with an Area Under the Curve (AUC) of 0.9 or higher with two of the three datasets [1].

Wan and Niu [17] used a knowledge-based approach with an underlying self-organization method to propose learning objects. This improved accuracy but suffered from increased time of computations and stacking of multiple algorithms.

2.5 Hybrid Methods

Certain hybrid methods have been experimented with including the combination of content-based and collaborative filtering to counter the cold start problem [12]. Hussain et al. [18] opted to use a collection of models including an artificial neural network, decision tree, logistic regression, and support vector machine to predict the troubles students face during an online learning course.

Similarly, Karga and Satratzemi [19] used a similarity matrix to create the relations between learners and their respective learning paths according to their needs and preferences. This methodology allows both content-based and collaborative filtering methods to complement each others weaknesses while improving prediction accuracy [12].

Chapter 3

Methodology

3.1 Data Collection

3.2 Flowchart of Methodology

3.3 Explanations of Techniques

Chapter 4

Results and Discussion

Chapter 5

Conclusion and Future Work

Bibliography

[1] Hongwei Wang, Miao Zhao, Xing Xie, Wenjie Li, and Minyi Guo. Knowledge graph convolutional networks for recommender systems. In *The world wide web conference*, pages 3307–3313, 2019.

[2] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568, 2020.

[3] Richa Sharma and Rahul Singh. Evolution of recommender systems from ancient times to modern era: a survey. *Indian Journal of Science and Technology*, 9(20):1–12, 2016.

[4] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

[5] Anuja Priyam, Rahul K. Gupta, Anju Rathee, and Saurabh Kr. Srivastava. Comparative analysis of decision tree classification algorithms. 2013.

[6] Bahzad Taha Jijo and Adnan Mohsin Abdulazeez. Classification based on decision tree algorithm for machine learning. *evaluation*, 6:7, 2021.

[7] Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25, 11 2015.

[8] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S. Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016. Recent Developments on Deep Big Vision.

[9] M.W Gardner and S.R Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14):2627–2636, 1998.

[10] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroSP)*, pages 372–387, 2016.

[11] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[12] Shristi Shakya Khanal, PWC Prasad, Abeer Alsadoon, and Angelika Maag. A systematic review: machine learning based recommendation systems for e-learning. *Education and Information Technologies*, 25:2635–2664, 2020.

[13] Yung-Hui Chen, Chun-Hsiung Tseng, Ching-Lien Huang, Lawrence Y Deng, and Wei-Chun Lee. Recommendation system based on rule-space model of two-phase blue-red tree and optimized learning path with multimedia learning and cognitive assessment evaluation. *Multimedia Tools and Applications*, 76:18237–18264, 2017.

[14] Jiangbo Shu, Xiaoxuan Shen, Hai Liu, Baolin Yi, and Zhaoli Zhang. A content-based recommendation algorithm for learning resources. *Multimedia Systems*, 24(2):163–173, 2018.

[15] Xiuju Liu. A collaborative filtering recommendation algorithm based on the influence sets of e-learning group's behavior. *Cluster Computing*, 22(Suppl 2):2823–2833, 2019.

[16] Moushir M El-Bishouty, Ahmed Aldraiweesh, Uthman Alturki, Richard Tortorella, Junfeng Yang, Ting-Wen Chang, Sabine Graf, et al. Use of felder and silverman learning style model for online course design. *Educational Technology Research and Development*, 67(1):161–177, 2019.

[17] Shanshan Wan and Zhendong Niu. An e-learning recommendation approach based on the self-organization of learning resource. *Knowledge-Based Systems*, 160:71–87, 2018.

[18] Mushtaq Hussain, Wenhao Zhu, Wu Zhang, Syed Muhammad Raza Abidi, and Sadaqat Ali. Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review*, 52:381–407, 2019.

[19] Soultana Karga and Maya Satratzemi. A hybrid recommender system integrated into lams for learning designers. *Education and Information Technologies*, 23:1297–1329, 2018.