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

In this chapter, the following main topics to be discussed are as follows:

1.1 Chapter Contents

- Recommendation Systems
- Supervised Learning Techniques
- Deep Learning Applications and Challenges
- Model Interpretability

Then the motivation and problem statement for this thesis will be stated. The thesis outline encapsulates the following chapters to come.

1.2 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 [2]. Today, there is a plethora of recommender systems that are created and modified to cater to a

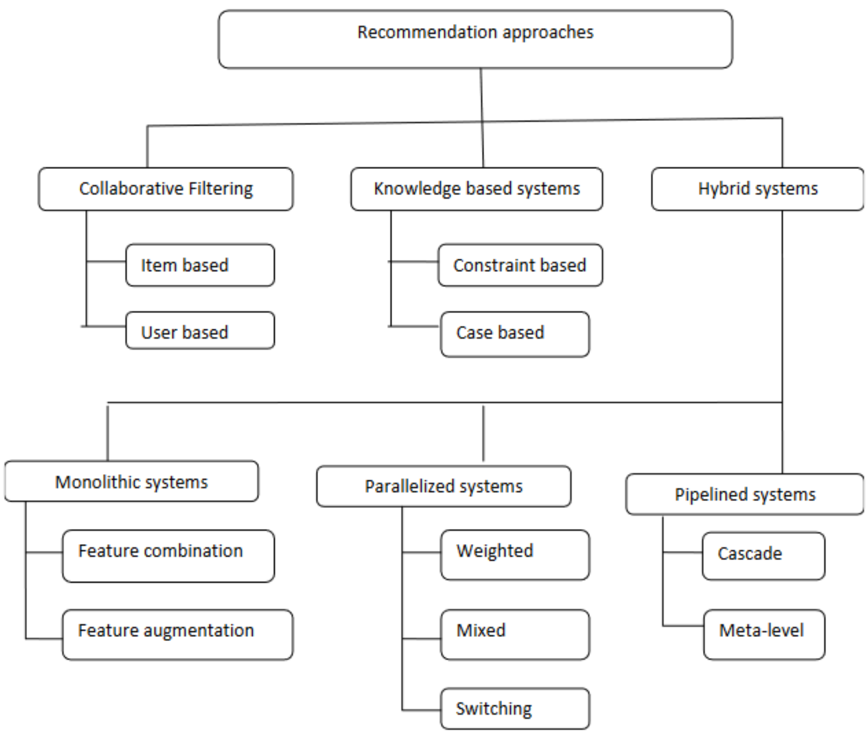

Figure 1.1: Classification of Recommendation Approaches

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 representing 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 [3].

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 [3].

1.3 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 outcomes [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).

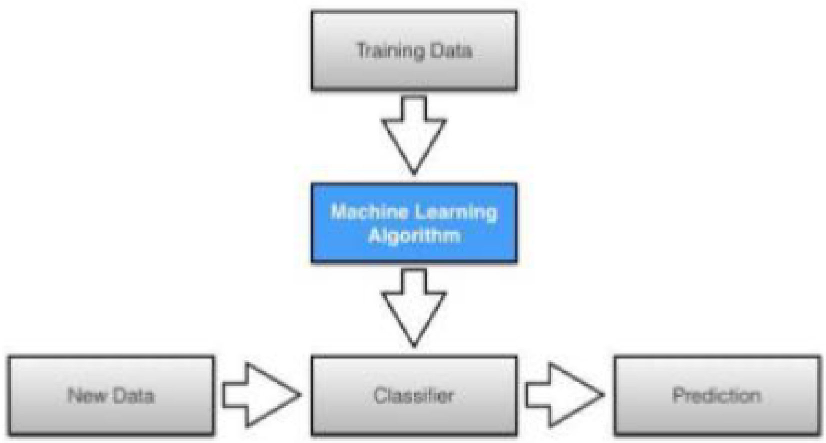

Figure 1.2: Supervised Learning Process

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].

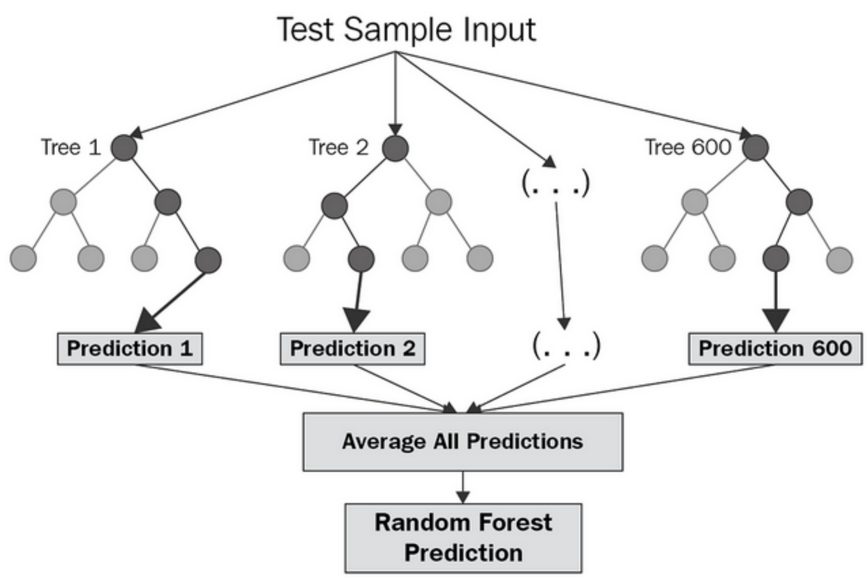

Figure 1.3: Random Forest Procedure

1.4 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].

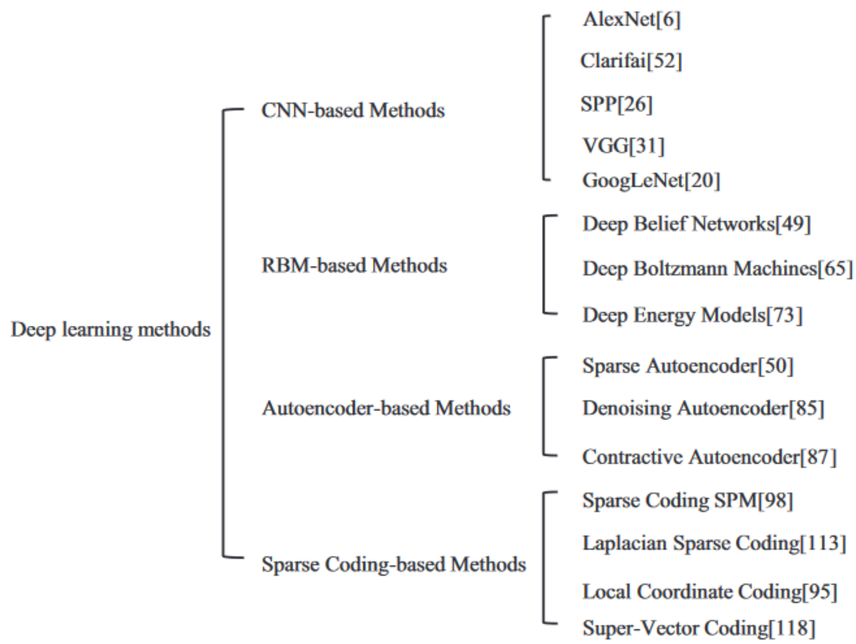

Figure 1.4: A categorization of the deep learning methods and their representative works.

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].

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].

1.5 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.

1.6 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] [3]. 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.

1.7 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 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 three sections: Collaborative Filtering, Knowledge-based Systems, and Hybrid Systems.

2.1 Collaborative Filtering

2.2 Knowledge-Based Systems

2.3 Hybrid Systems

Chapter 3

Methodology

3.1 Data Collection

3.2 Data Analysis and Methods

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] 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.

[3] 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.

[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.