**People's Democratic Republic of Algeria**

**Ministry of Higher Education and Scientific Research**

**Mohamed Boudiaf University of M'Sila**

**Faculty of Mathematics and Computer Science**

**Department of Computer Science**

**N°:……..**


**FIELD: Mathematics and Computer Science**

**BRANCH: Computer Science**

**OPTION: NETWORKS AND INFORMATION AND COMMUNICATION TECHNOLOGIES**

**Thesis presented for obtaining**
**The Academic Master's degree**

**By**:

Wafa Barka and Mounya Bouarbi

**Entitled:**

# A University Recommender System based on Students Profiles

**Defended in front of the jury composed of:**

| | | |
|---|---|---|
| Dr. Roussafi Mahdjoubi | University of M'sila | President |
| Dr. Noureddine Amraoui | University of M'sila | Reporter |
| Dr. Rached Yagoubi | University of M'sila | Examiner |

**Academic Year: 2022 / 2023**

*I dedicate this dissertation to my beloved family whose endless support, and belief in my dreams have shaped the person I am today. I am forever grateful for your presence in my life. To my esteemed supervisor, thank you for your guidance and expertise, I am honored to have had the opportunity to learn from you. To my friends and colleagues, thank you for the camaraderie and shared experiences. Your friendship and support have made this journey more enjoyable and meaningful. Finally, this work is dedicated to all those who have played a part in shaping my academic path. Your support and belief in me have been instrumental, and I am deeply appreciative.*

**B. Mounya**

*I dedicate this humble work to the dearest soul to my heart who left me two decades ago... My dear Father. The greatest woman in my life and the whole world... My dear Mother. All my brothers and sisters, and their children, each in his name...I am grateful for all your efforts, and your unconditional support that you have provided me throughout my career and past years. Your love and faith in me and my abilities pushed me to progress, continue, and reach what i am today. May Allah bless all of them.*

**B. Wafa**

# Acknowledgements

First and foremost, we express our gratitude to the Almighty Allah, who has bestowed upon us the strength, patience, and courage needed to accomplish this objective.

We are immensely grateful to our supervisor, Dr. Noureddine Amraoui, for his invaluable advice, continuous support, and patience during our work.

Also, we would like to thank the Jury President and members for their time, expertise, and thorough evaluation of our work.

We extend our appreciation to the staff of our computer science Department at Mohammed Boudiaf University of M'Sila, who provide us with the stimulating academic environment to shaping the ideas presented in this dissertation.

We are grateful to our friends and families for their unwavering support and encouragement throughout this journey. Their belief in our abilities has been a constant source of motivation.

Finally, to everyone mentioned above and those not explicitly stated, completing this dissertation would not have been possible without the collective support and guidance of all you... We are deeply grateful for their contributions and the role you have played in our academic growth and achievement.

# Table of contents

# List of figures

# List of tables

# Chapter 1

# Introduction

In the past few years, universities have increasingly been generating vast amounts of data related to student performance. This data includes information on student grades, attendance, engagement with coursework, and other relevant metrics. With the rise of digital technology, universities now have access to even more data, which can be extremely valuable for them, as it can be used to identify patterns in student performance, which can help inform decisions around curriculum development, teaching methods, and student support services [38].

## 1.1 Problem Statement and Objectives

Although universities are known for their pursuit of knowledge and research, they do not fully exploit the potential of the vast amounts of data they generate and collect. The reason for this may be a lack understanding of the value of data in decision-making processes [16].

One consequence of such a lack of data exploitation is that future students face a daunting university selection process. On one hand, the selection process made by students requires careful evaluation of several factors including academic reputation, studies, faculty skill, research opportunities, location, cultural environment, financial constraints, career situation, and personal preferences. On the other hand, each prospective student has unique ambition,

academic qualifications and specific requirements that must be considered during the selection process.

In this work, we aim to devise a flexible and adaptable system that can automatically propose the best universities for students which can simplify their selection process and raise their chances of admission.

## 1.2 Methodology and Motivation

In recent years, the use of Recommender Systems (RS) has become increasingly popular in various fields, including higher education institutions. In the context of higher education, these systems can help students and faculty members discover relevant academic resources, such as courses, research papers, and academic events. Additionally, they can assist administrators in making data-driven decisions regarding student retention, engagement, and academic success [22].

Despite advances in RS technology, there is little research on university recommender systems. Further research is needed on university recommendation systems.

## 1.3 Contribution and Results

This project aims to pave the way for research to help address this gap by proposing a University Recommender System that maximize a student's chances of admission by carefully evaluating a student's profile and matching it with the universities deemed most suitable.

To do so, our proposed university recommender system is developed using a methodology that includes gathering pertinent information about universities and programs, preprocessing the information, choosing appropriate classification algorithms, training and evaluating the models, and creating personalized recommendations based on student profiles.

The proposed RS has many features namely:

- It takes into account many different aspects of their university performance such as final grades (GPA) and TOEFL test scores.
- It offers personalized recommendations based on the individual requirements.
- Making the student selection process simpler: Facilitate the selection process by reducing available options and presenting an organized list that matches the student's background and with the possibility to discover new opportunities which the pupil has not considered.
- Flexible and adaptable: The capacity of the system to adapt to changing student histories and adapt recommendations accordingly, making suggestions realistic and more appropriate.

We conduct an experimental study to evaluate the performance of our RS and its ability to generate accurate and useful recommendations. In particular, we collect relevant data from various sources including student profiles, academic qualifications, preferences, and other factors.

Our obtained results show that our University RS is in average effective in generating personalized recommendations based on three powerful algorithms, namely, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM).

## 1.4 Dissertation Structure

This dissertation is organized into the following chapters:

In Chapter 2, we review the state of the art related to Recommender Systems in higher education field by providing the necessary theoretical background, and presenting our classification of existing RS approaches in this field.

In Chapter 3, we present the design of our proposed recommender system specifically tailored for the higher education domain.

In Chapter 4, we present the implementation and experimental study of our proposed University Recommender System to show its efficiency including evaluation datasets, protocol, and obtained results discussion.

Finally, in Chapter 5, we draw conclusions and suggest some directions for future works.

# Chapter 2

# Recommender Systems: Background and Existing Approaches

This Chapter provides an overview of the recent advancement of Recommender Systems (RS) pertaining to higher education. We first provide a theoretical background to understand RS, their advantages, application domains, techniques, and challenges. Then, we present our classification of existing RS approaches in the field of higher education.

## 2.1 Recommender Systems

Recommender Systems (RSs) can be defined as any system capable of providing customized recommendations to guide the user to interesting and useful resources within a large data space [6].

In recent years, the use of (RSs) has become increasingly popular in various fields, including higher education institutions. In the context of higher education, these systems can help students and faculty members discover relevant academic resources, such as courses, research papers, and academic events. Additionally, they can assist administrators in making data-driven decisions regarding student retention, engagement, and academic success.

### 2.1.1 Application Domains

Recommender systems are an integral part of our daily lives, often operating in the background to help us make decisions. They are used in a wide range of application domains to suggest personalized recommendations to users based on their past behavior, preferences, and other factors.

| Application Domains | Examples | Description |
|---|---|---|
| Education | Coursera, edX / Khan Academy | Platforms offering personalized course recommendations for online learning |
| E-commerce | Amazon, eBay / Alibaba | E-commerce platforms providing personalized product recommendations based on user preferences and history |
| Healthcare | WebMD, Fitbit /MyFitnessPal | Recommender systems in healthcare providing personalized health recommendations and wellness tips |
| Entertainment | Spotify, Netflix / YouTube | Music and video streaming platforms offering personalized recommendations based on user preferences |
| Finance | Robinhood, Mint/ Acorns | Financial platforms providing personalized investment recommendations and financial management advice |

Table 2.1 Application Domains of Recommender Systems

#### Education

Recommender systems are becoming increasingly important in the education domain. it have the potential to significantly enhance the educational experience for students and teachers . However, it is important to ensure that these systems are designed and implemented with care to ensure that they are accurate, effective, and ethical. Additionally, it is crucial to

consider the privacy implications of collecting and analyzing student data and to ensure that appropriate measures are taken to protect student privacy.

### E-commerce

In this filed recommender systems help improve customer experience and increase sales by providing personalized recommendations to customers. By using various approaches, these systems can help customers discover new products, making shopping more efficient and enjoyable.

### Healthcare

Healthcare recommender systems have the potential to improve patient outcomes by providing personalized recommendations that are tailored to each patient's unique needs and circumstances. However, it is important to note that healthcare recommender systems must be carefully designed and implemented to ensure that they are accurate, reliable, and compliant with relevant regulations and ethical standards.

### Entertainment

Entertainment is a vast and diverse domain that includes various forms of media and content, such as movies, TV shows, music, books, and games. Entertainment recommender systems have proven to be successful in increasing user engagement and satisfaction. By providing personalized recommendations, users are more likely to continue using the service and discover new content they enjoy.

### Finance

Recommender systems are also widely used in the finance domain it help customers make informed investment decisions and manage their finances more effectively. By providing

personalized recommendations based on the customer's investment history, risk appetite, and financial goals, these systems can help customers achieve their financial objectives and increase customer satisfaction and loyalty.

### 2.1.2 Techniques

In general, recommender systems can be classified into two main categories: personalized and non-personalized. Personalized recommender systems use information about the user's preferences and past interactions to generate recommendations, while non-personalized recommender systems generate recommendations based on the popularity or characteristics of the items.

Within these categories, there are many different types of recommender systems that use different algorithms, techniques, and data sources to generate recommendations. In this section, we will explore some of the most commonly used types of recommender systems.

#### 2.1.2.1 Collaborative filtering

Recommender systems based on collaborative filtering leverage the preferences of other users to provide a recommendation to a particular user [11].


Fig. 2.1 Collaborative filtering model [28]

The main idea of the collaborative filtering is to gather the information regarding the old user's behavior and opinions so as to correlate with the new user whether they are similar or not [42].

Researchers have devised a number of collaborative filtering algorithms that can be divided into two main categories: Memory-based (user-based) and Model-based (item-based) algorithms.

**Memory-based** collaborative filtering algorithms make use of the entire user-item database to generate predictions. These algorithms rely on statistical techniques to identify a group of users, called neighbors, who have a history of agreeing with the target user. This agreement can be reflected in similar ratings for different items or a tendency to purchase a similar set of items. Once the system forms a neighborhood of users, it applies various algorithms to combine their preferences and create a prediction or top-N recommendation for the active user. These techniques are also referred to as nearest-neighbor or user-based collaborative filtering and are widely used in practice due to their popularity [34].

**Model-based** collaborative filtering algorithms use a probabilistic approach to create a model of user ratings. This model is constructed using different machine learning techniques such as Bayesian network, clustering, and rule-based approaches. The algorithm then utilizes this model to generate item recommendations based on the expected value of a user's prediction.

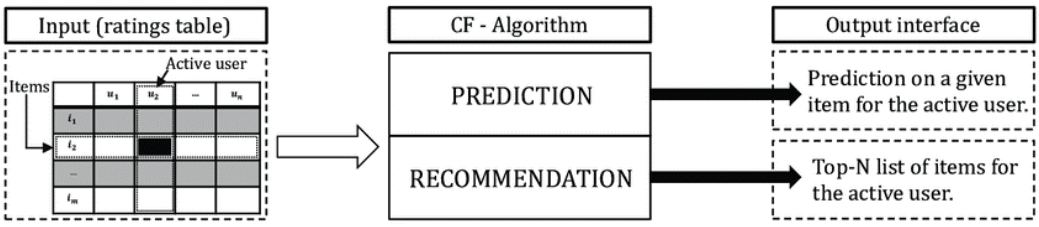


Fig. 2.2 Collaborative filtering Process [39]

#### 2.1.2.2 Content based filtering

Content based filtering (the cognitive filtering) method is based on the semantic search that is the information retrieval. This technique suggests items on the basis of customer's profile and the customer's item profile. The user's profile is created when the user starts the system. It collects the interest of the users and recommend the items after analyzing the features of the items and the users. The recommended items are identical to the things that were liked by the customer earlier or previously and they also match the attributes of the user. This technique works well only when the attributes are presented in a clear and proper way[21].

This technique is used to provide personalized recommendations for learning materials, such as textbooks, videos, or articles, based on the content of the materials and the interests of the student. The system analyzes the characteristics of the materials, such as the topic, level of difficulty, and language, to provide recommendations that match the student's interests and level of proficiency.

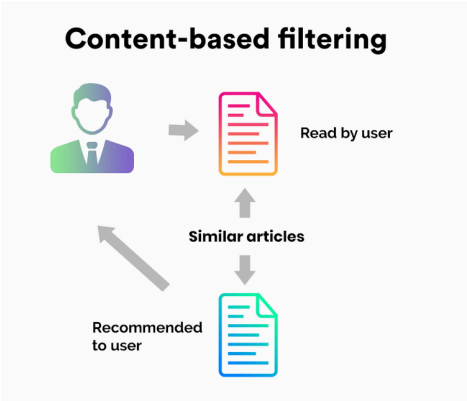


Fig. 2.3 Content based filtering model [28]

#### 2.1.2.3 Hybrid recommendation

To extract useful information and make recommendations, collaborative filtering and content-based filtering techniques are utilized. Each of these methods has its own strengths and weaknesses. Hybrid recommendation systems combine the two approaches to overcome their

limitations. By merging the filtering methods, the hybrid approach can improve the accuracy and efficiency of the recommendation process. There are various ways to implement hybrid recommendation approaches, including:

**Weighted Hybrid Approach** In this approach, both collaborative and content-based filtering techniques are combined by assigning weights to each method. The weights can be determined based on the accuracy or effectiveness of each technique. For example, if collaborative filtering has a higher accuracy rate for a particular recommendation task, it can be given a higher weight.

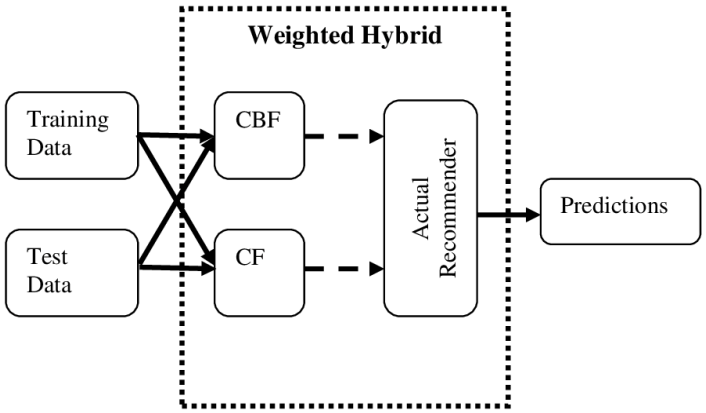


Fig. 2.4 Weighted Hybrid Approach [30]

**Cascade Hybrid Approach** This approach involves using one recommendation technique to generate a list of recommended items, which is then used as input for another recommendation technique. For example, collaborative filtering can be used to generate a list of items that are most similar to the user's preferences, and this list can be used as input for content-based filtering.

**Switching Hybrid Approach** This approach uses a switch to choose between collaborative and content-based filtering techniques. The switch can be based on factors such as the availability of data, user preferences, or the nature of the recommendation task. For instance,

if there is insufficient data to apply collaborative filtering, the system can switch to content-based filtering.

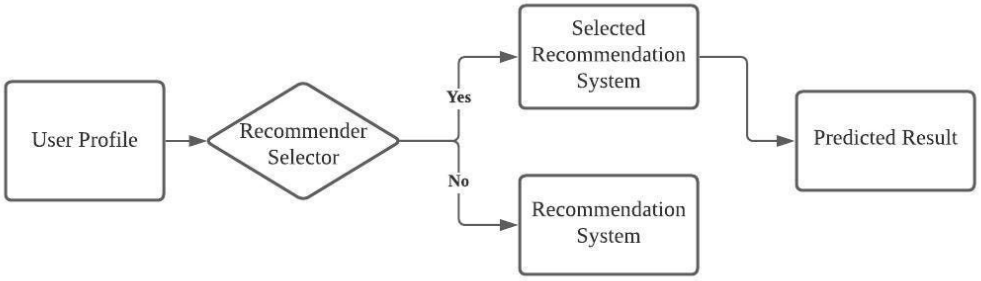


Fig. 2.5 Switching Hybrid Approach [9]

**Feature Combination Hybrid Approach** In this approach, features from both collaborative and content-based filtering techniques are combined to create a new set of features. These features can be used to generate recommendations. For example, features such as user ratings and item attributes can be combined to create a new set of features that can be used to make recommendations.

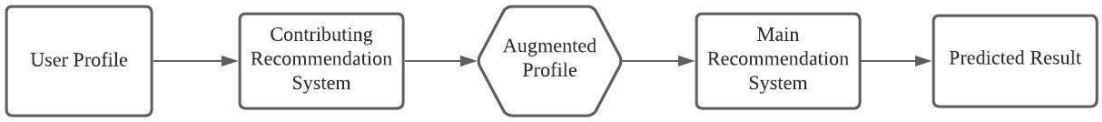


Fig. 2.6 Feature Combination Hybrid Approach [9]

**Meta-Level Hybrid Approach** In this approach, meta-level information about the recommendation process is used to combine different recommendation techniques. For instance, the performance of different recommendation techniques can be evaluated using meta-level information such as their accuracy, diversity, or novelty. The best performing techniques can then be combined to generate recommendations.

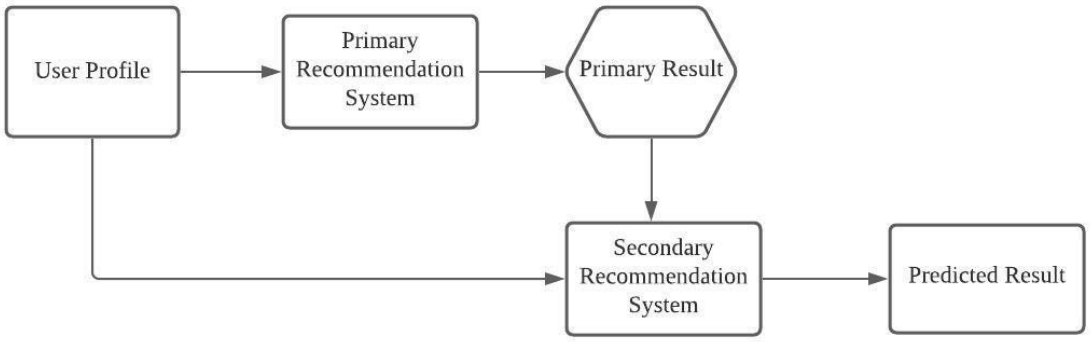


Fig. 2.7 Cascade Hybrid Approach [9]

### 2.1.3 Challenges

There are several challenges that face Recommender Systems including:

**Data quality and availability** One of the primary challenges facing recommender systems in higher education is the quality and availability of data. In some cases, the available data may be incomplete or inaccurate, which can affect the accuracy of the recommendations[12].

**Heterogeneity of users and items** Another challenge is the heterogeneity of users and items in the education domain. For example, students have different backgrounds, interests, and learning styles, and courses have different levels, requirements, and outcomes. This makes it difficult to develop a one-size-fits-all recommendation model that can satisfy all users and items[17].

**Cold start problem** Recommender systems in higher education often face the cold start problem, which occurs when there is not enough data available to make meaningful recommendations for new users or items. This can be particularly challenging for new courses or programs that have limited historical data[5].

**Contextual information** In higher education, contextual information such as the student's academic background, learning objectives, and preferences can play an important role in making relevant recommendations. However, this information is often not readily available or difficult to obtain [7].

**Ethical and privacy concerns** Finally, recommender systems in higher education must address ethical and privacy concerns related to the collection and use of student data. This requires careful consideration of issues such as data ownership, consent, and transparency to ensure that the recommendations provided are fair and unbiased [14].

## 2.2 Existing Recommender Systems Proposed in The Filed of Higher Education Institutions

In this section, we present a classification of existing RS proposed and with examples of papers in each category. But before that, let review some advantages of recommender systems in higher education institutions.

### 2.2.1 Advantages

Recommender systems have the potential to improve the quality of education, enhance the student experience, and increase the efficiency of resource allocation in universities. In the following section, we will explore some of the key advantages that recommender systems can offer in the context of higher education:

#### Personalized Learning

Recommender systems can provide students with personalized learning experiences that are tailored to their interests, learning style, and pace. By analyzing student data such as grades,

attendance, and participation, these systems can identify knowledge gaps and recommend relevant courses, textbooks, and resources.

### Improved Student Retention

Recommender systems can help improve student retention by identifying at-risk students and providing targeted interventions. For example, if a student is struggling in a particular course, the system can recommend additional resources, tutoring, or personalized feedback to help them improve.

### Enhanced Course Recommendations

Recommender systems can help students select courses and programs that are aligned with their career goals and interests. By analyzing student data such as past course selections and grades, these systems can recommend courses and majors that are likely to lead to success.

### Efficient Resource Allocation

Recommender systems can help universities and colleges allocate resources more efficiently by identifying areas of high demand and recommending solutions. For example, if a particular course or program is in high demand, the system can recommend additional faculty, resources, or funding to meet the needs of students.

### Improved Student Engagement

Recommender systems can increase student engagement by recommending extracurricular activities, events, and programs that are aligned with their interests and goals. This can help students feel more connected to the institution and improve their overall satisfaction with the college or university.

### 2.2.2 RS For University Selection For Secondary Students

| Reference | Techniques | Datasets | Results |
|---|---|---|---|
| Zayed et al. (2022) | CF, ML algorithms | Grad. student data | Improved accuracy in program selection for secondary students |
| Elahi et al. (2022) | Hybrid approach, user profiling | Univ. student data | Enhanced personalization and accuracy in program recommendation |

Table 2.2 Research Studies on Program Selection

Table 2.2 presents a summary of recommender systems (RS) for university selecton targeted at secondary students. The references include the work by Zayed et al. (2022), which employed collaborative filtering (CF) and machine learning (ML) algorithms on graduate student data, resulting in improved accuracy in program selection. The study by Elahi et al. (2022) utilized a hybrid approach and user profiling with university student data, leading to enhanced personalization and accuracy in program recommendation [13] [43].

### 2.2.3 RS For Courses Selection

| Reference | Techniques | Datasets | Results |
|---|---|---|---|
| Lynn & Emanuel (2021) | Literature review, comparative analysis | N/A | Insights into existing recommender systems for course selection |
| Lahoud et al. (2022) | CF, content-based filtering | Course enrollment data | Comparison of different recommender systems for course selection |

Table 2.3 Research Studies on Course Selection

Table 2.3 provides an overview of recommender systems (RS) for course selection. The references include the literature review and comparative analysis conducted by Lynn

and Emanuel (2021), which aimed to provide insights into existing RS for course selection. Additionally, the study by Lahoud et al. (2022) implemented collaborative filtering (CF) and content-based filtering techniques on course enrollment data to compare different recommender systems [27] [24].

### 2.2.4 RS For Study Program (Specialty) Selection

| Reference | Techniques | Datasets | Results |
|---|---|---|---|
| Smith et al. (2019) | Association rules mining, user preferences | Student specialty data | Identification of preferred specialty areas for students |
| Chen & Yang (2017) | Hybrid approach, student performance analysis | Student academic records | Improved accuracy in specialty recommendation for students |

Table 2.4 Research Studies on Course Selection

Table 2.4 focuses on recommender systems (RS) for study program (specialty) selection. The references include the work by Smith et al. (2019), which utilized association rules mining and user preferences on student specialty data to identify preferred specialty areas for students. The study by Chen and Yang (2017) employed a hybrid approach and student performance analysis using student academic records to achieve improved accuracy in specialty recommendation for students [36] [8].

## 2.3 Conclusion

Despite advances in RS technology, there is little research on university recommender systems. Further research is needed on university recommendation systems. This project aims to pave the way for research to help address this gap.

# Chapter 3

# A University Recommender System

This chapter focuses on the design of our proposed recommender system tailored for the higher education domain. In particular, it applies data processing and ML algorithms training to recommends most suitable universities to help student in their selection process. We will first present and overview of different stages involved. Then, we give more details on each stage.

## 3.1 Overview

Figure 3.1 shows a high-level overview of different stages involved in our proposed system.

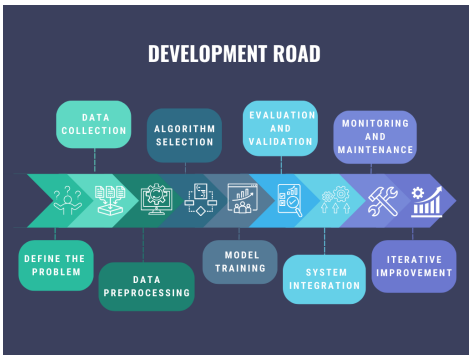


Fig. 3.1 Stages Involved in RS

## 3.2 Low-Level Operation

In this section, we present the operation of our proposed RS in detail.

### 3.2.1 Data Collection

This process includes obtain data on universities, such as their standings, academic programs, faculty profiles, evaluations and reviews by students, required for admission, and campus benefits.

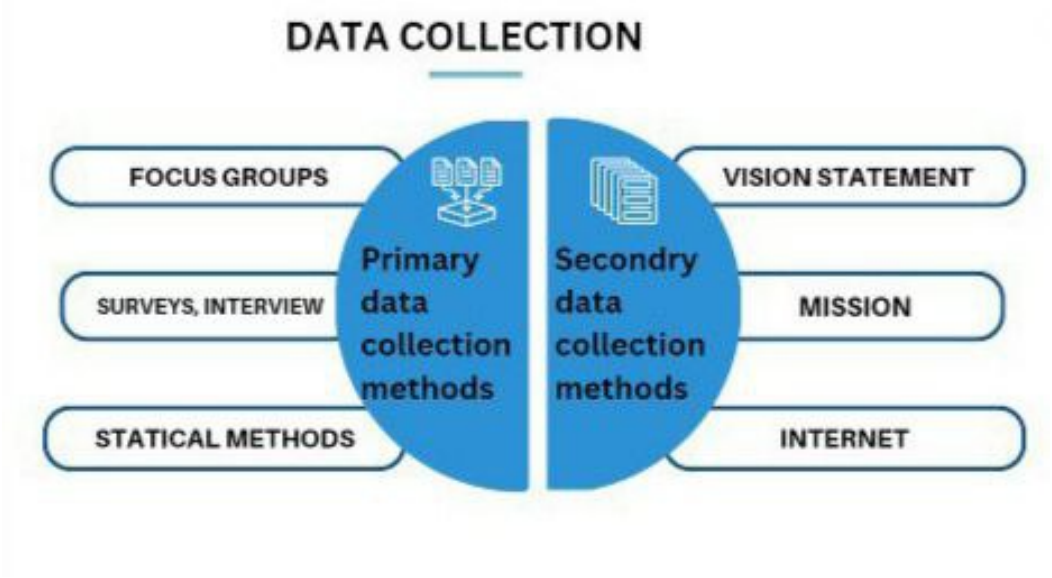


Fig. 3.2 Data Collection Methods
[QuestionPro]

It also includes gathering data about student profiles, such as their academic successes, interests, chosen study locations, location choices, costs, and any additional relevant information [20].

In most cases, data collecting occurs after the experiment or observation. Planning and estimating benefit from both primary and secondary data. Either qualitative or quantitative data is collected [emb].

### 3.2.2 Data Preprocessing

Once the data is collected, it needs to be preprocessed to prepare it for the recommendation algorithms. So Data preprocessing is a part of data preparation and describes any type of processing performed on raw data in order to prepare it for another data processing operation [23].

Although various data preprocessing techniques, the total work can be decomposed into several general and important steps: data cleaning, data integration, data reduction, and data transformation [ProjectPro].


Fig. 3.3 Data Preprocessing Steps
[V7 Labs]

#### 3.2.2.1 Data Cleaning

Data cleaning is a essential step in data preprocessing.this process detects and eliminate errors and dissimilarity in data and enhance its quality. Data quality issues arise due to typos, missing values, or other invalid data during data entry. Basically, "dirty" or "unclean" data is transformed into clean data. "Dirty" data does not give accurate and good results. Therefore,

processing these data becomes very important. Professionals spend a lot of time on this step [18].

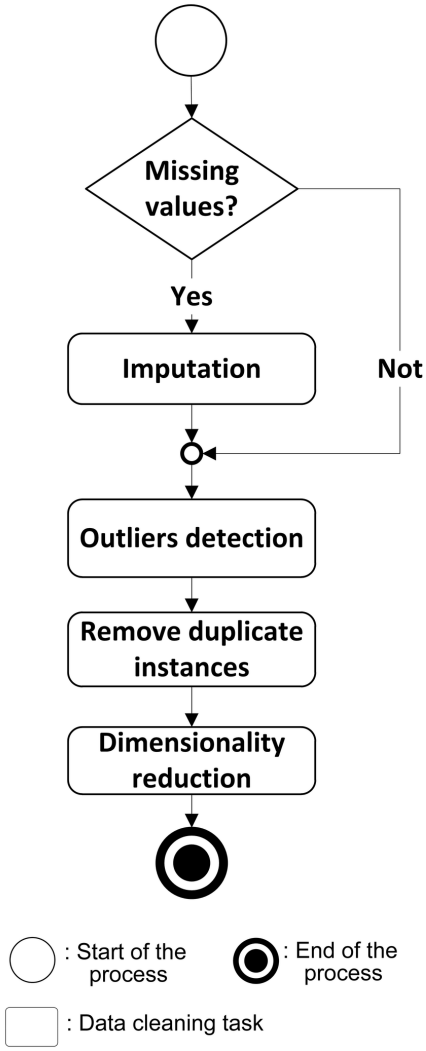


Fig. 3.4 Data Cleaning Process [10]

#### 3.2.2.2 Data Integration

Data integration combines data from multiple sources into a connected and same view. This process includes identifying and accessing various data sources, mapping data into a common format, and reconciling any inconsistencies or differences between data sources. The goal of data integration is to more easily access and analyze data distributed across multiple systems or platforms to gain a more complete and accurate understanding of the data [4]

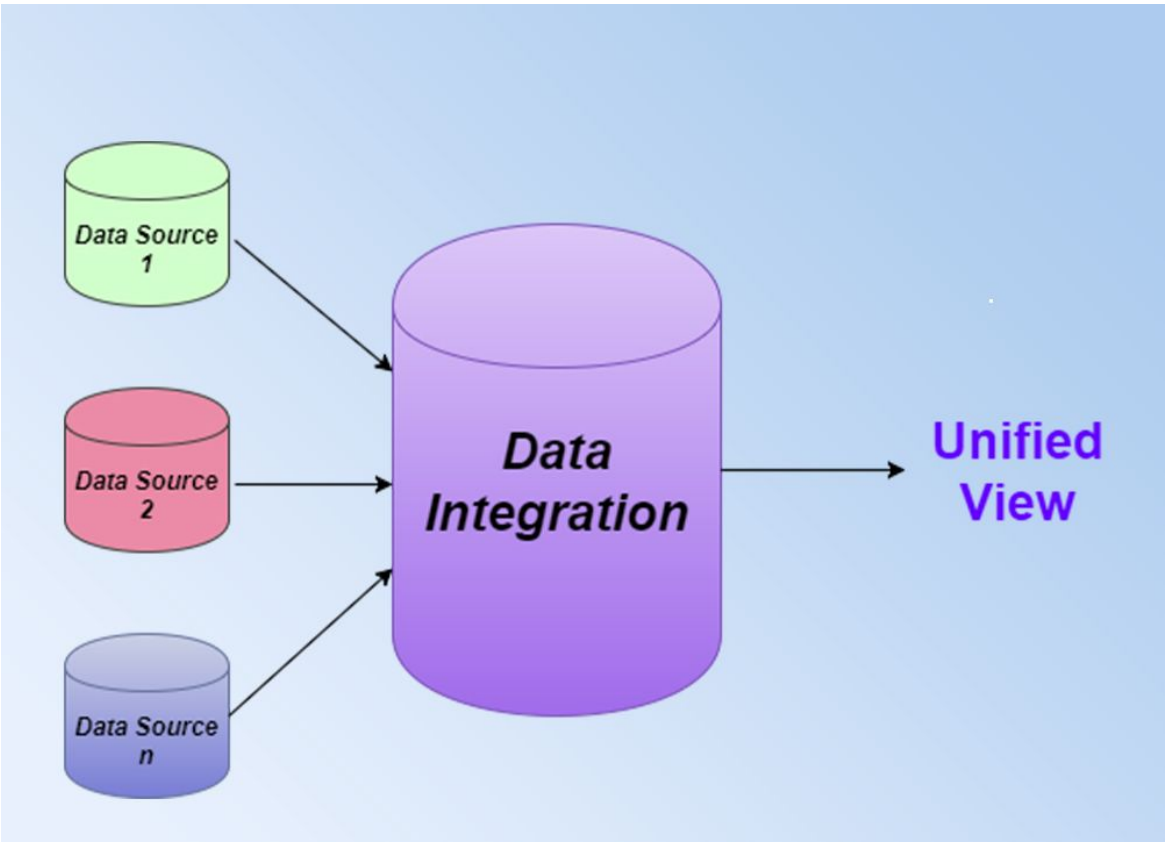


Fig. 3.5 Data Integration Process
[app]

#### 3.2.2.3 Data Transformation

Data is converted into a format suitable for analysis. Common data transformation techniques include Data Smoothing, Attribute Construction, Data Aggregation, Data Normalization, Data Discretization, Data Generalization as shown in this figure: [19].

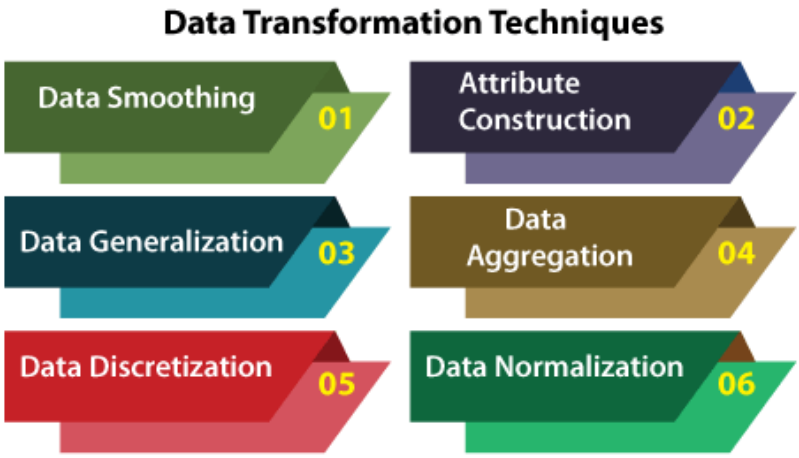


Fig. 3.6 Data Transformation Techniques
[19]

**Data smoothing** techniques, such as moving averages and median filtering, are used to eliminate noise and irregularities from the data.

**Attribute construction** involves creating new features based on existing ones.

**Attribute aggregation** data aggregation combines multiple data points into a single representation.

**Data normalization** scales numerical features to a common range.

**Data discretization** transforms continuous variables into categorical ones.

**Data generalization** replaces specific values with more general ones.

These techniques are essential for data preprocessing, enhancing data quality, and facilitating various analytical tasks.

The entire data transformation process is called ETL (Extract, Load, and Transform). The ETL process allows analysts to transform data into their desired format. Following are the steps of the data conversion process:

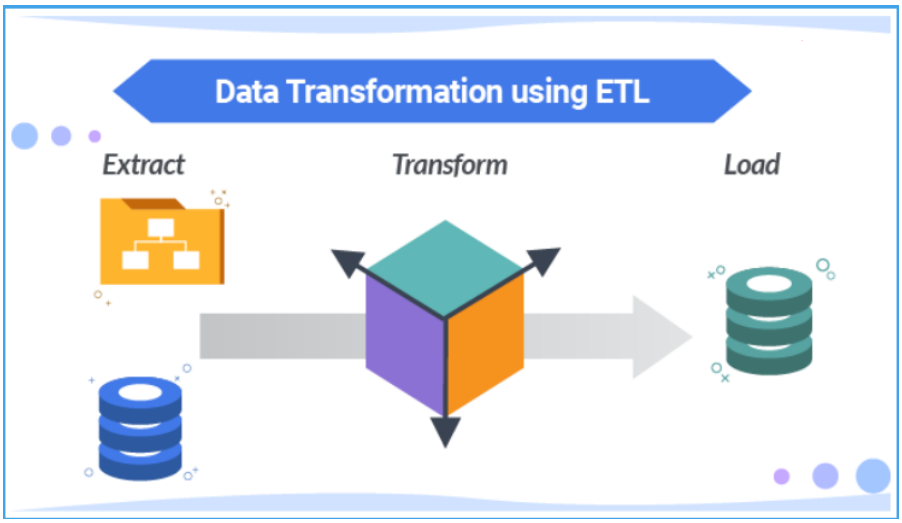


Fig. 3.7 Data Transformation Process
[41]

#### 3.2.2.4 Data Reduction

In order to reduce the amount and complexity of datasets, data reducing provides more effective and efficient data analysis without losing the quality or validity of the results by intelligently reducing data while maintaining its significant patterns and qualities In order to reduce the amount and complexity of datasets, data reduction is an important step in preprocessing. The main goals of data reduction are to increase storage effectiveness, boost computing speed, make data exploration easier, and speed analysis processes [25].

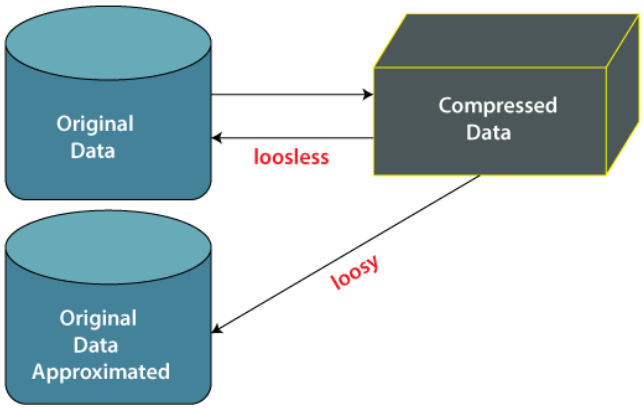


Fig. 3.8 Data Reduction Process
[19]

### 3.2.3 Algorithm Selection

The next step is to determine an appropriate recommendation algorithm for the problem at hold. There are different types of algorithms, including collaborative filtering, content-based filtering, hybrid methods, and machine learning models. The choice depends on available data, system requirements, and expected accuracy.

In what follows, we give a brief presentation of some well-known and powerful ML algorithms:

#### K-Nearest Neighbors

The K-Nearest Neighbors algorithm is a non-parametric approach for regression and classification. The result of K-NN classification is a class membership. An item is categorized by a majority vote of its neighbors, and is then put into the class that has the most members among its k closest neighbors (k is a positive number that is often small). The item is simply put into the class of its one nearest neighbor if k = 1.

#### Random Forest

A collection of decision trees is called Random Forest. Unlike single decision trees, which, depending on how they are adjusted, are likely to have significant bias or high variance. Averaging is used in Random Forests to achieve a natural equilibrium between the two extremes. They are basically non-parametric since they have a small number of parameters to adjust and may be utilized fairly well with default parameter values. When you don't know the underlying model or you need to generate a solid model quickly, Random Forests are a fantastic first cut.

**Support Vector Machine**

Support An sophisticated machine learning method called vector machines is used to categorize both linear and nonlinear issues. A linear kernel is employed when the training patterns can be separated linearly. By initially mapping the input pattern into a higher dimensional space using a kernel function, the linear SVM may be converted into a nonlinear classifier. The resulting nonlinear SVM classifier is nonlinear with respect to the original data but linear with respect to the converted data.

### 3.2.4 Model Training

When developing a university recommender system, the model training phase involves training a chosen recommendation algorithm using preprocessed data. The algorithm learns patterns and connections from the data to make personalized college rankings. The training process often includes optimization techniques, such as gradient descent, to minimize error or maximize recommendation accuracy [35].

### 3.2.5 Evaluation and Validation

It is essential to evaluate the model's performance after training. This can be done by analyzing how well the recommender system predicts user preferences or makes recommendations for products. The efficacy of the system is frequently evaluated using measures like precision, recall, accuracy, and mean average precision. To validate the generalizability of the model can used validation approaches like cross-validation or holdout methods.

Evaluation metrics include but not limited to:

- Mean Absolute Error (MAE): The average absolute difference between the real and anticipated values is what the MAE calculates. It provides a hint as to the average size of the errors.

- Mean Squared Error (MSE): MSE calculates the squared difference between the true and predicted values. In comparison to MAE, it assigns greater weight to larger mistakes.

- Root Mean Squared Error (RMSE): The average size of the mistakes in the original scale of the data is measured by RMSE, which is the square root of MSE.

- R-squared (R2) Score: The R2 score indicates the percentage of the dependent variable's variation that can be predicted from the independent variables. It evaluates a model's goodness-of-fit and shows how effectively the model accounts for data variability.

### 3.2.6 System Integration

To put the recommender system into the target application or platform when it has been created and approved. This includes putting in place the required interfaces, APIs, or SDKs to enable the system to communicate with users and offer tailored suggestions in real-time.

### 3.2.7 Monitoring and Maintenance

The recommender system has to be regularly updated and checked after implementation. This entails keeping track of user comments, gathering fresh data, frequently retraining the model, and upgrading the system to reflect any advancements or modifications.

### 3.2.8 Iterative Improvement

Iterative refinement and improvement are possible as the recommender system gains more user interactions and input. To improve the quality of the recommendations and the user experience, this entails returning to earlier phases and doing fresh data collection, algorithm updates, or system feature enhancements.

## 3.3 Conclusion

The chapter a clear structured framework for the development and improvement of our proposed university recommender systems. It describes key steps and components such as data processing and algorithm selection.Within this groundwork, we carefully selected three algorithms: k-Nearest Neighbors (kNN), Random Forest, and Support Vector Machines (SVM) to improve the recommendation process, These algorithms belong to different kinds of machine learning algorithms.

# Chapter 4

# Implementation and Experimental Study

This Chapter presents the implementation and experimental study of our proposed University Recommender System. We first present implementation technologies. Then, we introduce our conducted experimental study to show its efficiency including evaluation datasets, protocol, and obtained results discussion.

## 4.1 Implementation Technologies

Following is the list of used technologies to implement our university RS and conduct the experimental study.

### 4.1.1 Python

Developed by Guido van Rossum in 1989, holds the distinction of being one of the oldest programming languages. It serves as a fundamental language for individuals aspiring to venture into the realms of programming. Python is an open-source and object-oriented programming language. One of its prominent strengths lies in its extensive library and the unwavering support offered by its vibrant community. These factors contribute significantly to the appeal and advantage of using Python in various domains [26].

### 4.1.2 Google Colab

Practicing on projects becomes a constraint since you need high-end PCs for such workloads. The answer to this issue is Google Colab, or Collaboratory is a cloud-hosted version of Jupyter Notebook. To use Colab, you do not need to install and runtime or upgrade your computer hardware to meet Python's CPU/GPU intensive workload requirements [15].

Colab allows you to write and execute Python in your browser, with:

- Zero configuration required
- Access to GPUs free of charge
- Easy sharing

With Colab you can import an image dataset, train an image classifier on it, and evaluate the model, all in just a few lines of code,all you need is a browser. Google Colab makes data science, deep learning, neural network, and machine learning accessible to individual researchers who can not afford costly computational infrastructure [15].

### 4.1.3 NumPy

NumPy can be used to perform various mathematical operations on arrays. It adds powerful data structures to Python, guarantees efficient computation of arrays and matrices, and provides a large library of high-level mathematical functions that can manipulate these arrays and matrices [29].

### 4.1.4 Pandas

Pandas is a Python library for working with datasets. It has features for analyzing, cleaning, exploring and editing data. The name "Pandas" refers to both "Panel Data" and "Python Data Analysis", and was created in 2008 by Wes McKinney [40].

### 4.1.5 Scikit-learn

The most popular Machine Learning (ML) library in the Python environment is Scikit-learn, an open-source data analysis package. Important ideas and characteristics include: The several algorithms for making decisions, such as: Data can be classified by recognizing patterns in them [3].

## 4.2 Experimental Study

### 4.2.1 Evaluation Dataset

The identification of the dataset is the initial stage in the creation of a recommendation system. The academic information and background data from the dataset that were submitted throughout the application procedure are relevant to this specific issue. These data must be arranged with the proper labels in order to construct the classification model for the reference system.

In this work, we use a dataset scraped from the "Edulix" forum. It includes about 53,500 samples in the raw data file. Each sample comprises 26 attributes that match a student's profile. The attributes extracted include GPA, undergraduate institution, GRE verbal, quantitative, and analytical writing scores, number of journal and conference publications, professional experience, research experience, internship experience, and major being pursued [33].

Figure 4.1 depicts the data attributes sorted by importance.

### 4.2.2 Experimental Protocol

#### 4.2.2.1 Data Preprocessing

We need to do the preprocessing and cleaning, as there are lots of anomalies in the dataset. For this we use pandas and numpy frameworks. Cleaning the data was done by :

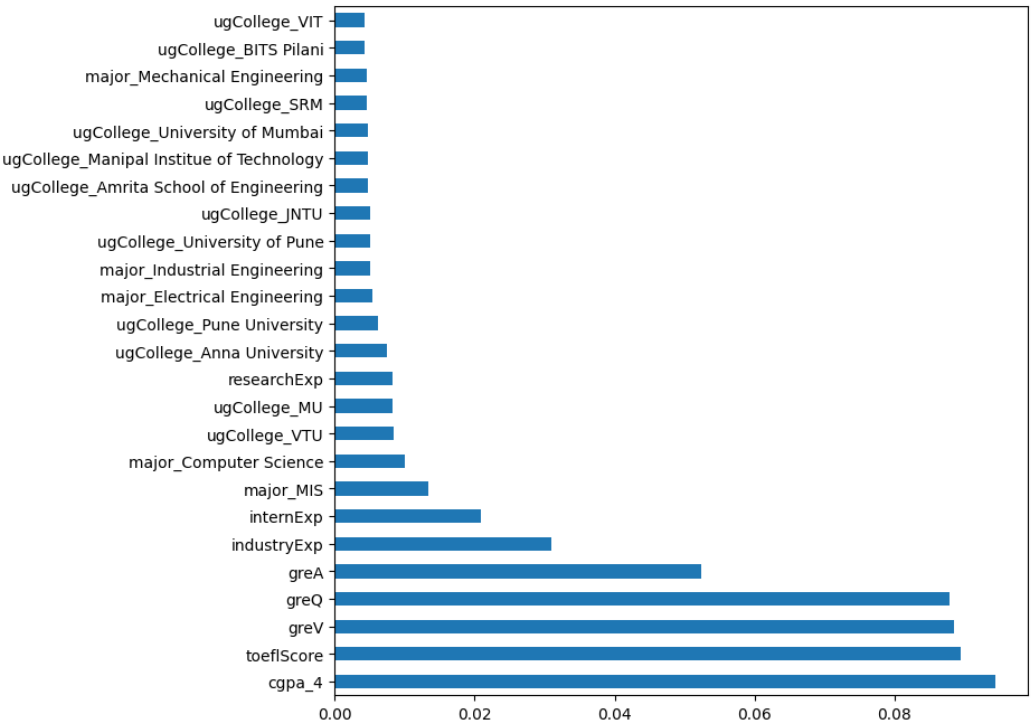


Fig. 4.1 features importances in pourcentage

- Dropping the redundant columns by using the drop column feature
- Filling the null values with the appropriate value or deleting the row containing null values.
- Removing the spaces in the data and reducing the size of the dataset.

All of the GPA scores in our dataset were equally scaled to a 4-point scale by utilizing normalize functions after the GRE scores were also cleaned because they contained scores.

$$X' = (X - min(X))/(max(X) - min(X)) \tag{4.1}$$

Where X is the value of the GPA, and X' the new one.

**Exploratory Data Analysis** This process involves using a number of techniques including:

- Plotting raw data

- Displaying straightforward statistics like mean, standard deviation, etc.
- Placing these plots in a way that maximizes our capacity for natural pattern detection.

An example of exploratory data analysis is shown in Figure 4.2

```
[ ] df.describe()
```

| | researchExp | industryExp | toeflScore | internExp | greV | greQ | greA | cgpa_4 |
|---|---|---|---|---|---|---|---|---|
| count | 11734.000000 | 11734.000000 | 11734.000000 | 11734.000000 | 11734.000000 | 11734.000000 | 11734.000000 | 11734.000000 |
| mean | 0.521817 | 6.402165 | 105.317539 | 0.715016 | 152.790097 | 162.711522 | 3.599857 | 3.088968 |
| std | 3.185130 | 14.478641 | 7.520210 | 2.928754 | 5.463499 | 4.287230 | 0.541725 | 0.650407 |
| min | 0.000000 | 0.000000 | 65.000000 | 0.000000 | 134.000000 | 140.000000 | 1.500000 | 0.000000 |
| 25% | 0.000000 | 0.000000 | 101.000000 | 0.000000 | 149.000000 | 160.000000 | 3.000000 | 2.933053 |
| 50% | 0.000000 | 0.000000 | 106.000000 | 0.000000 | 153.000000 | 163.000000 | 3.500000 | 3.200000 |
| 75% | 0.000000 | 0.000000 | 111.000000 | 0.000000 | 156.750000 | 166.000000 | 4.000000 | 3.436000 |
| max | 53.000000 | 132.000000 | 120.000000 | 96.000000 | 170.000000 | 170.000000 | 6.000000 | 4.000000 |

Fig. 4.2 Data Description

The distribution of data in various classes was shown using the number of admissions per university. This distribution hints at an unbalanced dataset.

#### 4.2.2.2 Algorithm Selection

The techniques that was used for generating recommendations in our system are K-Nearest Neighbors, Random Forest, and Support Vector Machine which were previously described in Chapter 3.

### 4.2.3 Results and Discussion

#### 4.2.3.1 Prediction Results

Table 4.1 show the obtained results of prediction:

Table 4.2 show the obtained accuracy values for each applied algorithm. We can see that SVM outperforms Random Forest and K-Nearest Neighbor in terms of ranking the top 3 universities.

| Prediction for | Results |
|---|---|
| Best prediction | ['Arizona State University'] |
| Top three | ['Arizona State University', 'SUNY Stony Brook', 'University of Texas Dallas'] |

Table 4.1 Prediction results

| algorithm | accuracy |
|---|---|
| K-Nearest Neighbours | 52% |
| Random Forest | 55% |
| Support Vector Machine | 56.3% |

Table 4.2 accuracy results

#### 4.2.3.2 Parameters Tuning

The number of neighbors utilized in the K-NN model was varied, and it was discovered that when the number of neighbors was equal to 90, the accuracy was best (around 52). Fig.4.3 depicts how accuracy varies with the number of trees built.

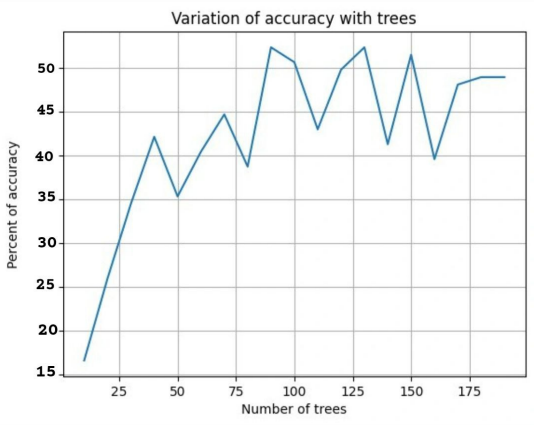


Fig. 4.3 K-Nearest Neighbours Model

When the number of trees employed in the Random forest model was varied, it was discovered that 150 trees produced the model with the greatest accuracy of 55. Fig.4.4 illustrates how accuracy varies with the number of trees built.

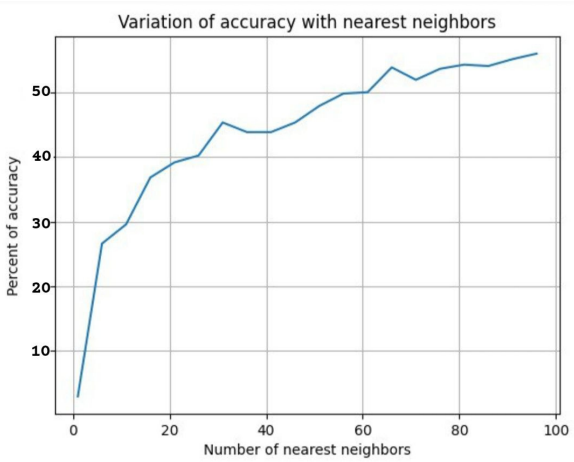


Fig. 4.4 Random Forest Model

## 4.3 Conclusion

In summary, the implementation of our University Recommender System demonstrated its average effectiveness in predicting the best universities and generating personalized recommendations and the successful integration of three powerful algorithms, K-Nearest Neighbors (KNN), Random Forest and Support Vector Machine (SVM) .

# Chapter 5

# Conclusion

## 5.1 Concluding Remarks

This works presented the design, implementation, and experimental validation of a university recommender system (RS). The proposed RS helps students to choose the best university according to the background and performance of the student compared to a ready-made template for students who were previously accepted at these universities.

The proposed RS has many features namely:

- It takes into account many different aspects of their university performance such as final grades (GPA) and TOEFL test scores.
- It offers personalized recommendations based on the individual requirements.
- Making the student selection process simpler: Facilitate the selection process by reducing available options and presenting an organized list that matches the student's background and with the possibility to discover new opportunities which the pupil has not considered.

- Flexible and adaptable: The capacity of the system to adapt to changing student histories and adapt recommendations accordingly, making suggestions realistic and more appropriate.

## 5.2 Future Work

It is necessary and important to complete the work and recommendations with more research, and we hope that there will be an opportunity in the future to apply it in Algerian universities and provide opportunities for students in Algeria to make informed decisions about their university's future.

# References


[emb] Data collection methods. https://www.embibe.com/exams/data-collection/.

[app] Data integration in data mining guide. Appsierra.

[3] ActiveState (Accessed 2023). What is scikit-learn in Python? https://www.activestate.com/resources/quick-reads/what-is-scikit-learn-in-python/. Accessed: [Insert Date].

[4] Alasadi, S. A. and Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, 12(16):4102–4107.

[5] Bharadhwaj, H. (XXXX). Meta-learning for user cold-start recommendation. *Journal Name*, X(X):X–X.

[6] Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, 46:109–132.

[7] Boliver, V., Gorard, S., and Siddiqui, N. (XXXX). Using contextual data to widen access to higher education. *Journal Name*, X(X):X–X.

[8] Chen, S. and Yang, J. (2017). An intelligent recommender system for assisting students in choosing their specialty in higher education. *International Journal of Information and Education Technology*, 7(2):107–111.

[9] Chiang, J. (Year). 7 types of hybrid recommendation system.

[10] Corrales, D., Corrales, J., and Ledezma Espino, A. (2018). How to address the data quality issues in regression models: A guided process for data cleaning. *Symmetry*, 10.

[11] Deschênes, M. (XXXX). Recommender systems to support learners' agency in a learning context: A systematic review. *XXX*, XXX(XXX):XXX.

[12] Dimensional Insight (2019). Data quality in higher education. https://www.dimins.com/blog/2019/09/04/data-quality-higher-education/. [Accessed: Month Day, Year].

[13] Elahi, M., Starke, A., El Ioini, N., Lambrix, A. A., and Trattner, C. (2022). Developing and evaluating a university recommender system. *Frontiers in Artificial Intelligence*, 4:212.

[14] Florea, D. and Florea, S. (XXXX). Big data and the ethical implications of data privacy in higher education research. *Journal Name*, X(X):X–X.

[15] Geekflare (2023). Google colab: Free cloud-based jupyter notebooks. Website. Accessed on May 25, 2023.

[16] Hannah, A. (2020). The importance of data-driven decisions in higher ed. https://www.othot.com/blog/higher-ed-data-decision-making.

[17] Happ, R., Zlatkin-Troitschanskaia, O., Beck, K., and Förster, M. (XXXX). Increasing heterogeneity in students' prior economic content knowledge – impact on and implications for teaching in higher education. *Journal Name*, X(X):X–X.

[18] Jain, D. (YEAR). *Data Preprocessing in Data Mining*. PUBLISHER, ADDRESS.

[19] javatpoint (2023). Data transformation in data mining. https://www.javatpoint.com/data-transformation-in-data-mining.

[20] JothiLakshmi, S. and Thangaraj, M. (2021). Design and development of recommender system for target marketing of higher education institution using edm. *Journal of Education and Learning Systems*, XX(X):XX–XX.

[21] Kaur, H. and Bathla, G. (2019). Techniques of recommender system. *International Journal of Innovative Technology and Exploring Engineering*.

[22] Kelly, R. (2013). A data-driven approach to student retention and success. *Publication Name*.

[23] Kumar, D. (Year Published). Introduction to data preprocessing in machine learning: Beginners guide for data preprocessing. *Journal or Website Name*. Accessed: <june 03,2023>.

[24] Lahoud, C., Moussa, S., Obeid, C., Khoury, H. E., and Champin, P.-A. (2022). A comparative analysis of different recommender systems for university major and career domain guidance. *Education and Information Technologies*, pages 1–27.

[25] Lawton, G. (2021). Data preprocessing.

[26] Logicrays (2020). What is python programming language? learn basic programs in python. https://www.logicraysacademy.com/blog/what-is-python-programming-language/. Accessed on [Insert Date].

[27] Lynn, N. and Emanuel, A. (2021). A review on recommender systems for course selection in higher education. In *IOP Conference Series: Materials Science and Engineering*, volume 1098, page 032039. IOP Publishing.

[28] Maruti Techlabs (Accessed 2023). Recommendation engine benefits. https://marutitech.com/recommendation-engine-benefits/.

[29] NumPy Contributors (2006–). NumPy. https://numpy.org/. Accessed: [Insert Date].

[30] Okaka, R. A. (2018). A hybrid approach for personalized recommender system using weighted term frequency inverse document frequency. In *Proceedings of the International Conference on Computer Science and Software Engineering (CSSE)*.

[ProjectPro] ProjectPro. Data preprocessing techniques and steps. https://www.projectpro.io/article/data-preprocessing-techniques-and-steps/512. Accessed: <insert date accessed>.

[QuestionPro] QuestionPro. Data collection methods. https://www.questionpro.com/blog/data-collection-methods/#types_of_data_collection_methods. Accessed: <insert date accessed>.

[33] Ramkishore, S., Manley, J., Gnanasekaran, Krishnakumar, S., and Suresh Kumar, A. (2023). University recommender system for graduate studies in usa. Student project report, University of California San Diego.

[34] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th International Conference on World Wide Web (WWW '01)*, pages 285–295.

[35] Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. *Recommender systems handbook*, pages 257–297.

[36] Smith, J., Johnson, M., and Anderson, K. (2019). A recommender system for selecting specialty areas in higher education. *Journal of Educational Technology*, 23(4):123–140.

[V7 Labs] V7 Labs. Data preprocessing guide. https://www.v7labs.com/blog/data-preprocessing-guide#h1. Accessed: <june 03, 2023>.

[38] Vukicevic, M., Jovanovic, M., Delibasic, B., and Suknovic, M. (2013). Recommender system for selection of the right study program for higher education students. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*, page 145.

[39] Vuong, N. L., Hong, M.-S., Jung, J. J., and Sohn, B.-S. (2020). Cognitive similarity-based collaborative filtering recommendation system. *Applied Sciences*, 10:4183.

[40] W3Schools (Accessed 2023). Pandas Introduction. https://www.w3schools.com/python/pandas/pandas_intro.asp. Accessed: [Insert Date].

[41] XenonStack (2023). Data transformation. https://www.xenonstack.com/insights/data-transformation/.

[42] Yanxiang, L., Deke, G., Fei, C., and Honghui, C. (2013). User-based clustering with top-n recommendation on cold-start problem. In *2013 Third International Conference on Intelligent System Design and Engineering Applications*, pages 1585–1589. IEEE.

[43] Zayed, Y., Salman, Y., and Hasasneh, A. (2022). A recommendation system for selecting the appropriate undergraduate program at higher education institutions using graduate student data. *Applied Sciences*, 12(24):12525.

# Abstract

Although universities are known for their pursuit of knowledge and research, they do not fully exploit the potential of the vast amounts of data they generate and collect. One consequence of this is that future students face a daunting university selection process. This work aims to devise a Recommender System (RS) that can automatically propose the best universities for students which can simplify their selection process and raise their chances of admission.

To evaluate the effectiveness and feasibility of our proposed RS, we collect relevant data from various sources. This data includes student profiles, academic qualifications, preferences, and other factors. We analyze this data to evaluate the performance of our recommender system and its ability to generate accurate and useful recommendations.

Our obtained results show that our University RS is in average effective in generating personalized recommendations based on three powerful algorithms, namely, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM).



**Keywords:** University Selection, Recommender Systems (RS), Machine Learning (ML), Data Processing.


---


## ملخص

على الرغم من أن الجامعات معروفة بسعيها للحصول على المعرفة والبحث، إلا أنها لا تستغل بشكل كامل إمكانات الكميات الهائلة من البيانات التي تولدها وتجمعها. إحدى نتائج ذلك هي أن طلاب المستقبل يواجهون عملية اختيار جامعية شاقة. يهدف هذا العمل إلى ابتكار نظام التوصية (RS) الذي يمكنه اقتراح أفضل الجامعات تلقائيًا للطلاب والتي يمكنها تبسيط عملية اختيارهم وزيادة فرص قبولهم.

لتقييم فعالية وجدوى نظام التوصية المقترح، نقوم بجمع البيانات ذات الصلة من مصادر مختلفة. تتضمن هذه البيانات ملفات تعريف الطلاب والمؤهلات الأكاديمية والتفضيلات وعوامل أخرى. نقوم بتحليل هذه البيانات لتقييم أداء نظام التوصية الخاص وقدرته على تقديم توصيات دقيقة ومفيدة.

تظهر نتائجنا التي تم الحصول عليها أن نظام التوصية فعال بشكل معتبر في إنشاء توصيات مخصصة بناءً على ثلاث خوارزميات قوية، وهي K-Nearest Neighbors (KNN) و Random Forest و Support Vector Machine (SVM).



**كلمات مفتاحية:** اختيار الجامعة، أنظمة التوصية، التعلم الآلي، معالجة البيانات.