

A
Project Report
On

“UNIVERSITY RECOMMENDATION SYSTEM”

By
PABBA SATHVIKA(21STUCHH010407)

Under the guidance of

Dr. PAVAN KUMAR

Sr.Assistant Professor



Department of Data Science and Artificial Intelligence

Faculty of Science and Technology

ICFAITECH, ICFAI Foundation for Higher Education, HYDERABAD

CERTIFICATE

This is to certify that the project report entitled **UNIVERSITY RECOMMENDATION SYSTEM** submitted by Pabba Sathvika to the ICFAI Institute For Higher Education, in partial fulfillment for the award of the degree of B. Tech in DSAI is a *bonafide* record of project work carried out by him/her under my/our supervision. The contents of this report, in full or in parts, have not been submitted to any other Institution or University for the award of any degree or diploma.

Internal Guide

Head of the Department

Name:

Name:

(Deemed to be University under section-3 of UGC act 1956)

DECLARATION

We declare that the work contained in the Project Report is original and it has been done by us under the supervision of Dr.Pavan Kumar. The work has not been submitted to any other University for the award of any degree or diploma.

Date: 01/05/ 2024

Signature of the Student

P.Sathvika

ACKNOWLEDGEMENT

It gives me immense pleasure to acknowledge with gratitude the help and guidance rendered to me by a host of people to whom I owe a substantial completion of the mid term seminar work.

I would like to express gratitude to Dr.Pavan Kumar ,Sr Assistant Professor ,Department of DS&AI, ICFAI Institute For Higher Education for all the timely support and valuable suggestions during the period of my seminar. I am extremely thankful to sir for their valuable suggestions and constant support throughout the seminar.

Finally ,thanks to the ones who help me directly or indirectly ,parents and friends for their cooperation in completing the seminar.

ABSTRACT

One of the major issue many students are facing now a days is choosing a right university to purse their postgraduate course.Choosing a right university required a lot of research which will be difficult for some students.To reduce the burden to students while choosing a university many researchers have developed recommendation system model using various machine learning algorithms.This recommendation system is built based on past information of students who are admitted into the university.It takes various scores of students like GRE,TOFEL,CGPA etc as input and recommend a best university suitable for them as output.Machine learning algorithms like Random Forest Classifier, Naive Bayes Classifier ,Support Vector Classifier,K Nearest Neighbour, XG Booster etc.

Keywords: Random Forest,Support Vector, K Nearest Neighbour.

TABLE OF CONTENTS

DECLARATION	2
ACKNOWLEDGEMENT	3
ABSTRACT	4
INTRODUCTION	9
LITERATURE REVIEW	10
CHAPTER -1 METHODOLOGY	15
1.1:Basic steps in constructing a Machine Learning model	15
1.1.1 - Data Collection	15
1.1.2 - Data Preparation	15
1.1.3 - Choose a Model:	15
1.1.4 - Train the Model	16
1.1.5 - Evaluate the Model	16
1.1.6 - Make Predictions	16
Chapter 2: Data Preparation and Cleaning	17
2.1: Data Loading and Overview	17
2.1.1: Importing Libraries:	18
2.1.2 Loading Data:	18
2.1.3 Exploring Dataframes:	19

2.2: Data Cleaning:	21
2.2.1: Handling Missing Data:	21
2.2.2: Dropping Columns:	22
2.2.3: Imputing Missing Values:	22
CHAPTER 3: DATA VISUALIZATION	24
3.1: Exploratory Data Analysis :	24
3.1.1: Analysis of Categorical Variables:	24
3.1.2: Analysis of Numerical Variables:	25
3.1.3: Visualization of University Counts Using Pie Chart:	25
3.1.4: SCATTER PLOT FOR GRE SCORES:	26
3.2. Correlation Analysis	27
3.2.1: Correlation Matrix	27
3.2.2: Scatter Matrix plot:	28
3.2.3:Heatmap Visualization:	29
3.3:Word Cloud Analysis:	29
CHAPTER-4 FEATURE EXTRACTION	31
4.1: Pre-Processing Categorical Variables:	31
4.2: Conversion of GRE Scores:	31
CHAPTER-5:MODELING	32
5.1:Random Forest Classifier:	32
5.2: Support Vector Machine:	33

5.3:K - NEAREST NEIGHBOURS:	34
RESULTS:	35
CONCLUSION:	36
REFERENCES	38

LIST OF FIGURES

Figure 1 : Project Outline	9
Figure 2 :Libraries	17
Figure 3 : Sample DataSet	19
Figure 4 : Dataset Info	20
Figure 5 : Pie Chart	26
Figure 6 : Scatter Plot	27
Figure 7 : Correlation Matrix	28
Figure 8 : Scatter Matrix Plot	28
Figure 9 : Heatmap	29
Figure 10 : Word Cloud	30
Figure 11 :RF Model	32
Figure 12 : RF prediction	33
Figure 13 :SVM Model	Error! Bookmark not defined.
Figure 14 : SVM Prediction	34
Figure 15 : KNN Model	Error! Bookmark not defined.
Figure 16 : KNN Prediction	35
Figure 17 : Results	36

INTRODUCTION

University Recommendation System plays a very important role in student's career to choose best university to pursue postgraduate Study based on their interest. But Building a University Recommendation System involves various steps. It includes Data Collection, Data Analysis, Data Pre-Processing, Data Visualization, Feature Extraction and Model Building. This recommendation system is built based on past information of students who are admitted into the university. It takes various scores of students like GRE, TOFEL, CGPA etc as input and produces a best university suitable for them as an output. Machine learning algorithms like Random Forest Classifier, Naive Bayes Classifier, Support Vector Classifier, K Nearest Neighbour, XG Booster etc. The datasets we used here to build recommendation system is a US based datasets. Finally, objective of this recommendation system is to recommend a university based on students past historical data like GRE, TOEFL and CGPA scores to pursue their postgraduate study.

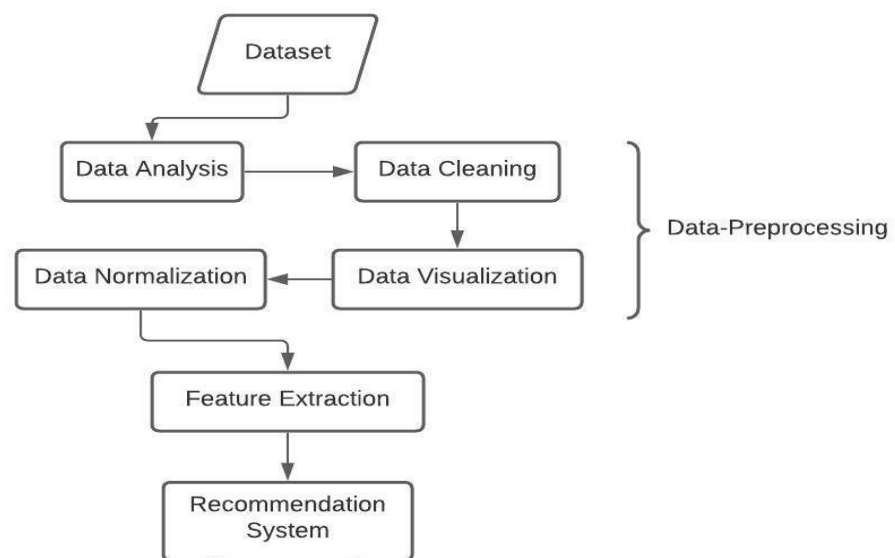


Figure 1 : Project Outline

LITERATURE REVIEW

Many kinds of research have been done on recommendation systems for various products throughout the universe. As demand for higher education keeps on increasing many researchers have designed a university recommendation system using various machine learning algorithms. Alcina Judy, Kesha D’cruz Janhavi Kathe, and Kirti Mot-wine in their paper “Recommendation System for Higher Studies Using Machine Learning” said that a Recommendation System generates suggestions for appropriate programs, courses, and universities based on the interests of students. The main goal of the System is to provide students to choose higher education universities to succeed in their goals. They have developed a system using machine learning algorithms. By using ML algorithms we can improve the quality and make recommendations to choose the best universities. To know whether a student is admitted or not, we have created a prediction model by using ML algorithms[2].

Aishwarya Nalawade and Bhavana Tiple in their article “The University Recommendation System for Higher Education” suggestions for course selection, Job development, and other academic activities are discussed along with its creation and evolution. By utilizing data mining and machine learning methods the model is created for a better experience. These methods will have a huge impact on how well recommendations are granted to universities for higher education[1]. Vandit Manish Jain and Rihaan Satia in their paper “College Admission Prediction using Machine Learning Model” said that the model takes inputs such as Scholastic history and standardized test results like GRE, and TOEFL. They also provided the creation and assessment of the model along with algorithms used to build the model, feature selection strategies, and Model evaluation metrics. making accurate predictions of students' chances of admission into particular universities makes them easy to approach for college applications[3].

“Recommendation System for Higher Studies Abroad via Machine Learning Techniques” By Pooja Bhatt, Manali Shah and Priyan Sheesonu. This paper proposes recommendation system that uses machine learning methods to provide recommendations for pursuing higher education. This model takes input as student background, preferences, and goals to recommend the most suitable university. By using this model students can choose university with less effects. The model can also save a lot of time by choosing the best universities for their success. Overall model solves the problem of many students by recommending the best universities and courses to study[4].

The article “Recommender Systems Challenges and Solutions Survey” by Marwa Hussien Mohamed, Mohamed Helmy Khafagy, and Mohammed Hasan Ibrahim explained about challenges and problems of the recommendation system. They also provided solutions for those challenges up to some extent. Problems like data quality, scalability, diversity, privacy, and security are faced while building the model. To overcome these problems various pre-processing techniques, ML algorithms, and privacy-preserving methods are used to improve the performance of the recommendation system. Overall this article provides a blueprint for developing a recommendation system[5].

A research article titled “Graduate School Recommender System: Assisting Admission Seekers to Apply for Graduate Studies in Appropriate Graduate Schools”.In this, the creation of a recommendation system based on students' academic profiles such as their GPA, GRE scores, and TOEFL Clearly Explained. By using ML algorithms to suggest graduate institutions. There are many institutions to select in the Admission procedure, students sometimes get confused about choosing the best university then the recommendation system comes into the picture and makes it easy to choose the best university among all. The algorithms used will provide suggestions with high accuracy. Overall, it adds an expanding body of knowledge on recommendation systems to help students choose graduate universities[6].

In the paper named “An Autonomous Courses Recommendation System for Undergraduate Using ML Techniques”.In this, the model takes inputs of averages

of various scores as well as their grades. The inputs are taken Via API. The model is built using Linear Regression Model, Naïve Bayes, SVM, KNN, and Decision Tree. Among all these models Naïve Bayes and SVM have the highest recommendation accuracy[7].Another paper titled “ An Effective Recommendation System to Forecast the Best Educational Program Using Machine Learning Classification Algorithms”.In this, the recommendation system takes input data from 10th-grade pass-out students and 12th-performance to suggest universities using various algorithms like Random Forest, XGBoost, Gradient Boosting, Gaussian Naïve Bayes, logistic Regression, Decision Tree, and K nearest neighbor [8].

Paper Name	Author Names	Algorithms
Selection of the right undergraduate major by students using supervised leaning techniques	Alsayes et al.[9]	DTC,ETC,RF,GBC,SVM
A Recommender System for Predicting Students' Admission to a Graduate Program using Machine Learning Algorithms	El Guabassi et al. [10]	DNN, LR, SVM, KNN, RF, DT, GNB
A College Major Recommendation System	Stein et al.[11]	KNN
Recommendation of Branch of Engineering using machine learning	Roshan et al.[12]	KNN

Students' Orientation Using Machine Learning and Big Data	Ouatik et al.[13]	NB
University Selection Model Using Machine Learning Techniques	Mostafa & Beshir[14]	SVM
Using data mining techniques to predict student performance to support decision making in university admission systems	Mengash[15]	RF
Machine Learning Models to Predict Students' Study Path Selection	Dirin&Saballe[16]	LR,RF,DT
CMRS: Towards Intelligent Recommendation for Choosing College Majors	Meng&Fun[17]	RF
Adaptive Recommendation System Using Machine Learning Algorithms for Predicting Student's Best Academic Program	Ezz&Elshnawy[18]	SVM,LR,RF
A graduate school recommendation system using the multi-class support	Baskota & Ng[19]	SVM,KNN

vector machine and KNN approaches		
Contributions of machine learning models towards student academic performance prediction: A systematic review	Balaji et al.[20]	NN,KNN,DT

CHAPTER -1 METHODOLOGY

1.1:Basic steps in constructing a Machine Learning model

1.1.1 - Data Collection

- The quantity & quality of your data dictate how accurate our model.
- The outcome of this step is generally a representation of data which we will use for training
- Using pre-collected data, by way of datasets from Kaggle, UCI, etc., still fits into this step

● 1.1.2 - Data Preparation

- Wrangle data and prepare it for training
- Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
- Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
- Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
- Split into training and evaluation sets

1.1.3 - Choose a Model:

Different algorithms are for different tasks; choose the right one

1.1.4 - Train the Model

- The goal of training is to answer a question or make a prediction correctly as often as possible • Linear regression example: algorithm would need to learn values for m (or W) and b (x is input, y is output)
- Each iteration of process is a training step

1.1.5 - Evaluate the Model

- Uses some metric or combination of metrics to "measure" objective performance of model
- Test the model against previously unseen data
- This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not)
- Good train/eval split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc.

1.1.6 - Make Predictions

- Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world

Chapter 2: Data Preparation and Cleaning

In Data Preparation and Cleaning, datasets play a very important role in building a model for a recommendation system.

DATA SOURCE:

URL: <https://www.kaggle.com/datasets/nitishabharathi/university-recommendation/data.csv>.

2.1: Data Loading and Overview

Library	Functionality
pandas	Data manipulation and analysis
<u>NumPy</u>	Scientific computing with Python
Collections	Specialized container <u>datatypes</u>
Warnings	To issue warning messages to the user
<u>Seaborn</u>	Data Visualization
<u>wordcloud</u>	Generates word clouds from text data
<u>sklearn</u>	Tools for machine learning
<u>xgboost</u>	Gradient boosting library
<u>Matplotlib.</u> `	Functions for controlling axis ticks
<u>Matplotlib. cm</u>	<u>Colormap</u> manipulation functions
<u>Mpl. toolkits</u>	Additional tools and utilities for <u>matplotlib</u>

Figure 2:Libraries

2.1.1: Importing Libraries:

We have imported a lot of libraries to access the pre-written functionalities. Libraries generally contain functions, classes, and modules. These libraries are well-tested and optimized for performance. Libraries abstract away low-level details. For example, we have imported pandas library for data manipulation and analysis.

2.1.2 Loading Data:

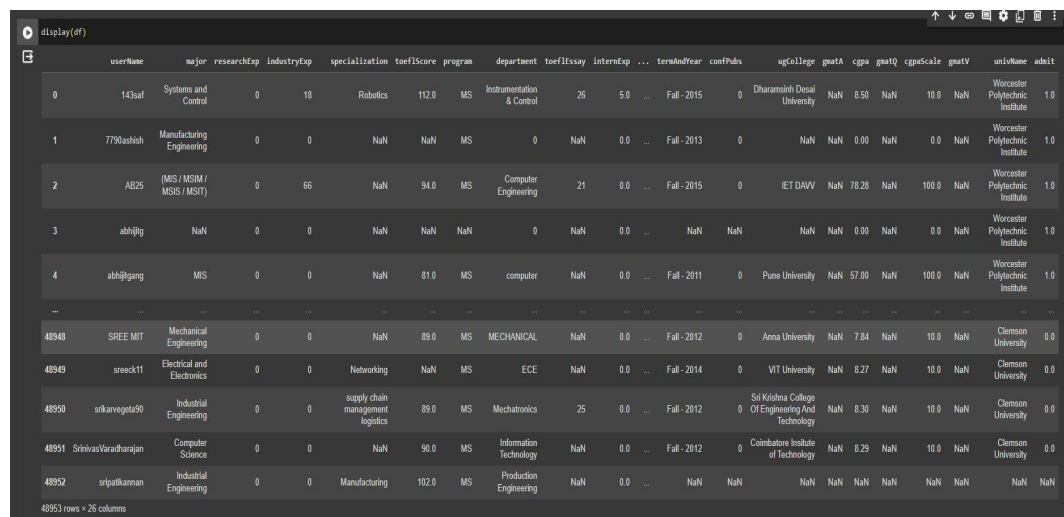
In this recommendation system, we have used two datasets. The dataset we used first is 1.csv which contains pieces of information about student profiles. Each profile has attributes like GRE score, TOEFL scores, Undergraduate college names, University names, Intern Experience, etc. There's also a score.csv dataset that contains new GRE scores.

We use the `read_csv` function to load the dataset. We have stored the dataset in a variable called `df`. To access data each time we use variable `df` instead of importing the dataset each time. Similarly score dataset is stored in a variable called `score_table`. To use the `read_csv` function we have to import a library called pandas. The pandas' library is imported as `PD` for easy access. Similarly, the numpy library is imported as `np`-numerical python, from the collections library we have imported a default dict then finally seaborn is imported as `sns`. In this dataset, many null values need to be handled by various techniques, their columns which are redundant to build the model. Those columns need to be dropped from the dataset. we will be having

columns with uncleared names those names should be adjusted. while loading the data not only csv we can load various types of data formats like JSON, excel, XSL, and many more.

2.1.3 Exploring Dataframes:

We have used various data analysis functions to know about our data. firstly, we used the display function to display the datasets, to know the number of columns and rows in the dataset.



	userName	major	researchExp	industryExp	specialization	toeflScore	program	department	toeflEssay	internExp	...	termAndYear	confPubs	ugCollege	gmatA	gpa	gmatQ	cpaScale	gmatV	univName	admit
0	143saf	Systems and Control	0	18	Robotics	112.0	MS	Instrumentation & Control	26	5.0	...	Fall - 2015	0	Dharamsinh Desai University	NaN	8.50	NaN	10.0	NaN	Worcester Polytechnic Institute	1.0
1	7790ashish	Manufacturing Engineering	0	0	NaN	NaN	MS	0	NaN	0.0	...	Fall - 2013	0	NaN	NaN	0.00	NaN	0.0	NaN	Worcester Polytechnic Institute	1.0
2	AB25	(MIS / MSIM / MSIS / MSIT)	0	66	NaN	94.0	MS	Computer Engineering	21	0.0	...	Fall - 2015	0	IET DAVV	NaN	78.28	NaN	100.0	NaN	Worcester Polytechnic Institute	1.0
3	abhijlg	NaN	0	0	NaN	NaN	NaN	0	NaN	0.0	...	NaN	NaN	NaN	NaN	0.00	NaN	0.0	NaN	Worcester Polytechnic Institute	1.0
4	abhijlgang	MIS	0	0	NaN	81.0	MS	computer	NaN	0.0	...	Fall - 2011	0	Pune University	NaN	57.00	NaN	100.0	NaN	Worcester Polytechnic Institute	1.0
...
48948	SREE MIT	Mechanical Engineering	0	0	NaN	89.0	MS	MECHANICAL	NaN	0.0	...	Fall - 2012	0	Anna University	NaN	7.84	NaN	10.0	NaN	Clemson University	0.0
48949	sreeck11	Electrical and Electronics	0	0	Networking	NaN	MS	ECE	NaN	0.0	...	Fall - 2014	0	VIT University	NaN	0.27	NaN	10.0	NaN	Clemson University	0.0
48950	srikarvegada90	Industrial Engineering	0	0	supply chain management logistics	89.0	MS	Mechatronics	25	0.0	...	Fall - 2012	0	Sri Krishna College Of Engineering And Technology	NaN	8.30	NaN	10.0	NaN	Clemson University	0.0
48951	SrinivasVaradharajan	Computer Science	0	0	NaN	90.0	MS	Information Technology	NaN	0.0	...	Fall - 2012	0	Colubatore Institute of Technology	NaN	8.29	NaN	10.0	NaN	Clemson University	0.0
48952	sripalkannan	Industrial Engineering	0	0	Manufacturing	102.0	MS	Production Engineering	NaN	0.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

48953 rows x 26 columns

Figure 3: Sample DataSet

Then we used the shape function which returns a tuple of dimensionality of the dataframe. This dataset contains 48953 rows and 26 columns. Here rows represent a list of students and columns represent the features of each one.

The dtypes function is used to know the datatype of each feature, we used the head function to display the first five rows of the dataset.

later we used the describe function to know the statistical description of each feature like mean, standard deviation, count, minimum, and maximum value of each feature.

In this, there are 3 columns named old, newQ, and newV. The data type of the old column is an integer, newQ is an integer and newV is also an integer. The describe function tells the description in terms of count, mean, standard deviation, minimum, maximum, and quartile information of each feature in the dataset. count indicates the total number of non-null values, mean indicates the average value of data in each feature, and standard deviation measures the spread of data around the mean. maximum and minimum values of each feature and quartile information of 25th, 50th, and 75th percentiles. The 25th Percentile tells about data values that fall below 25%, the 50th Percentile tells about the median of the data values and the 75th percentile value is the value where 75% of data falls.

Next comes The columns function which returns the list of feature names in the data set. The info function gives a summary of the data and also the presence of missing values.

```
df.columns
Index(['major', 'toeflScore', 'greV', 'greQ', 'greA', 'ugCollege', 'cgpa',
      'cgpaScale', 'univName', 'admit', 'univ_College_code', 'major_code'],
      dtype='object')

df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53644 entries, 0 to 53643
Data columns (total 26 columns):
#   Column                Non-Null Count  Dtype
---  -
0   userName              53644 non-null  object
1   major                 53644 non-null  object
2   researchExp           53644 non-null  int64
3   industryExp           53644 non-null  int64
4   specialization        31949 non-null  object
5   toeflScore            49230 non-null  float64
6   program               53322 non-null  object
7   department            53643 non-null  object
8   toeflEssay            11874 non-null  object
9   internExp             53630 non-null  float64
10  greV                  52388 non-null  float64
11  greQ                  52424 non-null  float64
12  userProfileLink        53644 non-null  object
13  journalPubs           53322 non-null  object
14  greA                  50786 non-null  float64
15  topperCgpa            53641 non-null  float64
16  termAndYear           53322 non-null  object
17  confPubs              53322 non-null  object
18  ugCollege             51366 non-null  object
19  gmatA                 119 non-null   float64
20  cgpa                  53644 non-null  float64
21  gmatQ                 123 non-null   float64
22  cgpaScale             53644 non-null  int64
23  gmatV                 114 non-null   float64
24  univName              53644 non-null  object
25  admit                 53644 non-null  int64
dtypes: float64(10), int64(4), object(12)
memory usage: 10.6+ MB
```

Figure 4: Dataset Info

The columns function displays a list of column names. columns are major, username, research experience, industry experience, specialization, TOEFL score, program, and department. TOEFL essay, intern Experience, greV, greQ, user profile link, journalPubs, greA, topper Cgpa, term, and year, confPubs, undergraduate college, gmatA, cgpa, gmatQ, cgpaScale, gmatV, university name and admit features in the dataset. count of non-missing values is given in the info function along with names and data types.

2.2: Data Cleaning:

The very first step after obtaining the dataset is data cleaning. Data cleaning is an important step because the quality of your data improves which will in turn improve the overall productivity of the model. Data cleaning usually consists of three steps as follows:

1. Analyze the data
2. Identify where changes are to be made (Ex: removing null values or unnecessary columns)
3. Making those necessary changes to the dataset.

2.2.1: Handling Missing Data:

It is a very important step in data preprocessing to ensure the integrity and accuracy of the data. Handling Missing Data involves various steps such as identifying missing values, determining the effect caused by missing values on analysis, and techniques to handle missing values. Missing values may occur due to faulty instrumental use, data entry errors, and negligence of the user.

In this we have created a Boolean variable named `missing_data`, where each cell in `df` shows their missing value or not using the `isnull` function. True indicates there is a missing value and False indicates it is not a missing value in that cell. Then we iterated on each column to see the distribution of missing values using the `value_count` function. This step plays an important phase in handling missing values and improving the accuracy of the data.

2.2.2: Dropping Columns:

We usually drop columns that are redundant, irrelevant, and columns with too many missing values. It is a very important step to improve the quality and focus of the dataset. Dropping irrelevant columns simplifies the analysis process, but it is very essential to observe while dropping columns because if we drop important columns then it will affect the integrity and effectiveness of the analysis.

In this, we have created a list of columns that are redundant, irrelevant, and stored in a variable called `del_col_list`. Later the variable passed into a drop function for dropping the columns from the dataset. After that, we dropped the universities which resulted in fewer instances followed by dropping all the students who were not admitted into the university. Because based on past information of admitted students we are going to build the model. For better accuracy of the model, we have to drop the students who are not admitted into any university.

2.2.3: Imputing Missing Values:

Imputing Missing Values involves replacing nan values with mean mode or median values of corresponding column values. Replacing

the nan values with mean involves calculating averages for numerical columns and then imputing missing values with those averages.

In this, we have computed the average of the numerical columns such as 'TOEFL score', 'internExp', 'greV', 'greQ', 'greA', and 'topperCgpa'. First, we converted each column into a float datatype using the type function and then calculated the mean using the mean function along the specified axis. Averages are important to impute missing values. After computing the average values, missing values in each column are replaced with the corresponding average using the replace function. while imputing the missing values the dataset becomes complete for analysis. Then later we filled categorical values with the mode of the respective column using the fillna function and mode function.

overall data loading and cleaning functions play an important role in building the model with more accuracy. By handling missing values we are ensuring data consistency, and enhancing the quality and reliability of the dataset, leading to meaningful full insights from the dataset.

CHAPTER 3: DATA VISUALIZATION

Data Visualization is the process of converting information into visual contexts like graphs, maps, and charts, to understand useful insights from the data easily. The main goal is to make it easier to identify patterns, trends, and outliers in large data.

3.1: Exploratory Data Analysis :

It is a very crucial step in the data analysis process, where analysts examine and summarize the characteristics of a dataset using various visualization methods.

3.1.1: Analysis of Categorical Variables:

Categorical Variables play an important role in understanding the composition and characteristics of a given dataset. In this, we focused on two categorical variables namely major and university name. The major variable consists of different streams present in the university like computer science, cyber security so on and the university variable consists of all the university names present in the dataset. In the dataset for major variables by using the value count function we are counting the frequency of each stream. By this, we can generalize what stream most of the students are choosing. It is very essential for contextualizing datasets and gaining useful insights from the datasets.

Similarly we used the value count function to know the frequency of each university present in the datasets.so we can know where the majority of students got admitted and where less no of students are admitted.

3.1.2: Analysis of Numerical Variables:

Numerical variables provide us a quantitative insights of the datasets. In this we have done mainly grouping and aggregating numerical data and calculation of averages.

In this,the dataframe is grouped by the greA column and we have used the aggregate function to know the aggregate statistics of each group. We have also calculated count of records in each group to provide useful insights into distribution of 'greA' scores.we additionally, also calculated averages of numerical columns such as 'toeflScore', 'internExp', 'greV'. to know the central tendency of the data.it helps us to understand values and variability within these variables.

3.1.3: Visualization of University Counts Using Pie Chart:

In this we have generated a pie chart to visualize the distribution of 'greA' scores across different categories.Each slice represents unique greA score and size of slice represents no of records with that score.it enables us for easy comparison and identification of score ranges.

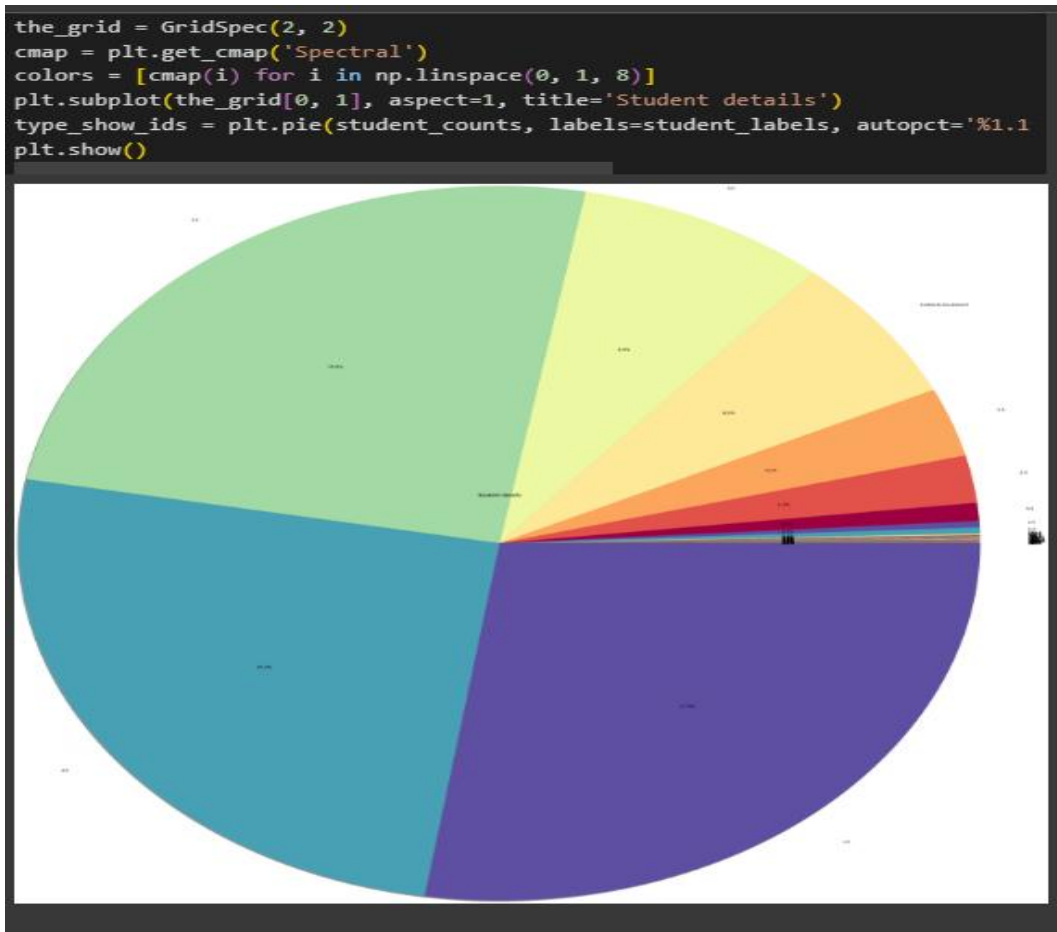


Figure 5: Pie Chart

It plots a graph for values of two variables of a set of data. It is useful in understanding the relationship between each of the columns.

3.1.4: SCATTER PLOT FOR GRE SCORES:

Scatter plot is generated to visualize the relation ship between greQ and greV scores.it allow us to identify correlation between variables and gre scores and also to observe patterns ,trends an outliers of the data.

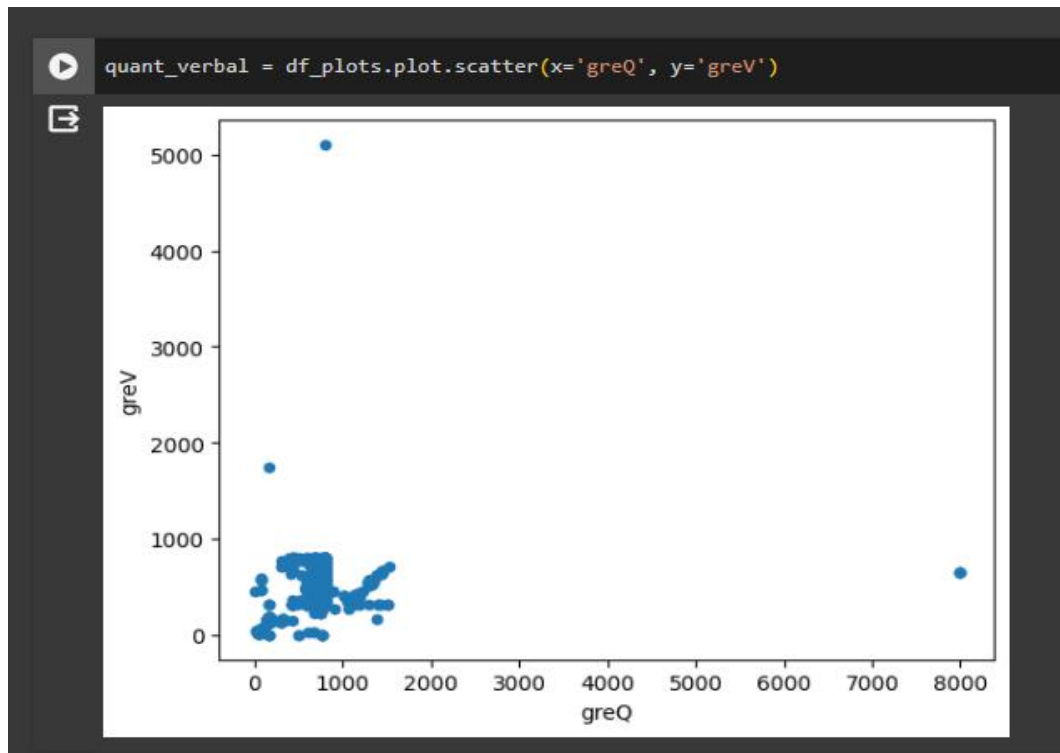


Figure 6: Scatter Plot

Scatter plots a graph for values of two variables of a set of data. It is useful in understanding the relationship between each of the columns.

3.2. Correlation Analysis

3.2.1: Correlation Matrix

Correlation Matrix is used to know the relationship between Numerical variables in the dataset. As in our dataset major, ugCollege and univName are non-numerical variables we can't find relationship of them with other variables.

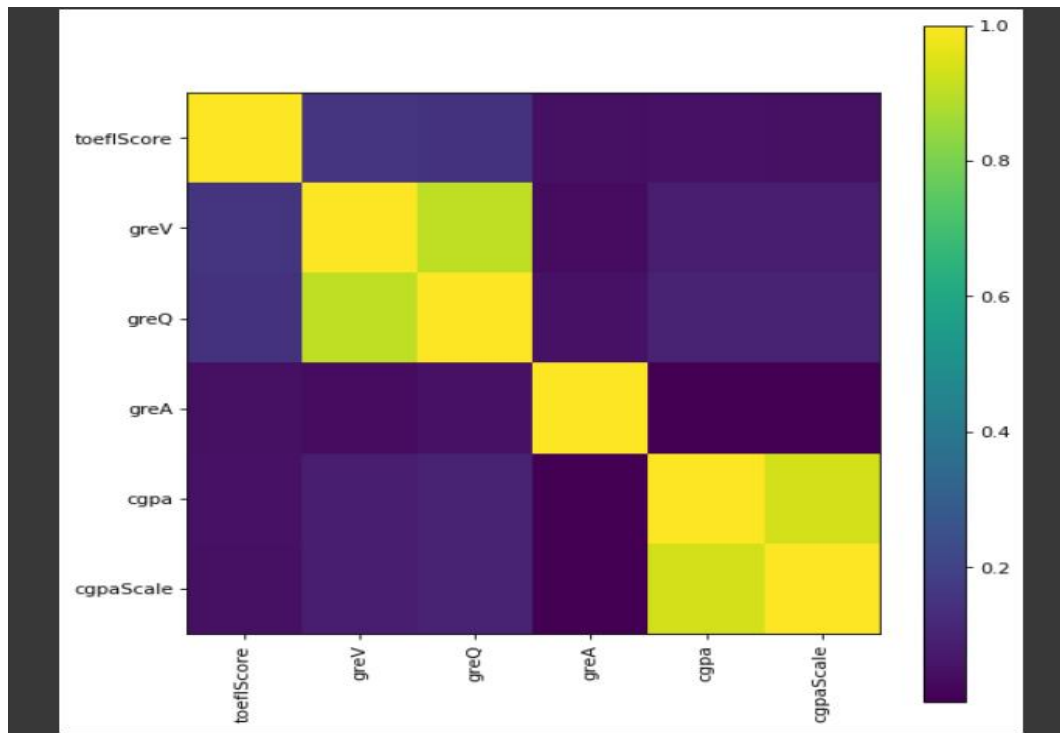


Figure 7: Correlation Matrix

3.2.2: Scatter Matrix plot:

It is a graph of relationship between any two numerical variables in the dataset. It tells us the correlation between variables. If they are highly correlated, moderately correlated, or lowly correlated.

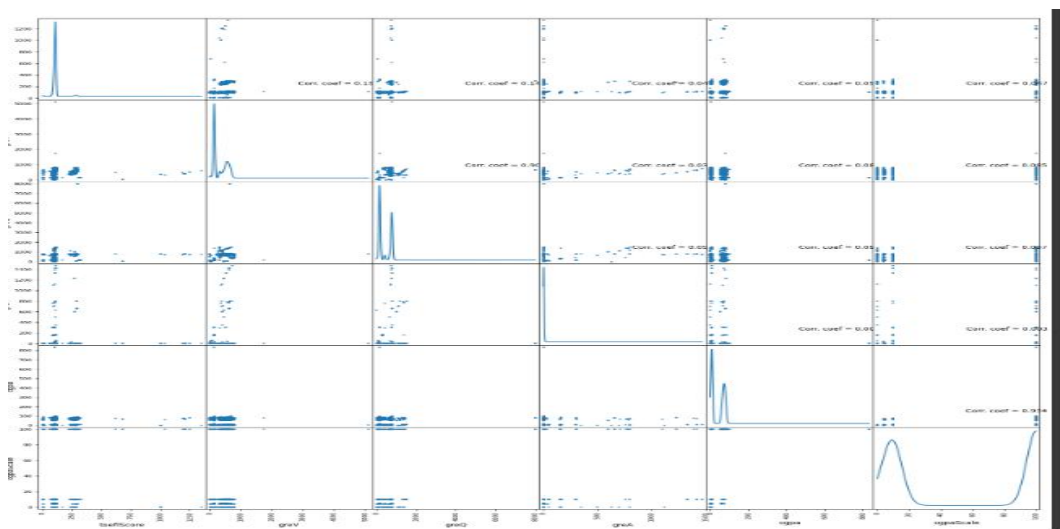


Figure 8: Scatter Matrix Plot

3.2.3:Heatmap Visualization:

Heatmap is created by correlations values of different numerical variables. correlations values lies between -1 and 1. if correlation values close to 1 high positive correlation and close to -1 the high negative correlation.

Heatmap is used in cross-examining multivariate data. it shows variance across each numerical variables .

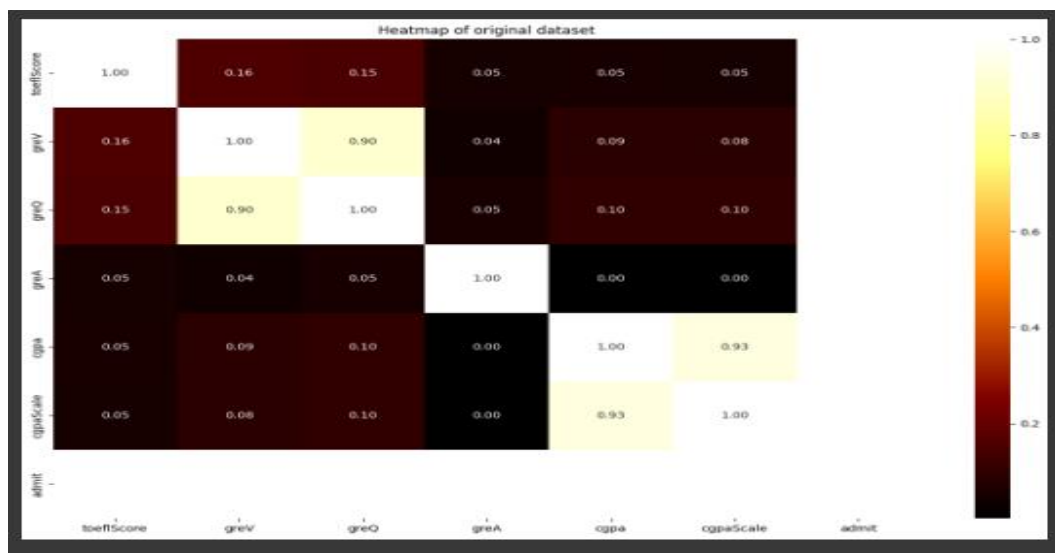


Figure 9: Heatmap

3.3:Word Cloud Analysis:

It is the simplest and most commonly used text visualization .it is visual representation of words with different sizes, colors, or categories. Word cloud is used to find high frequency words in the datasets. Larger the size of the word, the larger is its frequency in dataset.

CHAPTER-4 FEATURE EXTRACTION

To extract the features from the datasets, the data needs to be in same range. To convert the data into same range we have apply data normalization.

4.1: Pre-Processing Categorical Variables:

In this, I have cleaned and preprocess the categorical data by removing hyphens, periods and any special characters present in the datasets. In this first we converted the data in dataframe into strings and stored in a variable. Then this variable is went under Pre-Processing and the converted back into dataframe.

Overall the function takes a dataframe and a categorical feature, cleans the values and then return back to dataframe. In additionally we have preprocessed univName and major variables using the feature extraction function.

4.2: Conversion of GRE Scores:

In this, we have converted old GRE scores into New GRE scores based on score_table. The input taken is either greV or Q. Then I have extracted gre scores from the datasets and stored them in gre_scores. Next we iterated each value over a gre_score. if value is greater then 170 and feature is 'greV' then we will retrieve corresponding value from score_table and update the value in dataframe. Overall, it prepares the categorical variables in dataframe for further analysis .

CHAPTER-5:MODELING

Our model is built to shortlist the best universities, out of the 54 universities from the dataset, based on the student's profile. In this project we are using two classification algorithms: Random Forest Classifier and K-Nearest Neighbor algorithm. These algorithms were built using a combination of all the features mentioned above, to classify a student's profile to the best university among the 54 universities.

5.1:Random Forest Classifier:

Random Forest algorithm is a supervised machine learning algorithm. It uses ensemble learning technique, which means it combines the output of various classifiers to provide solution to complex problems. Random forest algorithm consists of huge number of decision trees and each decision tree gives out a class prediction and the class with the most number of votes automatically becomes the model's prediction.

The data was split as 70% training set and 30% testing set.

```
Accuracy: 0.8681948424068768
Confusion Matrix:
[[633  0  0 ...  0  0  0]
 [ 0 267  4 ...  0  0  0]
 [ 0  4 113 ...  0  0  0]
 ...
 [ 0  0  0 ... 82  0  0]
 [ 0  0  0 ...  8 43  0]
 [ 0  0  0 ...  2  2 22]]
Cross-validation scores: [0.87209005 0.87234587 0.86953185 0.87359263 0.86693961]
Mean CV Accuracy: 0.8709000016496065
```

Figure 11:RF Model

By using Random forest in modelling I got accuracy of 86 %. and validation set I got accuracy of 87%.

```
import pandas as pd

# Sample feature values
sample_values = [[150, 140, 6.8, 10, 50, 100]]

# Predict using the classifier
predicted_univName = clf.predict(sample_values)

# Create a DataFrame with the predicted values
result_df = pd.DataFrame(predicted_univName, columns=['Predicted_univName'])

# Print the DataFrame
print(result_df)
```

	Predicted_univName
0	Wayne State University

Figure 12: RF prediction

While predicting on test data I got output recommendation as Wayne State University.

5.2: Support Vector Machine:

Support Vector Machine (SVM) is an algorithm that belongs to machine learning as well. SVMs are known as high performance pattern classifiers. While Neural Networks aim at minimizing the training error, SVMs have as goal to minimize the “upper bound of the generalization error” . The learning algorithm in this technique is based on classification and regression analysis.

Support Vector Machine is used for classification of both linear and non-linear problems. We plot each data item as a point in n-dimensional space (where n is the number of features we have) with

the value of each feature being the value of a particular coordinate. Later on the classification is performed by finding the hyper-plane that differentiates the two classes.

```
import pandas as pd

# Sample feature values
sample_values = [[150, 140, 6.8, 10, 50, 100]]

# Predict using the classifier
predicted_univName = clf.predict(sample_values)

# Create a DataFrame with the predicted values
result_df = pd.DataFrame(predicted_univName, columns=['Predicted_univName'])

# Print the DataFrame
print(result_df)
```

	Predicted_univName
0	University of Texas Dallas

Figure 14: SVM Prediction

By giving the above test data to the model ,the model predicts the university based on test data and produces output as University of Texas Dallas.

5.3:K - NEAREST NEIGHBOURS:

K-Nearest Neighbor algorithm is a supervised algorithm that assumes similar things exist in close proximity. A object is classified by majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

```

import pandas as pd

# Sample feature values
sample_values = [[150, 140, 6.8, 100, 0, 166]]

# Predict using the classifier
predicted_univName = clf.predict(sample_values)

# Create a DataFrame with the predicted values
result_df = pd.DataFrame(predicted_univName, columns=['Predicted_univName'])

# Print the DataFrame
print(result_df)

```

	Predicted_univName
0	Northeastern University

Figure 16: KNN Prediction

By using this model on test data I got recommendation of Northeastern University as prediction of university.

The Random Forest model as highest accuracy compare to other two models .The Random Forest model predicts the best and accurate university based on students data.The Random Forest model has given 87% training accuracy and it is decreased to 86% on validation data. Followed by SVM model has given 19% accuracy on testing data and 20 % on validation data .Finally KNN model has given 22% on testing data and 21% on validation data. Overall RF model will do best university prediction among all models.

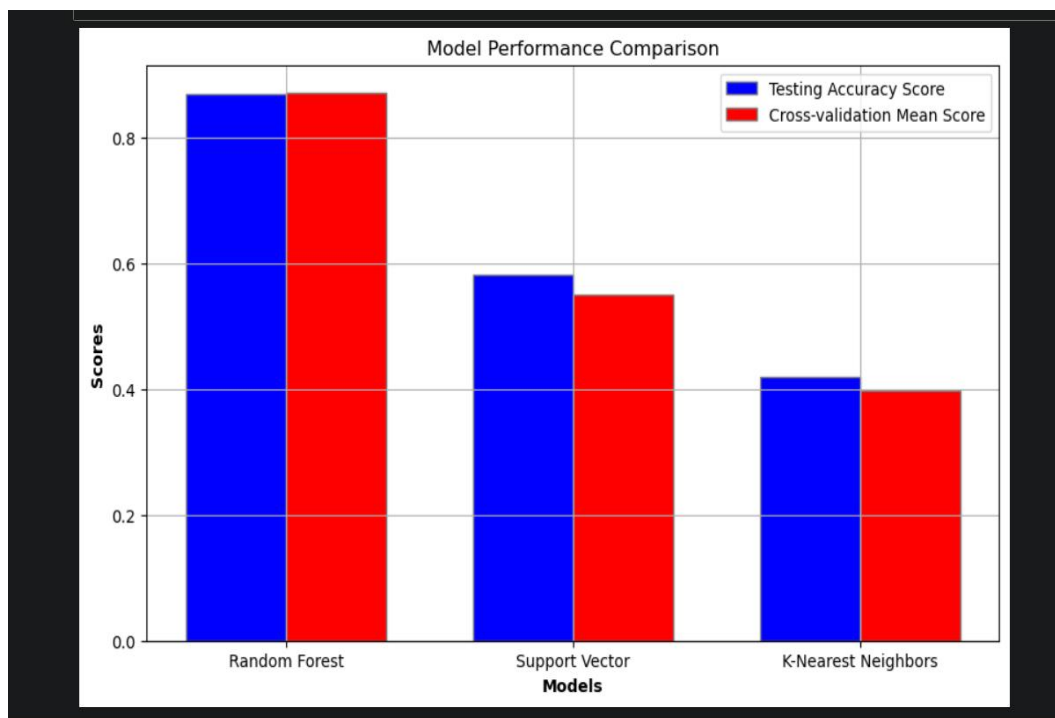
RESULTS:

In this project we have considered Random Forest, Support Vector Machine and K-Nearest Neighbor classification algorithms for recommending universities to students. The performance of the models was compared using the accuracy evaluation metric. The Random Forest algorithm was found to be most accurate with

an accuracy of 86% whereas the SVM and KNN have an accuracy of 19% and 22% respectively.

Model	Accuracy
Random Forest Model	86%
Support Vector Machine	19%
K Nearest Neighbour	22%

Figure 17: Results



CONCLUSION:

This project will help students in shortlisting the best university based on their profile and will save a lot of time for the applicants. The academic data of previously accepted students have been taken into account to recommend the best university for current admission seekers. Random Forest, Support Vector Machine and K-Nearest Neighbor models have been successfully used for building the university recommender system. The Random Forest is found to be the most accurate model among the 3 considered models. The proposed system will ask the student to enter their GRE score (Quant, Verbal & AWA) and CGPA and it will recommend a list of 5 best universities to the applicants.

REFERENCES

- [1]. M. Hassan and M. Hamada, "Smart media-based context-aware recommender systems for learning: A conceptual framework," 2017 16th International Conference on Information Technology Based Higher.
- [2]. Judy, D'cruz, Kathe, Motwani. (2020, April). Recommendation System for Higher Studies using Machine Learning. International Research Journal of Engineering and Technology (IRJET), 07(04), Article e-ISSN:2395-0056.
- [3]. Education and Training (ITHET), Ohrid, 2017, pp. 1-4. Jain, Satia. (2021, December). College Admission Prediction using Ensemble Machine Learning Models. International Research Journal of Engineering and Technology (IRJET), 08(12), Article e-ISSN: 2395-0056. I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4]. Bhatt, Shah, Soni. (2020, July). Recommendation System for Higher Studies at Abroad via Machine Learning Techniques. International Journal of Advanced Research in Science Technology (IJARST), 07(03), Article ISSN (Online) 2581-9429.

[5].M. H. Mohamed, M. H. Khafagy and M. H. Ibrahim, "Recommender Systems Challenges and Solutions Survey," 2019 International Conference on Innovative Trends in Computer Engineering (ITCE), Aswan, Egypt, 2019, pp. 149-155.

[6]. Mahamudul Hasan, Shibbir Ahmed,Deen Md.Abdullah, and Md.Shamimur Rahman, Graduate School Recommender System: Assisting Admission Seekers to Apply for Graduate Studies in Appropriate Graduate Schools, by 978-1-5090-1269-5/16/\$31.00 ©2016 IEEE

[7]. M. Isma'il, U. Haruna, G. Aliyu, I.Abdulmumin, and S. Adamu, "An Autonomous Courses Recommender System for Undergraduate Using Machine Learning Techniques," 2020 Int. Conf. Math. Comput. Eng. Comput.Sci. ICMCECS 2020, no. March, pp.6–11,2020,doi:10.1109/ICMCECS47690.2020.240882

[8]. J. Dhar and A. K. Jodder, “An effective recommendation system to forecast the best educational program using machine learning classification algorithms,” *Ing. des Syst. d’Information*, vol. 25, no. 5, pp. 559–568, 2020, doi:10.18280/ISI.250502

[9]. Alsayed, A.O.; Rahim, M.S.M.; AlBidewi, I.; Hussain, M.; Jabeen, S.H.; Alromema, N.; Hussain, S.; Jibril, M.L. Selection of the right undergraduate major by students using supervised learning techniques. *Appl. Sci.* 2021, 11, 10639.

[10]. El Guabassi, I.; Bousalem, Z.; Marah, R.; Qazdar, A. A Recommender System for Predicting Students’ Admission to a Graduate Program using Machine Learning Algorithms. *Int. J. Online Biomed. Eng.* 2021, 17, 135–147. [CrossRef]

[11]. Stein, S.A.; Weiss, G.M.; Chen, Y.; Leeds, D.D. A College Major Recommendation System. In *Proceedings of the 14th ACM Conference on Recommender Systems, Virtual, 22–26 September 2020*; pp. 640–644

[12]. Roshan, M.; Bhanuse, S.; Yenurkar, G. Recommendation of Branch of Engineering using machine learning. *Int. Res. J. Eng. Technol.* 2020.

[13]. Ouatik, F.; Erritali, M.; Ouatik, F.; Jourhmane, M. Students' Orientation Using Machine Learning and Big Data. *Int. J. Online Biomed. Eng.* 2021, 17, 111–119. [CrossRef]

[14]. Mostafa, L.; Beshir, S. University Selection Model Using Machine Learning Techniques. In *The International Conference on Artificial Intelligence and Computer Vision*; Springer: Cham, Switzerland, 2021; pp. 680–688.

[15]. Mengash, H.A. Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access* 2020, 8, 55462–55470. [CrossRef]

[16]. Dirin, A.; Saballe, C.A. Machine Learning Models to Predict Students' Study Path Selection. *Int. J. Interact. Mob. Technol.* 2022, 16, 158–183. [CrossRef]

[17]. Meng, Y.; Fun, M. CMRS: Towards Intelligent Recommendation for Choosing College Majors; ACM: New York, NY, USA, 2020; Volume 6.

[18]. Ezz, M.; Elshenawy, A. Adaptive Recommendation System Using Machine Learning Algorithms for Predicting Student's Best

Academic Program; Springer Science Business Media: Berlin, Germany, 2020; Volume 25, pp. 2733–2746.

[19]. Baskota, A.; Ng, Y.K. A graduate school recommendation system using the multi-class support vector machine and KNN approaches. In Proceedings of the 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, Salt Lake City, UT, USA, 6–9 July 2018; pp. 277–284.

[20]. Balaji, P.; Alelyani, S.; Qahmash, A.; Mohana, M. Contributions of machine learning models towards student academic performance prediction: A systematic review. *Appl. Sci.* 2021, 11, 10007.

[CrossRef]