VISA PREDICTION FOR HIGHER STUDIES USING MACHINE LEARNING

BY

Md Tipu Sultan
ID: 162-15-7758

Sk Hasibul Islam Shad
ID: 162-15-7748

AND

Asif Ahmmed
ID: 162-15-7862

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Jueal Mia
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Shah Md. Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JULY 2020

APPROVAL

This Project titled “**Visa Prediction for Higher Studies using Machine Learning**”, submitted by **Md Tipu Sultan, Sk Hasibul Islam Shad**, and **Asif Ahmmed** ,ID No:162-15-7758,162-15-7748 and 162-15-7862 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 8/10/2020.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain **Chairman**

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

ফিজার আহমেদ

Dr. Fizar Ahmed **Internal Examiner**

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Abdus Sattar **Internal Examiner**

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Dr. Mohammad Shorif Uddin **External Examiner**

Professor

Department of Computer Science and Engineering

Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Md. Jueal Mia, Senior Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

9.10.20

Md. Jueal Mia
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Tsiddique

Shah Md. Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Tipu Sultan

Md Tipu Sultan
ID: -162-15-7758
Department of CSE
Daffodil International University

Sk Hasibul Islam Shad
ID: -162-15-7748
Department of CSE
Daffodil International University

Asif Ahmmed

Asif Ahmmed
ID: -162-15-7862
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Md. Jueal Mia**, **Senior Lecturer**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Md Juel Mia, Shah Md. Tanvir Siddiquee, and Dr. Syed Akhter Hossain, Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Computer science is arguably one of the most common fields across both Bangladesh and the world today. It is obvious that a statistically significant percentage of learners struggle to achieve the peak of this discipline due to the lack of skill in this discipline. Without a doubt, one of the most popular studies is going abroad for higher studies. It is really necessary for students to choose the correct path before applying for a higher education visa in order to succeed. In this work, we predict the visa for higher studies based on student's information. Then we process those data (like; cleaning, transformation, integration, standardization, feature selection). Later we used different classification techniques i.e. C4.5 (j48), K-NN, Naive Bayes, Random Forest, SVM, Neural Network to classify these profiles. Based on the result analysis, it has been found that accuracy and other factors of a confusion matrix for Random Forest classifier are more cogent than others. We also find out the attributes upon which a student's visa accepted depends mostly. Therefore, GRE score, Undergraduate CGPA, are two of the most important factors to determine success in the visa approving for higher studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Nowadays getting a visa is very tough for a student who is applying abroad for studying. Many students in Bangladesh apply for a visa but many of them are rejected because of lacking analysis of their previous academic or non-academic work. Higher Studies abroad is a dream most of the students in a developing country. Bangladesh Government has already taken many challenges for the Scholarship Programs for Higher studies in abroad.

Bangladesh Government recently interconnecting many developed countries to help Bangladeshi Student get visas for higher studies. The latest UNESCO Institute for Statistics data shows that 60,390 Bangladeshis were pursuing higher education abroad in 2017.UNESCO data shows that in 2017 a total of 34,155 Bangladeshis enrolled at universities in Malaysia, 5,441 in the United States, 4,652 in Australia, 3,599 in the United Kingdom, 2028 in Canada, 2008 in Germany, 1099 in India, 870 in Saudi Arabia, 810 in Japan and 637 in the United Arab Emirates [1].

The purpose of this work is to build a classification model to classify which students have a chance to accept their visa and which doesn't. We evaluate their profiles before jumping into this program using their academic results, job experience, research papers, and IELTS - TOFEL, GRE - GMAT scores, etc. To predict their visa it's always important to know about their previous academic result. IELTS is the internationally recognized test most generally used and approved for that [2]. Besides academic results, to make our proposed classifier more accurate in this work we have collected data from different universities in Bangladesh which includes student's academic results, job experience, and research experience they have already faced. Firstly, we preprocessing our data set in different steps, such as cleaning, transformation, integration, standardization, and feature selection. Afterwards, we further labelled the dataset and built a classification model and apply different types of algorithms to predict the student's visa YES or NO for Higher studies.

Selecting the right study track for an abroad study is very significant for every student because this is something that determines the academic and professional achievement of a student. Nowadays getting a visa is very tough for a student who is applying abroad for studying.

1.2 Motivation

Many students wish to go overseas to seek higher education however, they will and sometimes struggle to do this several times. Since each nation has a certain number of criteria, heading out of the country relies on its processing [2]. They will go to the nation if they can fill up their recruitment otherwise, they can't go. Most of the times they can't fulfil the criteria because of a lack of careful study, and their results, so they are denied rather than accepting visas much of the time. Many researchers collaborated through their data, university data, and a lot of knowledge to help students choose the best choice [1]. We aim to build a framework between those who earn visas and those who do not obtain visas. So that a statement regarding visa acceptance and visa rejection can be made available to the public.

1.3 Rationale of the Study

The main purpose of our work is to focus on visa approved or not for higher studies outside the country, but the problem is that this is not an easy task at all, it requires the use of a variety of algorithms. Data mining will be used here. Through this, we can easily understand how a person can easily get a visa for higher studies? And there is nothing according to the requirement and if something is wrong, visa rejection can happen. Classification and frequency generation algorithms are used here, through which the tasks can be filled very easily. Everyone should keep pace with the present time That requires the right path [3]. People can go abroad for higher education in two ways One way is the scholarship. The other is Without Scholarship. Here will be given ideas about the Full Scholarship and Partial Scholarship. Currently, people in the Middle East are heading abroad for higher education, here you will find the complete idea of why you are getting visas and why you are not getting it. What is the reason a person has a visa and what is the reason a person is not getting a visa?

It basically depends on a lot of things. As like, GRE SCORE, IELTS/TOEFL result, academic result Someone's might or research paper Someone again depends on the subject, there are many more reasons for this.

A successful result can be easily achieved by using this data mining technique model, so it is easy to get full results using algorithms. And a great outcome is described in it in a very unique way.

1.4 Research Question

1. How can we define higher studies visas for scholarships?
2. What are the processes of data cleaning and data levelling in machine learning?
3. Which classification algorithm will we use to predict higher studies visas in data mining technique?
4. How can we use it in data collecting and pre-processing in the KDD process?
5. Why do we use Weka for best accuracy for data training and predicting?
6. How many tools do we use on better accuracy of predictive models and descriptive models in this research?

1.5 Expected Output

The main purpose of our work is to create a model to predict data that can easily get a higher studies visa. In this model case, we collect different types of data such as GRE, Undergraduate result, IELTS/TOEFL, scholarship, and more real-time data we use. We use different types of algorithms such as Random Forest, Decision Tree, Naive Bayes, *K*-NN, SVM, Neural Network for finding the accuracy, and after that, we found the best accuracy algorithm to average all algorithms. With a lot of data mining process can generate a model that could give us accurate detection of the higher studies student visa prediction every time.

This will help students to improve themselves according to the requirement of the university which they chose for their higher studies. As there is no project like this before, this will give a great revolution on higher studies for students.

1.6 Report Layout

We have decorated the whole paper with six chapters. For the specific implementation and extended description, each chapter was properly explained to make the concept easier in terms of understanding.   The first chapter represents the knowledge behind the study's motivation and also draws a picture of the study's reasoning for depicting the anticipated outcome. In the second chapter, we clarified the context of our study and listed some related studies. We have tried to explain some problems with previous study and the challenges we faced to solve those problems. After the second chapter, the third chapter has been made with the elaborate methodology of our work. We have tried to explain everything from the scratch. We explained the pre-processing of data, the architecture of models and classifiers. Then the statistical analogy for the research and also the re-equipment for the work. We have outlined the outcome in the fourth chapter. Finally, an overall brief but descriptive description of our research work was reflected at the end and a conclusion was drawn. The implication for future studies has been described with an accurate and rightful explanation to exploit the findings of the research work and technique for further usability.

CHAPTER 2

BACKGROUND

2.1 Preliminaries/Terminologies

To travel from our country to foreign countries, we must have a visa. There are eight kinds of visas: Business/Tourist Visa, Work Visa, Student Visa, Exchange Visitor Visa, Transit/Ship Crew Visa, Religious Worker Visa, Domestic Employee Visa, Journalist and Media Visa [4]. We must head in the direction in which we can use our talents to their fullest and contribute to society and work in securing a scholarship after discovering and learning the properties of our abilities and work. Here, we are going to talk about various data mining and Machine learning methods that are used to help students make decisions about applying for a scholarship abroad.

Over the years, researchers have concentrated on guiding students to the right track from various regions of Bangladeshi undergraduate students before selecting any higher type of education and scholarship before applying for a visa. It is possible to categorize the current terms in two approaches [5].

1. Data Mining for different research tracks.

2. Data Mining approaches to applying for higher studies and predicting student visas.

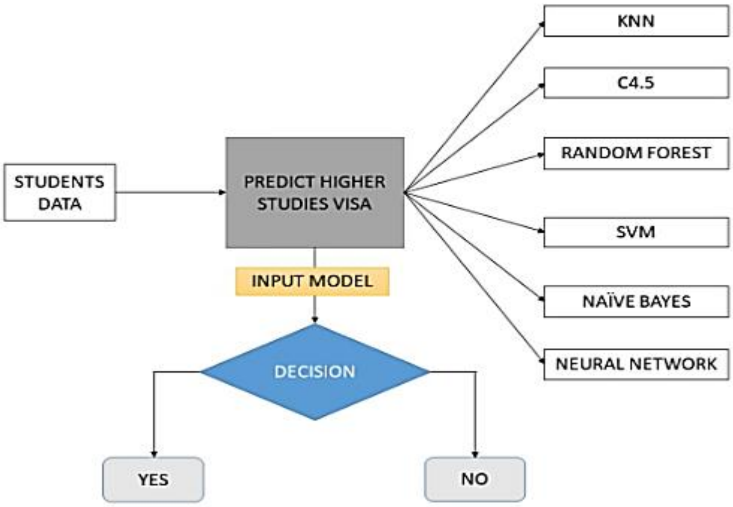


Figure 2.1.1: Operational Phase

2.2 Related Works

A few papers have been published in recent years 2017-19 in which some work like ours has been performed using machine learning. But in the last few years, a lot of methods have been developed using machine learning and classification.

Very recently, Md. Jueal Mia, et al. they proposed a recommendation system registration Status Prediction of Students Using Machine Learning. They have applied seven classifiers rules, including SVM, Naive Bayes, Logistic, JRip, J48, Multilayer Perceptron, and Random Forest [6]. They considered SVM to be the highest at 85.76% accuracy, while Random Forest reached the lowest at 79.65% [6].

Al Amin Biswas, et al. they proposed a recommendation system which, Predict the Enrollment and Dropout of Students using Machine Learning Classifier. They have applied seven classifiers rules, including Naïve Bayes, Multilayer Perceptron, Logistic, Locally Weighted Learning (LWL), Random Forest, Random Tree, and Part are applied in this context [7]. They considered LWL to be the highest at 86.36% accuracy, while Random Tree reached the lowest at 74.24% [7].

Mehrbakhsh Nilashi, et al. They proposed a recommendation system for the tourism industry using cluster ensemble and prediction machine learning techniques. As prediction techniques, they use Adaptive Neuro-Fuzzy Inference Systems (ANFIS) and Support Vector Regression (SVR), Principal Component Analysis (PCA) as a strategy for minimizing dimensionality, and Self-Organizing Map (SOM) and Expectation-Maximization (EM) as two well-known techniques of clustering [8]. Their studies conclude that cluster sets can have greater predictive precision in comparison to methods that focus entirely on single clustering techniques for the proposed recommendation system [8].

Md Aref BILLAH, et al. they proposed a recommendation system which are the factors of contributing programming skill and CGPA as a CS graduate by using Mining Educational Data [1]. They have applied Decision tree, Support Vector Machine and Naive Bayes Classifier algorithm on student’s academic results. They considered LWL to be the highest at 86.36% accuracy, while Random Tree reached the lowest at 74.24% [1].

D Kurniadi, et al. they proposed a recommendation system of predicting scholarship recipients using k-Nearest algorithm [1]. They implement the algorithm model of k-Nearest Neighbor (k-NN) for predicting. The k-NN algorithm with the highest accuracy score of 95.83% in predicting students who have the greatest chance of receiving the scholarship [1].

2.3 Comparative Analysis and Summary

The comparative performance of all the work is shown in table 4.2.8. To estimate the performance of the machine vision-based expert prediction method, we must compare the related research recently reported in this context. We disclose from the literature review that much of the research work has some weaknesses and fails to explain the method of data analysis without appropriate results and dataset. Most of them run on very few datasets. That's why evaluating the success of work with appropriate research is a challenging problem for them. We made an effort to review all the output of the paper related to our work. That is why it's a difficult problem for them to calculate the Work efficiency with appropriate analysis.

2.4 Scope of the Problem

Especially for someone who is not ready to take those challenges is sure to suffer visa approval and scholarship from abroad. Bangladesh Different department students of undergraduate don't know properly how to get a scholarship for going abroad and how to manage funds. Our building classification model which can help them make decisions which track they should follow for the higher study visas.

2.5 Challenges

We considered IT, Engineering, Business, Pure science, and other departments as our purpose of getting an approved visa for its immeasurable popularity and the kind of challenges it encounters. This research deals with Four Hundred Data in Undergraduate Student's many information and many of graduate student’s information. This Data requires pre-processing before feed them into a Data Mining and Machine Learning technique for training [7]. To sum it up a strong hardware-based computational system is required for this research. Our computer was able to keep up with the hardware but barely. As a result, each iteration took a huge amount of time while training. The biggest challenge for us was to learn Machine learning and Data Cleaning because we were completely new in this sector as we did not have any prior knowledge about it. We started from collecting data in various ways and spend a lot of time-solving our questions which kept generating on our mind as we started learning. We study more and more before implementing and we learn lots of things.so now after doing hard work we can solve any problem on this issue.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

Our research subject is an approved visa for student scholarship. We have used a few data mining techniques and existing algorithms and tested their efficiency through various techniques. To do this we have used a few tools [1].

To implement and verify our proposed work Orange which is available in Anaconda and Weka is employed. These tools are widely used as a predicting platform of intelligence and decision making. These tools contain a large number of data mining and machine learning advanced level algorithm which helps us to predict and make good decisions. We also used IBM SPSS statics to take data and process statistical analysis of data taken from students [5].

Orange

Orange is a component-based data mining software. Distribution is one of the best widgets to identify important features for the dataset [10],[3].

Advantages of Orange include:

1. Orange reads Google Sheets.
2. SQL data can stay on the database server
3. Colour Your Data.
4. Improved data pre-processing.
5. Focus on interactive visualization.
6. NumPy based data storage.

Weka

Weka is a widely used machine learning tool developed in java for desktop. It gives a lot of opportunities to use various types of machine learning algorithms [11].

Advantages of Weka include:

1. It's free software available to download
2. It's implemented in java and it has the portability
3. Suitable in any platform
4. A numerous collection of data mining tools and techniques

Easy to use

IBM SPSS Statistics

It supports different data mining tasks including Data preprocessing, various classification, automatic and manual clustering, feature selection, regression analysis and visualization. SPSS Statistics is mainly used for batches and various statistical analysis. It's a widely used tool for measurement of the statistical overview of data. IBM gained it back in 2009. [3] It is now used by government, social scientists, marketing analyst, educational researcher and data miner. Descriptive data mining model like summarization can be used to produce an automated report on dataset [12].

It contains

1. Descriptive static overview
2. Predictive tools
3. Cross-validation
4. Identifying groups i.e. k-means clustering
5. Regression analysis and other numerical outcomes

3.2 Data Collection Procedure/Dataset Utilized

The most difficult and significant method of this research is data collection. Data is collected through a questionnaire using various tools i.e.: IBM SPSS Statistics. After this process, all the available data is fetched in a data set and saved in a recognizable format. Training data is collected from existing students of undergraduate students. We have collected data from the students of different universities doing their different departments who applied their scholarship for getting a visa for abroad in different countries. For collecting data, we have divided our student's full information into several criteria using our domain knowledge i.e.: Undergraduate Departments, City, CGPA, Job experience etc. [13],[14],[15],[16].

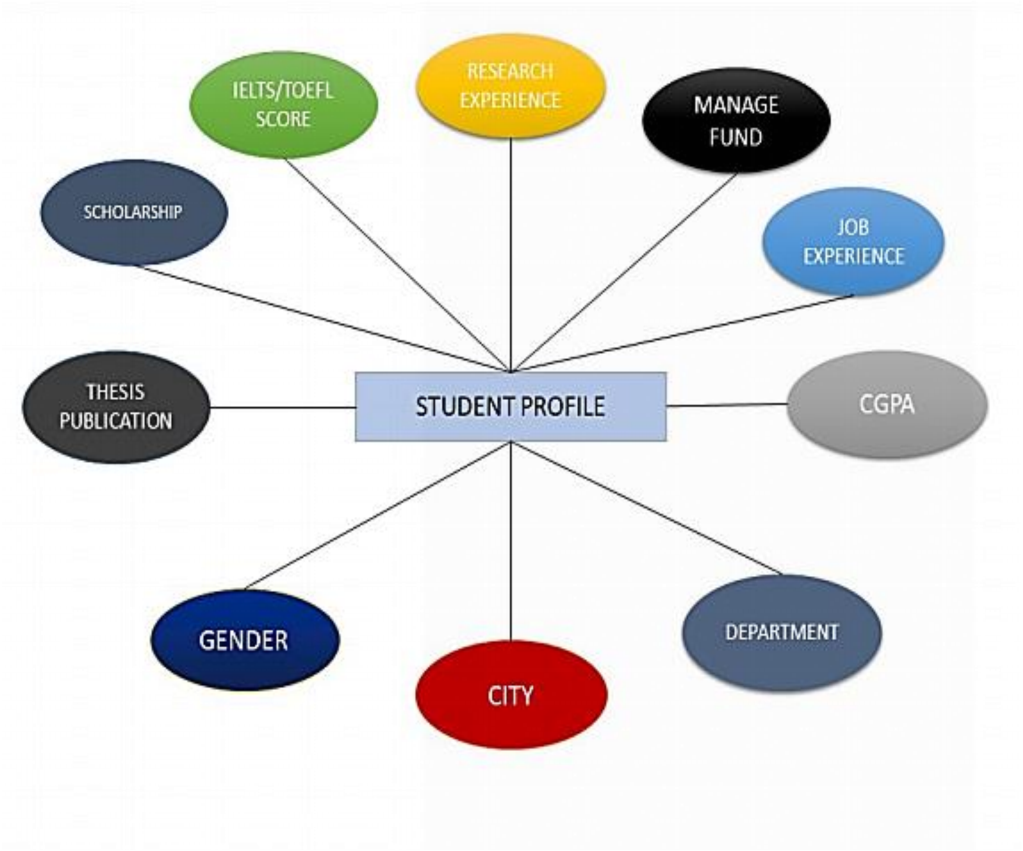


Figure 3.2.1 shows the overview of building student's profile.

3.3 Statistical Analysis

Here most of the student's Applied for Higher Study university location is the USA which is 65%. 12% is from Canada. 9% is from Australia, 6% from the UK, and 8% from Germany [17],[18].

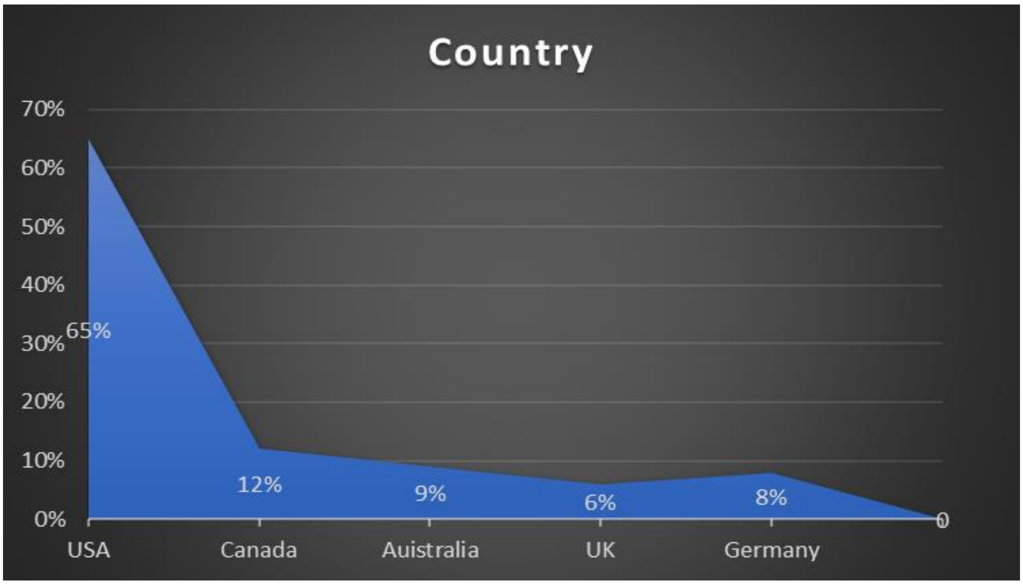


Figure 3.3.1: Area Statistic for Country

.

306 of 400 students are male and 94 is female. That is 76% students are male and 24% students are female [17],[18].

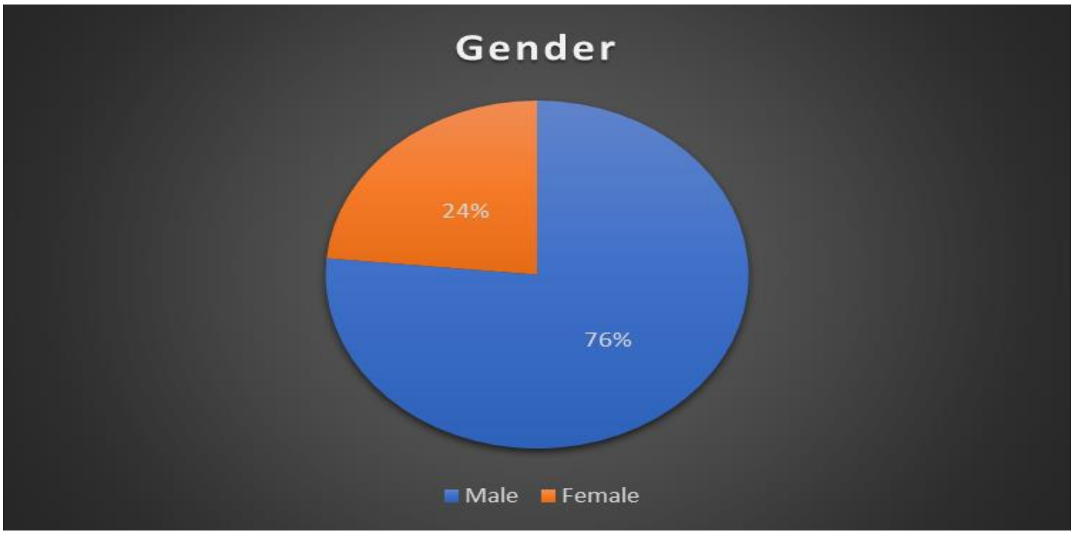


Figure 3.3.2: Gender Pie Chart

Here, Most Students from the Engineering Department.IT and Business Department are the Second and Third highest Department. Pure Science Department is 53 out of 400 students and Medical is 25 and Another Department is 34 out of 400 students [17],[18].

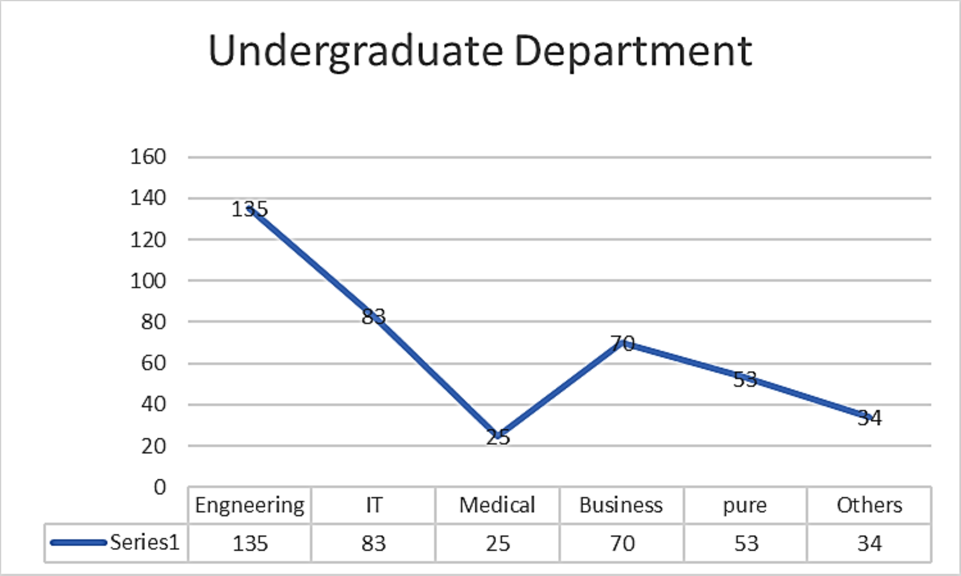


Figure3.3.3: Undergraduate Department Line Graph

3.4 Proposed Methodology/Applied Mechanism

As a machine learning approach our proposed method has mainly two phases. One is the “Build Phase” and the other is “Operational Phase” [5],[19].

The KDD process build phase of our proposed method has seven stages.

1.Data collection

2.Pre-processing

i)Data cleaning

ii)Transformation

iii)Integration

iv)Standardization

v)Feature selection

3.Data mining, model generation and Performance measurement of algorithms

4.Finally we will get a model to use.

Shows the overview of the build phase of our proposed model [3].

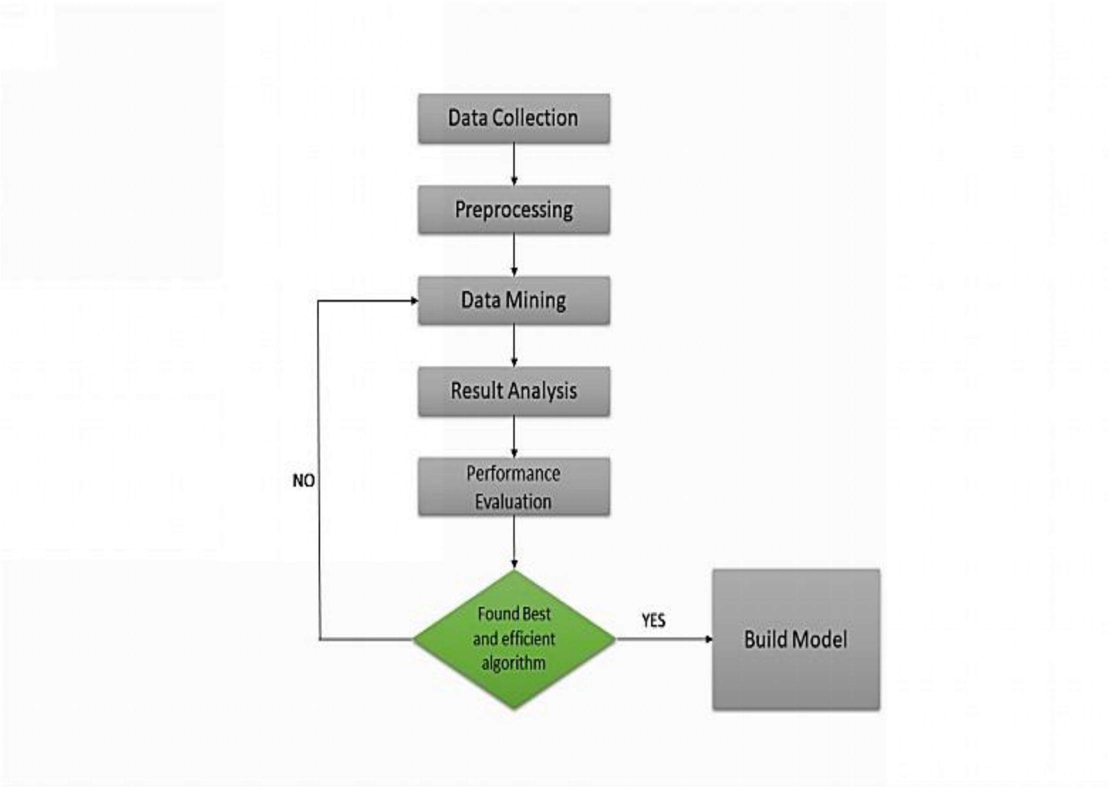


Figure 3.4.1: Overview the Build-up process

Data Collection

Preprocessing

Often collected data is not understandable, inconsistent, lacking in important criteria or can contain various errors. Preprocessing makes data understandable by various process and solve those issues [2],[5].

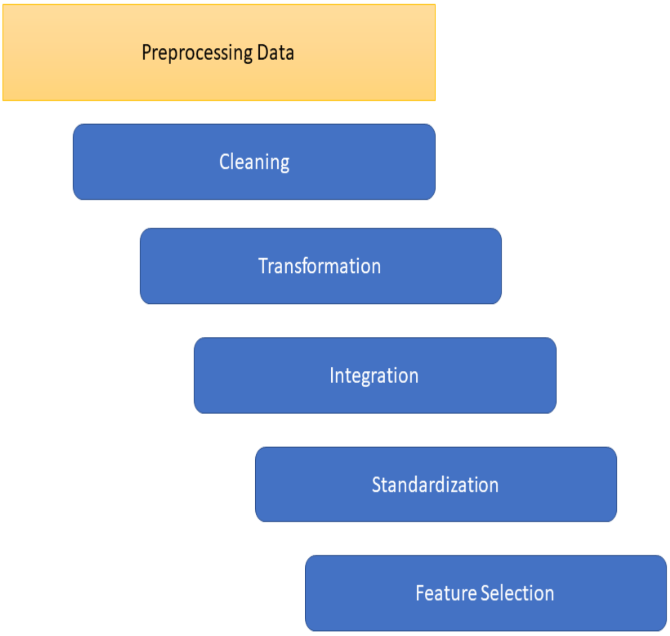


Figure 3.4.2 shows the steps of data preprocessing

Data cleaning

To improve the quality of data, data clearing is important. Sometimes redundant data takes places in the dataset. Or there can be inconsistent data [5].

Transformation

Data is aggregated by various methods. Also normalized and generalized to use the data Efficiently [3].

Integration

There can be a conflict between data in different places in the data set. This problem should be solved. It's known as the data integration process [3],[5].

Standardization

Standardization is the process to bring a dataset into a common format which is needed for cross-checking, research and large-scale analytics. After preprocessing pure and error-free data is ready for mining and further processing to create a model and predict. Our proposed algorithm is *K*-NN, Naive Bayes, Random forest, SVM for classification. **Figure 3.4.3** shows proposed algorithms for data mining process [3],[5],[7].

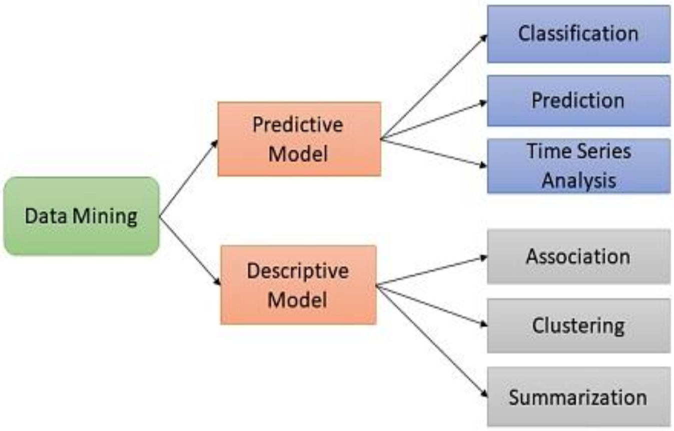


Figure 3.4.3: Data mining techniques

3.5 Implementation Requirements

In this research, many data mining implementations requirements are needed for us. Data mining techniques are also used to discover hidden patterns from large volumes of data. It is mostly applied to computer decision supporting systems, AI, business intelligence, and information processing. The data mining techniques are featured to create a model that will help to find new data using unknown data [3]. Data mining can be basically of two types- Predictive and Descriptive. Predictive techniques use a known data set for analysis and gather detailed information about that database. Classification, regression, time series analysis, the prediction is predictive. The descriptive technique finds patterns and relations in datasets. Clustering, sequence analysis, summarization, and association rules included in descriptive techniques.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

We use the training dataset to get better boundary conditions that could be used to determine each target class. Once the boundary conditions are determined, the next task is to predict the target class. The whole process is known as classification.

K-NN

K-nearest neighbour is the simplest supervised learning algorithm. It's a not parametric and lazy learning algorithm. Normally datasets are separated in several classes and the work of *K*-NN is to learn from these training datasets and predict future data [12].

Naive Bayes

Naive Bayes lies in a probabilistic class algorithm. Suppose there are 20 independent variables in a model. Naive Bayes takes into account only one variable at a time. It's not only an algorithm but it refers to a full set of algorithms [3],[12].

Random Forest

Random forest is an assembling method. It works with a multitude of decision trees. It uses the decision tree algorithm and the tree bagging but the difference is they use overlearning. Figure 5.3 below shows sample figure output of random forest trees [12].

C 4.5

It's a decision tree, based algorithm for classification of both numeric and nominal classes. It was written by J. R. Quinlan [3],[5].

SVM

Support Vector Machine (SVM) is an algorithm for supervised machine learning that can be used for both classification and regression problems. It uses a technique called the kernel trick to transform your knowledge and then finds an optimal boundary between the possible outputs based on these transformations [3],[5],[7].

Neural Network

A neural network is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial [12].

4.2 Experimental Results & Analysis

Performance of classification models usually evaluated by a confusion matrix. Confusion matrix contains information about original and predicted classification done by a classifier [3]. Table 4.2.1 shows the context of confusion matrix-

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Table 4.2.1: Confusion Matrix

Here,

TP = True Positive = Predicted as positive and originally member of positive class

FP = False Positive = Predicted as positive but originally member of negative class

FN = False Negative = Predicted as negative but originally member of positive class

TN = True Negative = Predicted as negative and originally member of negative class

Several standard terms for evaluating by confusion matrix for two class-

Accuracy	$ACC = (TP + TN) / (P + N)$
Sensitivity or Recall or True Positive rate	$TPR = TP / (TP + FN)$
Specificity or True Negative rate	$SPC = TN / (FP + TN)$
False Positive Rate	$FPR = FP / (FP + TN)$
False Negative Rate	$FNR = FN / (FN + TP)$
Precision	$PPV = TP / (TP + FP)$
F1 Score or F-Measure	$F1 = 2TP / (2TP + FP + FN)$

Performance Evaluation of Classification Algorithms [17],[18].

Table 4.2.2: confusion matrix of different classifier algorithms.

Classifier	YES	NO	
C4.5(J48)	192	8	YES
	10	190	NO
Naive Bayes	180	20	YES
	8	192	NO
Neural Network	188	12	YES
	9	191	NO
	194	6	YES

Random forest	6	194	NO
SVM (Poly Kernel)	187	13	YES
	8	192	NO
K-NN	187	13	YES
	7	193	NO

Table 4.2.3: Accuracy rate of various classifier algorithm

Classifier	Precision	Recall	F-Measure	ROC Area
K-NN	0.950	0.950	0.950	0.959
Naive Bayes	0.932	0.930	0.930	0.965
Neural Network	0.948	0.984	0.947	0.980
Random forest tree	0.970	0.970	0.970	0.994
SVM (Logistic)	0.948	0.948	0.947	0.948
C4.5/J48	0.955	0.955	0.955	0.947

Table 4.2.4 Accuracy rate of various classifier algorithm

Algorithm	Accuracy
K-NN	95%
Naïve Bayes	93%
Random Forest	97%
Neural Network	94.75%
SVM (Logistic)	94.75%
C4.5/J48	95.5%

According to the graph above *K*-NN 95%, Naive Bayes 93% Which is less than Random Forest 97% which is above all, Neural Network and SVM are same 94.75% and C4.5 95.5%. So, the best is the Random Forest Model, since it is above all [17],[18].

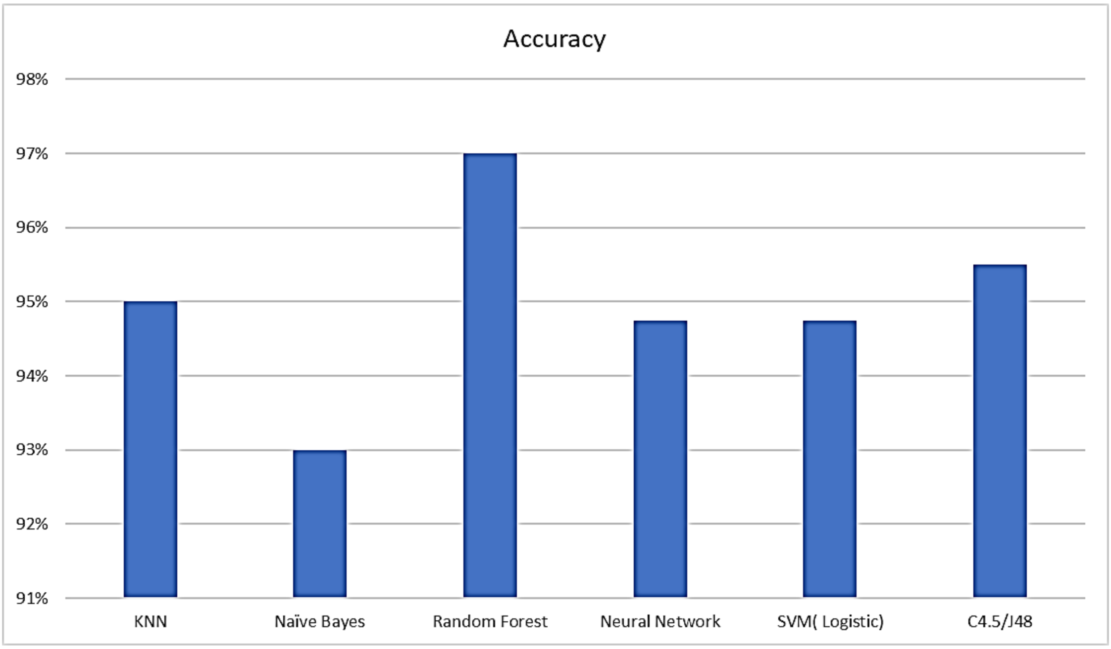


Figure 4.2.1: Best Accuracy Bar Chart

We have collected data, first of all, it was hard work to because Visa data cannot be obtained very quickly. For this task, we had to do an online survey and we had to look for a lot of Facebook sites [13],[14],[15].

Then the information needs to be manually analyzed. Because it was all supervised by our results, we followed the rules of classification. For this, we used Wake and we used several rules [20].

These are Naive Bayes, c4.5, Random Tree, *K*-NN, SVM and Neural Network, and we took the best six. Of all the rules, the best accuracy and the best model is Random Forest [5].

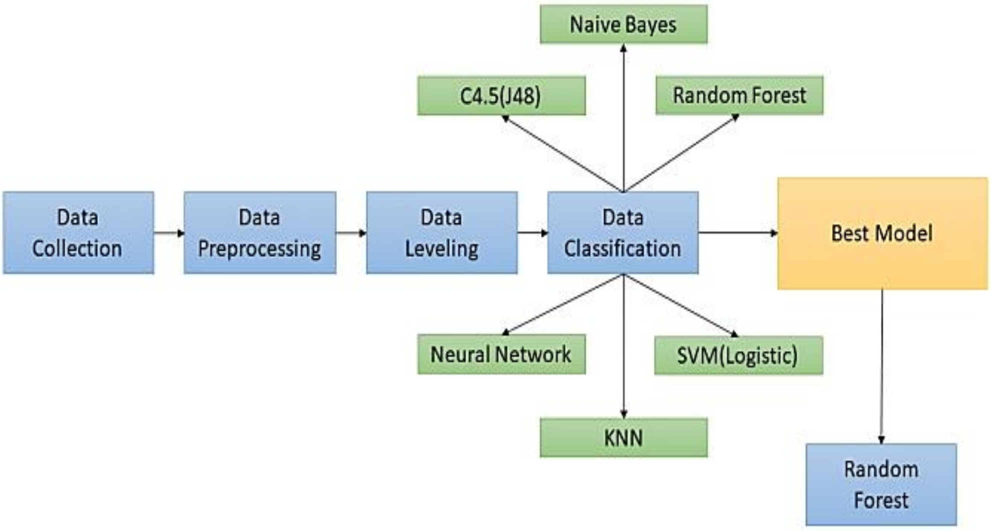


Figure 4.2.2: Finding the Best Model

ROC Curve

A ROC curve shows the relationship between clinical sensitivity and specificity for every possible cut-off. The ROC curve is a graph with: The x-axis showing 1 – specificity (= false-positive fraction = FP/(FP+TN)) [12],[20].

A model with perfect skill is represented by a line that travels from the bottom left of the plot to the top left and then across the top to the top right [3]. An operator may plot the ROC curve for the final model and choose a threshold that gives a desirable balance between the false positives and false negatives [5].

Given below our ROC Curve in Predicting model for Higher Studies Visa Prediction

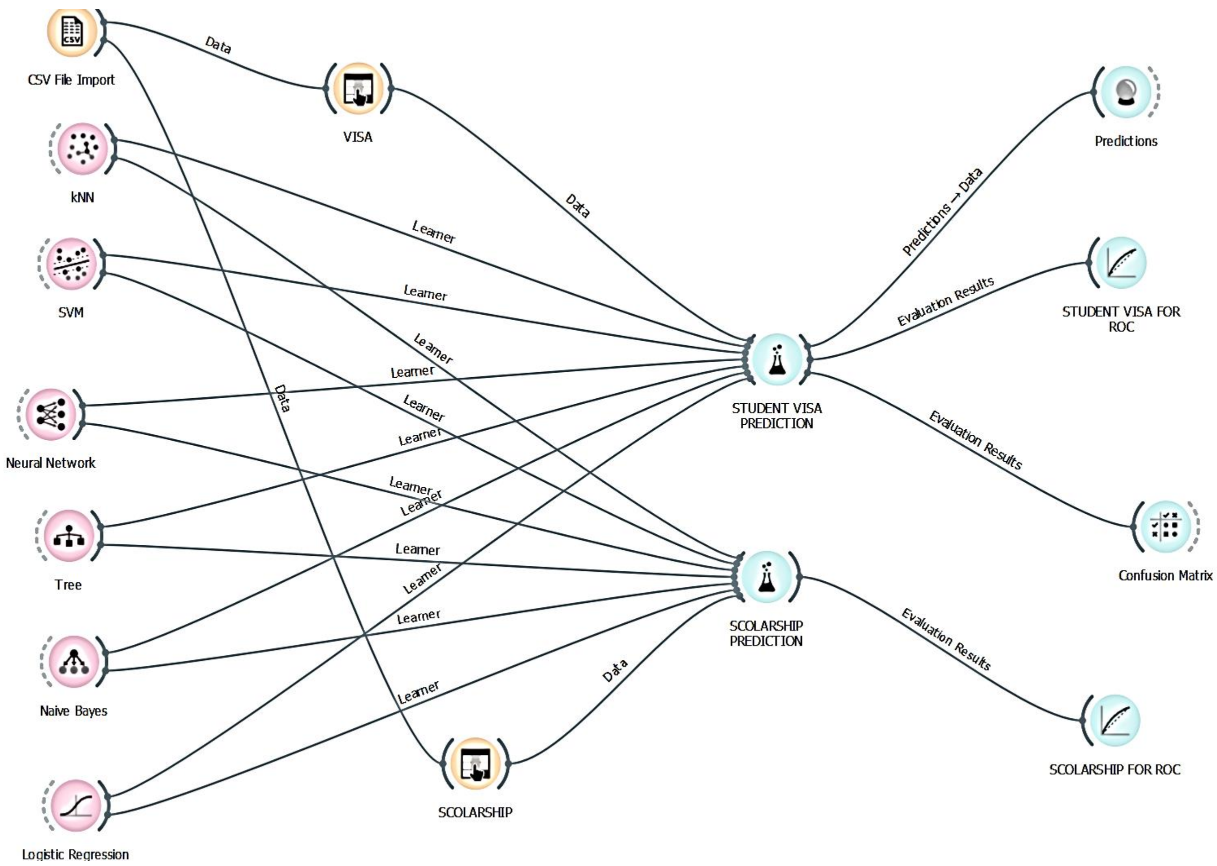
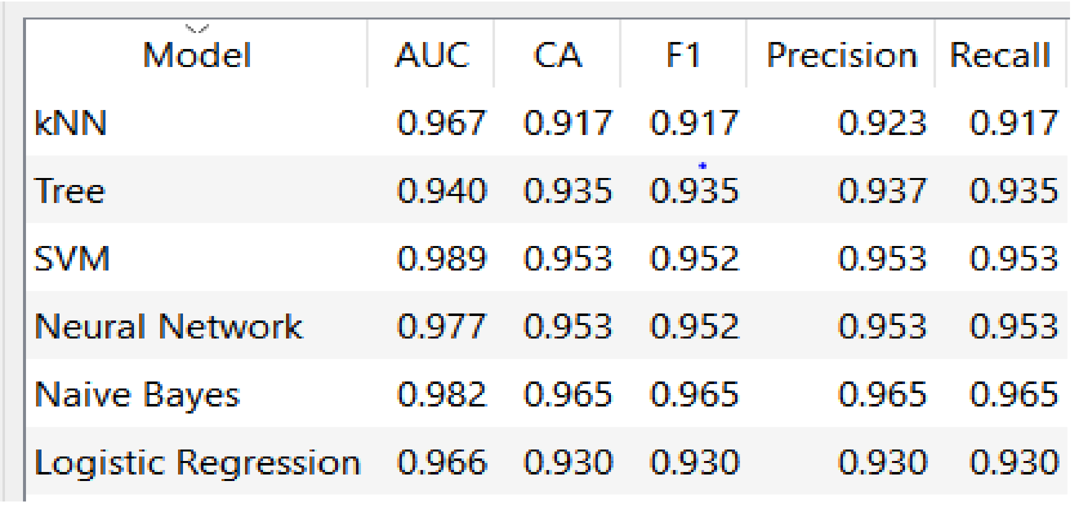


Figure 4.2.3: Prediction Model for Predicting Student Visa and Scholarship



Model	AUC	CA	F1	Precision	Recall
kNN	0.967	0.917	0.917	0.923	0.917
Tree	0.940	0.935	0.935	0.937	0.935
SVM	0.989	0.953	0.952	0.953	0.953
Neural Network	0.977	0.953	0.952	0.953	0.953
Naive Bayes	0.982	0.965	0.965	0.965	0.965
Logistic Regression	0.966	0.930	0.930	0.930	0.930

Figure 4.2.4: AUC Result Evaluation

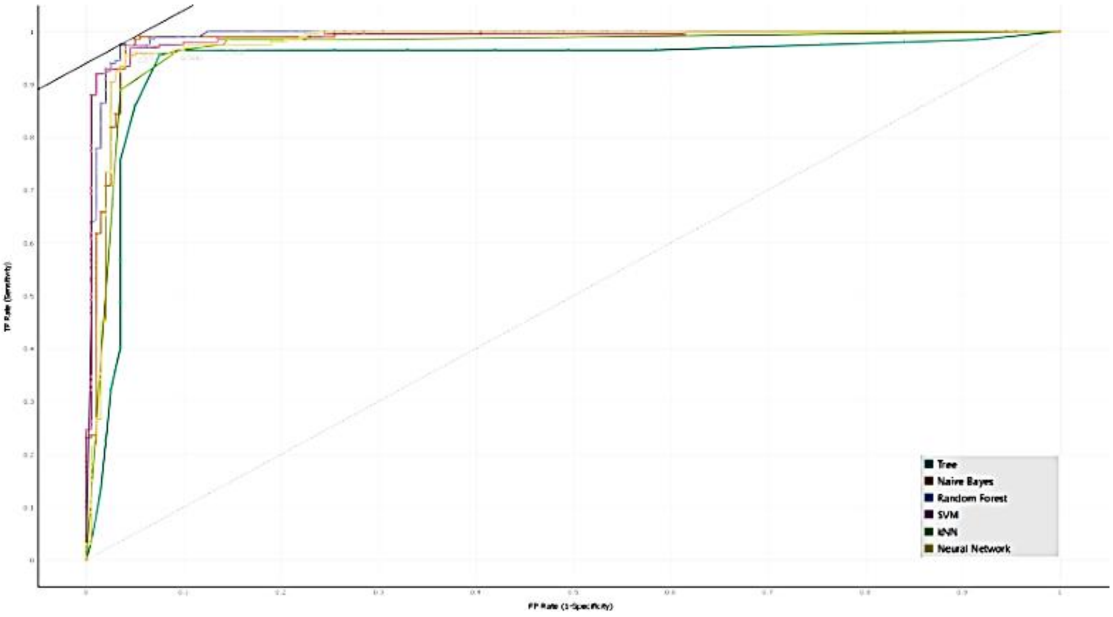


Figure 4.2.5: ROC Curve for Student Visa

LIST OF SELECTED FEATURES

In our model, we have selected important attributes by Gain Ratio (GR) setting the threshold at 2.0 algorithm and searching by ranker algorithm [12].

When analyzing data from a dataset whose origin or 'source' could be a database, raw file information, logs, spreadsheet data, etc., correlations are one of the most effective tools for concluding [12],[19].

Gain Ratio Calculation:

- Amount of knowledge obtained by understanding the attribute value.
- Gain Ratio for attribute P =$\frac{Gain(A)}{Split\ (A)}$

Where,

- Gain= (Entropy of distribution before the split)–(entropy of distribution after it)
- Entropy (P_1,P_2,..,P_n)=–P_1log(P_1)–P2log(P_2)–⋯–$P_n$log(P_n) = -$\sum_1^n P(n)\ logP(n)$
- Split Info(A) = $f(x) = \sum_1^n \left(\frac{|An|}{|A|} \times log2\,\frac{|An|}{|A|}\right)$, where A = Attribute

Correlation Calculation:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

r_{xy}–the correlation coefficient of the linear relationship between the variables x & y

- xi – the values of the x-variable in a sample
- x̅ – the mean of the values of the x-variable
- yi – the values of the y-variable in a sample
- ȳ – the mean of the values of the y-variable

Selected features are shown in Table 4.2.5 We can see we have attributed by selection 21 to 10 which is very efficient for our model.

Table 4.2.5: List of selected features

Feature Selector	No.
Gain Ratio	12,16,10,9,6,1,19,8,4,14
Correlation	12,16,10,9,20,6,11,14,19,4
CFS Subset	6,9,12,14,16,19,20

Table 4.2.6: MOST IMPORTANT FEATURES

Features	Frequency
12 GRE SCORE	3
16 UNDERGRADUATE CGPA	3
9 IELTS/TOEFL SCORE	3
6 SCHOLARSHIP	3
19 PUBLICATION THESIS	3
14 UNDERGRADUATE DEGREE	3
10 GRE/GMAT	2
4 HIGHER ADMISSION DEGREE	2
20 JOB EXPERIENCE	2
1 COUNTRY	1
8 IELTS / TOEFL	1
11 GMAT SCORE	1

The Shortlist factors and their ranks by various algorithms for dimension reduction

Table 4.2.7: LIST OF SELECTED FEATURES AND THEIR RATIO

Selected Features (visa prediction)	Gain Ratio	Correlation
GRE SCORE	0.58978	0.6543
UNDERGRADUATE CGPA	0.30756	0.6333
IELTS/TOEFL SCORE	0.23036	0.3905
SCHOLARSHIP	0.13503	0.3671
PUBLICATION THESIS	0.1159	0.3292
UNDERGRADUATE DEGREE	0.12536	0.2501
GRE/GMAT	0.23745	0.5624
HIGHER ADMISSION DEGREE	0.10827	0.2522
JOB EXPERIENCE	0.15568	0.2458
COUNTRY	0.09301	0.353
IELTS / TOEFL	0.0497	0.2537
GMAT SCORE	0.13012	0.1129

From the frequency table, we came to know that for some variable values are almost same for all. Also figured out the important attributes by feature selection process with the help of gain ratio and Correlation [3].

We must compare the related research recently published in this context to predict the efficiency of the machine vision-based expert prediction method.

Table 4.2.8: Results of the comparison of our work and others' works

Method	Object (s)	Data Size	Technique Used	Algorithm	Best Classifiers	Accuracy
This Work	Students	400	*Machine Learning*	SVM, Random Forest, Neural Network, etc.	Random Forest	97%
Billah et al [1]	Students	501	*Machine Learning*	SVM, Random Forest, Neural Network, etc.	SVM	86%
Kurniadi et al [2]	Students	434	*Machine Learning*	*K*-NN	*K*-NN	95.83%
Mia et al [8]	Students	344	*Machine Learning*	SVM, Naïve Bayes, Logistic, J48, etc.	SVM	85.76%
Ahmed et al [6]	Students	455	*Machine Learning*	SVM, Random Forest, Neural Network, etc.	SVM	89%

4.3 Discussion

Here we have discussed our questionnaire, shorted variable list for the transformation, and preprocessing of data. We have found that attributes in the section Personal Experience are mostly different for all. But there is no class difference between males and females for various attributes.

From the frequency table, we came to know that variable value is not the same. Also figured out the important.

CHAPTER 5

Impact on Society, Environment, and Sustainability

5.1 Impact on Society

In this modern age of science and technology students are applying for higher studies, but many of them are rejected. Students go to the agency for information about higher studies, but many agencies don't give the right information. Students are suffering from this kind of issue [3]. The standard of education in our country is very low compared to other countries, which is why we need higher education. Our country is very poor financially Because that's why we are always looking for scholarships for higher education. The prediction here will help us in this regard. This report impacts eligibility for higher study Scholarship programs, a need-based grant for higher study students, on student's CGPA, IELTS, TOFEL results by using a regression discontinuity approach. For any student in our country, the effect is significant [5]. If they use this project then they will predict the percentage of their chance of selection.

5.2 Ethical Aspects

At a very young age, many take their children overseas to learn. But in the sense of Bangladesh, going to study abroad typically exists after crossing the higher secondary caps. Of course, many also go abroad to do postgraduate or doctoral studies (PhD), not just at the undergraduate level. All know, though, that the world of education is forever expanding. The educational sector is seldom limited to those limits. Basic area and boundary of education are known to be the extension between country-time-nation-culture, etc. There is no alternative to seeking higher education outside the country if one wishes to enhance oneself with a large outlook and varied skills.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

The computer data science industry is rising tremendously. In recent years, the development in the data science industry has been massive. Our method in this research paper yielded a good outcome [3]. A lot of data required to make the project usable. After processing, we did Data mining, model generation, and Performance measurement of algorithms, and then finally we will get a model to use which can be the appropriate solution for achieving higher accuracy.

6.2 Conclusions

The journey has not been easy and swift for us. We have faced so many challenges in the collection of data. After collecting we have to process and select the best model for our further work. Though we have to work hard for this project, this will help students for their higher studies.

6.3 Implication for Future Work

For other significant graduation programs such as Corporate, Medical, Engineering, etc., classification models will be developed in the future. Many important characteristics should be taken into account, such as family background, the economic status of students, etc., to improve the study. Many individuals in future jobs can also use upcoming new strategies.

References

1.Kurniadi, D., E. Abdurachman, H. L. H. S. Warnars, and W. Suparta. "The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm." In IOP Conf. Ser. Mater. Sci. Eng, vol. 434, no. 1, p. 012039. 2018.

2.Chalmers, Denise. "Progress and challenges to the recognition and reward of the scholarship of teaching in higher education." Higher Education Research & Development 30, no. 1 (2011): 25-38.

3.Ahmed, Sheikh Arif, and Shahidul Islam Khan. "A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective." In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-6. IEEE, 2019.

4."APPLY FOR A U.S. VISA" Available: https://bit.ly/2I6CsJL. [Accessed: 20-Sep-2020]

5.BILLAH, Md Aref, Sheikh Arif AHMED, Shahidul Islam KHAN, and Bangladesh Chittagong. "Factors that contribute programming skill and CGPA as a CS graduate: Mining Educational Data." Database Systems Journal BOARD: 33.

6.Al Amin Biswas, Anup Majumder, Md Jueal Mia, Itisha Nowrin, and Nadia Afrin Ritu. "Predicting the Enrollment and Dropout of Students in the Post-Graduation Degree using Machine Learning Classifier."

7.Mia, Md Jueal, Abdus Sattar Al Amin Biswas, and Md Tarek Habib. "Registration Status Prediction of Students Using Machine Learning in the Context of Private University of Bangladesh."

8.Nilashi, Mehrbakhsh, Karamollah Bagherifard, Mohsen Rahmani, and Vahid Rafe. "A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques." Computers & industrial engineering 109 (2017): 357-368.

9.Nilashi, Mehrbakhsh, Karamollah Bagherifard, Mohsen Rahmani, and Vahid Rafe. "A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques." Computers & industrial engineering 109 (2017): 357-368.

10.Orange (software). Available: https://en.wikipedia.org/wiki/Orange_(software) [Last accessed on May 10, 2020].

11.Weka (machine learning). Available: https://en.wikipedia.org/wiki/Weka_(machine_learning) [Last accessed on May 10, 2020].

12.Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

13."BESSiG-Bangladeshi Expat & Student Society in Germany" Available: https://bit.ly/34qQGwa. [Accessed: 01-Oct-2020]

14."HigherStudyAbroad- Global Hub of Bangladeshis" Available: https://bit.ly/33tsom8. [Accessed: 25-Sep-2020]

15." Study In UK & Canada for Bangladeshi Students". Available:https://bit.ly/3ldvalC.[Accessed: 03-Oct-2020]

16."Foreign Scholarship Info for Bangladeshi Students" Available: https://bit.ly/3jy8SdN.[Accessed: 20-Sep-2020]

17."Approved Visa's Data from Higher Studies " Available: https://bit.ly/2EXJxLc.[Accessed: 1-Aug-2020]

18."Rejected Visa's Data from Higher Studies " Available: https://bit.ly/34n6Wyz.[Accessed: 10-Aug-2020]

19.Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

20.Chalmers, Denise. "Progress and challenges to the recognition and reward of the scholarship of teaching in higher education." Higher Education Research & Development 30, no. 1 (2011): 25-38.

Visa Prediction for Higher Studies using Machine Learning

ORIGINALITY REPORT

25%	21%	13%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.dbjournal.ro Internet Source	5%
2	Submitted to Daffodil International University Student Paper	3%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
4	"Computer Vision Based Local Fruit Recognition", International Journal of Engineering and Advanced Technology, 2019 Publication	1%
5	Sheikh Arif Ahmed, Shahidul Islam Khan. "A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective", 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019 Publication	1%
6	machinelearningmastery.com Internet Source	1%

7	repositorio.unesp.br Internet Source	1 %
8	Submitted to TechKnowledge Student Paper	1 %
9	www.investopedia.com Internet Source	1 %
10	www.confero.ep.liu.se Internet Source	1 %
11	eprints.utm.my Internet Source	1 %
12	Xin Huang, Feng Wu. "A novel topic-based framework for recommending long tail products", Computers & Industrial Engineering, 2019 Publication	1 %
13	dzone.com Internet Source	1 %
14	Submitted to University of Northumbria at Newcastle Student Paper	1 %
15	Submitted to Skyline High School Student Paper	1 %
16	acutecaretesting.org Internet Source	1 %
17	Submitted to University of Lancaster	

	Student Paper	1%
18	D Kurniadi, E Abdurachman, H L H S Warnars, W Suparta. "The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm", IOP Conference Series: Materials Science and Engineering, 2018 Publication	<1%
19	Md. Sabab Zulfiker, Nasrin Kabir, Al Amin, Partha Chakraborty, Md. Mahfujur. "Predicting Students' Performance of the Private Universities of Bangladesh using Machine Learning Approaches", International Journal of Advanced Computer Science and Applications, 2020 Publication	<1%
20	Fadliansyah Nasution, Elviawaty Muiza Zamzami. "Prediction of Vocational Students Behaviour using The k-Nearest Neighbor Algorithm", Journal of Physics: Conference Series, 2020 Publication	<1%
21	ustraveldocs.com Internet Source	<1%
22	faculty.daffodilvarsity.edu.bd Internet Source	<1%

23	www.ijitee.org Internet Source	<1%
24	Md. Kalim Amzad Chy, Sheikh Arif Ahmed, Ali Haider Doha, Abdul Kadar Muhammad Masum, Shahidul Islam Khan. "Social Media User's Safety Level Detection through Classification via Clustering Approach", 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 2019 Publication	<1%
25	"Registration Status Prediction of Students using Machine Learning in the Context of Private University of Bangladesh", International Journal of Innovative Technology and Exploring Engineering, 2019 Publication	<1%
26	Submitted to University of Westminster Student Paper	<1%
27	Nasrin Mottaghi, Mohammad Reza Keyvanpour. "Test suite reduction using data mining techniques: A review article", 2017 International Symposium on Computer Science and Software Engineering Conference (CSSE), 2017 Publication	<1%
28	Yuan Ma, Lifang Han, Huan Ying, Shouguo	<1%

	Yang, Weiwei Zhao, Zhiqiang Shi. "SVM-based Instruction Set Identification for Grid Device Firmware", 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 2019 Publication	
29	www.periyaruniversity.ac.in Internet Source	<1%
30	Submitted to Higher Education Commission Pakistan Student Paper	<1%
31	YaHua Lee, Fuchun Joseph Lin, Wei-Han Chen. "Chapter 24 Multiple User Activities Recognition in Smart Home", Springer Science and Business Media LLC, 2018 Publication	<1%
32	www.tci-thaijo.org Internet Source	<1%
33	research.ijcaonline.org Internet Source	<1%
34	"Second International Conference on Computer Networks and Communication Technologies", Springer Science and Business Media LLC, 2020 Publication	<1%
35	iopscience.iop.org	

	Internet Source	<1%
36	Submitted to Cathedral Vidya School Student Paper	<1%
37	Witten, Ian H., Eibe Frank, and Mark A. Hall. "Algorithms", Data Mining Practical Machine Learning Tools and Techniques, 2011. Publication	<1%
38	docplayer.net Internet Source	<1%
39	www.grin.com Internet Source	<1%
40	Ur-Rahman, N.. "Textual data mining for industrial knowledge management and text classification: A business oriented approach", Expert Systems With Applications, 201204 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		