A Counseling System to Predict the Study Path for Freshmen

by

Rowshni Tasneem Usha
15301082
Shiny Raisa Parvez
15301047
Fariha Sazid Sejuti
15101027
Maisha Hossain
15301096

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
April 2019

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Rowshni Tasneem Usha
15301082

Fariha Sazid Sejuti
15101027

Maisha Hossain
15301096

Shiny Raisa Parvez
15301047

Approval

The thesis titled "A Counseling System to Predict the Study Path for Freshmen" submitted by

1. Rowshni Tasneem Usha (15301082)
2. Fariha Sazid Sejuti (15101027)
3. Maisha Hossain (15301096)
4. Shiny Raisa Parvez (15301047)

Of Spring, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on April 25th, 2019.

Examining Committee:

Supervisor:
(Member)

Dr Mahbub Majumdar
Professor
Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Moin Mostakim
Lecturer
Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Md. Abdul Mottalib, PhD
Professor and Chairperson
Department of Computer Science and Engineering
BRAC University

Abstract

Now a days, dilemma related to one's career has been considered as a serious issue, specially among fresh graduates. Starting at the age of 18, the students usually fail to grasp the idea of which career path to pursue as they lack maturity and experience on the matter. Moreover, students suffer greatly in deciding which faculty would result the highest benefit for them due to the insufficiency of counselors in the pre-university education. The students do not have the sufficient knowledge to make themselves aware of the real life career related challenges, which is supported by academic majors. It is crucial for a student to make the proper decision on the matter of their career in order to avoid consequences that may be the result of wrong career selection. As a result, selecting an proper career with highest benefit has become one of the most difficult as well as challenging task for the students because wrong career selection may lead to a work field which was not meant for them. This paper presents a counselling system to predict study path for the freshmen by analyzing necessary attributes such as skills, interests, values and motivation, academic background. Moreover, the proposed freshmen counseling system helps the freshmen in their career choice as well as guides toward their respective appropriate career for future. We have used several different approaches for modeling and prediction such as Decision tree classifier, Random Forest, SVM, K-Nearest-Neighbors Classifiers etc. and differentiated between the resulting precision scores. The results were also cross-checked which determined the best parameters that is responsible for providing highest accuracy scores. Furthermore, some ranking algorithms were used to generate a ranked output for the student counseling system. In this paper, we have separated our work into different parts. Chapter 1 contains the overall idea about our work. Chapter 2 contains Related Works, followed by Chapter 3, where we have mentioned about the methodologies which we have used for the system. After that Chapter 4 contains the implementation of the system. Next in Chapter 5 result analysis has been mentioned. Lastly, Chapter 6 ended with conclusion, our limitations and future scopes of improvements related to our work.

Keywords: Machine Learning; Prediction; Decision tree; Random Forest; Ranking algorithm; Major Selection; AHP

Dedication

We would like to dedicate this paper to our caring parents for their unconditional guidance and support.

Acknowledgement

First of all, praise to the Great Allah for whom our thesis was fulfilled without any significant disruption.

In particular, we would thank our supervisor of the thesis, Prof. Dr. Mahbub Majumdar, for his continuous encouragement as well as creative guidance in this field of research. We are also grateful to Mr. Moin Mostakim, our co-advisor, for his kind assistance and guidance in our work.

We are very much grateful to Md. Nur Nahid Hasan, Senior Software Engineer of the IT Department of BRAC University for helping us with the necessary data regarding our thesis.

And finally, our gratitude extends to the faculties of the Department of Computer Science and Engineering, BRAC University, from whom we gained the knowledge, appreciation and help for the completion of our thesis work.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AHP	Analytic Hierarchy Process
AUC	Area Under the ROC Curve
CBR	Case Based Reasoning
CI	Consistency Index
CW	Criteria Weight
DM	Data Mining
KNN	K Nearest Neighbor
KPI	Key Performance Indicator
MAE	Mean Absolute Error
RF	Random Forest
RS	Ratio Scale
SVC	Support Vector Machine Classifier
SVM	Support Vector Machine

Chapter 1

Introduction

1.1 Introduction

Determining which path of career to pursue when there is a variety of options can be very challenging and a daunting task for freshmen, which will take them to their ambition in the end [21]. This matter dependably continues deceiving our psyches with respect to what our advantage lies in. This strain begins to mount as we grow up and understand that we have as of now achieved a phase where we need to choose what must be done next throughout everyday life, except still are in dilemma. Guardians who have youngsters concentrating in SSC/HSC or even O Levels and A Levels are strained about their future and end up passing their nervousness to their youngsters. Hence youngsters have no clue regarding which stream to pick, what occupation to take up and what life they need to lead [12]. The selection of a career which is not beneficial and in the wrong sector can decrease the potential productivity of human resource. The maintenance of effective graduation is of great suffering for advanced academic students (Aud and Wilkinson-Flicker, 2013).The basic need to create inventive ways to deal with assistance advanced education foundations to hold students, encourage their auspicious graduation, and guarantee they are well-prepared and workforce prepared in their field of study, is recognized by the 2001 National Research Council report (Council, 2001) [14]. Thus, selecting a study path or major subject could be overpowering while confronting many different options for undergraduate students. In other words, there are numerous new profession open doors in each field with the expansion in research and investigation in different spaces. As a result, more options create more dilemma to the freshmen, who have just been admitted to university [15]. Extensive number of various strategies which are associated with the training structure can prove to be advantageous in anticipating the capacity of the profession of students. As a result the idea of counselling in the sector of career selection has been widely utilized in the past and in several varieties.

Guidance or counselling system is a process or method that can aid a person to understand and build up his/her professional, academic prospective along with acquiring individual happiness and satisfaction. To be specific, career counselling is considered to be a training and motivating method that assist individuals and students to plan their academic and professional life ahead; take responsibilities of personal and vocational opportunities, managing support for further higher education, professional development and promotion. According to British Association Counselors (BAC), system of counselling includes skill and the rule utilization of relationship to aid in the development and ideal improvement of individuals [19]. This idea was also noted according to Ivan Prelovský (as cited by Arumugam Nagalingam, 2015). Furthermore, it helps labor profession and study systems achieve their goals. As the need to hire skilled workers is increasing immensely, it is very important for undergraduate students to know their available options and follow the reasonable path of study to prepare them well enough for the future. Remarkably, guidance and counseling is an integral practice in schools and colleges, the intent of which is to provide pieces of advice to students as a counseling system, in order to aid them in improving their performances in academic sectors, assist them to determine a career which will maximise their individual potentials. Additionally, being unaware of self-talent, self-personality trait and all the available options around themselves, one makes the mistake of choosing the inappropriate career option. Sometimes students may have confusion between two or more choices for selecting their future career. Under these circumstances, students are dissatisfied with their poor performance and sometimes suffer from anxiety, depression, stress and other mental sickness.

1.2 Motivation

Now-a-days there is a demand for a proper counseling system that will show a possible study path, future outcomes, course sequences and majors to select from, which better fits with his/her objectives. This is when students begin to search for other compatriots who have comparable foundations to perceive what their choices were for instance, professionals known as counselors, parents, teachers, other specialists [3]. This is when some may ask "what is the need of an automated career system?" It has become an absolute necessity because:

1. Unavailability of good counsellors in a good number of institutions.
2. Sometimes students do not feel comfortable enough to talk about their problems, thoughts and share their ideas with an unknown person.
3. Few number of counsellors attend the students with proper care and dedication.
4. Not having enough experience or past knowledge about the entire system to properly guide a newcomer towards the correct path.

Thus, rather than counseling just a couple of people or seniors, our proposed system exhibits an approach to encourage individuals to gain from a large number of comparative foundations and discover best study path that empower them to come to their objectives. Also with the help of proper design and implementation steps we believe that an automated career counseling system can provide effective solution than a real life counselor. Moreover, traditional counselors can be complemented by the system and as well as serve as a tool used by them [7].

As the interest for qualified work builds it turns out to be increasingly more essential to comprehend what inspires understudies to finish their program and how they select their majors. In parallel, colleges are persistently attempting to improve their projects and pull in more understudies. It would be useful for a college to have the ability to envision which major or specialization understudies will pick. In order to grasp the idea of which factors are most beneficial in both of these desires is significant as it might help grasp what drives college understudies through their academic job and in the long run prompting an appropriate profession. The objective of this paper is to ensure a pathway of finding a perfect classifying algorithm which has the most accuracy, but does not over fit the data, comparing to other classifiers and predict the outcomes of decisions, parameters of decision models, and decision errors directly from visual activity patterns. In other words, the significance of this counselling system can be briefly explained in two ways.

Firstly, it will assist the school administration to foresee the dimension of graduating understudies before they really graduate. It gives markers about what should be done as far as scholarly direction, future investigation plans, and advancement of instructive applications in various controls, particularly the instructive innovation. Secondly, it helps to identify the freshmen profiles who are believed to be often facing challenges in the very beginning, passing first academic year. Identifying potential failure in performance at an early stage by utilizing student information accessible at enlistment, for instance, school records, with a view to opportune and productively help the students, is the sole aim of career counseling [22]. In addition, the examination paper will help the administration of understudy directing focuses at colleges in giving scholarly guiding administrations, instructive administrations, and mental administrations. Every one of these administrations are important to make chances to raise the dimension of frail understudies, and therefore make the ideal progress in gathering the necessities of society and the profession advertise. We intend to control understudies by driving them into a progression of inquiries (fitness test) which will give them a thought how to begin and what to do further. The questions are a combination of academic skills as well as personal interests. As a result, the students will receive their result in a percentage form arranged in an ascending order which will state a certain career is better for them [19].

Our research contribution is mentioned below:

1. Analyzing related existing works
2. Preparing student dataset with relevant features for our research
3. Dataset manipulation and visualization
4. Fitting different machine learning models
5. Tuning dataset and comparison between different modeling algorithms
6. Data ranking process using random forest and AHP

1.3 Thesis Overview

Identifying factors that influence students' decision in choosing their university major is very important as these would definite give a clear scheme to a freshman, about how good their chances are given their educational profile. Machine learning strategies have been considered to address this issue, to get an unmistakable approach for finding the example of different attributes in understudy information and how they are influencing their decision of major. Additionally, we have exhibited a framework equipped for foreseeing, with sound achievement, which program understudies with certain arrangement of highlights would pick and how they would experience challenges and which will proceed to accomplish a few basic measurements of progress dependent on their CGPA.

In the second chapter we have discussed some similar works related to our research. Moving ahead to the Third chapter we talked about our system implementation where in the first half we briefly introduced our data-set and how we planned to work with it. We have taken into consideration two separate data sets to build and test the model. One was collected by us using Google Forms Survey and a more formal data-set containing student information was provided by BRAC University.

At the very beginning after visually examining the data-set, types of missing data were identified as well as handled. The data set is broken into the learning set containing known outputs and given the data-set, the model familiarizes itself in order to be generalized to unknown data in the future and utilizing the test group to analyze the modelling algorithms precision ability with the intention to generate an efficient counselling system.

As our data set had very large dimensions, 2462 rows and 12 columns, in order to avoid or reduce the chances of over-fitting and under-fitting three types of feature selection was done to select the most informative and important feature for the final model.

In the second half, we gave a brief description about the models we used to implement our system. A total of 8 machine learning algorithms were used to get an idea about which one of these are showing maximum accuracy when used on our data-sets. Among them three best performing ones that are Decision Tree, K Nearest Neighbors, SVM and Random Forest, were later used to see how much their accuracy can be increased through individual parameter tuning. After tuning and fitting different models on the data-set, the features which impact mostly were selected for ranking the outcome by applying AHP (analytical hierarchy process).

Chapter 2

Related Work

There are different sites and web applications available, for example, Career Horoscope, My Career Tools, Career Test and so on, which either help students know their suitable career path or gives suggestions to improve their performance. However, the majority of these frameworks just use identity attributes as the main factor to anticipate the profession, which may result in a conflicting answer [21]. Additionally, a few locales propose profession dependent on just the interests, pastimes of the students. These frameworks don't check whether the undergraduate students will almost certainly get by in these vocation ways.

One of the most frequent processes is Data mining, which encompasses exploring unique data. It likewise helps to create strategies so as to deal with the undeniably extensive scale information that rises up out of various academic divisions. These strategies are additionally used to more readily get understudies, and the examination way they pursue. Data mining is an apparatus which bolsters basic leadership procedure and which endeavors to guarantee required learning so as to achieve the arrangement of the issue [9]. Most of these websites used Data Mining algorithm for predicting the outcome. Data mining is about extracting data from a large set of data and predict the outcome. Perspective of investigating information from alternate point of view and condensing it into valuable data. Gorad et al [21]. Amrieh et al. [16] proposed another understudy's execution forecast model dependent on information mining methods with new information traits/ attributes, which are called understudy's social highlights. These kind of highlights are identified with the student's intuitiveness with the e-learning the board framework. The execution of understudy's prescient prototype is assessed by a range of computational models; Cognitive Platform, Naive Bayesian and Tree of Decision. They have utilized the calculation of Bagging, Bolstering and RF, which are the regular outfit techniques utilized in ML. The choice charts created by RF are focused on random information choice and random variables choice. These choice trees depend on Random determination of information and an arbitrary choice of factors. These two data is utilized all the while so as to get a progressively exact and increasingly fitting forecast. The outcomes uncover that there is a solid connection between student's practices and their scholarly accomplishment. The exactness improved to 22.1% thinking about the social highlights. Contrasted with the outcomes, by utilizing gathering techniques, when expelling such highlights and it accomplished about 25.8% precision improvement. By testing the model utilizing newcomer understudies, the accomplished precision is over 80%. This outcome demonstrates the unwavering quality of the proposed model [16].

Lykourentzou et al. [8] endeavored data mining tactics to think about understudies' course execution the individual took. Information mining gives different techniques and controls that we can follow so as to decide or investigate the understudy execution. In this paper Decision tree strategy has been utilized for the grouping task just as to assess understudy's execution in the courses. To be specific,the information utilized in the examination paper is confined to those understudies who took the C++ course in Yarmouk University in the year 2005.

In another paper, Cortez and Silva [7] investigated ongoing certifiable information from two Portuguese schools. They utilized two unique sources to gather their information. One is mark reports and another one is questionnaires. Since the former did not contain enough information to support the model as there were only the grades and student presence number available, some other demographic manner were included for instance social and school related characteristics (for example understudy's age, liquor utilization, mother's instruction). Their point was to anticipate understudy accomplishment just as to recognize the key factors that influence instructive achievement and disappointment. The two center classes (for example Science and Portuguese) were displayed under three Data Mining objectives. At first, parallel characterization (pass or flop); besides, arrangement with five dimensions ((I) awesome or phenomenal to (v) inadequate); Lastly iii) relapse, with a numeric yield ranging between 0% and 100%. Three data setups (e.g. with and without the college grade outcomes) and four DM calculations (e.g. Decision Trees, RF) were conducted for each of these methodologies. In addition, an illustrative investigation for determining the best models was applied, so as to distinguish the most important features.

Another strategy which is connected as far as anticipating profession decision is String Matching. This model analyzes the student's vocation objective to a huge number of other profession ways gathered through online studies and LinkedIn. At the info level, the student needs to sustain current training and his/her vocation objective. At the yield level, the different vocation ways he/she can select to accomplish his objective would be created. In every one of these cases, the yield depends on either identity characteristic or scholarly outcome [15].

Nawaz et al.[12] proposed another methodology as far as vocation forecast which is known as case-based thinking (CBR). In CBR, the framework's capacity is exemplified in a chronicle of past cases, rather than being encoded in conventional guidelines. Each case commonly includes a clarification of the issue, alongside an answer or the yield or result. The learning and thinking process shown by a specialist to understand the case is not noted, yet is contained in the arrangement. To tackle a one of a kind case, it is coordinated against the cases for the situation base or preparing set, and comparative cases are recaptured. The recovered cases are utilized to advocate an answer which is reused and tried for future questions. The arrangement is then changed. In conclusion, the new issue and the last arrangement are considered as a major aspect of another case. Reusing this case arrangement with regards to the new case depends on perceiving the differences between the recollected and the new case; and recognizing the piece of a recalled case which can be done to the new case [20].

Another examination depended on Classification Tree, Random Forest and Variable Importance. A grouping tree is an instinctive and incredible classifier and building an arbitrary woods of trees brings down the difference of the classifier and furthermore forestalls over-fitting. Irregular woodlands additionally take into account dependable variable significance estimations. These measures clarify what factors are valuable to both of the classifiers and can be utilized to more readily comprehend what is measurably identified with the understudies' decisions. The outcomes are two precise classifiers and a variable significance investigation that gives helpful data to the college. In this paper, two classifiers are developed utilizing arbitrary woodlands. In this examination, at first the courses which an understudy takes in the initial two semesters are utilized to foresee whether they will get a college degree. Besides, the understudies who have finished a program, their significant decision by and by is anticipated utilizing the initial couple of courses they have enrolled to [19].

From another effort, the University of Maryland has conducted wide-ranging studies on over 250,000 young students chosen at the university, 30,000 of which have been transferred from compliance network schools. The reason for the Predictive Analytic for Student Success venture was to total information over numerous foundations to follow the scholarly advancement and consummation of junior college exchange understudies, distinguish factors related with progress, and execute intercessions that advance understudy achievement. Utilizing strategic relapse a prescient model was worked to recognize the significant attributes prompting an assortment of fruitful results. The specialists recognized the bearing of progress in GPA after some time as a solid indicator of maintenance, a trait [13].

Furthermore, Beaulac and S. Rosenthal [23] applied machine-learning calculations for evaluation and early expectation of under grad student's achievement, with the aim of recognizing those understudy attributes and factors that demonstrate the most grounded prescient capacity concerning effective graduation, working on more than 30,000 understudy perceptions, comprising of green beans and Students from another university who registered at California State University.

Chapter 3

Methodology

Initially to begin with, we had run a total seven algorithms to get a brief idea about the model accuracy score. The idea behind this implementation was with the intend to determine and understand a rough idea about which algorithms perform or support the best in our data set providing us the maximum outcome in terms of accuracy score. A pie chart containing result is shown below:

Figure 3.1: Pie model of different algorithms

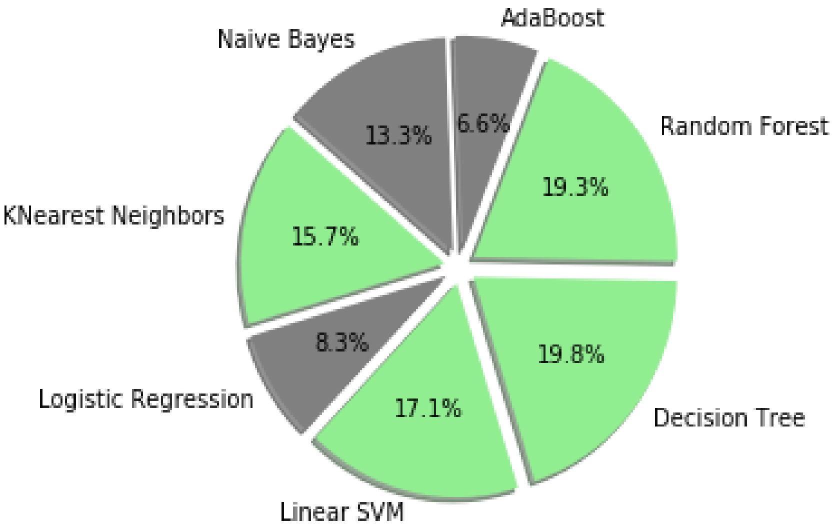


Table 3.1: Table containing Accuracy Score of different Modeling Algorithm

Model Name	Accuracy Score (%)
K Nearest Neighbour	78.90 %
Logistic Regression	41.58%
Linear SVM	86.00 %
Decision Tree	100.00 %
Random Forest	97.77 %
Ada Boost	33.27 %
Naive Bayes	66.94 %

From the above table 3.1, it is clear that some modeling algorithms performed well on our data-set whereas others were not suitable for our data-set. We decided to focus on four of the models that resulted in the highest accuracy score. We have continued further processing of data and implemented Decision tree along with Random Forest (RF). Moreover, we have considered Support Vector Machine tactics as well as K-nearest-neighbor algorithms for our research.

3.1 Decision Tree

It is a learning algorithm makes a tree like structure of the whole informational index as a prescient model. It comprises of numerous hubs or nodes containing element to a trait from the given data, otherwise known as an attribute. All of the tree branches interfacing the hubs entitles to a choice principle for the hubs and each leaf is the anticipated all out mark or ceaseless esteem. It is utilized for both grouping and relapse issues. Categorical Variable Decision Tree or Classification trees are models where the objective variable is a discrete arrangement of qualities. The leaves of trees address class names and branches address conjunctions of features that incited those class names or labels [11]. Decision tree offers offers numerous advantages, some are as per the following:-

1. Easily interpret-able by the end users.
2. It can deal with an assortment of information: Nominal, Numeric and Textual.
3. Ability to process mistaken data-sets or missing qualities.
4. Minimum effort, however, results in Higher performance.

Terminologies

Root Node: The top most hub is the Root Node. It relates to the best indicator. From the root node, we acquire at least two homogeneous sets by further isolation of elements. Toward the starting, the entire preparing set is considered as the root.
Decision Node: The root hub branches into potential results known as choice hubs.
Leaf or Terminal Node: These Nodes are without any kids, when no further split can happen. It conveys the anticipated class name.
Splitting: Partitioning a node into minimum of two separate nodes known as sub-nodes.

Pruning: is the measure of choice trees, decreased by evacuating hubs the procedure. It is the inverse of Splitting. Pruning is done to abstain from over-fitting.
Branch or Sub-Tree: Subsection of tree.

Tree Attribute Selection

Manufacturing decision trees more often work by following top-down approach, by selecting a variable or attribute at each progression that best divides the arrangement of elements. The process of finding the smallest tree that fits the data usually resulting in the least amount of cross-validated error. Gini index is one of attribute selection method. It is a measurement to quantify how frequently an arbitrarily picked component from the populace would be inaccurately distinguished. It means an attribute with lower gini index is preferred. The decision tree is generated from classification algorithm is always a binary approach based decision tree where each node will have only two child nodes [1].

3.2 K-Nearest-Neighbor

It is an algorithm based on the calculation of a non-parametric directed ML methodology utilized for both grouping and relapse. Non-parametric implies that it does not make any suspicions on the basic information appropriation and the model structure is resolved from the information. In other words, the algorithm does not make any assumptions on how the basic information is circulated and decides the last model structure from the basic information.

In other words, KNN is evaluated as a sort of lazy learning or instance-based learning is related to KNN. In these methods the capacity is only approximated locally and furthermore comprises of the k nighest training examples inside the element territory. When using KNN for classification issues, the statistics are dependent on pattern resemblance or how strongly the test elements mimic the learning set, this comparison determines which class should be categorized as a specified information level.

KNN classification results are class memberships. The anticipated class would be a discrete value. KNN classifies a data point by taking into account majority votes of its neighbors. By utilizing this the most common class label is selected and the data point is assigned it. The process is done by measuring a distance function. If number of nearest neighbor is 1, then there is only one neighbour and the data points should be assigned to that class label. Before implementing KNN, it is crucial to choose the most significant K point or the closest range of neighbors. In general, the larger the K value is, the more precise it will be due to the reduction of noises. This does not necessarily guarantee better outcome. Furthermore, cross-validation approach can aid in determining the K value. It can be enforced to independent data-set intending to validate the K value.

Calculating distance

As a part of majority voting, the interval or separation measured between the fresh unclassified data point and the entire classified information points are calculated. Distances are calculated with the help of Euclidean distance formula, there are other distance metrics available.

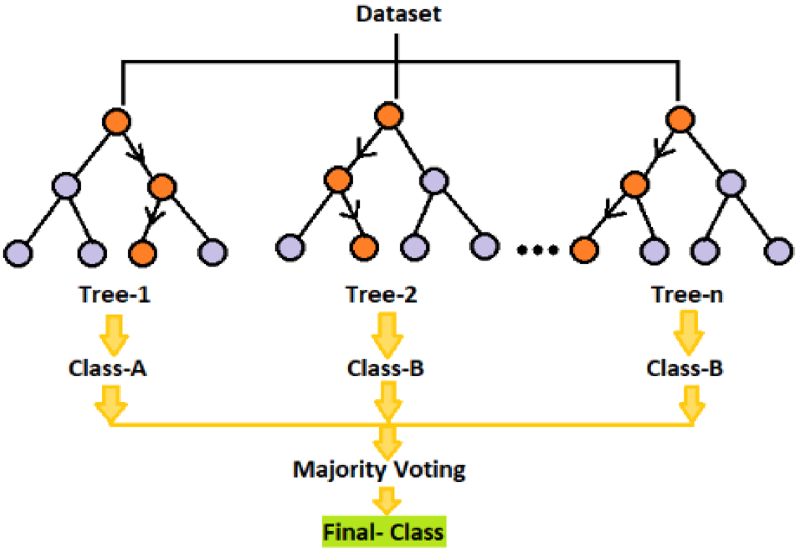
This distance is calculated against all the data points. The smaller the result of this calculation more the similarity between these two data. Later all these distances are sorted to select k smallest distances. Among the k smallest distances, the class that appears the most is the one the data is labeled as.

3.3 Random Forest

RF algorithm was utilized in our system to characterize and regress by making n-tree samples to boot from the source of the data set of undergraduates. One of RF's significant favorable position is that it appears to be assisting for order and recurrence concerns respectively, optimizing almost all of the latest Machine Learning system. Furthermore, while expanding the trees, Random Forest accessorize haphazardness. It scans for the best element among an irregular sub-set of underlines instead of trying to track down the most significant component while parting a hub. This yields in a wide decent variety that results in a superior alternative by and large.

In this way, even an arbitrary subset of features is perceived by the splitting computation in Random Forest. Another extraordinary nature of the RF calculation is that quantifying the overall significance of each element on expectation is anything but difficult. Sklearn gives this amazing device, which estimates a significance element by taking a closer look at how much the tree terminals, using the element, that seems to include decrease contaminating influence on all trees in the forests. It instinctively figures this score for each element subsequent to the preparatory work and measurement of the findings, with the aim that all of the significance is equivalent to 1. In the arbitrary RF algorithm, the hyper-parameters are either used to broaden the model's perceptive ferocity or to render the model faster.

Figure 3.2: Random Forest



3.4 SVM Classifier

Supervised learning, which is used for classification, is one of the primary purposes of SVM. The target of this calculation is to locate the ideal isolating hyper-plane which augments the margin of the training sub-set of the given data, subsequently ordering them. This hyper-plan segregates the classes for multi dimensional data, isolating them in two segments, although in two-dimensional range that distinguishes a plane into two portions, where there are two sections in each class. To properly generalize the data, a hyper-plane must be shortlisted from the variables of each class as far as applicable. The distance between the hyper-plane and the least furthest data points of all classes is assessed for a certain hyper-plane. SVM is the information indicates that lie nearest to this choice surface. When this separation is determined the edge is determined by multiplying it. In other words, after the distance is calculated the margin is calculated by doubling it. The interval or span between the hyper-plane and the support vectors are ”Margin of Separation”. Optimization techniques can solve the problem of finding this optimal hyper-plane as it is an optimization task [3].

Defining the separating Hyperplane

A hyper-plane is a type of condition characterizing the choice surface isolating the classes. Here, x is a vector of input, W is a vector of weight , b is bias.

$$W^T(x) + b = 0$$

The equation recreated as,

$$W^T(x) + b \geq 0 \; d_i = +1$$

$$W^T(x) + b < 0 \; d_i = -1$$

On one side of the hyper-plane, the shortest route to a point is d+.
The shortest distance to point on the other side of the hyper-plane is d-.

A segregated hyper-plane's margin (gutter) is,

$$(+d_i) + (-d_i)$$

3.5 AHP

Analytic Hierarchy Process, a viable instrument for which manages critical resolution making, and helps to set the needs by selecting the most significantly beneficial choice. The AHP grabs both abstract and trigger parts of a particular claim by diminishing complex choices to a progression of pairwise correlations, and after overseeing the findings. Furthermore, the AHP unifies another method for reviewing the continuity of the assessments of the leader, subsequently it decreases the inclination in the complex or critical determination process.

We have utilized the Analytic Hierarchy Process (AHP) for relative estimation of positioning estimations of properties and for making fitting decision. First it builds up the progressive structure with an objective at the top dimension, the characteristics at the following dimension and the traits at the further dimension. From that point forward, It decides the relative significance distinctive property or criteria regarding the objective capacity. To distinguish the standardized vital eigenvectors which are the loads of choices, we initially is standardized the eigenvector matrix [17]

Multiplication to invert summation of each column is shown in the equation below to generate normalized eigenvector matrix.

$$w = \sum_{j=n}^{j=1} a_{ij} = [w_1 \ \cdots \ w_n]$$

After that, we calculated the weight of alternatives by Equation shown below.

$$W_j = \frac{1}{n} \sum_{j=n}^{j=1} a_{1j}/w_j = \begin{bmatrix} a_{11}/w_1 & \dots & a_{1n}/w_n \\ \vdots & \ddots & \vdots \\ a_{n1}/w_n & \dots & a_{nn}/w_n \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \cdot n^{-1} = \begin{bmatrix} \frac{1}{n} \sum_{j=n}^{j=1} a_{1j}/w_1 \\ \vdots \\ \frac{1}{n} \sum_{j=n}^{j=1} a_{nj}/w_n \end{bmatrix}$$

We can approve through a consistency measurement from pairwise correlation. The assessment should be possible by estimating the Consistency Ratio (CR) with the assistance of the Consistency Index (CI) described in Equation 3.

$$(C.I.) = \frac{\lambda_{max} - n}{n - 1}$$

Finally, making an interpretation of subjective information to quantitative information settles on the establishment for basic selection process which is led by calculating precision.

Chapter 4

System Implementation

4.1 Dataset Analysis

4.1.1 Preparing Dataset

To begin with, we divided our work into subsections for efficiency. The purpose of our dataset is to specify the statistical measures of the attributes that we collected to analyze the behavioral pattern of choosing Major among the students. Such dataset containing student evaluation was not found from free online sources. As such, we have conducted a survey through which we gathered potential data necessary to support our research. At first, we considered the features or attributes such as- current stage of education, chosen group of SSC (Secondary School Certificate) and HSC (Higher Secondary Certificate) examinations, GPA of SSC and HSC examinations or O level and A level results, Optional subject chosen during SSC and HSC, types of extracurricular activities students are engaged in followed by their individual skills, hobbies, strengths, weaknesses. In addition weekly and daily study hours, whether they have strong verbal and written communication skills; interest in management or leadership, whether they are extrovert or introvert or likes to work indoors or outdoors along with their social skills.

The preparation of dataset is done in the following manner: We collected responses both by sharing the survey forms online as well as providing printed copies of the form to reach more students. A total of 300 responses were collected through this survey which had 42 features. Unfortunately the amount of responses we gathered from this survey was not enough for to implement our system efficiently. To develop our system we needed at-least 500 responses. Moreover, as we are the students of CSE department, we got the maximum responses we received was from CSE students. As a result, our dataset turned out to be very biased which is not suitable for further implementation of methods and models. There were lack of information of students from departments other than CSE department. In addition, some of the features such as weekly and daily study hours, whether students have strong verbal and written communication skills; interest in management or leadership, whether they are extrovert or introvert or likes to work indoors or outdoors along with their individual social skills, strengths, weaknesses were not helpful and directly related to our thesis. So, we decided to eliminate some of these features and create a more precise and efficient dataset.

As a result, to get more diversified data from students of other disciplines, eventhough it still lacked some of the information needed for our work due to insufficient data in the USIS database. We communicated with the head of IT department of BRAC University and collected available data of the student of BRAC University (BRACU USIS). This new dataset contained the following features:

1. 1st_choice (subject) of students during the admission test
2. 2nd_choicechoice (subject) of students during the admission test
3. 3rd_choice (subject) of students during the admission test
4. ssc or olevel result
5. hsc or alevel result
6. Medium (bangla/english/madrasah)
7. Ssc_board
8. Hsc_board
9. Registered program
10. CGPA of current semester

The dataset contained many faulty data as well as null data. We mapped the values of attributes into numerical values to run the algorithms. As a large numbers of features are involved, the dimension of the function gets increased which adversely impact scalability and accuracy of the approach. So, the next step we followed was to run a feature extraction algorithm which will be useful to predict and more accurate and important features.

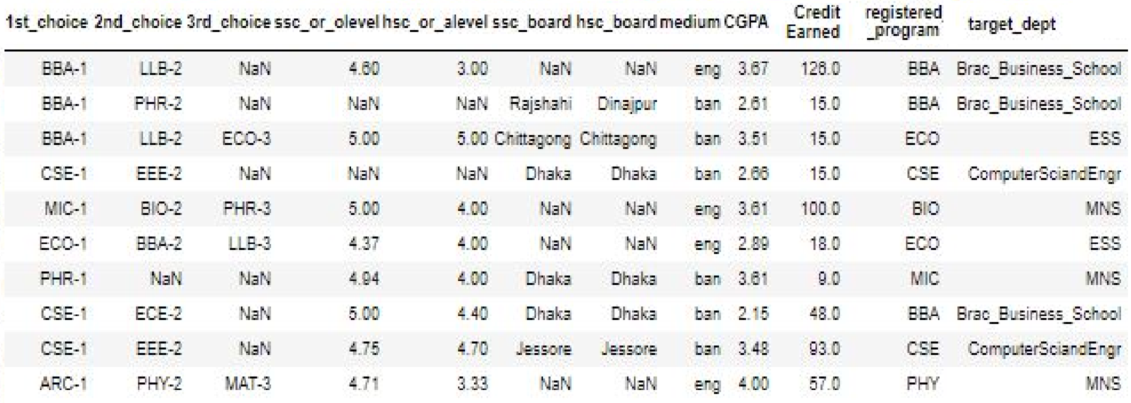
4.1.2 Preprocessing Data

Datatype Handling

The student information dataset contains several types of values. The dataset contains data types such as floats, integers and Strings. Python map() function is used to apply a function on all the elements of specified inerrable and return map object.

In the figure 4.1 there are 12 features. As shown above, each feature also contains several categories. Therefore, we mapped each subcategory to an integer value. Below we have shown summary of unique categories per features.

Figure 4.1: Data-set head from Jupyter notebook

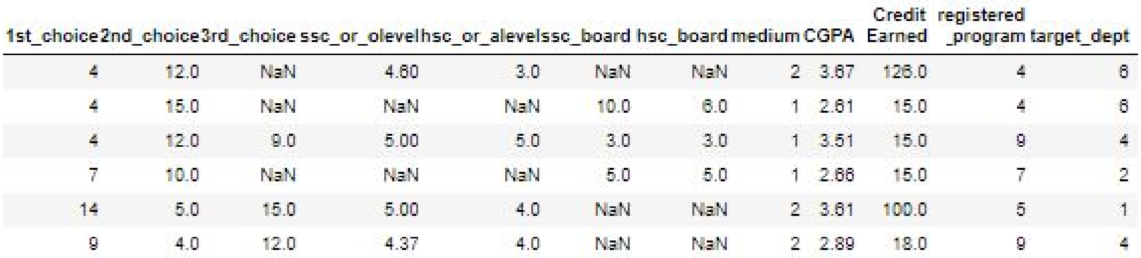


1st_choice	2nd_choice	3rd_choice	ssc_or_olevel	hsc_or_alevel	ssc_board	hsc_board	medium	CGPA	Credit Earned	registered _program	target_dept
BBA-1	LLB-2	NaN	4.60	3.00	NaN	NaN	eng	3.67	126.0	BBA	Brac_Business_School
BBA-1	PHR-2	NaN	NaN	NaN	Rajshahi	Dinajpur	ban	2.61	15.0	BBA	Brac_Business_School
BBA-1	LLB-2	ECO-3	5.00	5.00	Chittagong	Chittagong	ban	3.51	15.0	ECO	ESS
CSE-1	EEE-2	NaN	NaN	NaN	Dhaka	Dhaka	ban	2.66	15.0	CSE	ComputerSciandEngr
MIC-1	BIO-2	PHR-3	5.00	4.00	NaN	NaN	eng	3.61	100.0	BIO	MNS
ECO-1	BBA-2	LLB-3	4.37	4.00	NaN	NaN	eng	2.89	18.0	ECO	ESS
PHR-1	NaN	NaN	4.94	4.00	Dhaka	Dhaka	ban	3.61	9.0	MIC	MNS
CSE-1	ECE-2	NaN	5.00	4.40	Dhaka	Dhaka	ban	2.15	48.0	BBA	Brac_Business_School
CSE-1	EEE-2	NaN	4.75	4.70	Jessore	Jessore	ban	3.48	93.0	CSE	ComputerSciandEngr
ARC-1	PHY-2	MAT-3	4.71	3.33	NaN	NaN	eng	4.00	57.0	PHY	MNS

Table 4.1: Unique Catagory per Feature

Data types	
Feature Name	Unique Category count
1st_choice	16
2nd_choice	17
3rd_choice	17
ssc_board	12
hsc_board	12
medium	4
registered_program	16
target_dept	9

Figure 4.2: Dataset head after mapping



1st_choice	2nd_choice	3rd_choice	ssc_or_olevel	hsc_or_alevel	ssc_board	hsc_board	medium	CGPA	Credit Earned	registered _program	target_dept
4	12.0	NaN	4.60	3.0	NaN	NaN	2	3.67	126.0	4	6
4	15.0	NaN	NaN	NaN	10.0	6.0	1	2.61	15.0	4	6
4	12.0	9.0	5.00	5.0	3.0	3.0	1	3.51	15.0	9	4
7	10.0	NaN	NaN	NaN	5.0	5.0	1	2.66	15.0	7	2
14	5.0	15.0	5.00	4.0	NaN	NaN	2	3.61	100.0	5	1
9	4.0	12.0	4.37	4.0	NaN	NaN	2	2.89	18.0	9	4

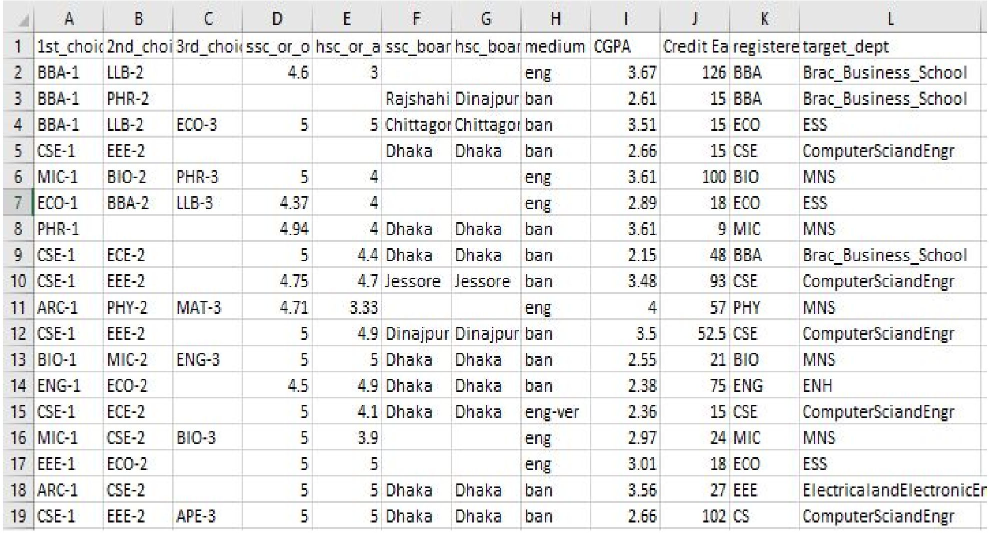
Managing missing entries

As our dataset has been constructed using student provided information, it is important comprehend the sources of missing information. Moreover, in the USIS student database, there are several blank, empty, invalid, corrupt, null or missing values in the dataset. Some common reasons why data might be absent are mentioned as follows.

1. User forgot to fill in a field
2. Data was lost while transferring manually from a legacy database
3. There might be a programming error
4. Users chose not to fill up that field

In the masking methodology, the veil may be an altogether isolated Boolean exhibit, or it might include apportionment of one piece in the information portrayal to locally demonstrate the invalid status of an esteem. To encourage this show, there are a few helpful techniques for identifying, expelling, and supplanting invalid qualities in Pandas information construction.

Figure 4.3: Missing Data in CSV File



	A	B	C	D	E	F	G	H	I	J	K	L
1	1st_choic	2nd_choi	3rd_choic	ssc_or_o	hsc_or_a	ssc_boar	hsc_boar	medium	CGPA	Credit Ea	registere	target_dept
2	BBA-1	LLB-2		4.6	3			eng	3.67	126	BBA	Brac_Business_School
3	BBA-1	PHR-2				Rajshahi	Dinajpur	ban	2.61	15	BBA	Brac_Business_School
4	BBA-1	LLB-2	ECO-3	5	5	Chittagor	Chittagor	ban	3.51	15	ECO	ESS
5	CSE-1	EEE-2				Dhaka	Dhaka	ban	2.66	15	CSE	ComputerSciandEngr
6	MIC-1	BIO-2	PHR-3	5	4			eng	3.61	100	BIO	MNS
7	ECO-1	BBA-2	LLB-3	4.37	4			eng	2.89	18	ECO	ESS
8	PHR-1			4.94	4	Dhaka	Dhaka	ban	3.61	9	MIC	MNS
9	CSE-1	ECE-2		5	4.4	Dhaka	Dhaka	ban	2.15	48	BBA	Brac_Business_School
10	CSE-1	EEE-2		4.75	4.7	Jessore	Jessore	ban	3.48	93	CSE	ComputerSciandEngr
11	ARC-1	PHY-2	MAT-3	4.71	3.33			eng	4	57	PHY	MNS
12	CSE-1	EEE-2		5	4.9	Dinajpur	Dinajpur	ban	3.5	52.5	CSE	ComputerSciandEngr
13	BIO-1	MIC-2	ENG-3	5	5	Dhaka	Dhaka	ban	2.55	21	BIO	MNS
14	ENG-1	ECO-2		4.5	4.9	Dhaka	Dhaka	ban	2.38	75	ENG	ENH
15	CSE-1	ECE-2		5	4.1	Dhaka	Dhaka	eng-ver	2.36	15	CSE	ComputerSciandEngr
16	MIC-1	CSE-2	BIO-3	5	3.9			eng	2.97	24	MIC	MNS
17	EEE-1	ECO-2		5	5			eng	3.01	18	ECO	ESS
18	ARC-1	CSE-2		5	5	Dhaka	Dhaka	ban	3.56	27	EEE	ElectricalandElectronicEr
19	CSE-1	EEE-2	APE-3	5	5	Dhaka	Dhaka	ban	2.66	102	CS	ComputerSciandEngr

Table 4.2: Summary of Missing Values

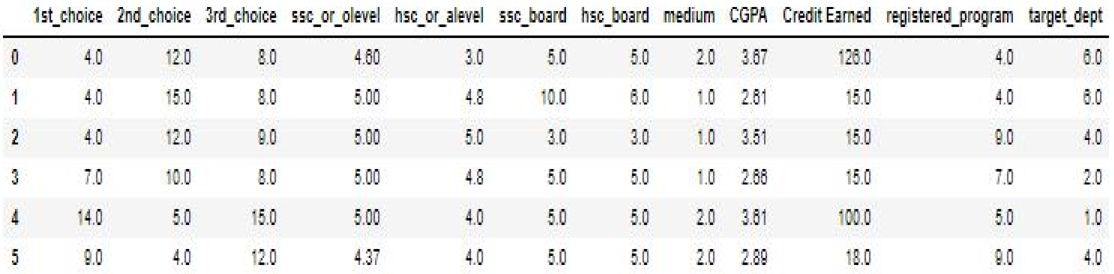
Descending order	
Feature name	Missing value count
3rd_choice	1111
hsc_board	590
ssc_board	581
hsc_or_alevel	330
ssc_or_olevel	233
2nd_choice	57
target_dept	0
registered_program	0

Handling null values

Pandas data construction library have many useful methods for the identification of missing or null values. To detect the null values in the data-set we used ”isnull()” and ”notnull()”. Either one of the syntax will return a Boolean result that will mask over the data indicating missing values. Using the isnull() method, both the empty value and “NA” will be recognized as missing values. Both boolean responses will be True. The method sum() will return the total number of missing values on each column or feature. There are many solutions for handling null values in the dataset.

One of the common solution is to Replace empty/null values with an integer value for example replace "NaN" with a zero. But if we replace 0 with a null value then we get a faulty dataset which is not suitable for our work. As a result, an alternative solution is to replace empty or null value with the median value. To Impute missing values we have used Imputer from sklearn.preprocessing imported from Sklearn library.

Figure 4.4: Datahead after handling missing values



	1st_choice	2nd_choice	3rd_choice	ssc_or_olevel	hsc_or_alevel	ssc_board	hsc_board	medium	CGPA	Credit Earned	registered_program	target_dept
0	4.0	12.0	8.0	4.60	3.0	5.0	5.0	2.0	3.67	126.0	4.0	6.0
1	4.0	15.0	8.0	5.00	4.8	10.0	6.0	1.0	2.61	15.0	4.0	6.0
2	4.0	12.0	9.0	5.00	5.0	3.0	3.0	1.0	3.51	15.0	9.0	4.0
3	7.0	10.0	8.0	5.00	4.8	5.0	5.0	1.0	2.66	15.0	7.0	2.0
4	14.0	5.0	15.0	5.00	4.0	5.0	5.0	2.0	3.61	100.0	5.0	1.0
5	9.0	4.0	12.0	4.37	4.0	5.0	5.0	2.0	2.89	18.0	9.0	4.0

Training the dataset

Next comes training the dataset. After we finished handling null values, we trained our dataset so that we can fit any model on our dataset and make predictions. After gathering and merging data from both survey forms and BRACU Usis database we got 2463 rows and 12 columns. On this dataset we applied train/test split method. The training set includes a established production and the framework learns to be summed up to other information. Then again, we utilize the test dataset or subset to test our model's forecast on this subset.We split our sample into training and test set to begin with. For example, we used 60% of our dataset and training set and remaining 40% we used for testing the model. We imported train_test_split() method of Scikit-learn Library. Next, to visualize our data we plotted some graphs, bar charts to get a clear idea about our dataset and relations of the variables with one another.

Table 4.3: Splitting the dataset

Split data	
Input	Output
dt.shape	(2462, 12)
X_train.shape	(1969, 11)
X_test.shape	(493, 11)
y_train.shape	(1969,)
y_test.shape	(493,)

4.1.3 Data Visualization

The dataset Received from BRAC University USIS system contains 12 columns and 2462 rows of student information. We have done some calculation to implement the graphs and plotting in order to better understand the data set and its features. For example, the following figure contains bar chart showing all the students from different backgrounds, Heatmap showing the relation between each feature etc. We also tried to plot which of the department did the students target most. For instance, the figure below shows us the frequency of the number of students in each department. It shows that students have been mostly enrolled in MNS department.

Figure 4.5: Number of students enrolled in each department

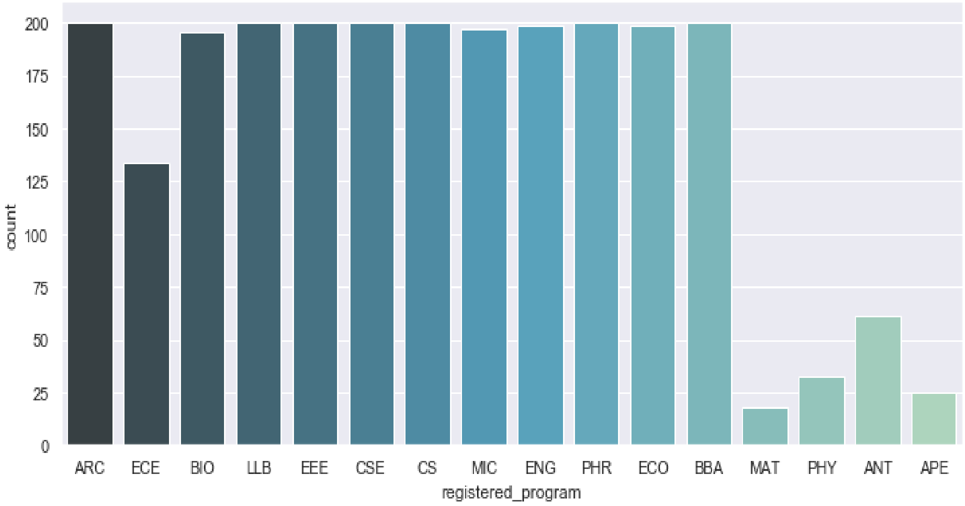
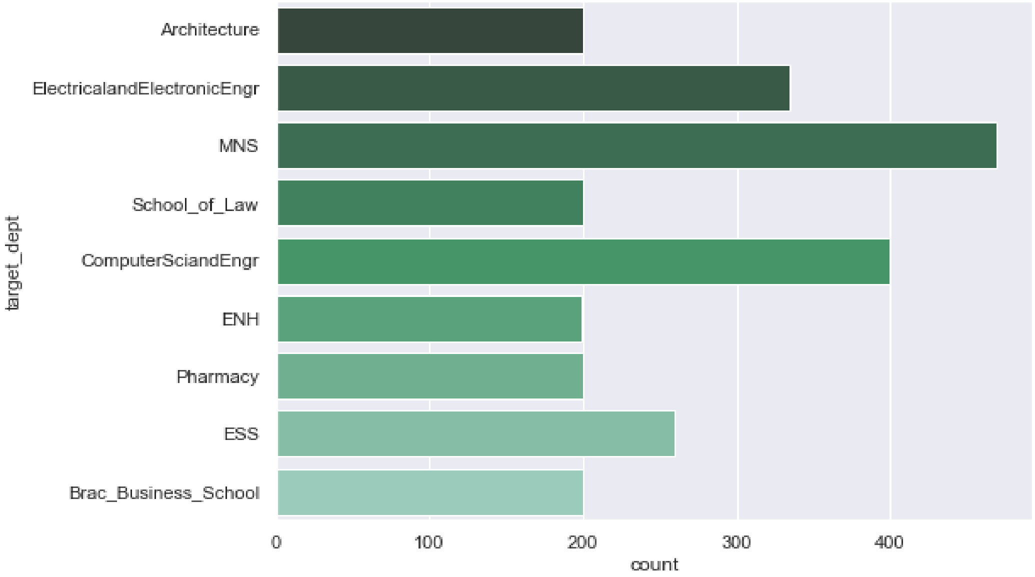
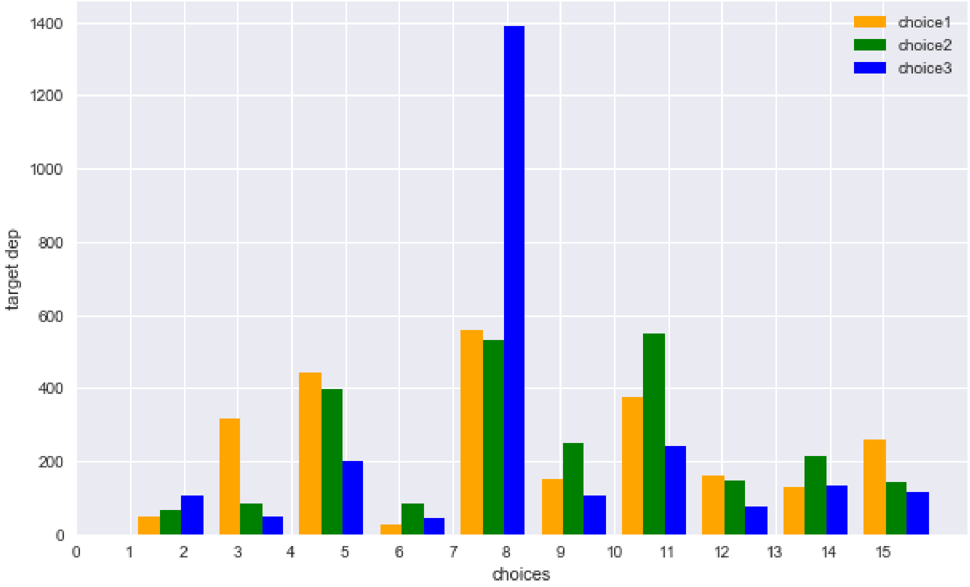


Figure 4.6: Number of students targeted the department most



We aimed to show how students differentiated between their choice of departments. The dataset we received from BRACU USIS contains three different choices which the students filled up during their admission test. These choices vary from one student to another. Below we plotted a histogram which tells us whether it was choice1 or choice2 or choice3 which the students chose for themselves.

Figure 4.7: Choices of students



4.1.4 Feature Selection

Increased dimension of data or additional features in dataset can add more complexity to the problem. High dimension is not always a good choice as it increases the risk of over-fitting. As such, reduction of the number of features is a good practice. For this purpose, two strategies were used to determine which features should be selected:

1. Univariate Statistics
2. Model Based Selection

Univariate Statistics

It is a process that shows statistically significant relation between each features and the output or target feature. In other words, it is the analysis of variants. In this process, features with Highest confidence with the target or output are selected. Each feat is considered in isolation or individually in determining its relation with the target. This process assist in reduction of complexity. How univariate statistics work is given below briefly: Firstly, the dataset was cleaned and processed. Then it was tested on both the original intact features and also by adding some noises to the data before applying the selection method. The resulting feature selection was given 50% of the length of dataset. In other words, feature space was reduced by 50%. After the selection of features a Boolean mask was applied on top of all features in order to make a graphical visualization of which features are selected out of all. Lastly, Logistic Regression was applied on the training subset with all features and as well as on the training set includes only those elected and we assess the production of the algorithm.

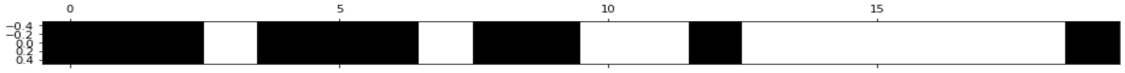
- Data with Noise

  X_train .shape is: (1231, 20)

  The score of Logistic Regression on all features: 0.768

  The score of Logistic Regression on the selected features: 0.786

Figure 4.8: Univariate Statistics



Model Based Selection

It is yet another operation that utilizes a controlled template to assess the significance of every other feature in the set of data and ultimately to preserve the most important features. Decision tree and random forest has the “feature_importance” attribute. We are applying this method on the same original dataset. The dataset will be tested both with and without 50 noise features. And then going to be compared with the performance of this method to univariate statistics, the feature selection method.

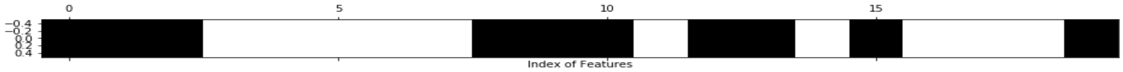
- With 10 noise feature added in dataset:

  The shape of X_train is: (1231, 20)

  The score of Logistic Regression with the selected features on the all set: 0.768

  The score of Logistic Regression with the selected features on the test set: 0.760

Figure 4.9: Model-based feature selection

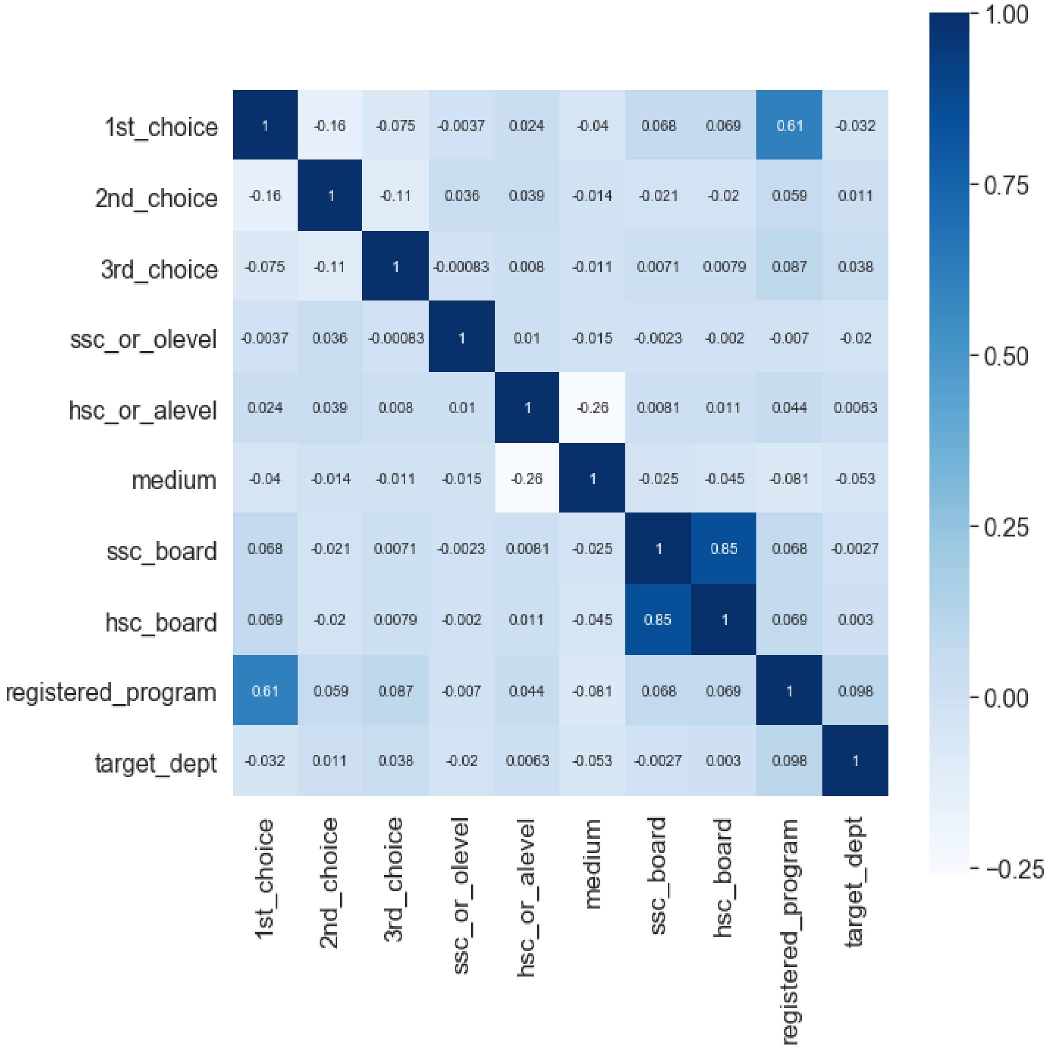


Comparison between results

From the above calculations, we concluded that Univariate statistical selection performs better in case of adding noise attributes to the data-set whereas Model-based feature selection performs better if no noise feature is introduced.

Furthermore, we utilized another component determination technique called ”Heatmap” which is a visual portrayal of information under which the discrete qualities in a frame are represented as tints. Heatmaps are ideal for investigating the relationship of attributes or features in an informational index. Figure 4.10 shows the Heapmap implemented on USIS data-set.

Figure 4.10: Heatmap representing correlation of features



4.2 Model Implementation

4.2.1 SVM Classifier

It is an administered AI calculation which can be utilized for both characterization or relapse difficulties. It can likewise be utilized as exception discovery. SVM works by evaluating the ideal choice capacity that can isolate the training data-set or as such distinguishing the hyper-plane that can partition the given preparing data-set in a productive manner dependent on number of classes [18].

SVM, not at all like customary calculations, for instance, Neural Networks, having their underlying foundations in Statistical Learning Theory (SLT) and improvement techniques, has turned out to be proficient in taking care of ML related issues with limited training data points and defeat normal troubles, for example, the ”scourge of dimensionality”, ”over-fitting” and so forth [10]. SVM plots each data as a point for n highlight quantities with each component being estimated as a specific coordinate, in a n-dimensional space. At that point, grouping is finished by finding the hyper-plane distinguishing the categories. The goal of this calculation is to accommodate the given information, reestablish a best fitting hyper-plan that isolates or categorizes the data. After that, this classifier could be supported by a few features to perceive what the expected class is [4].

SVM Parameter Tuning

For a given arrangement of preparing precedents or examples, each marked as having a place with any class, a computation prepping SVM constructs a paradigm that relegates new information focuses to a classification dependent on it's highlights. SVM has two hyper-parameters C and gamma which have colossal effect on the choice limit. SVM calculation builds a line, hyperplane or set of high- or wide-scale hyperplanes that can be used for grouping, recurrence or other initiatives, for example, outlier recognition. Instinctively, an incredible partition is practiced by the hyperplane that outcomes in the biggest separation against the nearest preparing information point or bolster vector of any class and the hyperplane itself. By thinking about the most extreme separation between the help vectors or closest purposes of each class and the hyperplane would produce an ideal isolating hyper-plane. This separation between the hyperplane and the help vectors is known as the edge, it's esteem is called utilitarian edge, this is one of SVM's hyperparameter, gamma. As a rule the bigger the margin, the lower the speculation and misclassification blunder of the classifier. The other hyper-parameter of SVM is C, it is the regularization parameter that controls the harmony between the misclassification blunder otherwise called slack variable punishment and width of the edge while augmenting the edge for an informational collection that can not be isolated by a line. A definitive objective of SVM classifier is to locate the ideal hyper-plane that has the biggest edge, that can isolate the information while making minimal number of misclassifications as it predicts class mark of new test information.

GridSearch

Grid search is one of the methods to perform hyperparameter tuning so as to decide the ideal hyperparameter values for a given calculation. The presentation of the whole SVM model essentially relies upon how well the hyper-parameter values determined. GridSearchCV is executed inside sklearn, python's ML library. A parameter lattice is indicated before utilizing the inherent grid search. This parameter matrix requires the rundown of parameters and a scope of qualities to test for every parameter of the predetermined estimator. For the estimator parameter of GridSearchCV requires the model that is being utilized for the hyper parameter tuning process, here Radial basis function kernel is being utilized to settle on the hyperplane choice limit between the classes. When working with the Radial basis0 work piece of the SVM model, the most affecting parameters required are c and gamma.

Cross Validation

A cross validation process otherwise called rotation estimation is performed alongside lattice seek on the predetermined parameters so as to decide the hyper-parameter esteem set which gives the best exactness levels. It partitions the information into couple number of folds while ensuring that each overlap is utilized as a testing set in the end. In the principal cycle, the main overlay is utilized to test the model with the predefined parameters and the rest are held off to prepare the model. In the second emphasis, second crease is utilized as the testing put while the rest kept aside as the preparation set. This procedure is rehashed until each overlap of the recently part overlays has been utilized as the testing set. Here, 10-Fold CV (n_splits=10) has been utilized, where the information will be part into 10 folds. In the wake of completing the framework seek over the k-overlay cross approval, the ideal qualities for c and gamma are found.

SVM With optimal parameters

After finding the optimal parameters for our data set, a SVM model is constructed from scratch. It resulted in a higher accuracy of 90.87%, than SVM accuracy without any parameter tuning.

Classification Report

Table 4.4: Table of classification report

	Precision	Recall	F1_score	Support
1.0	0.95	0.91	0.93	90
2.0	0.95	1.00	0.97	78
3.0	0.71	0.83	0.76	63
4.0	0.77	0.60	0.67	55
5.0	1.00	1.00	1.00	37
6.0	0.98	0.98	0.98	47
7.0	1.00	1.00	1.00	47
8.0	1.00	1.00	1.00	28
9.0	0.90	0.94	0.92	48
Micro (avg)	0.91	0.91	0.91	493
Macro (avg)	0.92	0.92	0.92	493
Weighted (avg)	0.91	0.91	0.91	493

To further analyze the goodness of the parameter tuned mode, a Classification report was generated in Table 4.3. It consisted of the following terms and equations.

Accuracy - Naturally, it is by far the most presentable way of measuring for any calculation and it is the percentage of the population of precisely anticipated interpretation to the whole number of perceptions. High score in exactness does not generally demonstrate the nature of a model rather high accuracy is appeared correspondence data collections in which most forecasts of false positives and false negatives are practically identical. Therefore,more parameters ought to be investigated alongside exactness to assess the presentation of the model.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Precision - It is the fraction of absolute expected positive expectations to positive notions that are effectively foreseen. This quantification speaks to as every one of the students are relegated a noteworthy name, what number of them are really concentrating upon that particular major. High accuracy is related with low false positive rate. We have a normal of 0.91 exactness which is quite great.

$$Percision = \frac{(TP)}{(TP + FP)}$$

Recall (Sensitivity) - It is the proportion of all real-class perceptions and precise positive perspectives intended. This measurement speaks to the proportion of all understudy concentrating a specific major and what number of the model could accurately name. We have a normal review of 0.92 which is useful for this model as it is above 0.5 .

$$Recall = \frac{(TP)}{(TP + FN)}$$

F1 score - It is the normal weighted outcome of Precision and Recall. It subsequently reflects both false positives as well as false negatives. Intuitively, it is not as clear cut as accuracy, however F1 is typically more beneficial than precision, especially even when the information index has an inconsistent appropriation of class. Precision performs best whenever there is comparable expense for adverse events and negative results. And if by some glimmer of hope that the inconvenience of false negatives and false positives is considerably different, it is more knowledgeable to glance at both Precision and Recall and not just Accuracy. For our situation, F1 score is 0.91.

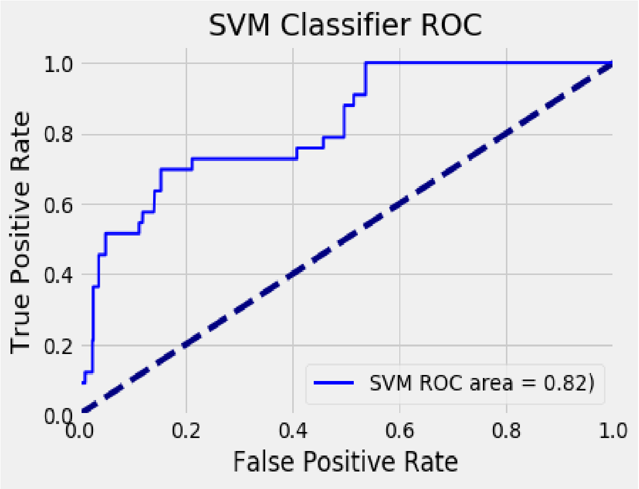
$$F1\ Score = 2 * \frac{(Precision * Recall)}{(Recall + Precision)}$$

Confusion Matrix

```
[[ 82  2  0  1  0  0  0  0  5]
 [  0 78  0  0  0  0  0  0  0]
 [  0  2 52  9  0  0  0  0  0]
 [  0  0 21 33  0  1  0  0  0]
 [  0  0  0  0 37  0  0  0  0]
 [  1  0  0  0  0 46  0  0  0]
 [  0  0  0  0  0  0 47  0  0]
 [  0  0  0  0  0  0  0 28  0]
 [  3  0  0  0  0  0  0  0 45]]
```

Confusion matrix is a detailed portrayal of what's going on with the labels and their relating forecasts. It demonstrates the appropriation of the anticipated qualities contrasted with every real outcomes. If there should arise an occurrence of a perfect classifiers with 100% exactness would create an unadulterated corner to corner framework which would have every one of the focuses anticipated in their real class. Which is not the situation here. For example, the top of the line had 90 points, 82 of them were accurately marked and 8 of them were mislabeled.

Figure 4.11: ROC Curve



Here, we have utilized a ROC curve to visualize and grasp the idea of how proficient our model while recognizing classes. ROC curves are commonly utilized in instances of parallel characterization to observe the yield of a classifier. It is otherwise called a probability curve. ROC curve normally include genuine real positive aspects on the Y axis and false positives on the X axis. This signifies that perhaps the plot's lower right corner corner is the "perfect" point with a true 0 advantageous density and a genuine 1. It indicates that while there is typically better a broader region under the curve (AUC). In this way, the ROC curve rate more like 1 is typically viewed as a well classifier.

4.2.2 KNN

KNN is a standout algorithm among-st easy interpretative data mining methods. It utilizes resemblance measures to classify new cases subsequent to storing all cases. The K in KNN is a hyperparameter that ought to be picked so as to get the most ideal model for the data-set. Hyperparameter K really governs the configuration of the decision limit.

KNN has one hyperparame that is the quantity of closest neighbors or K, to be considered in order to generate classification of another data point. K is viewed as a standout amongst the most significant components of the model that unequivocally impacts the standard of forecasts or predictions. The quantity of closest neighbors K goes about as a smoothing parameter.There is an adequate motivation for k, like any filtering configuration, to achieve the right balance between the orientation and the model's deviation.

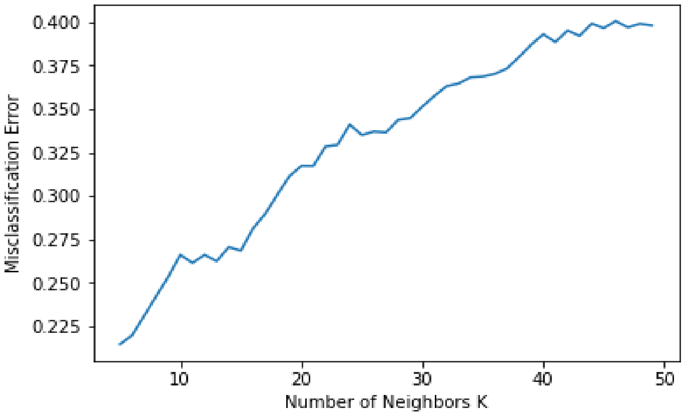
For some classification task, a small estimation of k will prompt an expansive change in predictions. Then again, an extensive estimation of k may prompt a vast model bias. In this regard k must be provided sufficient concern to narrow the odds of misclassification and adequately insignificant (concerning the quantity of cases in the fundamental informational collection) with the goal that the K closest indicates are close enough the inquiry point to impact it's class forecast. To gauge how precise the quantity of closest neighbor estimation is, we originally fit the implicit knn classifier utilizing subjective k and watch its exactness. Here k=3 was utilized, for that the model had demonstrated 81% accurate. Later parameter tuning of the cross verification has been used to discover the measured optimal k and its associated accuracy.

Parameter Tuning With Cross Validation

Cross-validation can be utilized to appraise the test error related with a learning calculation intended to assess its execution. It is an entrenched system that is utilized to acquire evaluations of obscure model parameters. Here we utilized this method to appraise k. First the data points is partitioned into v number of folds or area at arbitrary. For a set k estimate, we linked the KNN template to the vth fragment, which is the approval set while utilizing the v-1 portions to fit the technique. Toward the end after this is over and over done on every single other estimation of v, the registered misclassification blunders are arrived at the midpoint of to quantify the level of predictions of query points by the model. This is finished with cross_val_score() strategy for scikit-realize where the parameter number of folds cv=10 and scoring measurements 'accuracy' is referenced.

The above steps are then performed on different estimations of k and the k esteem that accomplished the most elevated characterization accuracy or the least misclassification error is then chosen as the ideal incentive for k. Here the ideal estimation of k is 5.

Figure 4.12: Number of neighbors k vs misclassification error plot



Distance Metrics

After calculating the optimal k, the whole algorithm was constructed from the beginning rather than using the built in order to test out the optimal k value. Data set is again split into test and train set. KNN is anticipated to make prediction conditional on the outcome of the closest K neighbours to that stage. Therefore, to evaluate the distances amongst all info points and fresh observations, a fresh prediction procedure is published. Here we characterized a measurement of metric for estimating the separation between the test set cases and the query points.

Euclidean separation or distance was selected as a measurement for estimating the span. It then picks k nearest data points or neighbors and performs a majority vote.

$$D(x,p) = \left\{ \begin{array}{lr} \sqrt{(x-p)^2} & \text{Euclidean} \\ (x-p)^2 & \text{Euclidean squared} \\ Abs(x-p) & \text{Cityblock} \\ Max(|x-p|) & \text{Chebyshev} \end{array} \right\}$$

Here, x and p, respectively, are the target point and test instance.

K-Nearest Neighbor Predictions

Next the function "KNearestNeighbor()" is written to loop over every test example and make a prediction. At last the algorithm is ran with the optimal number of neighbors we found. This way the accuracy is increased to 83% from previous 81% that was measured after using k=3 as a parameter for a built in KNeighborsClassifier method of sci-kit learn library.

4.2.3 Decision Tree

Among all four algorithms that performed well on the USIS dataset, Decision Tree is included. It is a supervised algorithm which is suitable to represent multiclass data as well as continuous and categorical data values. As shown previously, the accuracy score after fitting the model to the dataset is 100% which is the highest accuracy score of prediction in compared to other machine learning algorithms that was used on the dataset.

Each rows in the dataset are examples and the columns are attributes or features that describe the data. And the Last column is the label or the class that is to predict. The dataset is not perfectly separable, meaning that it contains some noise or faulty values such as having the same features resulting in different target labels. The root node will receive all of the dataset. Decision tree produces purest possible distribution at each node. It intends to purify or refine the data by growing tree.

Gini impurity and Information gain are two criteria that can be used to build a tree. These techniques are commonly used to measure the best splitting attribute or condition. Gini impurity describes how much uncertainty there is at a single node. These acts as a threshold for partitioning the data. The tree is build on these recursively on each node. It continues to grow and divide data until it reaches a level where gini impurity or entropy value is equals to zero. Gini index method is applied for the features that contains continuous values. On the other hand, Entropy calculation method for generating decision tree is considered for the attributes that occur in classes. By utilizing Cross-validation system, the calculation is evaluated for prescient execution of the model. Moreover, cross-validation was utilized to figure out how the decision tree model will perform on an example that is obscure or execution on another arrangement of test tests. Cross-validation method assisted to find the best attributes and criteria that can affect the model in resulting better prediction. Some of the parameters that are used in construction of a decision tree are: "criterion" (for splitting), "max_depth", "min_leaf_nodes", "min_sample_leaf", etc. These parameter are also known as hyper-parameters. The dataset contains 2462 samples, among which 25% was part of the test set (616 samples) and the rest 1846 samples for training the model [2]

Overfitting

Table 4.5: Overfitting

Sets	Criterion (best)	max depth	max features	min samples leaf	Accuracy Score
Train Set	Gini index	None	8	1	100% (Best)
Test Set	Gini index	None	8	1	65.09%

Previously, it was shown that the tuned decision tree parameters results in 100% accuracy score on prediction. But decision trees can often lead to over fitting the training data-set. To understand whether it over fits, the model is evaluated both on training data and on test data following by comparison for validation. The accuracy on both train and test data is compared. The score should be comparable as the train and sample blocks are arranged in the same way as they are component of the same data-set.

Here the decision tree performs very well on the training set with accuracy score 100% , but the score drops in case of the test set to 65.09% for the same best parameters decisions. This situation occurs because of growing a tree that is large. Growing big tree will lead to over fitting. As such the model will only work for that training data as it is very well known to the model. It will not perform well for any other unknown data set, for instance the test data-set. The accuracy increases while building the model and thus results in a highly accurate model. After a certain point, the model ends up with the entire data-set itself. In other words, all leaf nodes contains only single sample/value which is the observation itself.So, growing tree beyond a certain level of complexity causes over fitting.

Pruning:

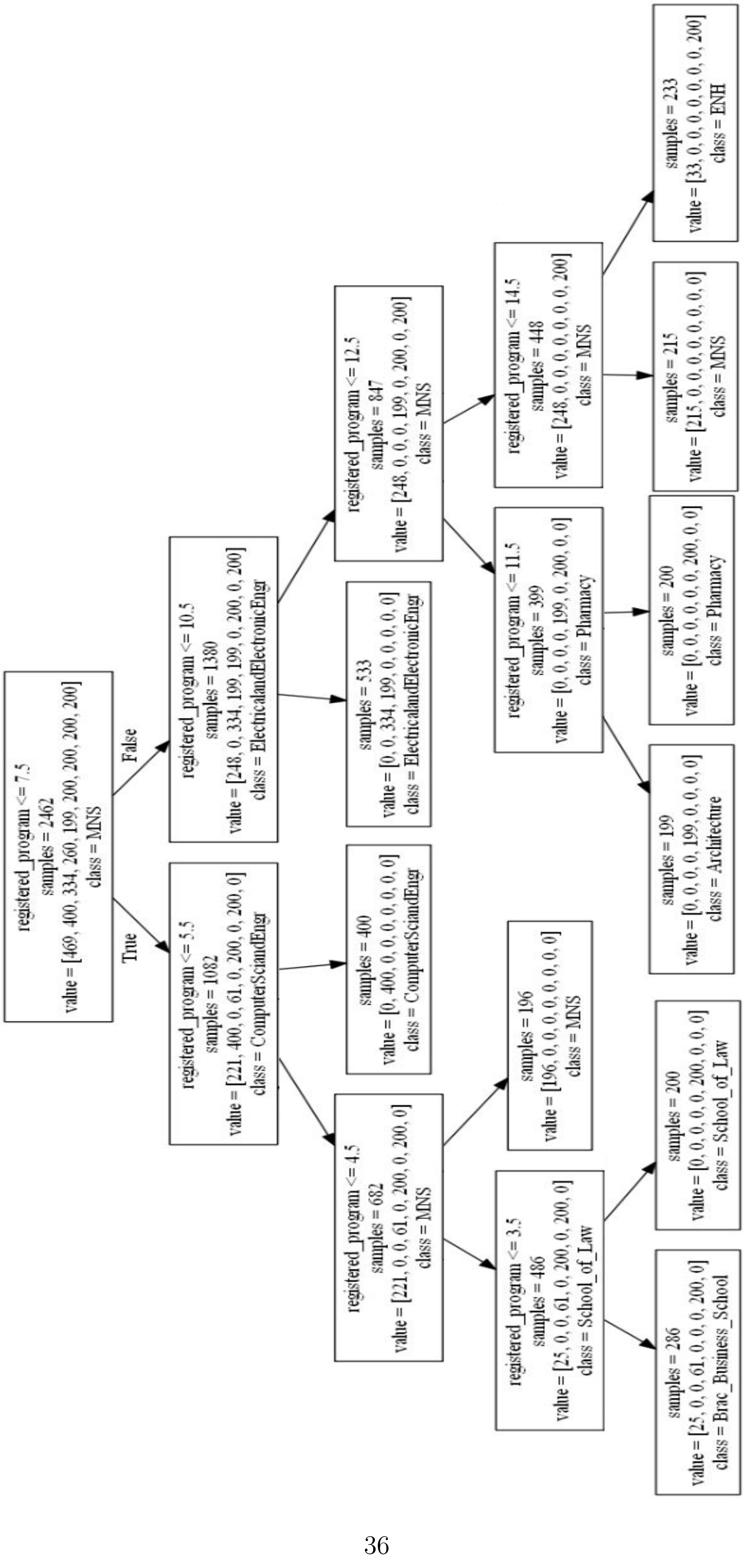
Pruning means trimming tree. It is technique that assists to diminish the growth of tree on the data set after a certain level to avoid over fitting. When the tree reacher a level where there is only a single feature at each node, it does not need to further split it into smaller samples. The hyper parameters can be used to perform pruning and stop the decision tree from growing too big and entirely matches the training set. To validate the limits for which the model will perform well on both train and test data set, the hyper parameter distribution was tested. The parameter affecting the model score is shown in the table 4.6.

Even though, the accuracy score for the training set decreased, it is now close to the score for the test set.It means the tree is no longer overfitted. Now among these two criteria, Gini index is performing comparatively well. For this reason, gini index has been utilized intended to build the proper decision tree on the model. So the final score is 86.85% , after fitting the model on the USIS dataset and handling overfitting using pruning technique.

Table 4.6: Parameter affecting Model Score

Criterion	max depth	max leaf nodes	min samples leaf	Accuracy Score
Gini index	3	9	8	Train_Set: 86.89% Test_set: 86.68%
Entropy	3	9	7	Train_Set: 86.89% Test_set: 85.87%

Figure 4.13: Decision Tree


registered_program <= 7.5
samples = 2462
value = [469, 400, 334, 260, 199, 200, 200, 200, 200]
class = MNS
True
False
registered_program <= 5.5
samples = 1082
value = [221, 400, 0, 61, 0, 200, 0, 200, 0]
class = ComputerSciandEngr
registered_program <= 10.5
samples = 1380
value = [248, 0, 334, 199, 199, 0, 200, 0, 200]
class = ElectricalandElectronicEngr
samples = 400
value = [0, 400, 0, 0, 0, 0, 0, 0, 0]
class = ComputerSciandEngr
registered_program <= 4.5
samples = 682
value = [221, 0, 0, 61, 0, 200, 0, 200, 0]
class = MNS
samples = 533
value = [0, 0, 334, 199, 0, 0, 0, 0, 0]
class = ElectricalandElectronicEngr
registered_program <= 12.5
samples = 847
value = [248, 0, 0, 0, 199, 0, 200, 0, 200]
class = MNS
samples = 196
value = [196, 0, 0, 0, 0, 0, 0, 0, 0]
class = MNS
registered_program <= 3.5
samples = 486
value = [25, 0, 0, 61, 0, 200, 0, 200, 0]
class = School_of_Law
registered_program <= 11.5
samples = 399
value = [0, 0, 0, 0, 199, 0, 200, 0, 0]
class = Pharmacy
registered_program <= 14.5
samples = 448
value = [248, 0, 0, 0, 0, 0, 0, 0, 200]
class = MNS
samples = 200
value = [0, 0, 0, 0, 0, 200, 0, 0, 0]
class = School_of_Law
samples = 286
value = [25, 0, 0, 61, 0, 0, 0, 200, 0]
class = Brac_Business_School
samples = 199
value = [0, 0, 0, 0, 199, 0, 0, 0, 0]
class = Architecture
samples = 200
value = [0, 0, 0, 0, 0, 0, 200, 0, 0]
class = Pharmacy
samples = 215
value = [215, 0, 0, 0, 0, 0, 0, 0, 0]
class = MNS
samples = 233
value = [33, 0, 0, 0, 0, 0, 0, 0, 200]
class = ENH

4.2.4 Random Forest

Random forests are groups of grading trees or gradients. We have used the random forest algorithm for both classification as well as regression in our system by making n-tree samples to boot from the original data set of students' information [6].For each sample of the boot models, an unclassified grading tree or gradient is developed. with the modification that selects a random selection of forecasters in each node instead of selecting the finest tree among all forecasters and selects the strongest tree among those characteristics [5].We have generated and prepared Mean Absolute Error(MAE) which shows the remaining for each data point, taking just the outright estimation of each with the goal that does not cancel negative and positive residuals out. We, at that point, took the average of every one of these residuals.

$$MAE = \frac{1}{n} \sum |y - \widehat{y}|$$

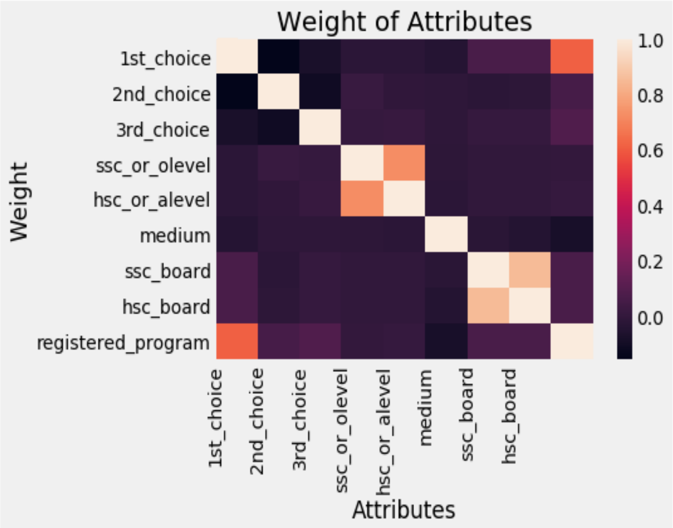
The importance of a variable is defined by its interaction with other variables which may become complicated. The RF algorithm appreciates the significance of the variable by considering the amount of error increase prediction. The necessary calculations are made from one tree to another while creating a random forest. For each t node and class probabilities computed as $p(k|t)$ $k = 1, 2, \cdots, Q$

$$G(t) = 1 - \Sigma_{k=1}^{Q} p^2(k|t)$$

Here, the quantity of classes is represented by Q, Gini index.

The randomForest function aid in returning an object from the classification "randomForest". By utilizing RF, we have characterized the joined data information. The estimations of properties are inspected from the result of the marginal distributions of the factors which are tested consistently from the hypercube containing the data by examining consistently within the scope of each attibutes. The genuine data indicates that are indistinguishable to each other will often be pointed in a similar terminal node of a tree which is estimated by the proximity matrix. Therefore, the proximity matrix can be taken as a closeness measure, and bunching or multi-dimensional scaling utilizing this comparability can be utilized to partition the original data points into sets for visual investigation. We can acquire the estimation of the error rate, in light of the training set information, by a proportion of the centrality of the desire factors and the estimation of the interior structure of the dataset which demonstrates the closeness of the various data to one another.

Figure 4.14: Attributes vs Weight

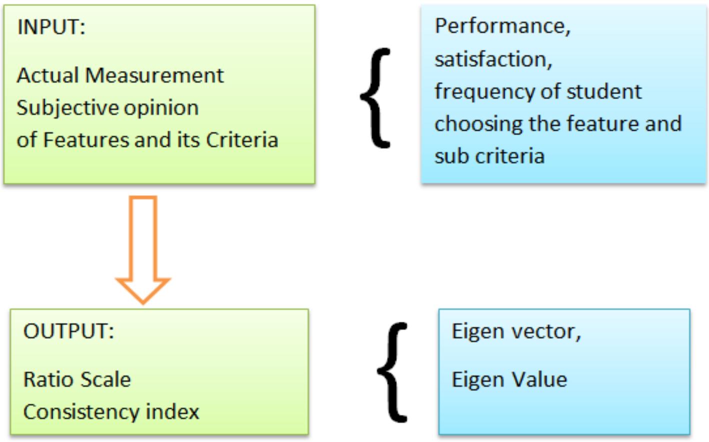


4.3 AHP Implementation

The intention of our paper is to originate a counseling system that will rank the choices of the students and help them decide which department or major would be beneficial for them. Based on 'Criteria Weight', we selected the features which impact mostly the outcome and then we applied AHP (analytical hierarchy process).

One of the initial intention of AHP process is deriving weight of features that can help in ranking alternatives. The Method derives ratio scales yielded from paired comparison. Furthermore, few inconsistency measurement in judgement can be drawn from this method for evaluation. As input we provided subjective opinion, for instance the frequency of students choosing or falling onto a particular criteria, in other words "feature", and sub-criteria. The output is ratio scales and consistency index.

Figure 4.15: Work flow of AHP method



The process is done in multiple paces.

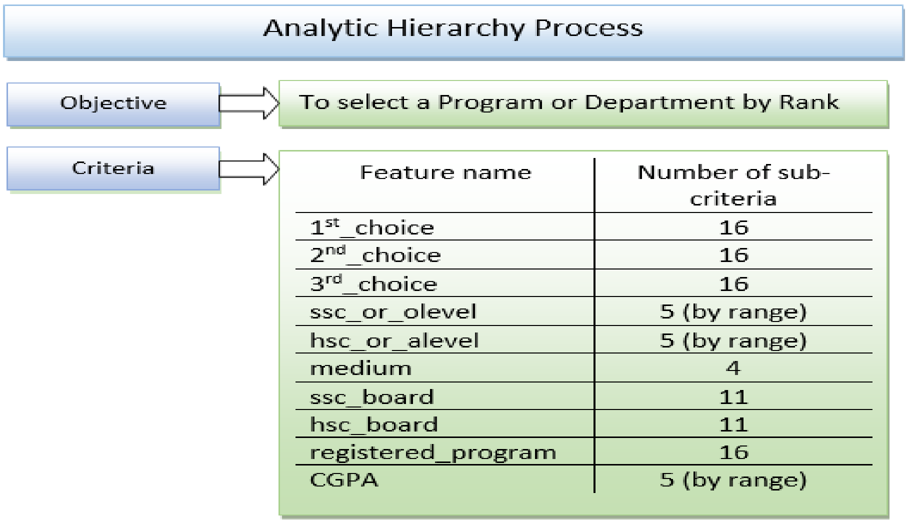
1. Clarify the objective for ranking
2. Combine elements in criteria organizations or key traits and sub-criteria and substitutes.
3. Generate a comparison in each organization between each of the elements.
4. Criteria Weight calculation.
5. Evaluate alternative or row value taken as a model from the data-set against weighting.
6. Ranking result.

AHP is viewd as a process that extracts ratio scale from comparisons of elements in pair form. It is a mathematical process with which we can combine individual performance indicator to one KPI and derive weights of features that which helps to generate ranking.

4.3.1 Define objectives and Structure elements in groups

For our data-set, the objective is to select a Program or Department by ranking from the most suitable for the students to the least suitable. The criteria are the features of the data-set that affects the outcome or target label. First step is to use requirements, sub-criteria and substitutes to group components.

Figure 4.16: Objective,Criteria and Sub-criteria



Analytic Hierarchy Process

Objective → To select a Program or Department by Rank

Criteria →

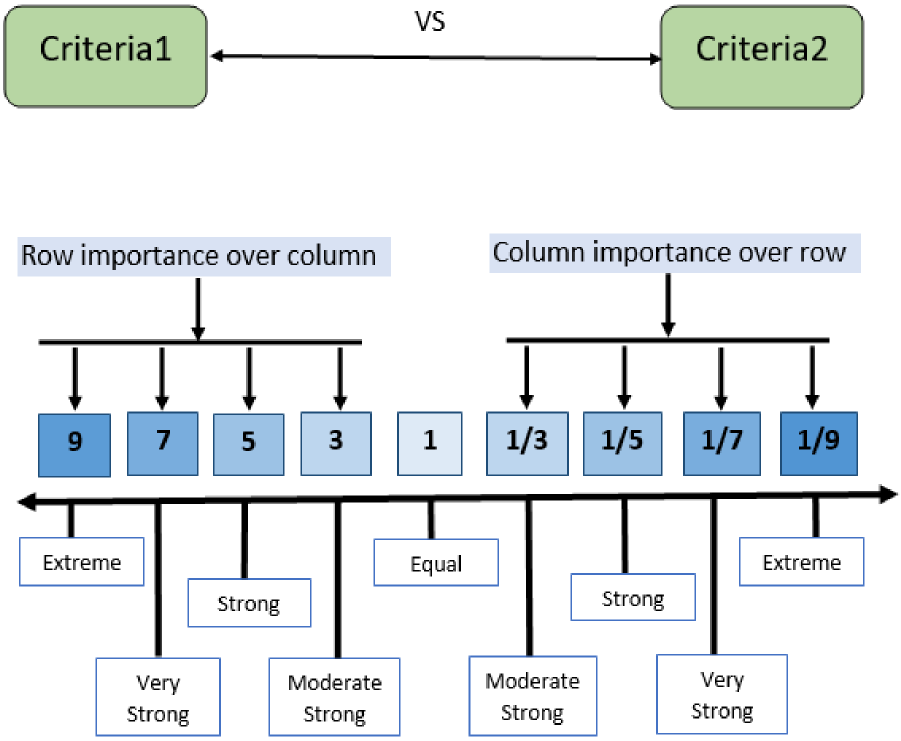
Feature name	Number of sub-criteria
1st_choice	16
2nd_choice	16
3rd_choice	16
ssc_or_olevel	5 (by range)
hsc_or_alevel	5 (by range)
medium	4
ssc_board	11
hsc_board	11
registered_program	16
CGPA	5 (by range)

4.3.2 Pairwise comparison matrix

For applying AHP method, we need to compare all elements pairwise with respect to the objective. The comparison of specifications is based on the significance of one criterion over another. For instance, if “Criteria1” is to be compared with “Criteria2” and the scale ranging from 1/9 to 9 then

1. 1 means ”Criteria1” and ”Criteria2” holds equal significance hence they are the same.
2. 9 means “Criteria1” is extremely important over “Criteria2”.
3. 1/9 interprets as “Criteria1” to be 9 times less significant than “Criteria2” (inverse comparison).
4. 2,4,6,8 are intermediate scale values.

Figure 4.17: Criteria Importance Comparison Scale



Now each of the criteria will be compared pairwise and their sub-criteria as well. For ease of understanding, a matrix representation is utilized. In order to generate comparing pair-wise elements of all the main criteria or features a 10x10 matrix is used to represent 10 criteria in the Student information data-set.

To determine which criteria is more important than another, we have calculated the preference or frequency of a student selecting that particular criteria. The importance of one feature above another feature in pair comparison was measured by the frequency of students selecting or falling onto that option.

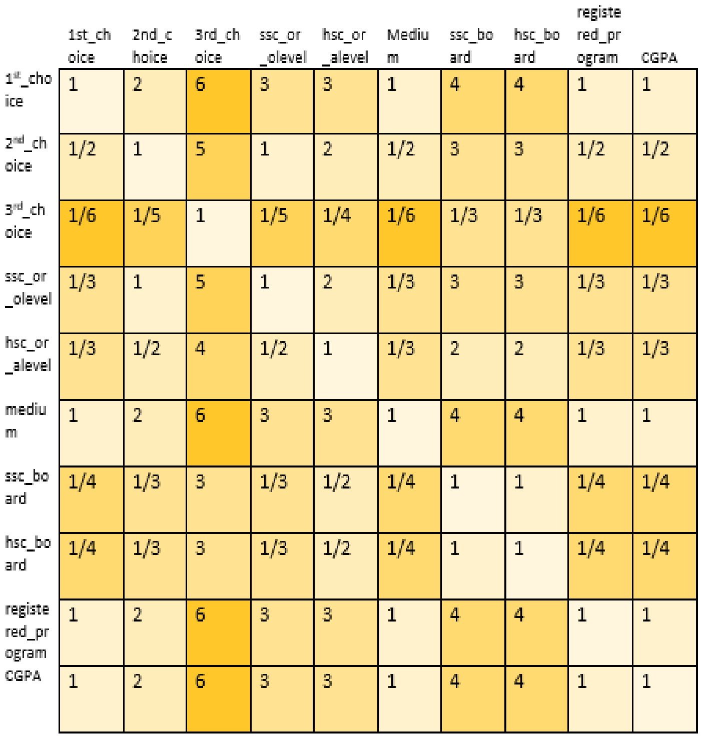
For example, a student is more likely to fill up the feature “1st_choice” in their admission form than “2nd_choice” or “3rd_choice” according to the USIS data-set. In the data-set we have 2462 individual entries. From the data-set, for each of the 10 features and their sub-criteria, we have to compare how important it is with the help of the criteria importance scale.

Table 4.7: Feature Preference Table

Features	Number of Entries	Total Null values	Preferences %
1st_choice	2462	0	100
2nd_choice	2435	27	98.9
3rd_choice	1351	1111	55.55
ssc_or_olevel	2229	233	90.53
hsc_or_alevel	2132	330	86.59
medium	2462	0	100
ssc_board	1881	581	76.40
hsc_board	1872	590	76.03
registered_ program	2462	0	100
CGPA	2462	0	100

From the calculation of preferences from the above table, we can see that the features "1st_choice", "medium", "registered_program" and "CGPA" has equal importance with one another. Furthermore, "1st_choice" has moderate-strong importance than "2nd_choice", extreme importance over "3rd_choice" as almost half of the students in the data-set did not fill up this field. Moreover "1st_choice" has moderate to strong importance over "ssc_or_olevel" and "hsc_or_alevel" and so on. Considering the above scenario based on preferences of the students, we produced the pair-wise correlation simulation for the criteria. Each feature is contrasted with each other in pairs. The next figure 4.18 demonstrates the comparison of the requirements or characteristics by pairs.

Figure 4.18: Pairwise Comparison



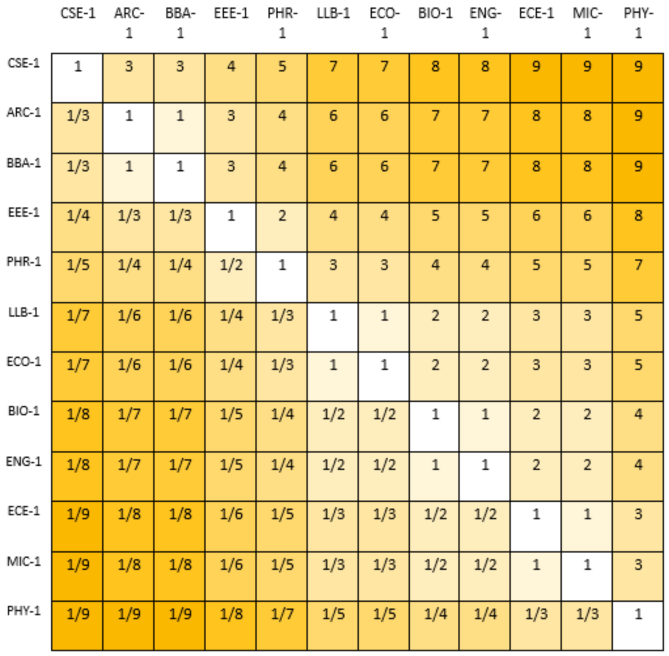
	1st_choice	2nd_choice	3rd_choice	ssc_or_olevel	hsc_or_alevel	Medium	ssc_board	hsc_board	registered_program	CGPA
1st_choice	1	2	6	3	3	1	4	4	1	1
2nd_choice	1/2	1	5	1	2	1/2	3	3	1/2	1/2
3rd_choice	1/6	1/5	1	1/5	1/4	1/6	1/3	1/3	1/6	1/6
ssc_or_olevel	1/3	1	5	1	2	1/3	3	3	1/3	1/3
hsc_or_alevel	1/3	1/2	4	1/2	1	1/3	2	2	1/3	1/3
medium	1	2	6	3	3	1	4	4	1	1
ssc_board	1/4	1/3	3	1/3	1/2	1/4	1	1	1/4	1/4
hsc_board	1/4	1/3	3	1/3	1/2	1/4	1	1	1/4	1/4
registered_program	1	2	6	3	3	1	4	4	1	1
CGPA	1	2	6	3	3	1	4	4	1	1

In case of sub-criteria for each feature, we calculated the frequency or the number of students choosing or falling under the sub-criteria. And then compared each sub-criteria of the feature. For “1st_choice” there is 16 different choices or sub-criteria between which pair-wise comparison is generated. The table shows the preferences of a student under a particular sub-criteria in descending order. Furthermore, Fig no 4.19 shows the pairwise comparison in sorted order of importance scale from equal to extreme.

Table 4.8: Preference: Sub-criteria of ”1st_choice”

Sub criteria for “1st _choice”	Number of entries from dataset	Preference (%)
CSE-1	439	17.83
ARC-1	315	12.794
BBA-1	301	12.2258
EEE-1	248	10.0731
PHR-1	222	9.017
LLB-1	159	6.458164
ECO-1	151	6.133225
BIO-1	143	5.808286
ENG-1	125	5.077173
ECE-1	119	4.833469
MIC-1	110	4.467912
PHY-1	37	1.502843
CS-1	28	1.137287
ANT-1	27	1.096669
APE-1	21	0.852965
MAT-1	17	0.690496

Figure 4.19: Pairwise Comparison of 1st_Choice



	CSE-1	ARC-1	BBA-1	EEE-1	PHR-1	LLB-1	ECO-1	BIO-1	ENG-1	ECE-1	MIC-1	PHY-1
CSE-1	1	3	3	4	5	7	7	8	8	9	9	9
ARC-1	1/3	1	1	3	4	6	6	7	7	8	8	9
BBA-1	1/3	1	1	3	4	6	6	7	7	8	8	9
EEE-1	1/4	1/3	1/3	1	2	4	4	5	5	6	6	8
PHR-1	1/5	1/4	1/4	1/2	1	3	3	4	4	5	5	7
LLB-1	1/7	1/6	1/6	1/4	1/3	1	1	2	2	3	3	5
ECO-1	1/7	1/6	1/6	1/4	1/3	1	1	2	2	3	3	5
BIO-1	1/8	1/7	1/7	1/5	1/4	1/2	1/2	1	1	2	2	4
ENG-1	1/8	1/7	1/7	1/5	1/4	1/2	1/2	1	1	2	2	4
ECE-1	1/9	1/8	1/8	1/6	1/5	1/3	1/3	1/2	1/2	1	1	3
MIC-1	1/9	1/8	1/8	1/6	1/5	1/3	1/3	1/2	1/2	1	1	3
PHY-1	1/9	1/9	1/9	1/8	1/7	1/5	1/5	1/4	1/4	1/3	1/3	1

Table 4.9: Mapping of Continuous Values

Sub-Criteria	ssc_or_olevel	hsc_or_alevel	CGPA
Excellent	4.00 to 5.00	4.00 to 5.00	cgpa>=3.7
Good	3.00 to 3.99	3.00 to 3.99	3.7>cgpa >=3.3
Average	2.00 to 2.99	2.00 to 2.99	3.3>cgpa >=2.7
Below-Average	1.00 to 1.99	1.00 to 1.99	2.7>cgpa >=2.0
Poor	0 to 0.99	0 to 0.99	cgpa<2.0

Similarly, pairwise comparison was generated for all other sub criteria as well. For the feature "ssc_or_olevel", "hsc_or_alevel" and "CGPA" there are no distinct sub-categories. The features contain continuous values. As such, we can divide the feature into five different sub-categories considering a range of numbers. The above Table 4.9 represents the ranges. Following these representation we can find out the pair-wise comparison between the sub-criteria of "ssc_or_olevel", "hsc_or_alevel" and "CGPA". A table below shows the pairwise comparison.

Table 4.10: Comparison of the features by pairs (Sub-Criteria)

Sub-Criteria	Excell-ent	Good	Average	Below-Average	Poor
Excellent	1	3	5	7	9
Good	1/3	1	3	5	7
Average	1/5	1/3	1	3	5
Below-Average	1/7	1/5	1/3	1	3
Poor	1/9	1/7	1/5	1/3	1

4.3.3 Calculating criteria weights

Till now we have pairwise comparison matrix for criteria (denoted as n) and for sub-criteria (indicated as m). For generating the weights of each criteria or feature and the sub-criteria, we need to calculate the Eigen vector value. After generating the matrix containing pair-wise comparison, the sum of columns is calculated and dividing all the elements of the matrix for pair comparisons with the sum of the column of pairwise matrix. Thus, normalized pairwise matrix is generated. By averaging all elements in the row of the normalized matrix for pair comparisons, we get the criteria weights for each features. Matrix N for criteria n = 3 is shown below.

$$N = \frac{\begin{bmatrix} 1 & a_{12} & a_{13} \\ a_{21} & 1 & a_{23} \\ a_{31} & a_{32} & 1 \end{bmatrix}}{S_{C1} \quad S_{C2} \quad S_{C3}}$$

Sum of columns,

$$S_C = \sum_{i=1}^{n} column_i$$

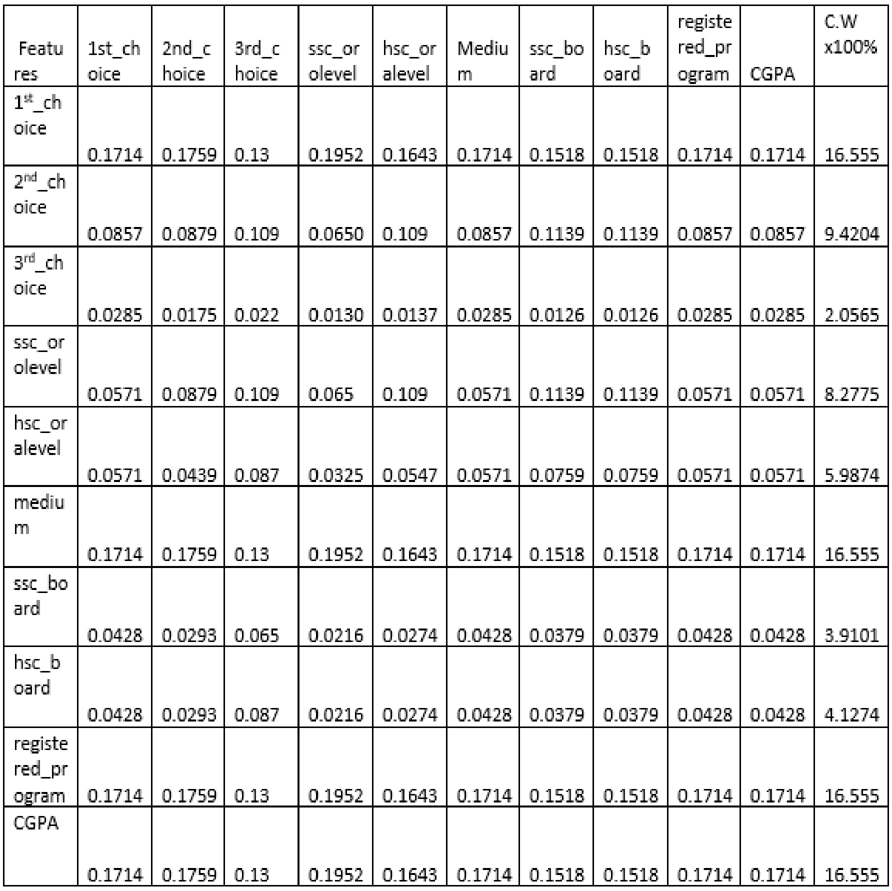
Normalized Matrix,

$$|N| = \begin{bmatrix} \frac{1}{S_{C1}} & \frac{a_{12}}{S_{C2}} & \frac{a_{13}}{S_{C3}} \\ \frac{a_{21}}{S_{C1}} & \frac{1}{S_{C2}} & \frac{a_{23}}{S_{C3}} \\ \frac{a_{31}}{S_{C1}} & \frac{a_{32}}{S_{C2}} & \frac{1}{S_{C3}} \end{bmatrix}$$

Eigen Vector,

$$X_1 = \begin{bmatrix} \sum_{i=1}^{n} row_1 \\ \sum_{i=1}^{n} row_2 \\ \sum_{i=1}^{n} row_3 \end{bmatrix}$$

Figure 4.20: Criteria Weight Calculation

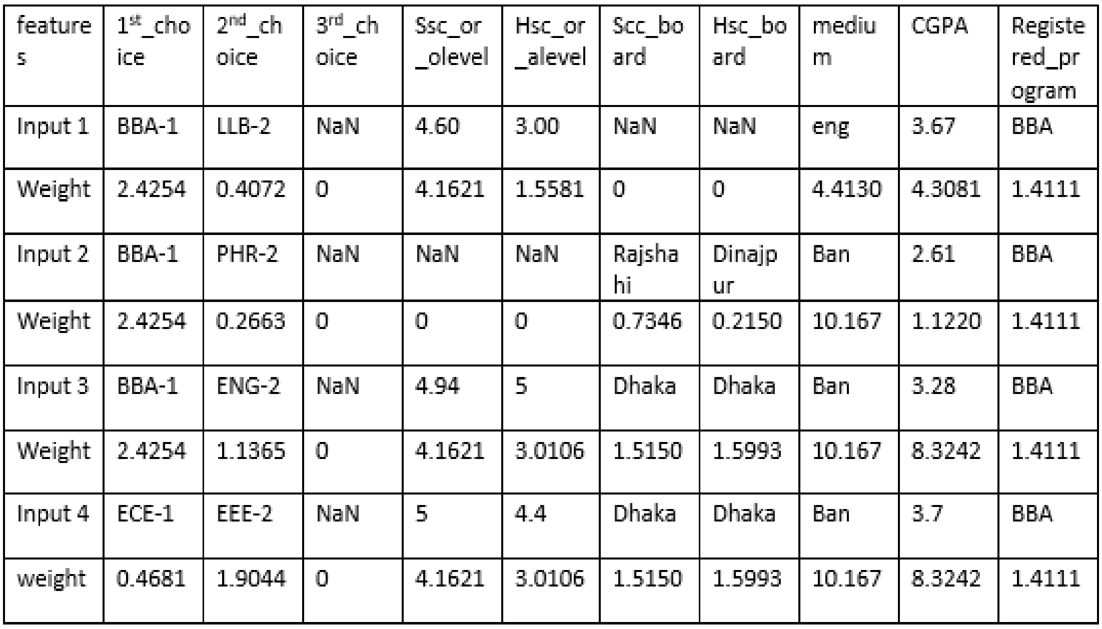


Features	1st_choice	2nd_choice	3rd_choice	ssc_orolevel	hsc_oralevel	Medium	ssc_board	hsc_board	registered_program	CGPA	C.W x100%
1st_choice	0.1714	0.1759	0.13	0.1952	0.1643	0.1714	0.1518	0.1518	0.1714	0.1714	16.555
2nd_choice	0.0857	0.0879	0.109	0.0650	0.109	0.0857	0.1139	0.1139	0.0857	0.0857	9.4204
3rd_choice	0.0285	0.0175	0.022	0.0130	0.0137	0.0285	0.0126	0.0126	0.0285	0.0285	2.0565
ssc_orolevel	0.0571	0.0879	0.109	0.065	0.109	0.0571	0.1139	0.1139	0.0571	0.0571	8.2775
hsc_oralevel	0.0571	0.0439	0.087	0.0325	0.0547	0.0571	0.0759	0.0759	0.0571	0.0571	5.9874
medium	0.1714	0.1759	0.13	0.1952	0.1643	0.1714	0.1518	0.1518	0.1714	0.1714	16.555
ssc_board	0.0428	0.0293	0.065	0.0216	0.0274	0.0428	0.0379	0.0379	0.0428	0.0428	3.9101
hsc_board	0.0428	0.0293	0.087	0.0216	0.0274	0.0428	0.0379	0.0379	0.0428	0.0428	4.1274
registered_program	0.1714	0.1759	0.13	0.1952	0.1643	0.1714	0.1518	0.1518	0.1714	0.1714	16.555
CGPA	0.1714	0.1759	0.13	0.1952	0.1643	0.1714	0.1518	0.1518	0.1714	0.1714	16.555

4.3.4 Ranking Model

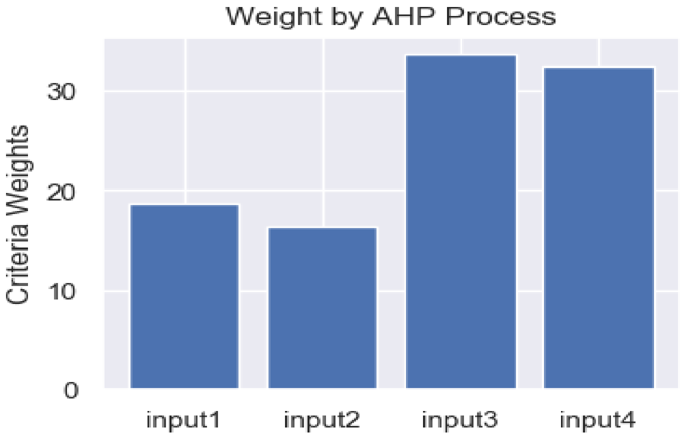
The last step of the AHP process is to evaluate alternative or row value taken as a model from the student information USIS data-set against weighting and get ranked models. We have generated weight for each feature and their sub-criteria through AHP process. Now, these weights can be used to determine ranks for the USIS student data-set. Each row or entries of the data-set will be considered as an input example or model for ranking the data-set. The idea is to add all the corresponding weights and generate ranked result. To show how the process works, we have taken 4 random raw data from our data-set as inputs and generated ranking between them. The figure below shows that the weights are added and the final weight summation is used to determine the ranking of the data entries.

Figure 4.21: Criteria Weight Evaluation



features	1st_choice	2nd_choice	3rd_choice	Ssc_or_olevel	Hsc_or_alevel	Scc_board	Hsc_board	medium	CGPA	Registered_program
Input 1	BBA-1	LLB-2	NaN	4.60	3.00	NaN	NaN	eng	3.67	BBA
Weight	2.4254	0.4072	0	4.1621	1.5581	0	0	4.4130	4.3081	1.4111
Input 2	BBA-1	PHR-2	NaN	NaN	NaN	Rajshahi	Dinajpur	Ban	2.61	BBA
Weight	2.4254	0.2663	0	0	0	0.7346	0.2150	10.167	1.1220	1.4111
Input 3	BBA-1	ENG-2	NaN	4.94	5	Dhaka	Dhaka	Ban	3.28	BBA
Weight	2.4254	1.1365	0	4.1621	3.0106	1.5150	1.5993	10.167	8.3242	1.4111
Input 4	ECE-1	EEE-2	NaN	5	4.4	Dhaka	Dhaka	Ban	3.7	BBA
weight	0.4681	1.9044	0	4.1621	3.0106	1.5150	1.5993	10.167	8.3242	1.4111

Figure 4.22: Bar Graph: Criteria Weight Evaluation



Chapter 5

Result Analysis

5.1 Comparison between Models

We have tuned parameters with the intend to delivering increased precision score utilizing Decision tree, Linear SVM, KNN and RF. Each classifier for modeling is distinguishable, so it encompasses innumerable tuning policies and calibrated parameters. With the desired constraints governed conditional on the ultimate gross classification precision, we have attempted to use a sequence of vital values for the tuning procedure. For every classifier, we utilized a progression of qualities for the adjustment procedure with the ideal dimensions decided dependent on the most elevated by and large grouping precision. We looked at the exhibition of classifiers by utilizing the grouped outcomes under the ideal parameters of every classifier.

For tuning of KNN algorithm, we modified the distance measuring metrics which analyses the mean distance value of neighboring nodes for a particular node. This increased the accuracy by 3% resulting 83%.

We got the accuracy of 85% by simulating SVM algorithm. Even if the data-set is significantly smaller than usual, SVM algorithm has the ability to work effectively as it does not need to depend on the entire data. After tuning, there was no change in the accuracy.

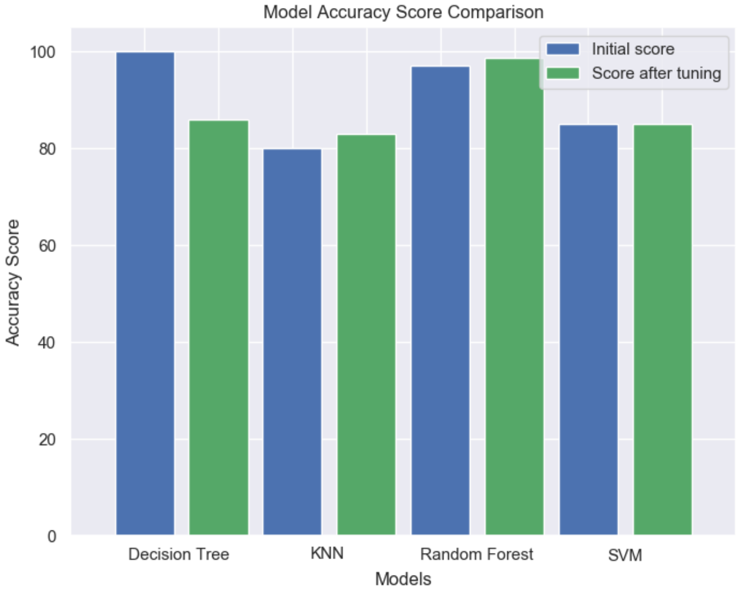
As, shown before, the decision tree model had an accuracy score of 100%, but in case of testing set the score dropped to 65%. The reason behind this is over fitting. The hyper parameters used in building the tree was tuned to overcome over fitting problem. After tuning the data accordingly the Accuracy score was 86%.

As, Random forest algorithm randomly generate decision trees the accuracy was quite high among all the algorithms we have been used which was 97%. Furthermore, we adjusted the values of estimator and random states for each iteration to get the best accuracy result for our current data set. Finally we obtained accuracy rate of 98.7%. To sum up the comparison we choose Random Forest as our classifier. The comparison between the original score and the adjusted score is shown in table 5.1.

Table 5.1: Model Accuracy Score Comparison

Models	**Decision Tree**	**KNN**	**Random Forest**	**SVM**
Initial accuracy score:	100%	80%	97%	85%
Tuned data accuracy score:	86%	83%	98.7%	85%

Figure 5.1: Bar Chart of Model Accuracy Score Comparison



5.2 Ranking by Random Forest

For ranking purpose, we used Random Forest as it is a robust, nonlinear, and doesn't require scaling to get the feature importance in order to get less over-fitting by minimizing features and lessen the computation time to improve precision. At first, the Model Based Ranking was performed so that, we can fit a classifier to each attribute and rank the predictive power which selects the most powerful attributes individually but ignores the predictive power when the attributes are combined. Besides, Recursive Feature Elimination recursively selects important subsets of features based on built-in attributes like coefficients or feature importance. We tried to map pair-plotting for the attributes and find out the correlation between the attributes using Random forest classifier but the points were to much scattered to define a correlation. For mean decrease accuracy, the decline in accuracy is measured when we need to exclude a single feature. It randomly sub-samples instances and features, selects good features on each subset and aggregates the results. We produced the Matrix of the considerable number of features alongside the individual model scores which we can utilize in generating our ranked positioning outcome.

To summarize, this pair plotting method serves to apply the feature determination on various pieces of the information and highlights over and over until the outcomes can be collected. Subsequently more grounded features will have higher scores in this strategy when contrasted with more fragile features. Finally, we created a function intended to almost certainly advantageously store our element rankings generated from the methods. As we could not relate the features clearly using Random forest classifier, we used AHP to find the correlation more accurately.

5.3 Results Generated by AHP

AHP process helps to determine weights for each criteria and its sub-criteria. by summation of these weight values in correspondence to the input, we can get a ranked data-set. The purpose is to develop a counseling system that will rank the choices of the students and help them decide which department or major is most beneficial for them.

The figures below shows the criteria weights bar chart of the main features and sub-criteria "1st_choice" and "2nd_choice" generated by AHP. The sub-criteria weights is converted according to the weights of main-criteria. By interpreting the bar charts we can see that the most important criteria are "1st_choice", "medium", "registered_program" and "CGPA".

Figure 5.2: Weights of Criteria

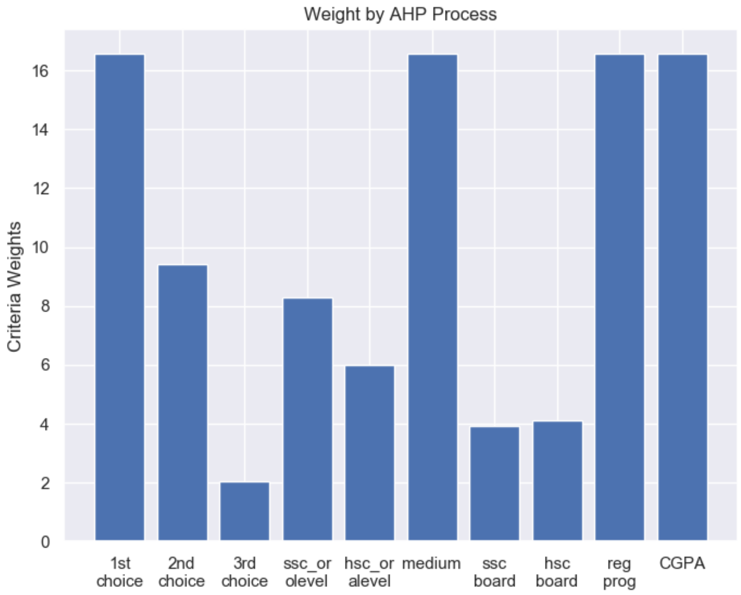


Figure 5.3: CW of "1st_Choice"

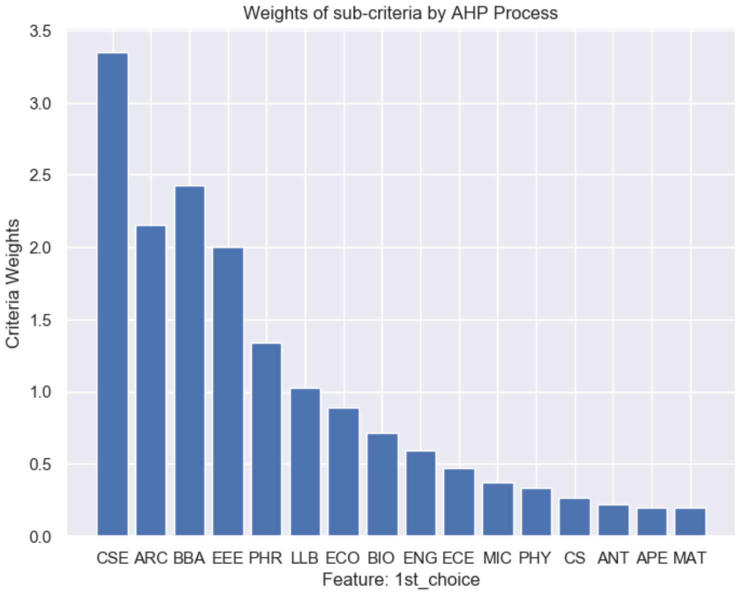
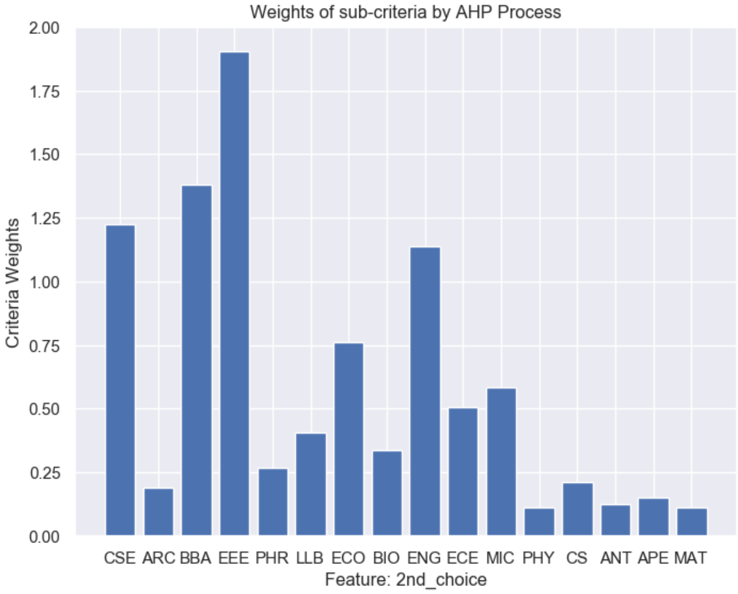
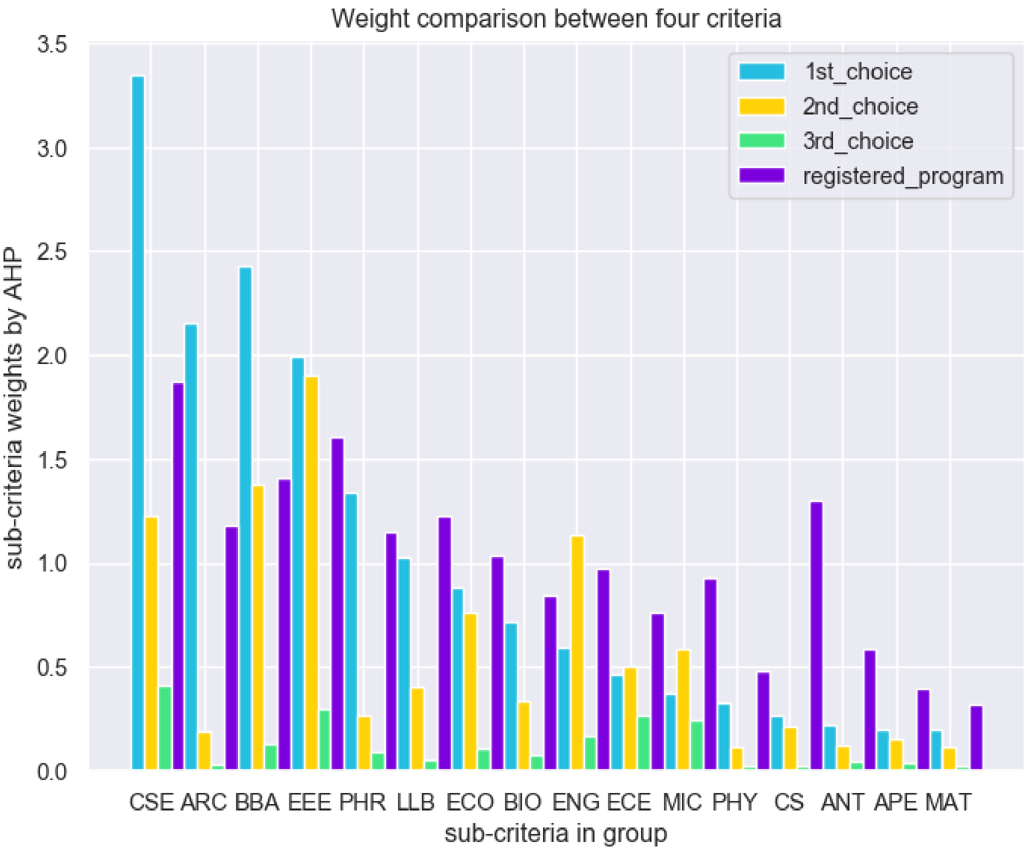


Figure 5.4: CW of "2nd_Choice"



The figure below represents a bar graph of the AHP weights that has been calculated for different sub-criteria for "1st_Choice", "2nd_Choice", "3rd_Choice" and "Registered_Program". The intention of showing the bar graph relation between these four criteria is to show how different each weighted result is according to the USIS student information data-set having similar sub-criteria.

Figure 5.5: Bar graph of AHP Weights



In order to complete the ranking mechanism, we have calculated weights of all the criteria along with their sub-criteria taken from the USIS student database. The weights of the sub-criteria have been transformed according to their primary or parent categories. The next three tables (Table 5.2, 5.3 and 5.4) contain the weighted value that was calculated using AHP method. These weighted results will be utilized in building the ranked outcome by which we will be able to evaluate our data-set and determine the best fitting option for counselling freshmen.

Table 5.2: AHP Weighted Table Part 1

1st_choice		2nd_choice		3rd_choice		Registered_ Program	
weight: 16.555%		weight: 9.420%		weight: 2.056%		weight: 16.555%	
sub-feature	weight%	sub-feature	weight %	sub-feature	weight %	sub-feature	weight%
CSE	3.3468	EEE	1.9044	CSE	0.4157	CSE	1.8987
ARC	2.1540	CSE	1.2257	ECE	0.2675	EEE	1.6095
BBA	2.4254	BBA	1.3801	EEE	0.3012	BBA	1.4111
EEE	1.9974	ENG	1.1365	MIC	0.2481	CS	1.2992
PHR	1.3403	ECO	0.7626	ENG	0.1664	LLB	1.2245
LLB	1.0295	MIC	0.5858	BBA	0.1278	ARC	1.1807
ECO	0.8868	ECE	0.5046	ECO	0.1101	PHR	1.1520
BIO	0.7157	LLB	0.4072	PHR	0.0889	ECO	1.0323
ENG	0.5943	BIO	0.3382	BIO	0.0738	ENG	0.9742
ECE	0.4681	PHR	0.2663	LLB	0.0581	MIC	0.9309
MIC	0.3749	CS	0.2133	ANT	0.0465	BIO	0.8442
PHY	0.3320	ARC	0.1889	APE	0.0412	ECE	0.7607
CS	0.2682	APE	0.1526	ARC	0.0333	ANT	0.5866
ANT	0.2229	ANT	0.1268	CS	0.0276	PHY	0.4826
APE	0.1976	MAT	0.1124	PHY	0.0245	APE	0.3948
MAT	0.2006	PHY	0.1141	MAT	0.0249	MAT	0.3211

Table 5.3: AHP Weighted Table Part 2

ssc_or_olevel		hsc_or_alevel		CGPA		medium	
weight: 8.277%		weight: 5.987%		weight: 16.555%		weight: 16.555%	
sub feature	weight %	sub feature	weight %	sub feature	weight %	sub feature	weight %
4.00 to 5.00	4.16212	4.00 to 5.00	3.0106	cgpa>= 3.7	8.324	ban	10.16
3.00 to 3.99	2.15408	3.00 to 3.99	1.5581	3.7>cgpa >=3.3	4.308	eng	4.413
2.00 to 2.99	1.11209	2.00 to 2.99	0.8044	3.3>cgpa >=2.7	2.224	eng-ver	1.264
1.00 to 1.99	0.56103	1.00 to 1.99	0.4058	2.7>cgpa >=2.0	1.122	madrasah	0.710
0.00 to 0.99	0.28823	0.00 to 0.99	0.20849	cgpa<2.0	0.576		

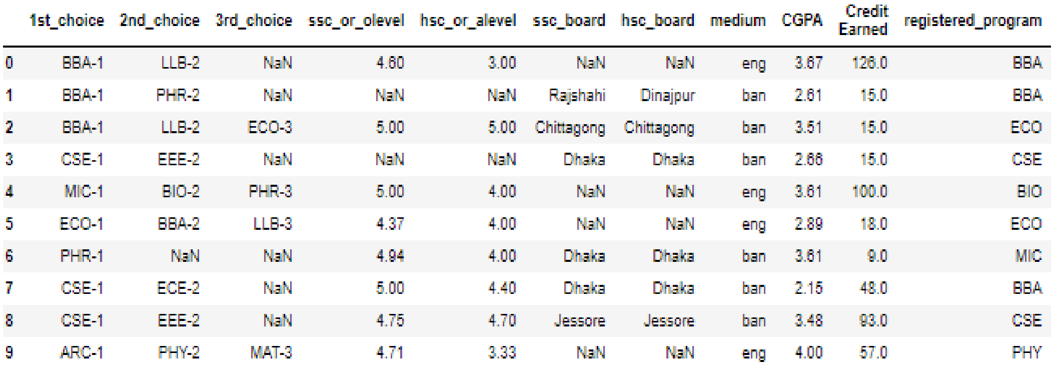
Table 5.4: AHP Weighted Table Part 3

ssc_board		hsc_board	
weight: 3.9191%		weight: 4.1274%	
sub-feature	weight %	sub-feature	weight %
Dhaka	1.515	Dhaka	1.599
Rajshahi	0.734	Chittagong	0.775
Chittagong	0.362	Rajshahi	0.382
Comilla	0.292	Jessore	0.309
Jessore	0.203	Dinajpur	0.215
Dinajpur	0.172	Comilla	0.182
Sylhet	0.143	Sylhet	0.151
Barisal	0.125	Barisal	0.132
Others	0.125	Others	0.132
Madrasah	0.115	CBSE	0.125
CBSE	0.119	Madrasah	0.125

5.4 Ranking Data-set by AHP

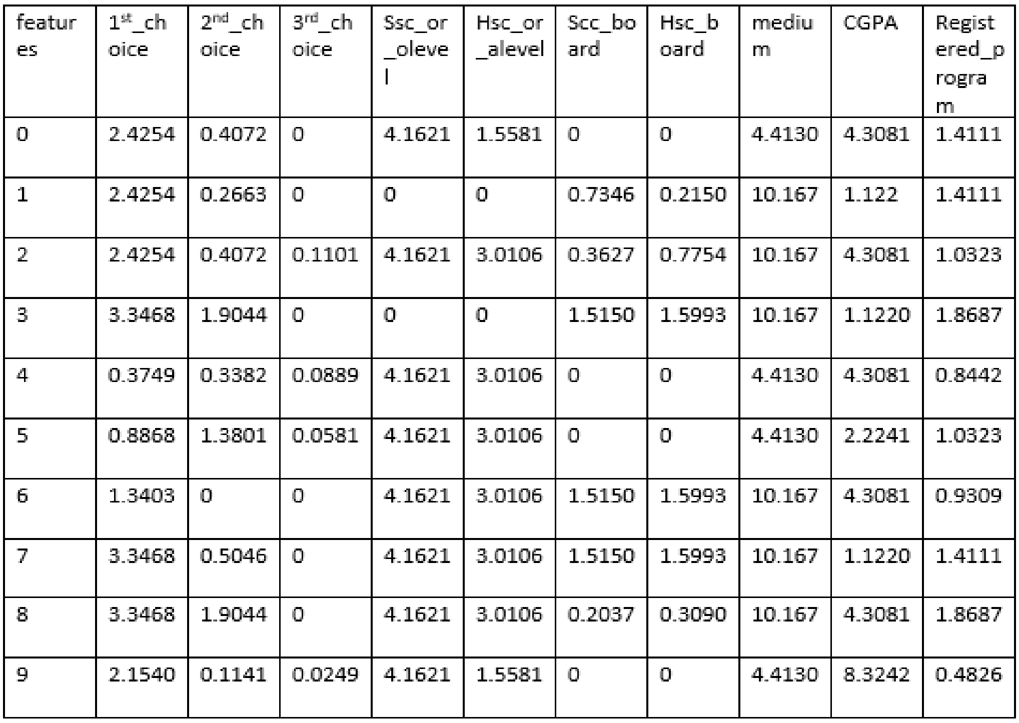
For analysing or evaluating the data-set, we have taken the first 10 data as input from our student information USIS data-set. For input we have considered each rows of the data-set to measure the final weights to determine the ranks of data. Note that the features ”Credit Earned” and ”target_dept” has been ignored. A logical reason behind it is that, one cannot simple determine a students performance by how many courses he/she has completed or how many credits he/she has earned. To each a certain level of CGPA or above average result, calculating ”Credit Earned” is irrelevant. Similarly the feature ”target_dept” is nothing but a generalised form of the department. As such, calculating particular individual program and general department as well is a redundant work. After calculating the weight and ranks, it shows that data entry no 8 has the highest rank. A data head showing the sorted data with respect to its ranking value has been shown below.

Figure 5.6: Raw Data head



	1st_choice	2nd_choice	3rd_choice	ssc_or_olevel	hsc_or_alevel	ssc_board	hsc_board	medium	CGPA	Credit Earned	registered_program
0	BBA-1	LLB-2	NaN	4.60	3.00	NaN	NaN	eng	3.67	126.0	BBA
1	BBA-1	PHR-2	NaN	NaN	NaN	Rajshahi	Dinajpur	ban	2.61	15.0	BBA
2	BBA-1	LLB-2	ECO-3	5.00	5.00	Chittagong	Chittagong	ban	3.51	15.0	ECO
3	CSE-1	EEE-2	NaN	NaN	NaN	Dhaka	Dhaka	ban	2.66	15.0	CSE
4	MIC-1	BIO-2	PHR-3	5.00	4.00	NaN	NaN	eng	3.61	100.0	BIO
5	ECO-1	BBA-2	LLB-3	4.37	4.00	NaN	NaN	eng	2.89	18.0	ECO
6	PHR-1	NaN	NaN	4.94	4.00	Dhaka	Dhaka	ban	3.61	9.0	MIC
7	CSE-1	ECE-2	NaN	5.00	4.40	Dhaka	Dhaka	ban	2.15	48.0	BBA
8	CSE-1	EEE-2	NaN	4.75	4.70	Jessore	Jessore	ban	3.48	93.0	CSE
9	ARC-1	PHY-2	MAT-3	4.71	3.33	NaN	NaN	eng	4.00	57.0	PHY

Figure 5.7: Corresponding Weights



features	1st_choice	2nd_choice	3rd_choice	Ssc_or_olevel	Hsc_or_alevel	Scc_board	Hsc_board	medium	CGPA	Registered_program
0	2.4254	0.4072	0	4.1621	1.5581	0	0	4.4130	4.3081	1.4111
1	2.4254	0.2663	0	0	0	0.7346	0.2150	10.167	1.122	1.4111
2	2.4254	0.4072	0.1101	4.1621	3.0106	0.3627	0.7754	10.167	4.3081	1.0323
3	3.3468	1.9044	0	0	0	1.5150	1.5993	10.167	1.1220	1.8687
4	0.3749	0.3382	0.0889	4.1621	3.0106	0	0	4.4130	4.3081	0.8442
5	0.8868	1.3801	0.0581	4.1621	3.0106	0	0	4.4130	2.2241	1.0323
6	1.3403	0	0	4.1621	3.0106	1.5150	1.5993	10.167	4.3081	0.9309
7	3.3468	0.5046	0	4.1621	3.0106	1.5150	1.5993	10.167	1.1220	1.4111
8	3.3468	1.9044	0	4.1621	3.0106	0.2037	0.3090	10.167	4.3081	1.8687
9	2.1540	0.1141	0.0249	4.1621	1.5581	0	0	4.4130	8.3242	0.4826

Figure 5.8: Data Ranking by AHP

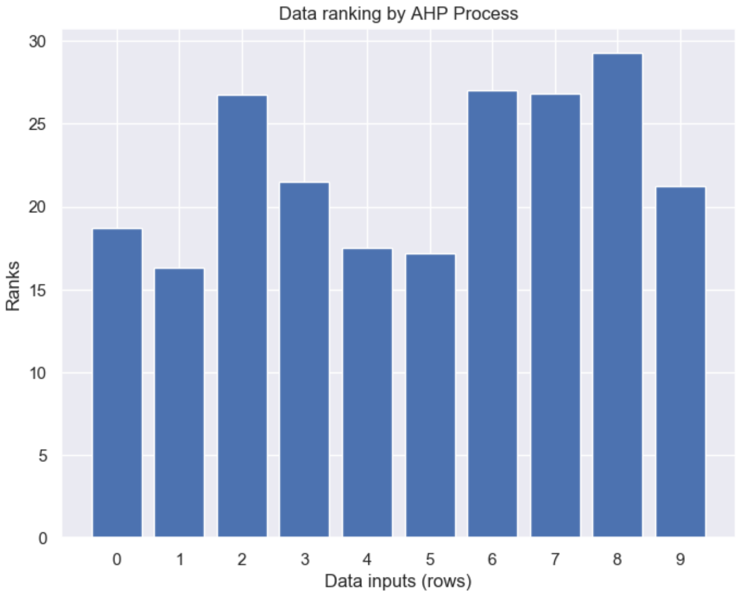
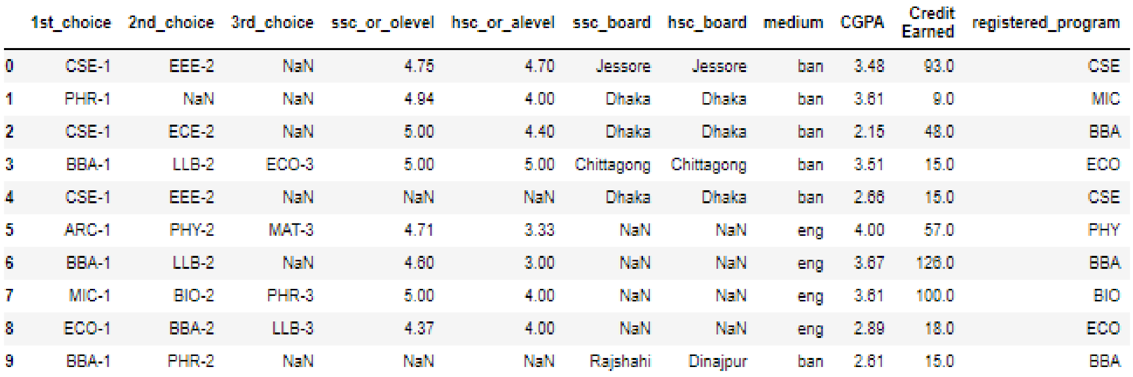


Figure 5.9: Data-head: Sorted by rank



	1st_choice	2nd_choice	3rd_choice	ssc_or_olevel	hsc_or_alevel	ssc_board	hsc_board	medium	CGPA	Credit Earned	registered_program
0	CSE-1	EEE-2	NaN	4.75	4.70	Jessore	Jessore	ban	3.48	93.0	CSE
1	PHR-1	NaN	NaN	4.94	4.00	Dhaka	Dhaka	ban	3.61	9.0	MIC
2	CSE-1	ECE-2	NaN	5.00	4.40	Dhaka	Dhaka	ban	2.15	48.0	BBA
3	BBA-1	LLB-2	ECO-3	5.00	5.00	Chittagong	Chittagong	ban	3.51	15.0	ECO
4	CSE-1	EEE-2	NaN	NaN	NaN	Dhaka	Dhaka	ban	2.66	15.0	CSE
5	ARC-1	PHY-2	MAT-3	4.71	3.33	NaN	NaN	eng	4.00	57.0	PHY
6	BBA-1	LLB-2	NaN	4.60	3.00	NaN	NaN	eng	3.67	126.0	BBA
7	MIC-1	BIO-2	PHR-3	5.00	4.00	NaN	NaN	eng	3.61	100.0	BIO
8	ECO-1	BBA-2	LLB-3	4.37	4.00	NaN	NaN	eng	2.89	18.0	ECO
9	BBA-1	PHR-2	NaN	NaN	NaN	Rajshahi	Dinajpur	ban	2.61	15.0	BBA

Criteria are features and each criterion have sub-criteria's. The idea is to compare all element pairwise with respect to the objective. Here 1 means that both criteria in the pair is of equal importance. Value 9 means that criteria1 is very very important than criteria 2 and 1/9 means the reverse. The importance of one feature above another in pair comparison was measured by the frequency of students selecting or falling onto that option.

After generating weights of each criterion from AHP method, we then tested it against our dataset. Each row of the dataset is considered as an individual Model and then the weight was calculated for that model. By the resulting weigh the models or rows are ranked. For instance, 4 rows from the dataset was taken as Models and weighted for ranking. As a result, we get a ranked dataset.

Chapter 6

Final Remark

6.1 Conclusion

This article aims at establishing an effective counseling scheme, which can not only predict data from the training set but also learn on basis of previous databases with the intend of creating new information that is beneficial for the students. The number of academic information stored in educational databases is expanding quickly. These databases contains student information which can help predict student performances. Furthermore, hidden data or information from these student databases can aid massively in data mining prediction models. It is clear however that, predicting from a large data-set is quite challenging when the information is constantly changing and increasing.

If we glance at the education system of our country Bangladesh, we can see that it entirely lacks the modern digitize system of analysing and monitoring student progress. As such, a students performance is not being properly addressed. Two situations can be considered as the reason behind this. Firstly, insufficient and unsuitable prediction methods that exist now are failing to identify student's performance. Secondly, there are no system available to monitor individual courses that affects a students outcome in academic sector. As a result, students are facing great consequences on preparing for higher education followed by struggles of their career life.

Therefore, to overcome these challenges, we proposed a system in this paper which includes the idea of creating a coherent counselling system. The system will analyze previous performances of students, their choices and their outcomes, upon that determine ranked result of selected major programs that a new student can choose from which will be the most beneficial to him/her.

6.2 Limitations and Future Work Plan

Though our system performed well on the data-set providing us a desired outcome, there are scopes of areas to be improved. Our plan is to identify limitations and develop the system for a large scale data-set so that it can generate more precise results. The limitations that we faced are as follows

Firstly, it is clear that the data-set provided from USIS database lacks important factors that can affect the prediction, As mentioned before, we have worked in two different sets of student data. Due to confidentiality, we were only able to get 12 featured information from the USIS student databases. We compared it with the data received from the students through survey and came to the conclusion that there are many other factors which is essential for the counselling system. For instance, grades of individual courses taken by the student, evaluation on class performance, attendance, demographic data, personal information of the students and so on.

Secondly, there are more algorithms and methods to apply for ranking. To generate more accurate results, we need to apply several algorithms and compare outcomes to find the best possible option applicable for individual students given their information.

In the future, we hope to explore more and expand our data-set with necessary features. We have tried applying different modeling and ranking algorithms for instance AHP to classify the decisions. However, we are looking forward to implement the proper User Interface of the system so that the students can use the system to analyses their background and be familiar with the possibilities towards his career path to study for.

Bibliography

[1] J. R. Quinlan, "Induction of decision trees", *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[2] J. Mingers, "An empirical comparison of pruning methods for decision tree induction", *Machine Learning*, vol. 4, no. 2, pp. 227–243, Nov. 1989, ISSN: 1573-0565. DOI: 10.1023/A:1022604100933. [Online]. Available: https://doi.org/10.1023/A:1022604100933.

[3] C. Cortes and V. Vapnik, "Support-vector networks", *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[4] E. Bredensteiner and K. Bennett, "Multicategory classification by support vector machines", *Computational Optimization and Applications*, vol. 12, Sep. 1999. DOI: 10.1023/A:1008663629662.

[5] A. Liaw and M. Wiener, "Classification and regression by randomforest", *Forest*, vol. 23, Nov. 2001.

[6] H. Pang, M. Holford, H. Zhao, A. Lin, B. E. Enerson, B. Lu, E. Floyd, and M. P. Lawton, "Pathway analysis using random forests classification and regression", *Bioinformatics*, vol. 22, no. 16, pp. 2028–2036, Jun. 2006, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl344. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/22/16/2028/548216/btl344.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btl344.

[7] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance", 2008.

[8] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques", *Computers & Education*, vol. 53, no. 3, pp. 950–965, 2009.

[9] A. Buldu and K. Üçgün, "Data mining application on students' data", *Procedia-Social and Behavioral Sciences*, vol. 2, no. 2, pp. 5251–5259, 2010.

[10] Y. Tian, Y. Shi, and X. Liu, "Recent advances on support vector machines research", *Technological and Economic Development of Economy*, vol. 18, Mar. 2012. DOI: 10.3846/20294913.2012.661205.

[11] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining", *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 6, 2013.

[12] M. Nawaz, A. Adnan, U. Tariq, J. F. Salman, R. Asjad, and M. Tamoor, "Automated career counseling system for students using cbr and j48", *Journal of Applied Environmental and Biological Sciences*, vol. 4, pp. 113–120, 2014.

[13] "1 predictive analytics for student success : Developing data-driven predictive models of student success", 2015.

[14] S. Ismail and S. Abdulla, "Design and implementation of an intelligent system to predict the student graduation agpa", *Australian Educational Computing*, vol. 30, Dec. 2015.

[15] A. Nagpal and S. P. Panda, "Career path suggestion using string matching and decision trees", *CoRR*, vol. abs/1505.06306, 2015. arXiv: 1505.06306. [Online]. Available: http://arxiv.org/abs/1505.06306.

[16] E. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict student's academic performance using ensemble methods", *International Journal of Database Theory and Application*, vol. 9, pp. 119–136, Sep. 2016. DOI: 10.14257/ijdta.2016.9.8.13.

[17] R. Karim and C. Karmaker, "Machine selection by ahp and topsis methods", *American Journal of Industrial Engineering*, vol. 4, no. 1, pp. 7–13, 2016.

[18] M. Peker, "A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and svm", *Journal of Medical Systems*, vol. 40, no. 5, p. 116, Mar. 2016, ISSN: 1573-689X. DOI: 10.1007/s10916-016-0477-6. [Online]. Available: https://doi.org/10.1007/s10916-016-0477-6.

[19] A. Daud, N. Aljohani, R. Abbasi, M. Lytras, F. Abbas, and J. Alowibdi, "Predicting student performance using advanced learning analytics", Apr. 2017. DOI: 10.1145/3041021.3054164.

[20] C. Ezenkwu, E. Johnson, and O. Jerome, "Automated career guidance expert system using case-based reasoning technique", vol. 8, pp. 81–88, Apr. 2017.

[21] N. Gorad, I. Zalte, A. Nandi, and D. Nayak, "Career counselling using data mining", 2017.

[22] A.-S. Hoffait and M. Schyns, "Early detection of university students with potential difficulties", *Decision Support Systems*, vol. 101, pp. 1–11, 2017.

[23] C. Beaulac and J. S. Rosenthal, "Predicting university students' academic success and choice of major using random forests", *Research in Higher Education*, Feb. 2018. DOI: 10.1007/s11162-019-09546-y.