THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

A CAREER RECOMMENDATION FRAMEWORK TO HIGH SCHOOL STUDENTS USING SIMILAR FRESHMAN STUDENTS

Master's Thesis

SAAD SHAHID SALEEM ANSARI

ISTANBUL, 2021

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

MASTER OF BIG DATA ANALYTICS AND MANAGEMENT

A CAREER RECOMMENDATION FRAMEWORK TO HIGH SCHOOL STUDENTS USING SIMILAR FRESHMAN STUDENTS

Master's Thesis

SAAD SHAHID SALEEM ANSARI

Supervisor: ASSIST.PROF. DR. ÖZGE YÜCEL KASAP

ISTANBUL, 2021

T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL

...../...../.....

MASTER THESIS APPROVAL FORM

Program Name:	Big Data Analytics and Management
Student's Name and Surname:	Saad Shahid Saleem Ansari
Name Of The Thesis:	A CAREER RECOMMENDATION FRAMEWORK TO HIGH SCHOOL STUDENTS USING SIMILAR FRESHMAN STUDENTS
Thesis Defense Date:	26-01-2021

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Assoc. Prof. Dr. Burak KÜNTAY
Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	**Signature**
Thesis Advisor's	Assist.Prof, Dr. Özge Yücel Kasap	
Member's		
Member's		

DEDICATION

To my family, for their endless support and for always being there for me

ACKNOWLEDGEMENTS

I would like to take a moment and thank the following people, without whom I would not have been able to complete this thesis, and without whom I would not have made it through my master's degree!

The Faculty of Engineering and Applied Sciences at BAHÇEŞEHİR UNIVERSITY, especially to my supervisor ASSIST.PROF, DR. ÖZGE YÜCEL KASAP whose insight and knowledge into the subject matter steered me through this research. And special thanks to my peers and friends who helped me in tough times and sailed me through rough waves.
The students, who took the time to return surveys and allowed me a couple of minutes from their busy lives without whom I would have no content for my thesis.

And my biggest thanks to my family for all the support you have shown me through this degree, the culmination of two years of distance learning. For my parents who are an inspiration for me and who make me strive harder in life. Thanks for all your support, without which I would have stopped these studies a long time ago, you have been amazing.

İstanbul, 2021 Saad Shahid Saleem ANSARI

ABSTRACT

A CAREER RECOMMENDATION FRAMEWORK TO HIGH SCHOOL STUDENTS USING SIMILAR FRESHMAN STUDENTS

Saad Shahid Saleem Ansari

Master of Big Data Analytics & Management

Thesis Supervisor: Assist.Prof,Dr. ÖZGE YÜCEL KASAP

November 2021, 80 Pages

The reason for this thesis is to explore the connections or similarities between high school students and university students. The features and their performance factors would include suggestions for high school students to seek the required further education when they reach the university. The data that were collected from the students in this thesis were high school valedictorians who were about to choose the field of profession and university students studying in various institutions. The main purpose of doing this is because many students don't pursue a career according to their liking, they either choose some other career due to social pressure (Which happens mostly in Asia) and many other reasons. Using Big Data, the uncertainty can be removed and students can pursue their dream careers or the ones in which they are most suited. The goal was to investigate the factors which impact the change of major and the determination of program of study and the effect of various external factors prompting the change of major and the choice of a program of study. The autonomous factors were: age, skills, academic performance score, extra-curricular activities participation, interests in various departments i.e law, or many others. Prediction models were developed to identify which parameters were indeed critical in making predictions and to illuminate which variables were often best utilized to help them find a professional path that is convenient for them.

Keywords: Big Data, Big Data in Education, Career Planning Big Data, Career Recommendation system, Career Prediction.

ÖZET

ÖĞRENCİLER İÇİN MÜMKÜN EN İYİ KARİYER SEÇİMİNİ TAHMİN ETMEK VE TAVSİYE ETMEK

Saad Shahid Saleem Ansari

Büyük Veri Analitiği ve Yönetimi Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğr. Üyesi, ÖZGE YÜCEL KASAP

Kasım 2021, 80 Sayfa

Bu çalışmada, lise öğrencileri ile üniversite öğrencileri arasındaki bağlantıların ve benzerliklerin keşfedilmesi hedeflenmiştir. Lise öğrencilerinin akademik başarıları, becerileri ve ilgi alanları dikkate alınarak, üniversiteye girdiklerinde ilgili çalışmaları sürdürmeleri için bir öneri sağlamaya çalışılmıştır. Bu tezdeki öğrencilerden toplanan veriler, meslek alanını seçmek üzere olan lise öğrencileri ve çeşitli kurumlarda okuyan üniversite öğrencileridir. Bu çalışmanın yapılmasının ana amacı, birçok öğrencinin beğenilerine göre kariyer planlaması yapmaması, aksine ya toplum baskısı (Asya bölgesinde sıklıkla görülmektedir) ya da başka sebeplerden ötürü farklı bir kariyer seçmeleridir. Büyük Veri analizi kullanılarak bu belirsizlik ortadan kaldırılabilir ve öğrenciler hayal ettikleri veya onlara en uygun olan kariyerlere ulaşabilirler. Hedeflenen amaç, anadal değişikliğini ve eğitim programının belirlenmesini etkileyen faktörleri ve ayrıca anadal değişikliğine ve eğitim programı seçimine neden olan çeşitli dış faktörlerin etkisini araştırmaktır. Çalışmada kullanılan bağımsız değişkenler yaş, beceriler, akademik performans puanı, müfredat dışı faaliyet katılımları, çeşitli bölümlerdeki ilgi alanları (örneğin hukuk) da dahil olmak üzere çeşitli verilerdir. Kullanılan modeller, tahminlerde bulunmada hangi parametrelerin gerçekten kritik olduğunu belirlemek ve öğrencilerin kendileri için uygun olan profesyonel yolda ilerlemelerine yardımcı olmak için hangi değişkenlerin genellikle en faydalı sonucu verdiğine karar vermek için geliştirildi.

Anahtar Kelimeler: Büyük Veri, Eğitimde Büyük Veri, Kariyer Planlama Büyük Veri, Kariyer Öneri Sistemi, Kariyer Tahmin

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

BNN	:	Bayesian neural network
CGPA	:	Cumulative Grade Point Average
DT	:	Decision Tree
GPA	:	Grade Point Average
ID	:	Identification
KNN	:	K-nearest neighbors
LOGIT	:	Logistic
LR	:	Linear Regression
MAE	:	Mean Absolute Error
ML	:	Machine Learning
N/A	:	Not Applicable
NN	:	Neutral Networking
RF	:	Random Forests
RMS	:	Root Mean Square
RMSE	:	Root Mean Square Error
SPSS	:	Statistical Packages for Social Sciences
SVM	:	Support Vector Machine
XG Boost	:	Extreme Gradient Boosting

1. INTRODUCTION

Choosing a career is a fundamental decision that a student must take in his life. Several valedictorians, selecting the higher major and career pathway ought to be a difficult process. This judgment will have a major influence on their future and will impact them throughout their lives. A handful of factors are involved in the career-defining decision such as the academic performance of the students, their interest in various fields, and the skillsets they have. Several students choose a career mainly due to the pressure of society and the scope level of a field. Some students consider the salary packages of various jobs and want to choose the one which is the highest-paid. However, this hampers the performance of the students when they choose the majors in university eluding proper guidance because they fail to excel in it. Each student acquires an exceptional history about their past and this focuses on how they view the world, this history is made of its academics, the student's personality, and the opportunity that rises will decide how valedictorians settle on professional decisions. After that point, follows that how the candidate perceives their condition, character, and opportunity additionally will decide the vocational decisions the understudies make (Sahin 2012).

1.1 OBJECTIVE

The main objective of this thesis is to provide a roadmap for the high school students that have just graduated. This roadmap is laid out by the students who are currently studying in the universities. As understudies are experiencing their academics and seeking their potential courses, it is significant for them to evaluate their abilities and recognize their interests so that they will become acquainted with it and in which profession territory their interests and capacities will place them. This will help them in improving their performance and increasing their interests so they will be coordinated towards their focused profession and get comfortable with it. Additionally, enrollment specialists while selecting the competitors in the wake of surveying them in every single diverse viewpoint, these sorts of professional recommender frameworks help them in choosing which work job the applicant ought to be kept independent on his/her exhibition and different

assessments. This thesis mainly focuses on the professional zone expectation of software engineering space competitors (Daud 2017).

For most valedictorians graduating secondary school, the possibility of entering university is viewed as a major exciting experience. Stepping out of the comfort bubble of secondary school and into a different world. For certain candidates, be that as it may, the cycle can be amazingly distressing and overwhelming. Considerably more significant than where you go now, is the module or major that you are going to study. The procedure for selecting a university major can be highly confusing. It is a reason for extraordinary stress because for most it will probably be one of the most significant life choices the students will have to make (Adekola 2011).

1.2 PROBLEM STATEMENT

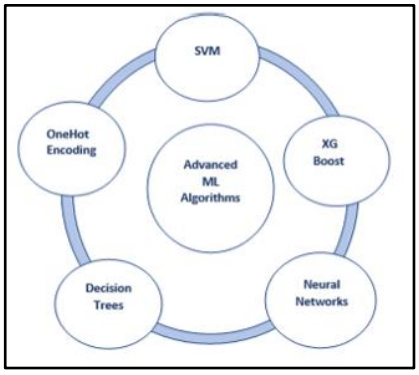
In the modern era, competition is increasing at a rapid pace. Particularly it is excessively substantial in the present day's specialized world. To contend and reach their maximum potential, the candidates should be arranged and sorted out from the beginning phases of their education according to their characteristics. This is the reason that prioritizing the evaluation of their academic performance, recognizing their skills, interests need to be done, this, in turn, helps them reach the peak point later in their careers. Not just that enrollment specialists while enlisting individuals into their organizations assess new graduates on various boundaries and reach the last conclusion whether to hire an employee or not and if chosen, finds the most appropriate job and professional territory for him. Every one of these jobs requires some pre-necessities in them to be put. In this way, selection representatives investigate these abilities, talents, and benefits and place the employee in the appropriate role for them. These sorts of prediction frameworks make the recruiters work simply because as the inputs are given, suggestions are done according to the inputs. As of now, they're all already kinds of different recommendation systems and employment job recommendations. They just take factors like specialized abilities and talents of candidates into consideration. These entries assess the candidates and propose to the students and organizations the best possible job roles suited to their performance. In any case, here different components remembering capacities of

understudies for sports, scholastics, and their diversions, interests, rivalries, aptitudes, and information are additionally thought about (Khousa 2014).

Another method to conduct this is to analyze and provide a recommendation based on different students in different departments and point out the similarities between high school students and university students and then recommend the former with the major according to it. The students can cluster together for a thorough analysis. Different Models like K-Nearest Neighbor, decision trees are used (Ere 2014).

Machine learning is a technique where the machines are prepared in a way that acquires the capacity to react to a specific input or situation based on the past information it has learned. In simple words, it is enabling computers to learn by utilizing statistical methods. Machine Learning encourages computers to act without unequivocally being customized. This helps in decreasing the human intercession in the machine dependable issues and situations. This aids in tackling exceptionally complex problems and issues effectively without including a lot of human work. Different aspects of machine learning include Natural Language Processing, Classification, Prediction, and many others. In this thesis, some of the same aspects of machine learning are used. Figure 1.1 shows a visualized outline of some of the most common algorithms that are used in machine learning. They are support vector machines, extreme gradient boost, decision trees, and neural networks (Ferry 2014).

Figure 1.1: The algorithms in machine learning



Source: (K. Sripath Roy 2018)

A larger part of issues in machine learning can be transpired by utilizing supervised and unsupervised learning. On the off chance that the last class marks are known beforehand, and the various data items are to be assigned with one of the accessible class labels, it is then known to be supervised learning (Bublitz 2013). On the other hand, if the last output classes and sets are not known and it is processed by distinguishing the similarities between data points, their qualities, and lastly, they are formed into groups based on these attributes then it is called unsupervised. Classifications are categorized in supervised. The data is given and according to their properties, a predefined class mark is allotted. There are different options like clustering and regression. Considering the kind of issue, a well-suited model is picked (Roy 2018).

Later, after training and testing the data with these, the next step is to mull over the most precise outcomes given algorithm for our further handling (Barnes 2005). Along these lines, the starting objective done is foreseeing the output by utilizing all calculations mentioned above and later investigating the outcomes and then proceed further with the most precise algorithm (Borchert 2002). Eventually, this thesis uses different progressive machine learning algorithms and comparisons and is used to improve the precision for better forecast, reliability, and breaking down these calculations' execution (Guruler 2010).

Profession represents one of the major fruitful milestones accomplished by the individual in his life (Adekola 2011). The career decisions are profoundly reliant on the student's character attributes, extra-curricular activities are taken part in, the mechanics of response towards confronting difficulties, and perspectives created in solving certain issues (Asanov 2011)

Especially, today there is a need to investigate the role of guidance and counseling in preparing students for work from high school students to university levels of training (Lulu 2002).

In the first world or third world countries, the education system is planned in a certain way which leaves very few options for the students to choose from (Firkola 2005).

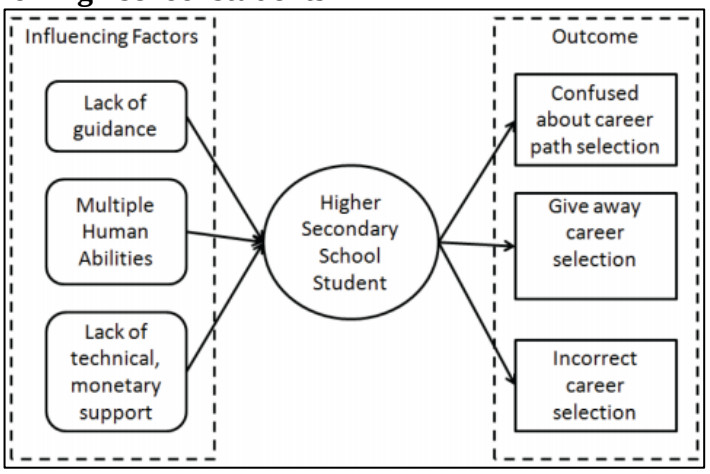
Furthermore, it is less helpful in the investigation of new professions or to pick the correct one for a certain candidate (Gardner 2011).

In the new age, picking an inappropriate career is one of the most significant measures of employment disappointments (Shen 2016). The goal of this thesis is to investigate and give a suggestion based on various candidates in various university major fields and point out the similarities between secondary school students and the university candidates and suggest the former with the major as per it. The students can group for a careful examination (Bauer 2005). As, the essential issue with high school students found to be puzzled while picking a college major, because of the presence of many different factors (Nel 2012).

Hence, a better analysis of carious college major choices will empower the students to increase the better perspective of various kinds of fields they would need to investigate in the future (Heckman 2012). It is fundamental for high school candidates to get driven and to create incredible knowledge paths through the cycle of appropriate studies followed by professional choice (Nowotny 2013). Moreover, High school students frequently might be ignited for incorrect decisions in choosing their major because of a lack of field-related experience, proper direction, and many other factors. Additionally, high school students often look flabbergasted while picking a college major or a career pathway, because of numerous uncontrollable factors. As technology is evolving rapidly, even information gathering is at the fingertips of humans (Lichtenberger 2013). Therefore, it is easy for the student to access the portal and gather information and investigate the recommendations of majors or career pathways on different platforms (Norma 2012).

It has become a favorable, financially efficient solution for those candidates who can't bear the cost of appropriate career counseling to pick the correct major because of the extremely high charges of career advocates (Elfaki 2015). A few hypotheses are created on how the career path preference is a key factor that further prompts career improvement, work fulfillment, professional commitment, and professional advancements as appeared in Figure 1.2.

Figure 1.2: Effecting parameters and their impact on the selection of career paths for high school students



Source: (Verma 2017)

It is to be noted that, the majority of the already conducted researches revolves around the suggestions of career pathways to the candidates of undergraduate levels at the very least. As talked about, not much researches and discoveries are completed to consider the viability of pathways and majors' suggestions at high school students' level. In this thesis in which a framework will be created for students is adjusted to tackle the goal that is centered around the issue for High school students. Choosing a university major or a career pathway is a complex decision-making process in which subjectivizes, and imprecisions are generally involved. Therefore, doing a data analysis through machine learning is proposed. As indicated by Holland's hypothesis, there exists a solid connection between the character attributes of a person with his major decision. However, if the candidate realizes which characteristic quality it lies in and has adequate information about the wide scope of professional plausibility, then as per his skills, personality, and academics, an individual will have the option to pick the correct major (Kalra 2017).

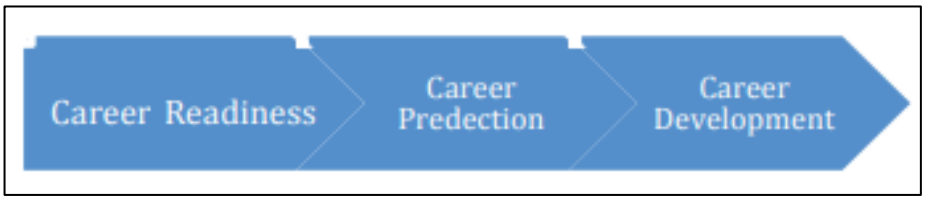
Information: the essential for everything explanatory. You can't be investigative without information and you can't be great at analyzing without having great information. The latest technologies supporting intuitive learning strategies and tools have opened novel chances to gather monstrous amounts, ranges, and sizes of data about students. However, the measure of information accessible to higher education departments outperforms their uses, because of sustained surges of information and the deficiency of customary databases to manage such consistent information streams. This huge spike of information

led to the introduction of the “Big Data” field to gather the abundance of data out of such consistent progressions of information streams. According to other perspectives, the use of computational methods for overseeing, preparing, and dissecting these huge volumes of information incited the foundation of the Learning Analytics field to improve learning measures (Gottfredson 1996).

The fusion among them is required to change the future of high school education (Eggerth 2005). This sector has generally been infamous for the inefficient utilization of information to improve the quality and the value of graduates in addressing market needs. This inability to use the obvious information to address market issues creates a massive gap between it and the industry also decreasing the opportunities for mediation to get ready the candidates ready for a fruitful vocation path and a prevalent expert performance. Adversely, the dynamic demands of local, global economies and the quest for competitions swiftly gave rise to a "War for ability" by business employers to effectively search out fresh job seekers who are clever, versatile, and flourish in an ever-changing environment. The weight initiated by education change and market needs require the coordination of a new system in the higher education sector to connect differing perspectives and build up a typical statement of being professionally prepared. Education and applied science research with a commitment to career pathway preparation with flexibility and a pledge to long-lasting learning, alongside the authority of basic information, key abilities, and expert depositions (Hoffait 2017).

The development of a candidate in their career starts on the foundation of their abilities and strengths as depicted in Figure 1.3. It evolves after predicting their best possible outcomes using their characteristics and develops further. The global spread utilization of technology permits catching exceptional measures of advanced information - depicted as Big Data about students' academics and skills, just as itemized sets of events and scenarios happening in educational settings. Machine learning combines a few existing procedures to dissect and recognize designs inside learning Big Data to analyze insufficiencies in the learning cycle and derive changes to improve learning results (Roy 2018).

Figure 1.3: Evolution of career advancement



Source: (K. Sripath Roy 2018)

After carefully studying over 20+ research papers, enough evidence was found the is a major problem in the modern world and this needs to be addressed quickly to avoid the worst.

This thesis is divided up into 5 major sections. The first one is the Introduction part followed by the literature review/background study section, then by looking at the data and the methodologies involved in analyzing the dataset that was created. Lastly, the following step will be looking at the results and then concluding at the end. This thesis will be more like a data analysis study. It is necessary to be working with data, do some analysis, and at the end will be making some suggestions that are concurrent to the findings.

1.3 SIGNIFICANCE OF THE STUDY

The Essentialness of the investigation was as per the following:

a) Some understudies didn't start to investigate 'genuine' professional prospects until after graduation. Scholarly universities, specialized schools, and legitimate direction or advising can give applicable data before their tutoring. They could be more frequent, giving candidates data they could test and use in their day-by-day studies and apply to their vocation decision.
b) Before graduating, a few understudies have not thought enough about other optional decisions in professional choice to legitimize settling on an educated choice. Influencing sources could be brought into a hover of guiding and discussion, to enable the understudy to shape a complete professional plan or framework.

c) The industry could look at where, why, and when it could be valuable for them to contribute assets to prepare and teach understudies. To select the ideal up-and-comer as indicated by your prerequisites
d) If vocation arranging were executed productively, understudies would at the least be following a vocation plan of educated dynamic, as opposed to one of agony.

1.4 LIMITATIONS

Impediments of the examination incorporated the accompanying:

a) While the size of the examination was huge enough to finish up sensible suspicions, the sample information was gotten from different understudies in a specific geographic area and might not have been characteristic of the bigger populace. The legitimacy of the study has depended on the understudies' real and insightful reactions**.**
b) Due to the pandemic situation, the reach for many students was hampered as institutions are still closed in the target region. Thus, the dataset consists of a mediocre number of students of both the high school and the university students
c) As the target region of students was from Pakistan, being physically present at the location to gather the data was also a hindrance as the flights were suspended due to the lockdown situation.

1.5 RESEARCH QUESTION

The main question in this research is that what are the suitable career fields for a high school secondary student mirroring similar university students. It is a comparison between high secondary school students and university students and their similar skill sets. If they resemble, then it is recommended for the high school student to follow in the footsteps of the university student and choose the college major that the latter is currently enrolled in.

2. LITERATURE REVIEW

As a massive quantity of information is accessible to organizations, machine learning methods rise as the main components to handle the situations. The utilization of machine learning in this type of problem isn't fresh. The advancement in this department with the machine learning research field can be followed in much previous research. These thesis or research papers give numerous instances of the use of the various machine learning strategies to support various types of perspectives concerning education. These works of literature cover, for instance, the dropout forecasting of students, the introductions of recommendation systems, and the candidate's academics exhibition forecast. Concerning this forecasting, Márquez-Vera et al_ had written a research paper in which it directed a few tests utilizing a record of data of over 400 understudies to forecast dropouts during various stages of a subject while selecting the feasible indicators for the exit. On the other hand, Bydzovska created up a recommendation model that established the candidate's abilities, information, interests, and free timetable schedule openings to help the students in their decision of choosing courses; this model is validated by utilizing a dataset of 1444 students (Patel 2017).

The forecasting of a candidate's performance in academics is one of the most mainstream and valuable uses of machine learning techniques. This comprises assessing the unknown marks of the candidate's academics performance. This is a highly prioritized issue that needs to be settled, owing to the enormous number of unforeseen circumstances that can affect students' performance are, including financial status, past researching experience, co-operations between partners, demographic attributes, psychological profile, and social foundation. Tinto presents a prevalent hypothetical structure for academic achievements. This famous system considers academics accomplishment as a cultural community, Among the participants' traits joining up with the university and the involvement with that establishment. For this model, it suggests that the allegiance for an organization with a pledge towards objective academic fulfillment is extremely related to the level of coordination (Chen 2013).

Evaluating the features stated earlier, the educational machine learning techniques examine that the focus of the candidate's performance has been utilizing a high number of characteristics to categorize the candidates and their surroundings. He re-evaluates the traits in foreseeing candidates' performance in a course. Internal evaluation, like assignment scores, tests, lab assessments, class tests, and participation, is quietly frequently utilized among the analysts to foresee the candidates' academics. Additionally, it also stresses that the candidate's demographic traits and outer assessment qualities as relevant statures. Demographic traits incorporate sexual orientation, age, family foundation, and disabilities. External markings are the corresponding ones to the marks the students got in the final term exam for a specific subject. Moreover, characteristics corresponding with a higher school foundation, social collaboration systems, and extra-curricular exercises are mostly monitored. There are additionally a few pieces of research that used psychometric elements to predict the candidate's overall academic performance. The quantitative coefficient is distinguished between the candidate's fascinations and commitment hours. Table 2.1 in the appendix section addresses most of the previous works that have been undertaken that deal with the student's review performance (Ferry 2014).

To the idea of candidate's attainment, Research is partly devoted to investigating the achievements in a course. For this situation, a few investigations characterize academic accomplishments as indicated by the grade point average value, while different investigations just think about whether a candidate fizzled or graduated in that course. There is another part of the writing that investigates scholastic execution at the degree level. For this situation, a few examinations mean to foresee whether an understudy will get a degree, while others intend to anticipate an understudy's last grade. A few different examinations center around deciding the understudies' academic success by the concluding primary educational term (Guruler 2010).

A few data machine learning techniques can be implemented to dissect the assembled educational knowledge. Regression & classification and others are commonly used when addressing a prediction or output problem. While the anticipated variable is a propositional one the classifier is used, on the other hand, the regression method can be

seen when the anticipated variable is stable. Under the category of classification and regression techniques, the models that come are artificial NN, KNN, decision trees, SVM and these are the most common models used. The method of decision trees is utilized, for instance, to anticipate the performance of 250 students who are currently enrolled in the 3rd semester of a master's program, and by Natek and Zwilling to investigate the utilization of a low amount of candidates of 2 datasets to foresee the candidate's performance in the course of informatics. On the other hand, concerning the neural networking model, by using the data obtained in the inception semester, Arsad introduced an approach to determine the academic success of over five hundred candidates in the concluding semesters. In this analysis, Maraboutic was using the model called Naive Bayes to classify and find the candidates who walk on thin ice in a course, the dataset consisting of around 1600 students (Miguéisa 2018).

Gray used the model of support vector machines to distinguish candidates in danger of flunking in the starting year of their examinations, a dataset that included 1074 candidates was utilized in this research. Likewise, to anticipate the candidates' accomplishments using a dataset that consisted of 391 records Strecht used the support vector machines.
The above table sums up the educational machine learning techniques for the academic performance of the candidates. The table revolves around the kind of research and the techniques utilized. It also contains the number of expectations used in the analysis and offers additional detail about the depending value used in the investigation.

Deep research and review of the studies show that in a few places there is still space for more development. Initially, a significant number of studies discuss success forecasting in a course, while others appear to result mainly from the GPA obtained. Along these lines, there is space for investigations of various measures and then use this opportunity to work on higher secondary school students and provide a career pathway or recommend a major to them (Miguéisa 2018).
The paper by Deitz-Ulher is anticipating the candidate's academic success depends on his likes, talents, and qualities. The examination work by "Ephrem" forecasts the terminal mark of the candidate by utilizing Pearson's relationship of factor strategy (Uhler 2013).

The researchers Baradwaj and Pal led an exploration to examine a candidate's academic performance depending on a cluster of 50 pupils. Their focus of the research was during a time of 4 years, with numerous pointers, including "Past academic numbers", "Grades of class tests", "Assignments Performance", "General Proficiency", "Participation", and " Concluding Semester Results ". Numerous variables are hypothetically expected to influence candidates' academic performance in advanced education, and they anticipated the pupil's execution depending on the related individual and social components. There are a handful of bearings of high-level machine learning, including the forecast of ability, aptitudes, and vocation development. Aptitude can be anticipated depending on reports like venture portrayals, human information databases (Baradwaj 2011).

Selecting a profession is a fundamental and extensive research point in professional research, beginning with the ability to coordinate methodology that was created by Parsons. The coordinating hypothesis was in this way formed into the quality and factor hypothesis of word-related decisions, which fundamentally estimated singular gifts and the traits of specific positions.

Even though it is like occupation suggestion systems, it set more accentuation on singular characters, interests, aptitudes, or other logical and quantifiable qualities. However, work suggestions endeavor to take in the attributes and factors from conduct observations, which may experience low interpretability. The reason for this is its analysis in an absence of adaptiveness with the difference in personal and word related conditions, formative hypotheses, professional investigation speculations, social learning/cognitive theory, and others were produced for deciding and clarifying professional decisions at various times of life expectancy. The objective of these speculations is to help individuals to gain developing self-idea and dynamic word-related conditions, and to additionally settle on vocation choices.

A few scientists have investigated the issue of career path suggestions. The larger part is zeroing in on work proposals, while just a modest bunch investigates career path suggestions. For instance, Paparrizos et al. prepared an AI model to anticipate competitors' next professional progress depending on their past activities with the data of the applicant and the other. A few theoretical kinds of research utilize metadata and outer

information to make suggestions. Chien et al. built up an information mining system depending on the decision trees and the rules of associations to produce valuable guidelines for choosing faculty highlights, like the experience of the work and monitoring performance of the job. Many examination techniques utilize a hybrid approach and the collaborative filtering way to produce work suggestions. Almalis et al. suggested this approach that expands and refreshes the distance of Minkowski to address the test of the coordinating individuals and the occupations. The suggested approach utilizes an organized type of occupation and the applicant's 2017 IEEE Third International Conference on Big Data Computing Service and Applications profile, created from the analysis of the content which is the unstructured type of the expected set of responsibilities and the new applicant's CV. Zhang et al. did something that used client-based and item-based suggestion systems consolidating data from the client's resume to create work suggestions. Fazel-Marandi and Fox joined diverse matchmaking procedures in a mixture approach for coordinating position candidates, occupations while utilizing logical with distinguishing based co-ordinations. Lu et al. suggest a crossover framework joining a procedure with content-based protocols. A marginally unique process is utilizing client grouping to improve the suggestion cycle. Lastly, Ting and Bakar proposed a Bayesian system-based answer for suggesting delicate abilities required for each activity, utilizing a dataset gathered through separating data from ads and through meeting with a couple of distinguished specialists (Baradwaj 2011).

V.L. Miguéisa, Ana Freitasb, Paulo J.V. Garciab, André Silvab created research about segmenting students as early as possible based on their educational achievement, using a predictive modeling technique, for this, they proposed a two-way approach. Firstly, they grouped the students using a binning algorithm based on their educational outcomes, to set up 5 degrees of academic success. The ID of five academic performance degrees was principally identified with the administrative inquiries. According to them, it was the greatest number of gatherings thought to be under control by the establishment and it simultaneously, empowers them to separate the understudies with sensible reasons. They also mentioned that considering the arithmetic values in general academic performance pointer, the equivalent-width binning calculation makes an already determined count of

clusters, by partitioning the scope of academic performance score into periods size (Miguéisa 2018).

On the other hand, it was very critical for them to note that their objective wasn't to decide the specific academic success estimation of an understudy, but instead to have the option to fragment them, as controlled by gathering the performance of the understudies' friends. Compared with the dependent variable of the forecast model built in the subsequent stage, the output parameters are acquired via the binning method. This meant that this was the predictor they intended to forecast at the end of the key year of the academic vocation of the understudies and, secondly, they built a machine learning projection model that used the data available as a non-dependent model during the first year and on the other hand, used the categorical variables as the dependent variables. The dependent variable was a categorical element that mirrored the degree of academic performance for the completion of each candidate's appointed degree. This model has subsequently been used since it involves a problem of multi-class classification. Following the writing, they recommended integrating autonomous variables associated with the high school base, psychometric variables, social elements, and socioeconomics into the expectation model. Besides, they proposed to incorporate the effects of inward and outward assessments of the courses that are important for the scholastic program's main semester. Other than the dual-method solution suggested, their paper was intended to separate the understudies as they watched the implementation and their usual success during their first year.

Consequently, they further proposed to utilize a grid that stands up to the understudies' anticipated performance levels with the diverse degree of scholastic achievement previously saw in the primary year of their academics. The pointers used to cluster the understudies as indicated by their scholarly accomplishment in the principal year was the similar technique used to gather understudies toward the conclusion of the degree. They believed that every understudy fragment prompts distinctive potential activities that can be utilized to advance higher academic achievement or to benefit the understudies with proved outcomes.

Their dataset comprised 7000 students which were accompanied by 500 academics staff and around 300 researchers. In this research, they used around 6 modeling techniques of classifications namely Naïve Bayes, Support Vector Machines followed by the decision

trees, bogged trees, random forests, and adaptive boosting trees. Their results revolved around the 96 percent accuracy that they were more than adequate, thus confirming their effectiveness of the predictive model. Their outcomes uncover that, in that specific contextual investigation, the random forest characterization procedure was the one introducing the best outcomes, and Naive Bayes was the one introducing the worst outcomes (Miguéisa 2018).

Rajalakshmi Krishnamurthi, Mukta Goyal created a system that predicts the careers of the students, their topic of research was “Automatic Detection of Career Recommendation Using Fuzzy Approach”. Their research used a whole new approach where they recommended student's careers using a fuzzy approach. A fuzzy approach according to them is “a generalized set of crisps, it is a set having the degree of membership between 0 and 1. The member of one fuzzy set can also be a member of other fuzzy sets in the same universe. For example, if a universe of intelligent people is classified into friend or enemy, then all intelligent people cannot be classified in these two categories, rather they can be classified somewhere other than these two extremes also such as less enemy, more enemy, less friendly, more friendly, etc.” Other than the Fuzzy approach, the main thing that was eye-catching in their research was Holland’s Code theory they were able to utilize in their surveys (Rajalakshmi Krishnamurth 2018). In this article, Holland's hypothesis is utilized to distinguish the applicant's character. Weights were labeled with every character trait to Figure out the scores are aligned manually with the assistance of research work done in this area previously and the criticism. An overview is a questionnaire that was used to test the precision of the proposed technique on 217 records. The outcome shows the 74.35 percent up-and-comers are happy with the suggested vocation and the normal rating for character attributes score was 3.58/5. This article additionally proposes a fuzzy approach to deal with programmed recognition of vocational suggestions depending on the character score.

A survey depending on the Holland hypothesis was introduced to the understudies. For each question, answers were gathered as Y(Yes) and N(No) where Y covered the answer, Yes and N covered the answer No. The information gathered for each answer was in string structure for example "YYYNNYNY". 50 inquiries represented the diverse attribute of the character. Every segment was representing various character attributes. On the other

hand, each row was representing as various inquiries. Every one of the inquiries was representing the distinctive space character. At first the score for each question was "Zero". The first inquiry had a place with the character attribute "imaginative" where the client's reaction was negative, represented as - 1.

Their technique has a better chance at working perfectly only when under the presumption that all the appropriate responses are submitted truly by the student and the student must have self-comprehension at least even a little bit. The benefits of the proposed strategies are it predicts the best professional choices for the understudies in which they can exceed expectations in the future. Multi-rules-based recommendation system and decision inferring device. Holland code hypothesis and fuzzy approach methodology display performance concerning professional suggestions for understudies toward the conclusion of Grade 12 and exposes higher levels of commitment to prior vocational choice. The proposed strategy establishes a base to create candidate goals, pinpoint and adjust their vocational approaches and afterward improving themselves to achieve the greatest fruitful profession.

In almost any scenario, with the aid of previous inquiries, the weights are taken physically in these techniques that can limit the information as various character variables are absent in this technique. In an easy-to-understand road, the use of the strategy should be possible by blended learning of the surveyor posing inquiries at an alternate time to get the impartial response to the inquiry that drives us to a decent suggestion. To further enhance the plan, greater character values and vocation domains can be deemed and weights for figuring can be aligned with machine learning assistance, Artificial insight (Krishnamurth 2018).

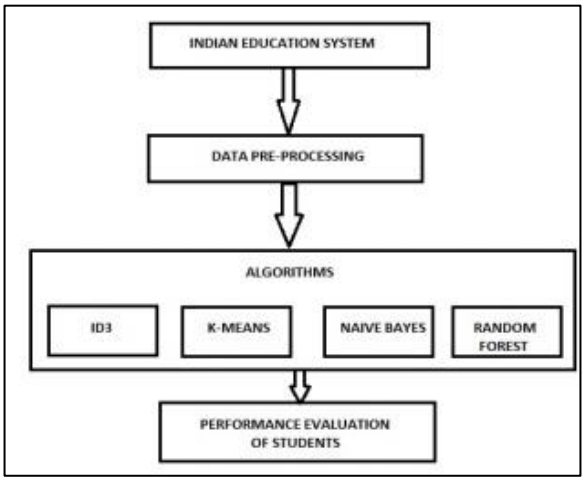
Subhalaxmi Panda, P.A Pattanaik, Tripti Swarnkar researched the Indian education system, their topic being “A Higher Education Predictive Model Using Data Mining Techniques”. Their paper audits a relative investigation of ID3, K-Means, Naïve Bayes, Random Forest calculation. In this paper, they have proposed the methodology of Random Forest to foresee the professional choice for the High school students that are passing out. The utilization of Random Forest had helped the understudies to accept the right suitable choice according to their advantage and abilities and acts as a career advocate (Panda 2017).

According to them, the Indian education system lacks an efficient and effective knowledge system that causes a ripple effect down the system thus hampering the benefits for the students. They also went on to state that the data mining technique acts as a bridge between the Indian education system and lacunas.

The researchers stated that the Information mining approach prompts some information mining strategies which will assist with improving the adequacy, productivity, and precision of the processes. Thus, this advancement helps in improving the Indian educational framework by expanding educational framework effectiveness, limiting understudies drop-out rate, continuously expanding understudy's advancement rate, understudies standard for retention, at the same time academics improvement rate, understudies achievement, increment in understudies learning rate. Along these lines, to accomplish the general quality improvement and needing some information mining methods in the framework that encourages the higher authorities to act swiftly.

The techniques known as the random forest is one of the dynamic group learning strategies which cause the understudies to take the right choice for their proper vocation decisions after final tests. This information mining procedure teaches the understudy a specific pathway to coordinate their shining profession in a viable way.

Figure 2.1: The data prediction workflow



Source: (Subhalaxmi Panda 2017)

Around 200 records were used in this process of students, these students were mostly of 10^{th} and 12^{th} standards. After performing the necessary methods and comparing the accuracies of all the techniques used and fortunately, random forests showed a total accuracy of 80 percent which was highly positive. Other than this, the 2^{nd} most close model with the higher accuracy was the Iterative Dichotomiser 3 model with 70 percent. Figure 2.1 depicts the process flowchart through which the predictions were done, the pre-processed data passed through ml algorithms and the output generated (Panda 2017).

Zun Hlaing Moe, Thida San, Hlaing May Tin, Nan Yu Hlaing, and Mie Tin, conducted technically different research, their research titled “Evaluation for Teacher’s Ability and Forecasting Student’s Career Based on Big Data”. Their paper strives to offer the assessment of the instructor's capacity and the determining of understudies' professional openings. The lecturer’s capacity is chosen depending on the understudy's returning remarks, dynamic part-taking in the class, understudies' outcome in the tests, and the instructor's competency. The returning remark is a basic component in the learning cycle. Understudies' input is an important instrument for educator assessment bringing about instructor improvement. The professional opportunity accessible for an understudy is a critical territory that decides the positioning of a college. This exploration likewise also Figures out the understudy's profession depending on their subject evaluation. Understudy vocation estimate depends on prescient logic. It involves an assortment of procedures that anticipate future results dependent on recorded and current information (Moe 2018).

The understudy marks from the time of 2015 to 2017 are gathered as a dataset, which is utilized to make a regression to predict the understudy's profession depending on their last grade of every year. This dataset incorporates highlights like alternate to practical numbers, lab numbers, test numbers, assignment numbers, class part-taking marks and mid-term numbers, and last test of the year numbers, which will be utilized to conjecture the profession of understudies.

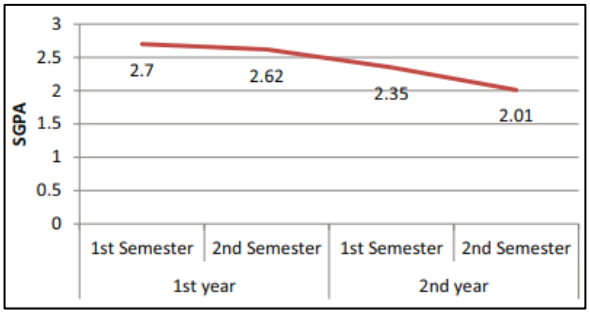
For their exploration, the information of college understudies was gathered from the Myanmar Institute of Information Technology. At first, around 360 understudy records

were gathered. In the University, there were two-degree programs, Computer Science and Engineering and Electronic and Communication Engineering. In their university, 100 and twenty understudies are added consistently. According to this hypothesis, the following ten years, understudy information was to be greater. And afterward, if our college increases the number of the programs, likewise the information was supposed to be increased as well build to an ever-increasing extent. Along these lines, the primary objective of exploration is to gauge the understudy's professional opportunity after graduation. Figure 2.2 outlines the grade point average value for students that are enrolled in the higher secondary classes.

Preparing Data: In this thesis, the entire dataset was slashed into two parts, one was the training dataset which was used to train the model with 80 percent and 20 percent is utilized as a testing dataset.

Testing Data: In the testing cycle, they isolated 20 percent of CSE understudies from the third year from the first dataset. It was supposed to be utilized to test with the model.

Figure 2.2: Higher secondary classes grade point average using SGPA



Source: (Zun Hlaing Moe 2018)

Taking everything into account, investigation, and conjecture on understudy academic performance had propelled them to complete further exploration to be applied. It will push the educational framework to screen the understudies' performance efficiently (Moe 2018).

M C B Natividad, B D Gerardo, and R P Medina also researched a topic titled “A fuzzy-based career recommender system for senior high school students in K to 12 education”. A vocational suggestion framework was introduced which was planned to help the career advocate as well as the senior secondary school understudies to direct them in considering various elements related to their choice on what profession they will seek after. To do this the attribute collection technique is best used to delete irrelevant highlights that impair the applicability fuzzy-based system which involves choosing the understudy characteristics from separate variables. Diverse channel methods have been used in this paper to pick the best attributes and then these properties are used as new sources of information. The final performance of the presentation illustrates a suitable result for verdicts. The proposed technical advice system for the understudies was assumed to be timely and will be one of the noteworthy study works in the Philippines' future training framework era (Natividad 2019).

The process of feature selection is successfully utilized along with numerous arrangement methods in an academic environment like in anticipating a candidate’s academic attainment algorithm as the studying effectiveness and execution precision are upgraded particularly that the intricacy of the study outcomes decreased. Therefore, it gives a few advantages, like the expulsion of low-level importance highlights, and with these a limited scale of highlights, despite everything, promising outcomes were accomplished. Moreover, by utilizing feature selection in recognizing noteworthy factors just as the understudies' attributes that impact the considering the academic performance of understudy to utilize in a system for an alert model, is exceptionally useful to the students just as to the higher authorities in assessing academic success for development since this is the essential objective of every educational association.

In particular, there are three aspect decision strategies: embedded, filter, and wrapper techniques, On the other hand, in this article, the filter techniques were used in pre-handling as widely as highlights are chosen based on the attribute of the data as it will certainly conform to grouping undertakings in spaces of immense density as found in their attributes.

In 1965, Lotfi Zadeh implemented the process of fuzzy logic to deal with processing relying upon levels covering honesty Except for the standard genuine or invalid Boolean rationale of 0 or the numerical value 1. It treats dubious, loses, or obscure knowledge that is faced in most real problems and uses it as a reason for constructing structures for Fuzzy deduction. Table 2.2 depicts the structure.

The vulnerability cannot be ignored about preferred structures, particularly when taking client inclinations. However, this dilemma may be taken care of with Ambiguous reasoning to help prescribe mechanisms and provide reliable recommendations of convincing and effective evidence, including in technical guidance, as well as to break down the academic exhibits of understudies.

Table 2.2: Description of Student Attributes

Value	Reason	Value	Reason
Sex	Sex of student	Salary of Parents	The monthly salary of parents
Years	Student's Age	Character	Temperament of candidate
Pursue	SHS track enrolled	Guardians impact	Impact of Guardians
Roots	Continued roots	Family members	Impact of family members
Profession of dad	Job of father	Friends	Impact of friends
Education attainment of father	Education of father	Social status	Social status
Occupation of mother	Job of mother	Proximity	Academic vicinity
Education achievement of mom	Academics of mother	Opportunities for Job	Employment credibility

Overall Result	Result	Not Applicable

Source: (M C B Natividad 2019)

The understudy information was sophomore collected through secondary institute understudies Out of the 2 divisions of the institute in the region of Isabela whose section information is collected from a coordinated survey in where the designated values were taken through separate sections whereas the understudies class marks were also collected with an earlier endorsement by the school authorities. A few characteristics were taken from the gathered understudy results. The dataset was split into three sets as 429 understudies were considered (60 percent) of the population as the train collection, and the remainder (40 percent) was equally broken down for the cluster that consisted of over one hundred understudies and the other around one hundred and forty-three understudies were supposed to be the test set, around an estimate of over seven hundred plus understudies. Before it changed to a .csv folder, all superfluous spaces were deleted and inaccurate knowledge writings were updated and Weka was used for programming and information digging devices to provide options using five correlation-based sifting techniques, Benefit Ratio, Data Value, Relief and Symmetrical, individually. Going through this analysis, the Fuzzy recommendation technique is suggested which will allow secondary understudies to distinguish and study the patterns relevant to their choice of vocation, however, because different variables have been considered, then provide feature selection procedures that are used to exclude unessential understudy characteristics so that the Fuzzy derivation model produces fair results. The suggested structure for additional works should be tested for further research sets with the diverse arrangement of criteria as the selected understudy features would be altered as the large components of legitimate details would also be altered. It is reasoned that a technical advice system based on Fuzzy is easy and would have been necessary for assisting secondary understudies under the current curriculum (Natividad 2019).

Min Nie, Lei Yang, Bin Ding, Hu Xia, Huachun Xu, and Defu Lian also researched this domain, and the research title being “Forecasting Career Choice for College Students Based on Campus Big Data” in the year 2016. in this paper, they proposed a directed professional decision prediction system depending on the understudies' conduct

information and vocational decisions of graduation. Inside this structure, they set forward conduct-based agent factors for influencing understudy vocational choice. These elements, upheld by the mental investigation, incorporate proficient aptitudes gained from course-taking records, conduct requests in the honesty of enormous five-character, premium and inclination for acquiring books, and family monetary status assessed by everyday utilization from smart-chip enabled card use. It is instinctive to project professional prediction into a multi-class characterization issue, with the goal that it calculates it. KNN, Decision Tree, or Multinominal logistic regression could be utilized to anticipate his potential vocation decision in a determinant or probabilistic way.

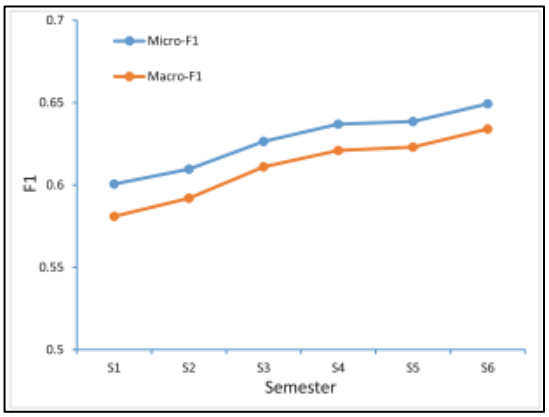
These multi-class grouping calculations catch every understudy's likeness/separation/dissimilarity with graduates over those relatable factors, appropriately concurring with the social correlation hypothesis in psychological science. The focal suggestion of the social examination hypothesis is the correspondence theory, which shows that human assesses their capacity and constraints by contrasting and comparative others, especially when objective and non-social methods for assessment are not accessible. Self-assessment for this situation most likely turns out to be steadier and more precise. More significantly, individuals are particularly prone to make upward correlations, that is to assess themselves against effective people, higher presumably prompting personal growth at last (Nie 2016).

On the contrary, supporting understudies to decide their vocation decision, in this system, they will lead a connection investigation between conduct portrayed factors and professional decisions, to find the influent agent factors for influencing understudy profession choice. Consequently, it is conceivable to use this information to assist understudies with accomplishing their initial expressed objectives. For instance, if they have watched the noteworthy impact of English courses at understudies' professional decision about traveling to another country for additional investigation. Subsequently, understudies, expressing this as their objective, ought to endeavor to procure the language abilities of English. At last, they assessed the proposed structure on conduct information and professional decision records of more than 4,000 understudies. Gathering vocation decisions of graduation into "abroad further examination", "looking for occupations", "homegrown further investigation", and "others", Micro F1-proportion of the best multi-

class arrangement calculation could accomplish 0.6 at the principal semester and improve with the expansion of semesters.

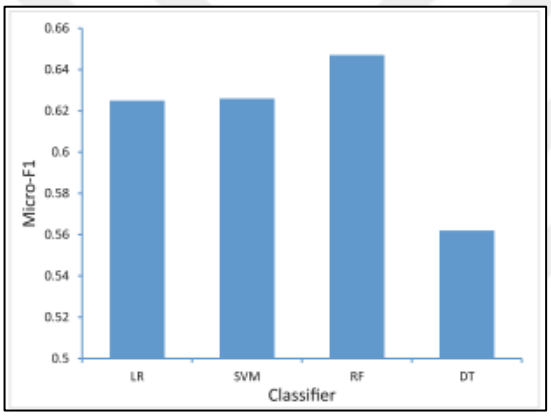
Hence, they arrange these proofs by semesters. As such, in every one of the initial six semesters, considering their conduct information inside the semester, they separate all presented highlights for every understudy. At that point in each semester, highlights of the procedure semesters will relate to the current one of sequential requests. As indicated by the connection examination between conduct-based factors and vocation decisions, they find that components like proficient aptitudes, conduct consistency, and monetary status could essentially associate with professional decisions. The assessment is led on a dataset from a 4,246 understudy of a similar evaluation. The absolute number of utilization records is 13,122,696, among which the records in the wreck are 6,875,698. Within four years in the college, these understudies have obtained 172,894 books, creating 336,238 book credit records, and taken 1,072 courses altogether, creating 276,588 course-grade records. For surveying the presentation of the multi-class arrangement, they generally utilized the Macro-F1 measure and Micro-F1 measure. The previous one gives equivalent load to each class, paying little mind to its recurrence, and is a for each classification normal of F1, and the last one gives equivalent load to each report and is a for every record normal of F1. Figure 2.3 gives a visualized image of this. For every one of six semesters, 5-fold cross approval is performed. For the multi-class grouping, they directed an examination between Linear Regression (LR), SVM, Random Forest (RF), and Decision Tree (DT). In this paper, they considered the professional decision forecast dependent on understudies' grounds conduct and proposed an information drive system for vocation decision expectation. Their results were significantly correlated. Figure 2.4 shows the comparison between different classification algorithms (Nie 2016).

Figure 2.3: Micro-F1 & Macro-Measure increasing semester



Source: (Min Nie 2016)

Figure 2.4: The comparison of different classification algorithms



Source: (Min Nie 2016)

Ankhtuya Ochirbat, Timothy K. Shiha, Chalothon Chootonga, Worapot Sommoola, W. K. T. M. Gunarathnea, Wang Hai-Huia, and Ma Zhao-Heng banded research together titled "Hybrid Occupation Recommendation for Adolescents on Interest, Profile, and Behavior" in the year 2017. The main purpose of that research work was to develop an occupation suggestion framework by utilizing the information mining technique. In the research, they revolved around testing Mongolian, Sri Lankan, Taiwanese, and Thai understudies. The framework can give subtleties of occupations and can help the understudies for significant decisions, just as the vocations to seek after. Besides, the examination objective joins a lot of results, which are suggested utilizing comparability estimations and proposal strategies. It is called a hybridization framework. These strategies fill in as a base for suggesting occupations that meet the interests and skills of understudies (Ochirbat 2017).

They included three arrangements of data including understudy's profiles, professional interests from the survey utilizing Holland code, and their practices. The understudy profile contains two sorts of information foundation and intrigue/interest recovered from Facebook. In the investigation, the understudies from four nations comprised of Mongolia, Sri Lanka, Taiwan, and Thailand utilized the OCCREC. Furthermore, five occupations were appeared to the understudies by utilizing five similitude estimates which are Euclidean, Intersection, Cosine, Jaccard, and Pearson. At last, OCCREC permits understudies to rate the outcomes appropriately dependent on the client's fulfilled scores and to share their encounters on Facebook.

The survey was planned dependent on Holland Code Model. They pick 30 things of the inquiry and each question had two options. The inquiries and answer decisions were given in four distinct dialects as language choices including Chinese, English, Mongolian, and Thai language. The entireties of the decisions were a subset of individual sorts that can be classified into six gatherings as follows:

i. Practical: an individual who is a practitioner,
ii. Insightful: an individual who is a scholar,
iii. Masterful: an individual who is a maker,
iv. Social: with an individual who trademarks like a partner,
v. Enterprising: an individual likes a persuader, and
vi. Conventional: an individual who is a coordinator.

The framework likewise includes client input about the suggested occupations. The Likert scale: the framework gives this capacity to clients and lets them give the fulfilled score on prescribed occupations from 5 to 1: 5 = extremely fulfilled, 4 = fulfilled, 3 = uncertain, 2 = disappointed, 1 = exceptionally disappointed. They planned on a scoring interface means to give an easy-to-use interface that lets the client give a fulfilled score by simplified the scoring circle into a container of each suggests occupation. What's more, after the client clicked the send input button, the framework will gather client criticisms to coordinate it for the development of the recommender framework.

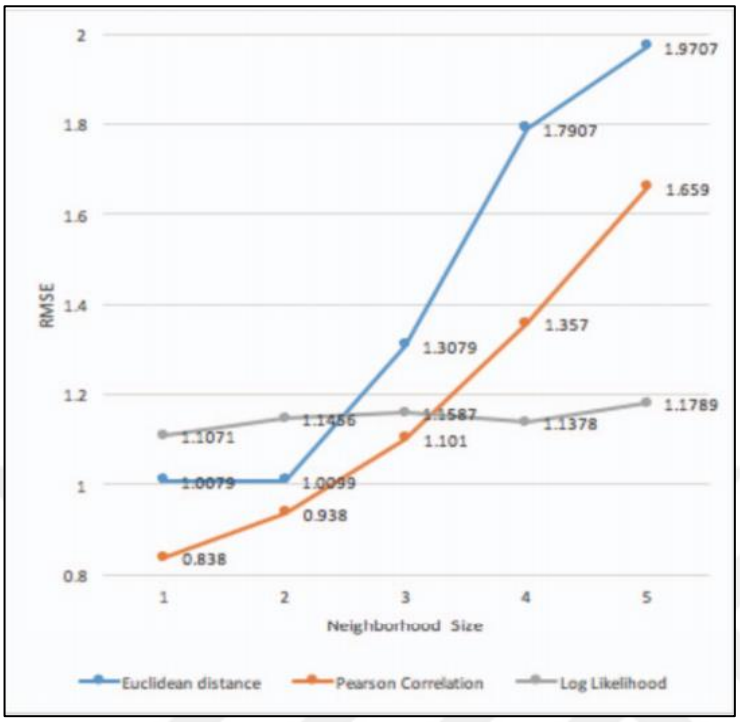
In this paper, they assembled the tests with 612 understudies from Mongolia, Sri Lanka, Taiwan, and Thailand in various examination fields. To prescribe the well-fitting occupation to understudies utilizing their professional intrigue and then utilizing the crossbreed proposal calculations (Ochirbat 2017).

Bharat Patel, Varun Kakuste, Magdalini Eirinaki of the State Jose State University came forward with research in the year 2017 titled “CaPaR: A Career Path Recommendation Framework”. They attempt to address the previously mentioned issues by introducing CaPaR, a Career Path Recommendation system zeroing in on understudies and youthful experts. CaPaR utilizes text mining abilities and joins them with community separating calculations, to perform two distinct kinds of suggestions to its clients, specifically work and expertise proposals, to such an extent that the clients can both get a new line of work, yet additionally distinguish regions and aptitudes that are as of now wanted in the activity market, however, they may, in any case, be missing, with the end goal that they can chip away at achieving them (Patel 2017).

They attempted to addresses such weaknesses. Utilizing text mining and cooperative separating methods the framework first sweeps the client's profile and resume, distinguishes the key aptitudes of the competitor, and creates customized work suggestions. Also, the framework prescribes extra abilities to understudies required for related employment opportunities, just as learning assets for every aptitude. Along these lines, the framework not just permits its clients to investigate a lot of data, yet also grow their portfolio and resume to have the option to propel their vocations further. They tested and assessed the different proposal calculations with certifiable information gathered from the San Jose State University vocation focus site.

The expected dataset was given by San Jose State University's professional place. The information was skimmed and included positions posted by organizations on the SJSU profession focus' gateway prior in the year, zeroing in on Computer Science and Engineering-related postings. For the investigations, they split the information into a training information set and test information set on a 70:30 proportion. They likewise gathered client resumes from SJSU understudies of the Computer Engineering division. The whole dataset comprises 10,000 sets of expectations and 100 clients. A few outcomes depend on 30 clients and 3000 sets of responsibilities. Figure 2.5 outlines the comparison of the nearest neighbor for skills. The results were significant and effective.

Figure 2.5: Nearest Comparison of Neighborhood Scale for Skills Algorithm Recommendation



Source: (Bharat Patel 2017)

Figure 2.6: Skill's questions and linguistic variable for professional recommendation in the career

Question	Skill	Linguistic Variable
Q1	Design and Develop a Web	{Weak, Medium, Good}
Q2	Handle whole web project from start to roll-out	{Weak, Medium, Good}
Q3	Skills and knowledge in PHP, HTML, CSS, Javascript and MySQL	{Weak, Medium, Good}
Q4	Good in problem solving, communication, interpersonal and organization skills	{Weak, Medium, Good}
Q5	Up to date with the latest web technology trends and programming techniques	{Weak, Medium, Good}
Q6	Writing reports, documentation and operating manuals	{Weak, Medium, Good}
Q7	Familiar with SQL database platform	{Weak, Medium, Good}
Q8	Use mathematics to solve problems	{Weak, Medium, Good}
Q9	Teach others how to do something	{Weak, Medium, Good}
Q10	Communicate information and ideas in writing so others will understand	{Weak, Medium, Good}
Q11	Perform a good end-user support	{Weak, Medium, Good}
Q12	Ability to solve problems creatively or strong analysis skills	{Weak, Medium, Good}
Q13	Type of person that results oriented / meet deadlines / maintain calm approach with multiple tasks	{Weak, Medium, Good}
Q14	Interpersonal skill or ability to get along with others & work well in a team	{Weak, Medium, Good}
Q15	Willing to work at home or outside of office to finish up the projects	{Weak, Medium, Good}

Questions	Skill	Scale	Career Recommendation
Q1	Design and Develop a Web	0-10	Web Programmer (WP)
Q2	Handle whole web project from start to roll-out	0-10	
Q3	Skills and knowledge in PHP, HTML, CSS, Javascript and MySQL	0-10	
Q4	Good in problem solving, communication, interpersonal and organization skills	0-10	
Q5	Up to date with the latest web technology trends and programming techniques	0-10	
Q6	Writing reports, documentation and operating manuals	0-10	Database Administrator (DA)
Q7	Familiar with SQL database platform	0-10	
Q8	Use mathematics to solve problems	0-10	
Q9	Teach others how to do something	0-10	
Q10	Communicate information and ideas in writing so others will understand	0-10	
Q11	Perform a good end-user support	0-10	System Analyst (SA)
Q12	Ability to solve problems creatively or strong analysis skills	0-10	
Q13	Type of person that results oriented / meet deadlines / maintain calm approach with multiple tasks	0-10	
Q14	Interpersonal skill or ability to get along with others & work well in a team	0-10	
Q15	Willing to work at home or outside of office to finish up the projects	0-10	

Source: (Bharat Patel 2017)

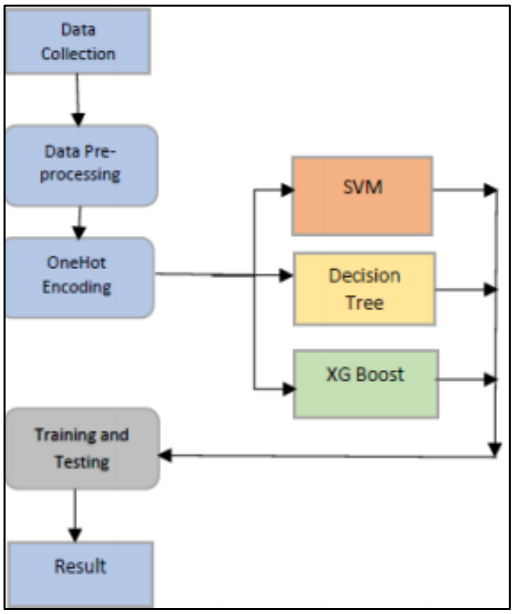
Tajul Rosli Razak ,Muhamad Arif Hashim ,Noor Faizal Mohd Noor ,Iman Hazwam Abd Halim ,Nur Fatin Farihin Shamsul of the UiTM of Malaysia conducted research together in the year 2014 titled "Career Path Recommendation System for UiTM Perlis Students Using Fuzzy Logic" in which they proposed a framework that suggests understudies about their profession depending on the academic outcome and their capacities by utilizing fuzzy rationale approach. Figure 2.6 describes the linguistic variable attained for each skill and the level asked (Razak 2014).

The students had to answer 15 questions that focused on their skill interests according to their program. These answers would lead to career recommendations whether Web Programmer (WP), Database Administrator (DA), and System Analyst (SA) which is the output for this study. Then this input was transforming into a linguistic variable as shown in Fig #. Taking everything into account, the results of this investigation will profit the UiTM understudies in their vocation determination in which the cycle will get simpler, adaptable, and quicker. This is because of the reality that self-testing should be possible

without the need for extensive coaching by the advisor. Besides this investigation likewise measure the understudy's aptitude qualities, capacities, and character aspects and suggest them with conceivable vocation decisions by utilizing a fuzzy rational approach. Besides, this framework additionally keeps the outline of the vocation test where understudy can do the test consistently and can contrast and the past professional test result. Figure 2.6 outlines the questions asked to test the skills levels (Razak 2014).

K. Sripath Roy, K. Roopkanth, V. Uday Teja, V. Bhavana, J. Priyanka from India the year 2018 banded together to construct research titled “Student Career Prediction Using Advanced Machine Learning Techniques “. According to them, their main purpose for this research was mostly focused on the vocation region forecast, and also this paper manages different progressed machine learning calculations that include characterization and expectation and is utilized to improve the precision for better forecast, unwavering quality and dissecting these calculations execution. Figure 2.7 shows the proposed workflow chart that was created to predict (Roy 2018).

Figure 2.7: Phase Flow Diagram of the proposed technique

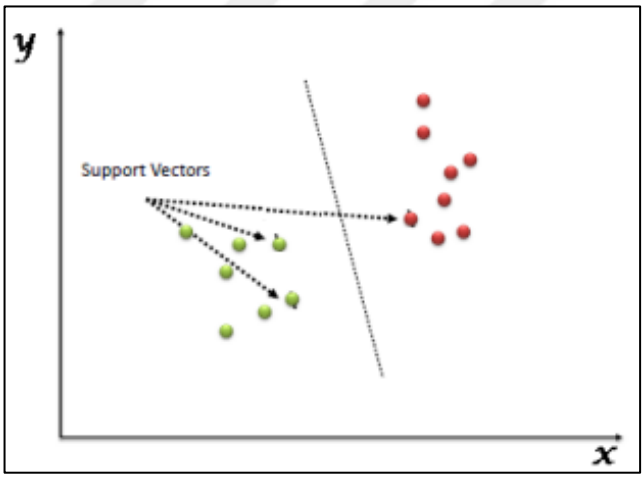


Source: (K. Sripath Roy 2018)

The data collection was extensively done as it is one of the most important and major processes. Information was gathered from numerous points. Some information was

gathered from representatives working in various associations, some measures of information are gathered through the LinkedIn programming interface, a few measures of information are haphazardly produced, and others from school graduated class information base. Gathering information is one of the major and most significant assignments of any AI venture. Since the information, is needed for the calculations is information. Thus, the calculation's productivity and exactness rely on the accuracy and nature of the information gathered. So as the information same will be the yield. For the understudy profession forecast, numerous boundaries are required like understudies scholarly scores in different subjects, specializations, programming, and diagnostic capacities, memory, individual subtleties like relationship, interests, sports, rivalries, hackathons, workshops, affirmations, books intrigued, and some more. As every one of these elements assumes an imperative part in choosing understudy's advancement towards a lifelong territory, all these are mulled over. Information is gathered from multiple points of view.

Figure 2.8: Support Vector Machines Example



Source: (K. Sripath Roy 2018)

Some information is gathered from representatives working in various associations, some measure of information is gathered through the LinkedIn programming interface, some measure of information is arbitrarily created, and others from school graduated class information base. Absolutely about 20 thousand records with 36 segments of information are gathered. Around 20 thousand records were taken into consideration with 36 columns

of data was collected. The Machine Learning Algorithms that were used are Decision Tree, XG Boost, and then Support Vector Machines also depicted as an example in Figure 2.8.

Eighty percent of the total information was used as testing data and the remaining twenty percent was used to analyze the data. Gathering the information is one errand and making that information valuable is another essential undertaking. Information gathered from different methods will be in a chaotic configuration and there might be the parcel of invalid qualities, invalid information esteems, and undesirable information. Cleaning all this information and supplanting them with suitable or inexact information and eliminating invalid and missing information and supplanting them with some fixed substitute qualities are the essential strides in pre-preparing for information. Indeed, even information gathered may contain trash esteems. It may not be in a careful arrangement or way that is intended to be. All such cases must be confirmed and supplanted with substitute qualities to make information meaning significant and valuable for additional handling. The information must be kept in a sorted-out arrangement. After completing the relevant processes, the results were highly efficient with the SVM algorithm producing a 90 percent accuracy level and then XG boost relaying an 88 percent accuracy level (Roy 2018).

Ahmad Slim, Gregory L. Heileman, Jarred Kozlick and Chaouki T. Abdallah of the University of New Mexico U.S.A studied together within this same domain and came up with the research topic “Predicting Student Success Based on Prior Performance” in the year 2014. According to them, this paper presents a structure that utilizes the algorithms of machine learning, more specifically, a Bayesian Belief Network (BBN), to foresee the academic performance of understudies right at the beginning in their academic career decisions. The outcomes acquired show that the proposed system can foresee understudy progress, more importantly, the candidate’s grade point average about the expected major, with negligible error after noting the performance of a solitary semester. Besides, as extra performance is noticed, the anticipated GPA in resulting semesters turns out to be progressively precise, giving the capacity to enable the candidates to likely get successful results right off the bat in their academic careers (Slim 2014).

In this paper, they proposed a probabilistic graphical model that permitted them to observe the candidate's academic progress. Specifically, they utilized a Bayesian Belief Network (BBN) model to visualize the educational plan graphs of academic degree programs. According to the academic performance of an understudy in each semester, it is speculated that the BBN model can anticipate the "future" academic performance of the understudy in ensuing semesters. The model created in this paper was applied to various diverse degree programs at the University of New Mexico (UNM) and had the option to foresee the GPA circulation of the understudies with an insignificant error. They gave foundation data and study some related work in this examination area and depicted the hypothesis behind the BBN procedure utilizing in the structure.

They utilized the information of 115,746 understudies for all the courses in the BBN. At that point, around 400 understudies are picked, who previously earned their degree programs, arbitrarily from various divisions like mechanical building office, substance designing office, electrical designing office, and atomic building office to test the system. The exhibition of the system was estimated utilizing mean squared mistake. The courses taken by these understudies are spread over more than 22 semesters.

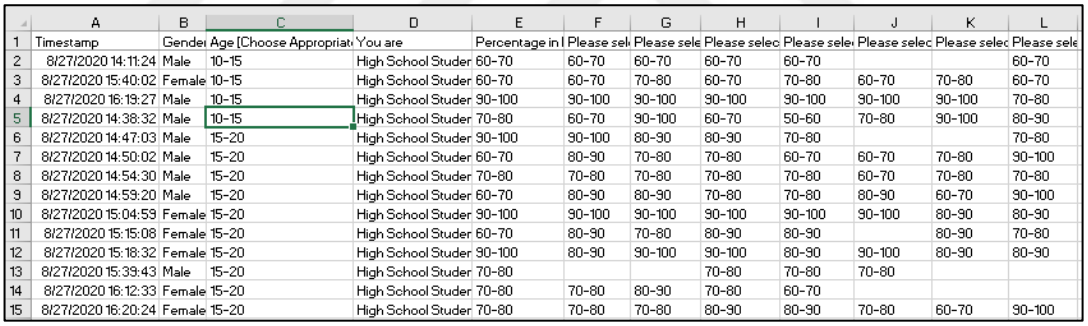
The outcomes that appeared in this paper show that BBN's can without much of a stretch model an educational program chart and can be utilized to foresee the future advancement of an understudy. It has been indicated that a negligible blunder of 0.16 can be accomplished after accepting the evaluations of the principal semester. Besides, the outcomes show that the MSE diminishes progressively after getting extra proof, exhibiting the reasonability of the proposed BBN model. This underlying work will be reached out later to show numerous factors in the BBN model with course variable.

3. DATA & METHOD

Information collection is one of the major and most significant aspects of any machine learning venture. Since the information is that is needed to undergo machine learning algorithms, this data must be smooth and as clean as possible.

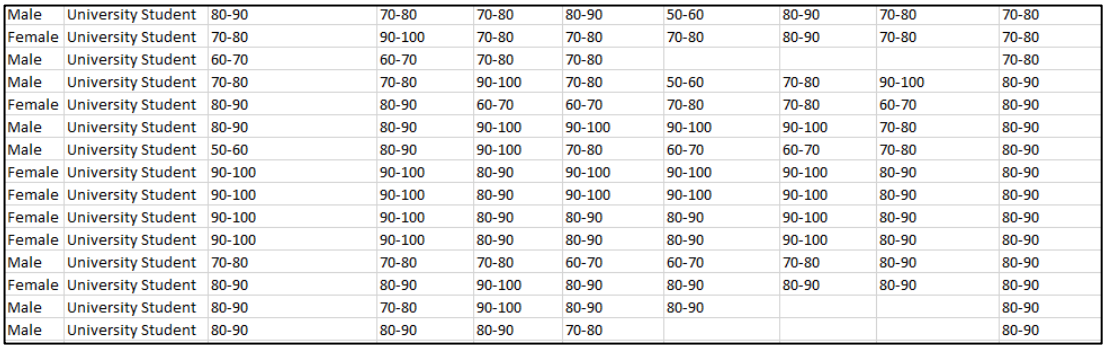
For the understudy career prediction, numerous answers to the questions are required like understudy's academic scores in different subjects, percentages, and interests or skills. As every one of these variables assumes an imperative function in choosing understudy's advancement towards a lifelong region, all these are thought about. Information is gathered in the most fitting manner of utilizing a questionnaire. Figures 3.1-3.3 shows the data that was collected.

Figure 3.1: High school students collected



	A	B	C	D	E	F	G	H	I	J	K	L
1	Timestamp	Gender	Age [Choose Appropriat	You are	Percentage in I	Please sel	Please sele	Please selec	Please sele	Please selec	Please selec	Please sele
2	8/27/2020 14:11:24	Male	10-15	High School Studer	60-70	60-70	60-70	60-70	60-70			60-70
3	8/27/2020 15:40:02	Female	10-15	High School Studer	60-70	60-70	70-80	60-70	70-80	60-70	70-80	60-70
4	8/27/2020 16:19:27	Male	10-15	High School Studer	90-100	90-100	90-100	90-100	90-100	90-100	90-100	70-80
5	8/27/2020 14:38:32	Male	10-15	High School Studer	70-80	60-70	90-100	60-70	50-60	70-80	90-100	80-90
6	8/27/2020 14:47:03	Male	15-20	High School Studer	90-100	90-100	80-90	80-90	70-80			70-80
7	8/27/2020 14:50:02	Male	15-20	High School Studer	60-70	80-90	70-80	70-80	60-70	60-70	70-80	90-100
8	8/27/2020 14:54:30	Male	15-20	High School Studer	70-80	70-80	70-80	70-80	70-80	60-70	70-80	70-80
9	8/27/2020 14:59:20	Male	15-20	High School Studer	60-70	80-90	80-90	70-80	70-80	80-90	60-70	90-100
10	8/27/2020 15:04:59	Female	15-20	High School Studer	90-100	90-100	90-100	90-100	90-100	90-100	80-90	80-90
11	8/27/2020 15:15:08	Female	15-20	High School Studer	60-70	80-90	70-80	80-90	80-90		80-90	70-80
12	8/27/2020 15:18:32	Female	15-20	High School Studer	90-100	80-90	90-100	90-100	80-90	90-100	80-90	80-90
13	8/27/2020 15:39:43	Male	15-20	High School Studer	70-80			70-80	70-80	70-80		
14	8/27/2020 16:12:33	Female	15-20	High School Studer	70-80	70-80	80-90	70-80	60-70			
15	8/27/2020 16:20:24	Female	15-20	High School Studer	70-80	70-80	70-80	80-90	80-90	70-80	60-70	90-100

Figure 3.2: University student



Male	University Student	80-90	70-80	70-80	80-90	50-60	80-90	70-80	70-80
Female	University Student	70-80	90-100	70-80	70-80	70-80	80-90	70-80	70-80
Male	University Student	60-70	60-70	70-80	70-80				70-80
Male	University Student	70-80	70-80	90-100	70-80	50-60	70-80	90-100	80-90
Female	University Student	80-90	80-90	60-70	60-70	70-80	70-80	60-70	80-90
Male	University Student	80-90	80-90	90-100	90-100	90-100	90-100	70-80	80-90
Male	University Student	50-60	80-90	90-100	70-80	60-70	60-70	70-80	80-90
Female	University Student	90-100	90-100	80-90	90-100	90-100	90-100	80-90	80-90
Female	University Student	90-100	90-100	80-90	90-100	90-100	90-100	80-90	80-90
Female	University Student	90-100	90-100	80-90	80-90	80-90	90-100	80-90	80-90
Female	University Student	90-100	90-100	80-90	80-90	80-90	90-100	80-90	80-90
Male	University Student	70-80	70-80	70-80	60-70	60-70	70-80	80-90	80-90
Female	University Student	80-90	80-90	90-100	80-90	80-90	80-90	80-90	80-90
Male	University Student	80-90	70-80	90-100	80-90	80-90			80-90
Male	University Student	80-90	80-90	80-90	70-80				80-90

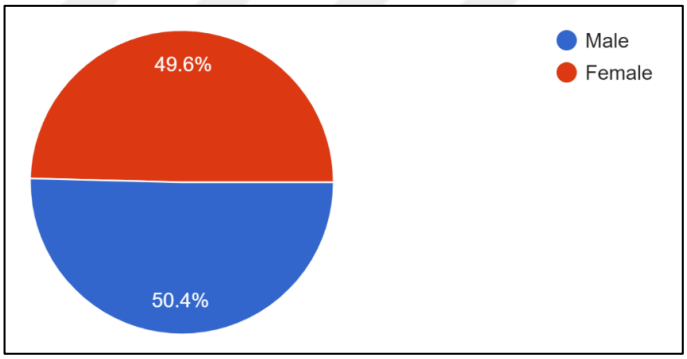
Figure 3.3: The skills answers that were collected

What is your Department in University	Please rate the following	Please rate the following sta	Please rate the following :	Please rate the following :	Please rate the following stat	Please rate the following s	Please rate the following sta	Please rate the followi	Please rate the following st	Please rate
	Strongly agree	Neutral	Strongly agree	Neutral	Agree	Neutral	Disagree	Neutral	Disagree	Agree
Engineering/Computer Science	Strongly agree	Strongly disagree	Agree	Neutral	Disagree	Strongly disagree	Strongly disagree	Agree	Neutral	Neutral
Medical/Health	Strongly disagree	Strongly agree	Strongly agree	Strongly agree	Disagree	Disagree	Neutral	Strongly agree	Disagree	Disagree
Engineering/Computer Science	Disagree	Disagree	Disagree	Disagree	Disagree	Strongly disagree	Disagree	Disagree	Disagree	Strongly dis
Medical/Health	Neutral	Agree	Strongly agree	Strongly agree	Neutral	Strongly agree	Agree	Neutral	Agree	Neutral
Business Administration	Agree	Strongly disagree	Strongly agree	Agree	Agree	Neutral	Agree	Agree	Agree	Agree
Engineering/Computer Science	Strongly agree	Neutral	Neutral	Agree	Neutral	Disagree	Disagree	Neutral	Neutral	Neutral
Engineering/Computer Science	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Neutral
Medical/Health	Agree	Agree	Agree	Agree	Neutral	Agree	Neutral	Neutral	Neutral	Neutral
Engineering/Computer Science	Strongly agree	Neutral	Strongly agree	Agree	Strongly agree	Neutral	Neutral	Strongly agree	Neutral	Neutral
Medical/Health	Agree	Neutral	Neutral	Strongly agree	Neutral	Strongly agree	Neutral	Neutral	Strongly disagree	Neutral
Engineering/Computer Science	Strongly agree	Neutral	Strongly agree	Agree	Strongly agree	Neutral	Neutral	Strongly agree	Neutral	Neutral
Physics	Neutral	Agree	Agree	Disagree	Disagree	Strongly disagree	Strongly disagree	Strongly disagree	Strongly disagree	Agree
Engineering/Computer Science	Strongly agree	Neutral	Neutral	Agree	Neutral	Disagree	Disagree	Neutral	Neutral	Neutral
Engineering/Computer Science	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Neutral
Politics	Strongly agree	Agree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Disagree	Neutral	Agree	Strongly agi

An original questionnaire was prepared with the target audience as students, and then it was distributed online accordingly to the people fill it, the targets included both high school students and university students. This process was done online because the COVID-19 pandemic prevented any face to face with the students as all the students were in lockdown and the only access was available through the internet

3.1 GENDER

Figure 3.4: The ratio of the Male and Female that attempted the questionnaire



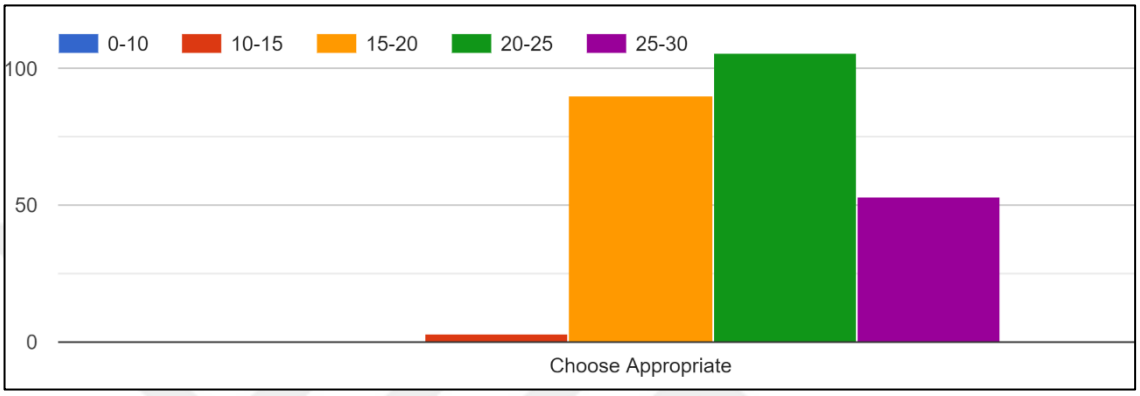
The total number of records collected was 253 responses and as shown in Figure 3.4, Around 49 percent of the data was filled by female students from high school and the university. Similarly, the male gender was 50.4 percent to be precise that filled the questionnaires from likewise universities and the high school students.

Age was the second question that was asked in the demographic section of the questionnaire, This gives us a clear distinction between the high school students and

university students, as most students that are aged between 15 and 20 are of the high school section and on the other hand, students that are aged over 20 are of the university.

3.2 AGE

Figure 3.5: Age classification using a bar-graph

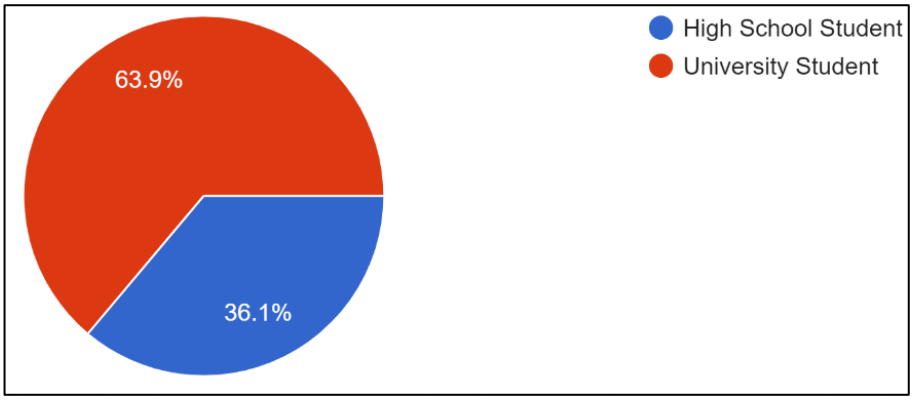


As depicted by the image in figure 3.5, it is visible that the greatest number of students that filled the questionnaire are aged between 20 to 25, while the 2nd number following the list is the aged between 15 to 20. According to the graph, university students were the higher number of students that filled the survey including those aged between 25 and 30. The lower number of students that filled the survey were high school students.

After breaking down the survey, it is seen that around 4 students that were aged between 10 - 15 of high school students filled the survey, furthermore precisely 88 students who attempted the survey were aged between 15 – 20 and were of higher secondary school students. Moving on, exactly 105 candidates that completed the survey were university students that were aged between 20 and 25. In conclusion, meticulously 52 students were aged between 25 and 30 submitted the questionnaire and were university students. After a thorough analysis, overall a dataset that consists of 253 records combined was considered.

3.3 STUDENT TYPES

Figure 3.6: Ratio of student types that answered the questionnaire

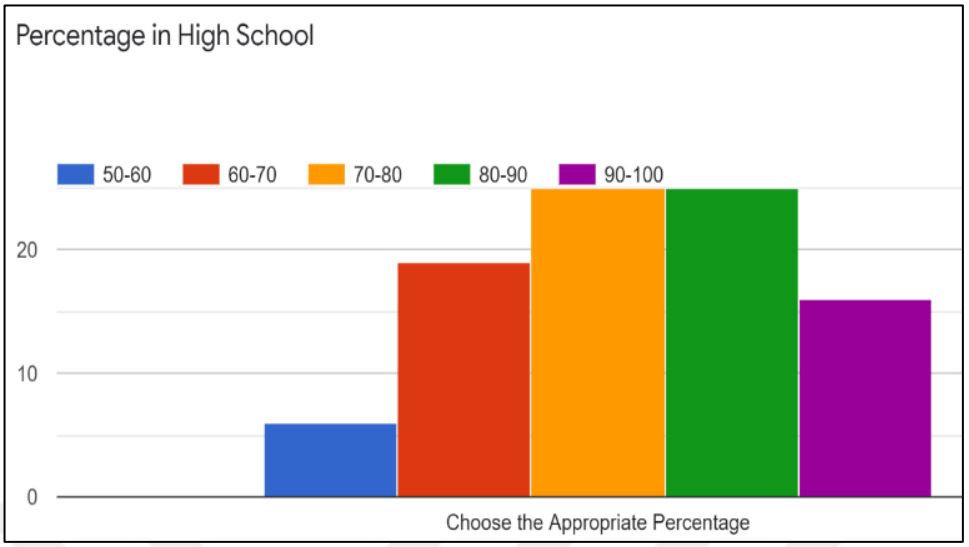


After studying the dataset overall, it was observed that 36.1 percent of students that completed the survey were university students and on the other hand, 63.9 percent of the candidates that completed the questionnaires were university students. A pie-graph is depicted for the student type ratio in figure 3.6.

3.4 PERCENTAGES IN HIGH SCHOOL OF HIGH SCHOOL STUDENTS

The next section in the questionnaire is the academics part that the candidates had to answer according to their academic background. The academics part was broken down separately for higher secondary school students and the other for the university students. However, students of both sections had to fill the high school academics. As shown in Figure 3.7 below, the high school students were asked for the high school percentage they scored.

Figure 3.7: Percentage scored in high school by high school students



After the careful analysis of this part, Figure 3.7 shows that:

i. Around 6 high school students scored 50 percent – 60 percent in high school
ii. Precisely 19 students of the same sections i.e. high school candidates scored between 60 percent – 70 percent
iii. Furthermore, around 25 high school students were able to pass out with a percentage between 70 percent - 80percent.
iv. Similarly, 25 students were also able to acquire a percentage of 80 percent – 90 percent in their high school finals
v. Lastly, around 16 high school candidates were able to graduate at the top of the percentage list between 90 percent – 100 percent.

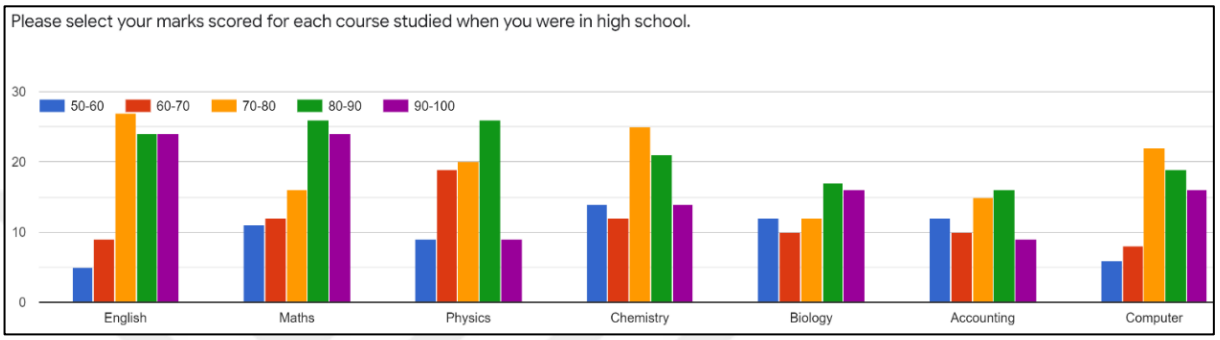
All in all, a total of 91 students (i.e. 36.1 percent) were of higher school secondary education.

3.5 SUBJECTS STUDIED IN HIGH SCHOOL BY HIGH SCHOOL STUDENTS

Moving on, Following the section after the percentages gained in high school. The academics courses and the marks obtained in them were asked by the audience. This is one of the most crucial sections in the questionnaire as this part was one of the fundamental aspects that determined the direction for the career. The main objective of this part of the questionnaire is that it determines the academic base that the student has

and from there it gives a clear pattern in which subject the student excels, hence using this to pave the way for future endeavors. As shown in Figure 3.8, The listed courses in the questionnaire were English, Math's, Physics, Chemistry, Accounting, Biology, and Computer. These courses are the main courses in the curriculum of the Pakistan education system.

Figure 3.8: Different type of subjects that were studied by students in high school



After analyzing thoroughly, Figure 3.8 shows diverse students who studied diverse courses in high school. It can be described as

a) English
   i. 50-60: 5 high school students
   ii. 60-70: 9 students
   iii. 70-80: 27 candidates
   iv. 80-90: 24 students
   v. 90-100: 24 students

It is important to note that out of 91 high school students, 89 students studied the English course

b) Math
   i. 50-60: 11 students
   ii. 60-70: 12 candidates
   iii. 70-80: 16 students
   iv. 80-90: 26 candidates

v. 90-100: 24 students

Similarly, out of 91 students, 89 candidates studied math in high school

c) Physics

i. 50-60: 9 students.
ii. 60-70: 19 candidates.
iii. 70-80: 20 students.
iv. 80-90: 26 candidates.
v. 90-100: 9 students.

Out of 91 students, 83 students completed this course in high school

d) Chemistry

i. 50-60: 14 students.
ii. 60-70: 12 candidates.
iii. 70-80: 25 students.
iv. 80-90: 21 candidates.
v. 90-100: 14 students.

From a total of 91 students, 86 candidates studied the course of chemistry.

e) Biology

i. 50-60: 12 students.
ii. 60-70: 10 candidates.
iii. 70-80: 12 students.
iv. 80-90: 17 candidates.
v. 90-100: 16 students.

67 students out of 91 candidates have studied this course.

f) Accounting

i. 50-60: 12 students.
ii. 60-70: 10 candidates.
iii. 70-80: 15 students.
iv. 80-90: 16 candidates.
v. 90-100: 9 students.

Of 91 student, 62 students took this course in high school

g) Computer

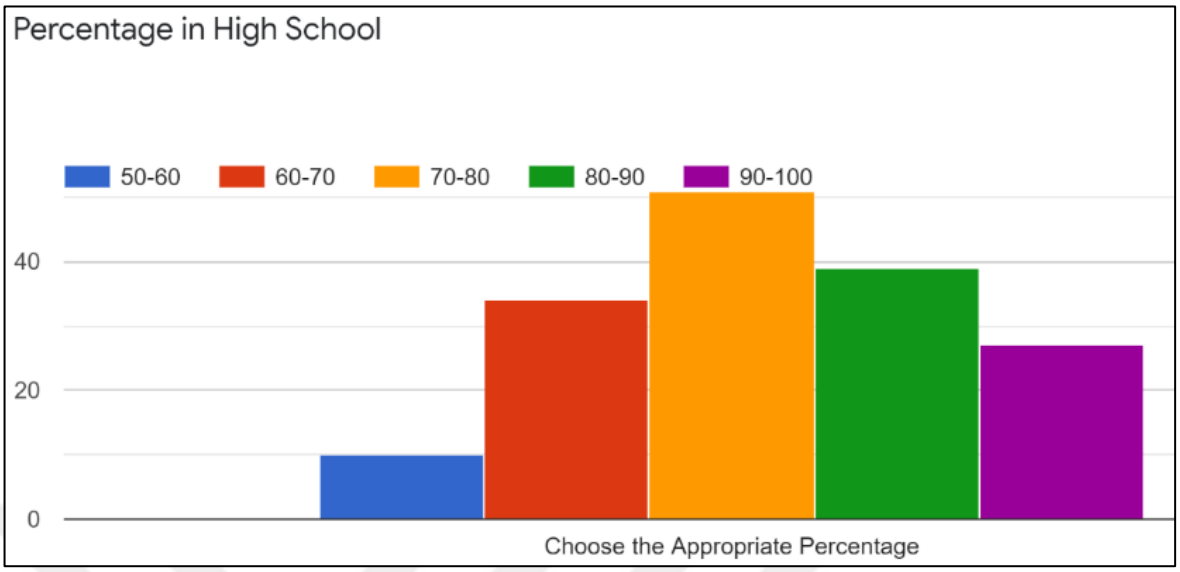
i. 50-60: 6 students.
ii. 60-70: 8 candidates.
iii. 70-80: 22 students.
iv. 80-90: 19 candidates.
v. 90-100: 16 students.

71 students out of 91 candidates have studied this course in high school
So, this closes out the section that was asked specifically for the high school students. Further in the questionnaire, the next section is the questions that are asked specifically for university students.

3.6 PERCENTAGES IN HIGH SCHOOL OF UNIVERSITY STUDENTS

The following area in the survey is the academic part that the university students needed to answer as indicated by their academic background. The academic part was separated independently for the secondary school understudies and the other for the university understudies. Although, candidates of both the segments needed to fill the secondary school scholastics. As appeared in Figure 3.9 beneath, the university candidates were requested to answer their higher secondary school numbers they scored when they graduated from high school. Proceeding onward, Following the segment understudies of secondary school. Proceeding onward to the University understudies and ask them similar inquiries and the outcomes they acquired when they moved on from secondary school.

Figure 3.9: Percentage scored in high school by university students



After the careful analysis of this part, Figure 3.9 shows that:

i. Around 10 understudies scored half – 60 percent when they were in secondary school
ii. Exactly 34 understudies when they were in secondary school scored between 60 percent – 70 percent
iii. Moreover, around 51 understudies had the option to drop with the rate between 70 percent – 80 percent when they were in secondary school.
iv. Likewise, 39 understudies were additionally ready to procure a level of 80 percent – 90 percent when they were in secondary school.
v. Finally, around 27 candidates had the option to graduate at the head of the rate list between 90 percent – 100 percent when they were in secondary school

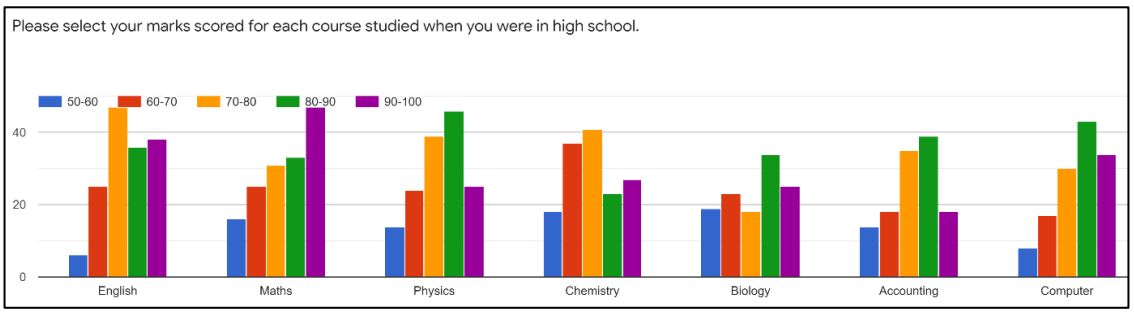
All in all, a total of 161 (i.e. 63.9 percent) students of universities were recorded.

3.7 SUBJECTS STUDIED IN HIGH SCHOOL BY UNIVERSITY STUDENTS

The scholastics courses and the numbers acquired in them were asked from the target audience. This is one of the most critical segments in the survey as this part was one of the major angles that decided the course for the careers. As appeared in Figure 3.10, The recorded courses in the survey were English, Maths, Physics, Chemistry, Accounting,

Biology, and Computer. These courses are the primary courses in the educational plan of the Pakistan training framework.

Figure 3.10: Different type of subjects that were studied by students in high school



After investigating completely, Figure 3.10 shows differing understudies considered different courses in secondary school. It very well may be portrayed as:

a) English
   i. 50-60: 6 students
   ii. 60-70: 25 students
   iii. 70-80: 47 candidates
   iv. 80-90: 36 students
   v. 90-100: 38 students

It is important to note that out of 161 high school students, 152 students studied the English course

b) Math
   i. 50-60: 16 students
   ii. 60-70: 25 candidates
   iii. 70-80: 31 students
   iv. 80-90: 33 candidates
   v. 90-100: 47 students

Similarly, out of 161 students, 152 candidates studied math in high school

c) Physics
    i. 50-60: 14 students.
    ii. 60-70: 24 candidates.
    iii. 70-80: 39 students.
    iv. 80-90: 46 candidates.
    v. 90-100: 25 students.

Out of 161 students, 148 students completed this course in high school

d) Chemistry
    i. 50-60: 18 students.
    ii. 60-70: 37 candidates.
    iii. 70-80: 41 students.
    iv. 80-90: 23 candidates.
    v. 90-100: 27 students.

From a total of 161 students, 146 candidates studied the course of chemistry.

e) Biology
    i. 50-60: 19 students.
    ii. 60-70: 23 candidates.
    iii. 70-80: 18 students.
    iv. 80-90: 34 candidates.
    v. 90-100: 25 students.

119 students out of 161 candidates have studied this course.

f) Accounting
    i. 50-60: 14 students.
    ii. 60-70: 18 candidates.
    iii. 70-80: 35 students.
    iv. 80-90: 39 candidates.
    v. 90-100: 18 students.

Of 161 student, 124 students took this course in high school

g) Computer

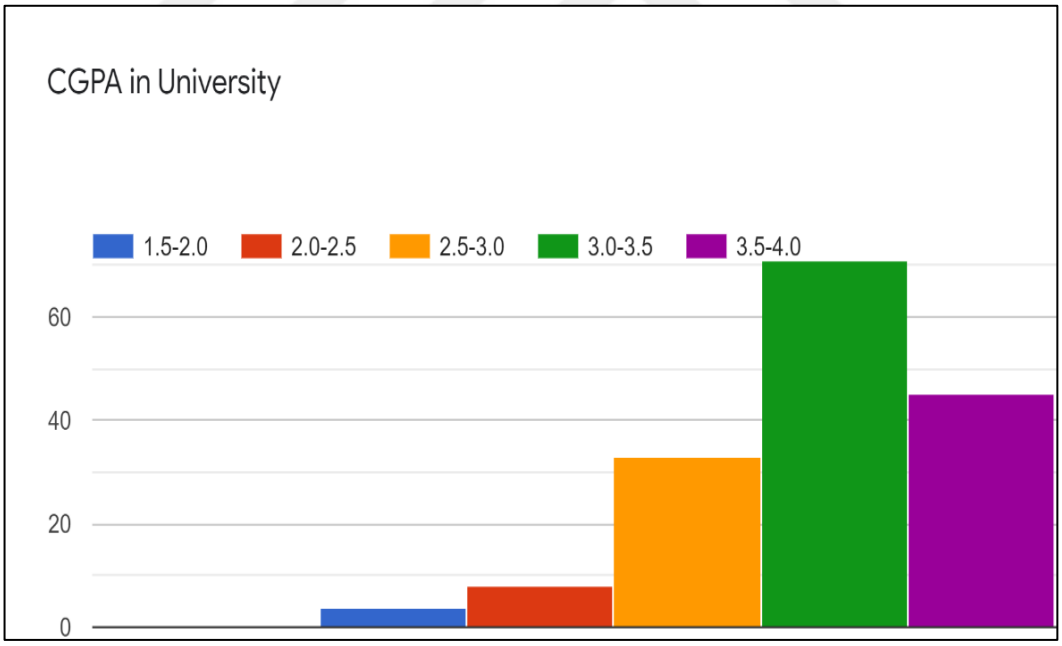
i. 50-60: 8 students.
ii. 60-70: 17 candidates.
iii. 70-80: 30 students.
iv. 80-90: 43 candidates.
v. 90-100: 34 students.

132 students out of 161 candidates have studied this course in high school

3.8 CUMULATIVE GRADE POINT AVERAGE SCORED

The next question that was asked by the university students was the cumulative grade point average they scored whilst they were studying in the university. This data was collected in, Likert scale type format. The students had to choose the GPA they obtained; Figure 3.11 depicts it.

Figure 3.11: Overall grade point average value scored by a university student



The CGPA is calculated out of 4, the CGPA that the students according to the questionnaire are as follows:

i. 1.5- 2.0: 4 students were recorded to have a CGPA between 1.5 and 2.0
ii. 2.0 – 2.5: 8 students had their CGPA between 2. And 2.5
iii. 2.5 – 3.0: 33 students had their CGPA between 2.5 and 3.0
iv. 3.0 – 3.5: 71 candidates acquired a CGPA between 3.0 – 3.5
v. 3.5 - 4.0: 45 students got a top-ranked CGPA between 3.5 – 4.0

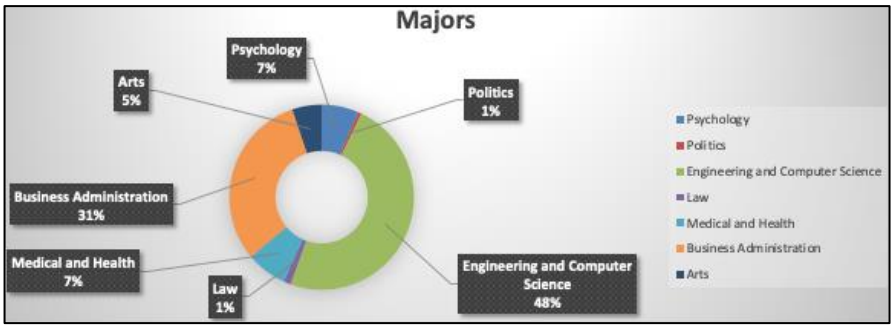
After analyzing the graph, most of the students acquired a cumulative grade point average value between 3.0 – 3.5, and the next nearly populated CGPA is between 3.5 – 4.0.

3.9 UNIVERSITY DEPARTMENTS/MAJOR

After the CGPA was asked in the questionnaire, the next question in line to answer for the university students was the major or the program in which they were enrolled in. There was a total of 7 majors to choose from and each student had to select the most appropriate one. The majors listed were,

i. Arts
ii. Psychology
iii. Politics
iv. Business Administration
v. Medical and Health
vi. Law
vii. Engineering and Computer Sciences

Figure 3.12: Percentage of students that were currently enrolled in the university within their major



According to Figure 3.12, the percent attending the majors out of 100 are as follows:

i. Arts: 5 percent
ii. Psychology: 7 percent
iii. Law: 1 percent
iv. Medical & Health: 7 percent
v. Engineering and Computer Science: 48 percent
vi. Business Administration: 31 percent
vii. Politics: 1 percent

So, after analyzing the image and the dataset carefully, it is to be noted that the highest number of candidates were registered in the major of engineering and computer science i.e. 48 percent and followed closely by business administration i.e. 31 percent.

On the other hand, the least number of students studying the selected majors were politics and law. Due to the very low number of students that were enrolled in these fields, it can cause a limitation as there are not enough records to use for predictions. It will be very hard, almost impossible.

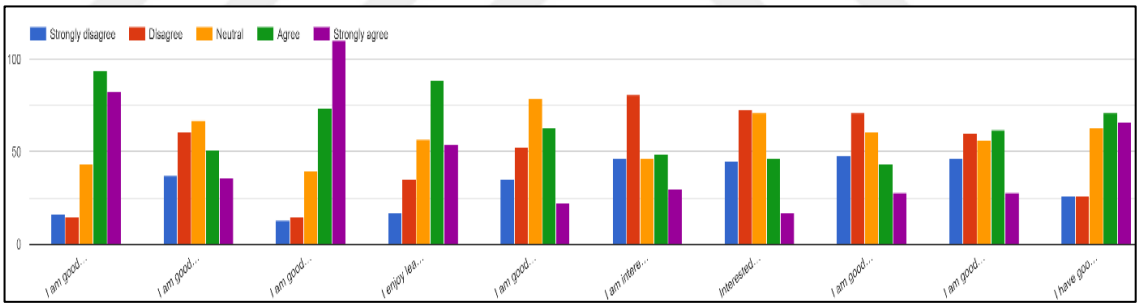
3.10 SKILLS

The final part of the questionnaire was the skills part, this was asked from both the university students as well as high school students. Each student had to select the most appropriate choice in each row. There were 10 questions in this part and their format were liker scale survey. The 5 options to choose from were:

i. Strongly Disagree
ii. Disagree
iii. Neutral
iv. Agree
v. Strongly Agree

The question was asked based on the student's skills and their interests and students had to rate them according to their appropriate judgment. As shown in Figure 3.13, the questions were answered differently by different students.

Figure 3.13: The skills questions that were asked with the overall students



3.11 STATISTICAL PACKAGES FOR SOCIAL SCIENCES (SPSS)

After the data is collected, the most important step that is needed to be done is to validate the questionnaire and it is only done after the data is collected. The validation technique can be done by using software called "Statistical Packages for Social Sciences" or SPSS. To validate the questionnaire using SPSS, the first thing that needed to be done is to convert the sheet in the sheet view of SPSS. The next step is to add all the data in the variable view and assign the appropriate type variable as in SPSS the string variables cannot be validated.

According to the questionnaire, the recorded answers of skills sections that are required to be validated were a string type variable which was an issue in SPSS. To proceed further it was required to convert them and fill them in numeric formats.

To convert the variable type, it was necessary to assign values, in the questionnaire the dummy variables had to be assigned the following values to the variables

i. Strongly Disagree = 1
ii. Disagree = 2
iii. Neutral = 3
iv. Agree = 4
v. Strongly Agree = 5

One of the main requirements of SPSS is that the data view is supposed to be filled with numeric values. Using the above method, the different values are assigned to different columns until the whole excel sheet nothing but numeric values only, excluding the top headings and labels. Figure 3.14 shows the variable view of the dataset in the SPSS after being converted. In other words, all the nominal values are assigned numeric values.

Figure 3.14: The variable view in SPSS

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Gender	Numeric	8	2	Gender	{1,00, Male}...	None	8	Right	Scale	Input
2	Age	Numeric	8	2	Age	{1,00, 0-10}...	None	8	Right	Ordinal	Input
3	StudentType	Numeric	8	2	StudentType	{1,00, High ...	None	8	Right	Scale	Input
4	HSHSPerce...	Numeric	8	2	HSHSPercentage	{1,00, 50-60...	None	8	Right	Ordinal	Input
5	EnglishMarks	Numeric	8	2	EnglishMarks	{1,00, 50-60...	None	8	Right	Ordinal	Input
6	MathsMarks	Numeric	8	2	MathsMarks	{1,00, 50-60...	None	8	Right	Ordinal	Input
7	PhysicsMarks	Numeric	8	2	PhysicsMarks	{1,00, 50-60...	None	8	Right	Ordinal	Input
8	ChemistryM...	Numeric	8	2	ChemistryMarks	{1,00, 50-60...	None	8	Right	Ordinal	Input
9	BiologyMarks	Numeric	8	2	BiologyMarks	{1,00, 50-60...	None	8	Right	Ordinal	Input
10	Accounting...	Numeric	8	2	AccountingMarks	{1,00, 50-60...	None	8	Right	Ordinal	Input
11	ComputerM...	Numeric	8	2	ComputerMarks	{1,00, 50-60...	None	8	Right	Ordinal	Input
12	USCGPA	Numeric	8	2	USCGPA	{1,00, 1.5-2....	None	8	Right	Ordinal	Input
13	UniversityD...	Numeric	8	2	UniversityDepar...	{1,00, Engin...	None	8	Right	Ordinal	Input
14	ComputerKills	Numeric	8	2	ComputerKills	{1,00, Stron...	None	8	Right	Ordinal	Input
15	SingingActi...	Numeric	8	2	SingingActingS...	{1,00, Stron...	None	8	Right	Ordinal	Input
16	GuidingPeo...	Numeric	8	2	GuidingPeople...	{1,00, Stron...	None	8	Right	Ordinal	Input
17	HumanBody...	Numeric	8	2	HumanBodySkills	{1,00, Stron...	None	8	Right	Ordinal	Input
18	PoliticsSkills	Numeric	8	2	PoliticsSkills	{1,00, Stron...	None	8	Right	Ordinal	Input
19	MedicalSkills	Numeric	8	2	MedicalSkills	{1,00, Stron...	None	8	Right	Ordinal	Input
20	LawSkills	Numeric	8	2	LawSkills	{1,00, Stron...	None	8	Right	Ordinal	Input
21	GraphicWe...	Numeric	8	2	GraphicWebSki...	{1,00, Stron...	None	8	Right	Ordinal	Input
22	BankingSkills	Numeric	8	2	BankingSkills	{1,00, Stron...	None	8	Right	Ordinal	Input
23	Leadership...	Numeric	8	2	LeadershipSkills	{1,00, Stron...	None	8	Right	Ordinal	Input

Figure 3.14 depicts the variable view in which all the dataset is converted into a numeric type even for gender and student type data. The measure is converted into ordinal for all the data types.

Figure 3.15 depicts the excel sheet after all the values have been assigned and converted. All the string values now show only the numeric value

Figure 3.15: The excel sheet in SPSS

	Gender	Age	StudentType	HSHSPercentage	EnglishMarks	MathsMarks	PhysicsMarks	ChemistryMarks	BiologyMarks	AccountingMarks	ComputerMarks	USCGPA	UniversityDepartment	ComputerSkills	SingingActingSkills	GuidingPeopleSkills	HumanBodySkills	PoliticsSkills	MedicalSkills	LawSkills	GraphicWebSkills	BankingSkills	LeadershipSkills
184	2,00	5,00	2,00	2,00	3,00	3,00	2,00	2,00	2,00	3,00	3,00	5,00	7,00	3,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00
185	1,00	5,00	2,00	1,00	.	.	.	.	.	.	.	3,00	1,00	3,00	3,00	3,00	3,00	2,00	2,00	2,00	3,00	2,00	3,00
186	1,00	5,00	2,00	3,00	3,00	3,00	3,00	2,00	2,00	3,00	3,00	4,00	6,00	5,00	3,00	3,00	3,00	3,00	3,00	3,00	3,00	2,00	3,00
187	2,00	5,00	2,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	4,00	7,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
188	2,00	5,00	2,00	1,00	3,00	1,00	1,00	1,00	1,00	1,00	3,00	5,00	5,00	5,00	5,00	5,00	5,00	1,00	1,00	1,00	1,00	1,00	1,00
189	1,00	5,00	2,00	3,00	5,00	2,00	4,00	3,00	5,00	2,00	1,00	4,00	6,00	2,00	5,00	5,00	5,00	5,00	2,00	1,00	1,00	2,00	5,00
190	1,00	5,00	2,00	3,00	2,00	3,00	2,00	3,00	2,00	3,00	2,00	3,00	6,00	5,00	3,00	5,00	5,00	5,00	2,00	3,00	3,00	5,00	5,00
191	2,00	5,00	2,00	3,00	3,00	4,00	3,00	4,00	4,00	3,00	2,00	4,00	5,00	5,00	3,00	5,00	5,00	3,00	5,00	2,00	5,00	3,00	3,00
192	2,00	5,00	2,00	3,00	3,00	4,00	3,00	4,00	4,00	3,00	2,00	4,00	5,00	5,00	3,00	5,00	5,00	3,00	5,00	2,00	5,00	3,00	3,00
193	1,00	5,00	2,00	3,00	3,00	3,00	2,00	2,00	3,00	4,00	4,00	5,00	6,00	5,00	3,00	5,00	5,00	5,00	1,00	2,00	1,00	3,00	3,00

SPSS shows them in a spreadsheet-like style as appeared in Figure 3.15. This sheet is called a data view continually showing our information values.

3.11.1 Validity Test

Pearson's correlation coefficient is the covariance of the two factors partitioned by the result of their standard deviations. The type of the definition includes a "product moment", that is, the mean (the primary second about the inception) of the result of the mean-balanced irregular factors.

The next part after assigning values to the records is to find out whether the questionnaire data is either valid or not. To find it out, A Pearson correlation-two-tailed test is performed on the data. Figure 3.16 depicts the results after performing the validity through bivariate correlation.

Figure 3.16: The Pearson's Correlation

Pearson Correlation	1	,153*	,463**	,215**	,207**	,148*	,140*	,294**	,371**	,455**	,576**
Sig. (2-tailed)		,015	,000	,001	,001	,018	,025	,000	,000	,000	,000
N	253	253	253	253	253	253	253	253	253	253	253
Pearson Correlation	,153*	1	,305**	,337**	,085	,139*	,214**	,299**	,197**	,103	,488**
Sig. (2-tailed)	,015		,000	,000	,180	,027	,001	,000	,002	,102	,000
N	253	253	253	253	253	253	253	253	253	253	253
Pearson Correlation	,463**	,305**	1	,514**	,385**	,267**	,298**	,206**	,301**	,516**	,710**
Sig. (2-tailed)	,000	,000		,000	,000	,000	,000	,001	,000	,000	,000
N	253	253	253	253	253	253	253	253	253	253	253
Pearson Correlation	,215**	,337**	,514**	1	,330**	,535**	,394**	,185**	,126*	,214**	,653**
Sig. (2-tailed)	,001	,000	,000		,000	,000	,000	,003	,046	,001	,000
N	253	253	253	253	253	253	253	253	253	253	253
Pearson Correlation	,207**	,085	,385**	,330**	1	,176**	,378**	,025	,195**	,434**	,547**
Sig. (2-tailed)	,001	,180	,000	,000		,005	,000	,688	,002	,000	,000
N	253	253	253	253	253	253	253	253	253	253	253
Pearson Correlation	,148*	,139*	,267**	,535**	,176**	1	,433**	,181**	,048	,093	,522**
Sig. (2-tailed)	,018	,027	,000	,000	,005		,000	,004	,451	,140	,000
N	253	253	253	253	253	253	253	253	253	253	253
Pearson Correlation	,140*	,214**	,298**	,394**	,378**	,433**	1	,133*	,363**	,338**	,635**
Sig. (2-tailed)	,025	,001	,000	,000	,000	,000		,035	,000	,000	,000
N	253	253	253	253	253	253	253	253	253	253	253
Pearson Correlation	,294**	,299**	,206**	,185**	,025	,181**	,133*	1	,296**	,187**	,486**
Sig. (2-tailed)	,000	,000	,001	,003	,688	,004	,035		,000	,003	,000
N	253	253	253	253	253	253	253	253	253	253	253
Pearson Correlation	,371**	,197**	,301**	,126*	,195**	,048	,363**	,296**	1	,489**	,588**
Sig. (2-tailed)	,000	,002	,000	,046	,002	,451	,000	,000		,000	,000
N	253	253	253	253	253	253	253	253	253	253	253
Pearson Correlation	,455**	,103	,516**	,214**	,434**	,093	,338**	,187**	,489**	1	,651**
Sig. (2-tailed)	,000	,102	,000	,001	,000	,140	,000	,003	,000		,000
N	253	253	253	253	253	253	253	253	253	253	253
Pearson Correlation	,576**	,488**	,710**	,653**	,547**	,522**	,635**	,486**	,588**	,651**	1
Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	
N	253	253	253	253	253	253	253	253	253	253	253

According to the document of SPSS, no single reliability list can be viewed as an ideal appraisal apparatus to comprehend the problem. Subsequently, at any rate, a few different methods ought to be utilized to guarantee the dependability of the survey. The ideal range for the Pearson Correlation is between -1 to 1 and Figure 3.16 depicts all the values that fall between hence confirming that the data is valid.

3.11.2 Reliability Test

Cronbach's alpha is a proportion of inside consistency, that is, how firmly related a lot of things are as a gathering. It is viewed as a proportion of scale dependability. "high" esteem for alpha doesn't infer that the measure is one-dimensional. Cronbach's alpha is certifiably not a factual test – it is a coefficient of dependability (or consistency).

Now that the values are converted, it's easier to apply the reliability analysis. The questionnaire was broken down into two sections, one section deals with the demographic information or facts of both high school and university students.

The skills/interest's part of the questionnaire contains unresolved data which is where Cronbach's Alpha is needed to be applied. The reliability of the Cronbach alpha is above 0.5, the higher it is then there is more internal consistency in the data. If the value of Cronbach's alpha is below 0.5 then it means, there's a higher internal inconsistency which makes the data and the questionnaire invalid. Table 3.1 shows the results of the Cronbach alpha

Table 3.1: Cronbach's alpha

Case Processing Summary		
Valid	253	100.0
Excluded	0	.0
Total	253	100.0
Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.783	.783	10

After applying Cronbach's alpha to measure the reliability of the questionnaire, the output states that the value of Cronbach's alpha is **.**783 which is significantly higher than 0.5 thus also confirming that there is a high internal consistency level in the dataset.

3.11.3 Descriptive Statistics

After completing the reliability and validity test and confirming that the data that was collected is reliable and valid, Descriptive analysis can be performed on the data to get an ideal perspective regarding the dataset.

Descriptive statistics help portray the essential highlights of information, for instance, the rundown measurements for the scale factors and proportions of the information. In an exploration concentrate on huge information, these insights may assist us in dealing with the information and present it in an outline table.

Figure 3.17: Descriptive statistics

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Gender	253	1,00	2,00	1,4941	,50096
Age	253	2,00	5,00	3,8221	,77407
StudentType	253	1,00	2,00	1,6364	,48200
HSHSPercentage	253	1,00	5,00	3,2530	1,15781
EnglishMarks	242	1,00	5,00	3,5248	1,15293
MathsMarks	242	1,00	5,00	3,4504	1,35105
PhysicsMarks	232	1,00	5,00	3,2155	1,20105
ChemistryMarks	233	1,00	5,00	3,0515	1,28885
BiologyMarks	186	1,00	5,00	3,2043	1,40316
AccountingMarks	187	1,00	5,00	3,1658	1,25694
ComputerMarks	204	1,00	5,00	3,5294	1,18866
USCGPA	161	1,00	5,00	3,9006	,95003
UniversityDepartment	161	1,00	7,00	3,3106	2,42702
ComputerKills	253	1,00	5,00	4,2055	1,29608
SingingActingSkills	253	1,00	5,00	3,1502	1,48042
GuidingPeopleSkills	253	1,00	5,00	4,2925	1,22856
HumanBodySkills	253	1,00	5,00	3,8656	1,39362
PoliticsSkills	253	1,00	5,00	3,1937	1,44682
MedicalSkills	253	1,00	5,00	2,9486	1,52275
LawSkills	253	1,00	5,00	2,8538	1,40801
GraphicWebSkills	253	1,00	5,00	2,9051	1,47697
BankingSkills	253	1,00	5,00	3,1028	1,54987
LeadershipSkills	253	1,00	5,00	3,7668	1,45466
Valid N (listwise)	106				

In this section, descriptive statistics are conducted on the whole dataset. The N column states that the number of users that filled each of the questions asked, for example, the University Department was ought to be filled but only the university students hence 161 out of 253 total students filled it indicating that the rest were high school students. In the minimum and the maximum columns, the lowest and the highest values entered like the values assigned between females and males in the gender column were “1” and “2” respectively. Thus, the minimum “1” value represents the male participants and on the other hand, “2” represents the female participants. The last two columns calculated the mean and the standard deviation respectively for each. Figure 3.17 represents the overview of the descriptive statistics performed on the dataset.

3.12 DATA PRE-PROCESSING

Information preprocessing is a significant advance in the information mining measure. The expression "trash in, trash out" is especially material to information mining and AI ventures. Information gathering strategies are regularly approximately controlled, coming about in out-of-go values, outlandish information blends, missing qualities, and so on. Examining information that has not been painstakingly screened for such issues can deliver misdirecting results. Subsequently, the portrayal and nature of information are above all else before running an examination. Frequently, information preprocessing is the most significant period of an AI venture, particularly in computational science. If there is a lot of unimportant and excess data present or uproarious and inconsistent information, at that point information revelation during the preparation stage is more troublesome. Information readiness and separating steps can take a significant measure of handling time. Information preprocessing incorporates cleaning, Instance determination, standardization, change, highlight extraction and choice, and so forth. The result of information preprocessing is the last preparing set.

Information pre-handling may influence how results of the last information preparation can be deciphered. This viewpoint ought to be deliberately viewed as when an understanding of the outcomes is a central issue, such as in the multivariate preparation of concoction information.

The likelihood of irregular information has expanded in the present information because of its humongous size and its cause for heterogeneous sources. Considering the way that top-notch information prompts better models and forecasts, information preprocessing has gotten indispensable and the key advance in the information science and machine learning pipeline.

While gathering information, one may go over three principle factors that would add to the nature of information:
Accurate: Erroneous qualities that go astray from the normal. The reasons for erroneous information can be different, which include:

a) Human/PC blunders during information passage and transmission
b) Clients purposely submitting erroneous qualities
c) Mistaken arrangements for input fields
d) Duplication of preparing models

Complete: Lacking attribute esteems or estimations of interest. The dataset may be inadequate due to:

a) Inaccessibility of information
b) Cancellation of conflicting information
c) Cancellation of information regarded insignificant at first
d) Consistency: Aggregation of information is conflicting.

To guarantee top-notch information, it's critical to preprocess it. To make the cycle simpler, information preprocessing is separated into four phases:

3.12.1 Data Cleaning

Information cleaning alludes to strategies to 'clean' information by eliminating outliers, supplanting missing information, smoothing noisy information, and remedying conflicting information. Numerous methods are utilized to play out every one of these undertakings, where every procedure is explicit to the client's inclination or issue set.
Missing Values – The incomplete values can be processed by either filling them in or removing the whole row.

i. Filling in missing worth physically: This methodology is tedious, and not suggested for gigantic informational collections.
ii. Utilizing a standard incentive to supplant the missing worth: The missing worth can be supplanted by a worldwide steady, for example, 'N/A' or 'Unknown'.
iii. Utilizing focal inclination (Mean, Median, Mode) for a trait to supplant the missing worth: Based on information circulation, mean or can be utilized to fill in for the missing worth or having a place in the same class as its predecessors.

However, due to the questionnaire being on a Likert scale, it is not feasible in the dataset, so to deal with the missing values, the removal of the entire record is done to make the

dataset as smooth as possible. The other method utilized was the mode method in which if only a single value was missing then the value that was occurring the most in the dataset was used, thus making it better.

3.12.2 Data Scaling

This implies the transformation of the information, so it fits inside a scale, like 0–100 or 0–1. This is needed to be done to scale information when utilizing techniques dependent on proportions of how far separated information focuses, like support vector machines or k-nearest neighbors. With these calculations, a difference in "1" in any numeric component is given a similar significance.

In this dataset, all the subject marks and the percentages of high schools along with the CGPA and skills were all arranged in a Likert scale format, in other words, it was a categorical value as the options were between ranges. This data was then scaled ad assigned values as follows:

The marks divisions across each course i.e. English, math, physics, chemistry, biology, accounting, computer were as follows

i. 50 – 60 = 1
ii. 60 – 70 = 2
iii. 70 – 80 = 3
iv. 80 – 90 = 4
v. 90 – 100 = 5

The total percentage divisions across the education system were as follows

i. 50 percent – 60 percent = 1
ii. 60 percent – 70 percent = 2
iii. 70 percent – 80 percent = 3
iv. 80 percent – 90 percent = 4
v. 90 percent – 100 percent = 5

The Cumulative Grade Point Average Value for university students was as follows:

i. 1.5– 2.0 = 1
ii. 2.0 – 2.5 = 2
iii. 2.5 – 3.0 = 3
iv. 3.0 – 3.5 = 4
v. 3.5 – 4.0 = 5

The University department in which the students were doing their major was as follows:

i. Engineering/Computer Science = 1
ii. Law = 2
iii. Medical/Health = 3
iv. Politics = 4
v. Psychology = 5
vi. Business Administration = 6
vii. Arts = 7

The Student type divisions across the dataset were done as follows:

i. High School Student = 1
ii. University Student = 2

The Gender column was assigned the following value:

i. Male = 1
ii. Female = 2

The Age group division across the dataset was as follows:

i. 0 – 10 = 1
ii. 10 – 15 = 2
iii. 15 – 20 = 3
iv. 20 – 25 = 4
v. 25 – 30 = 5

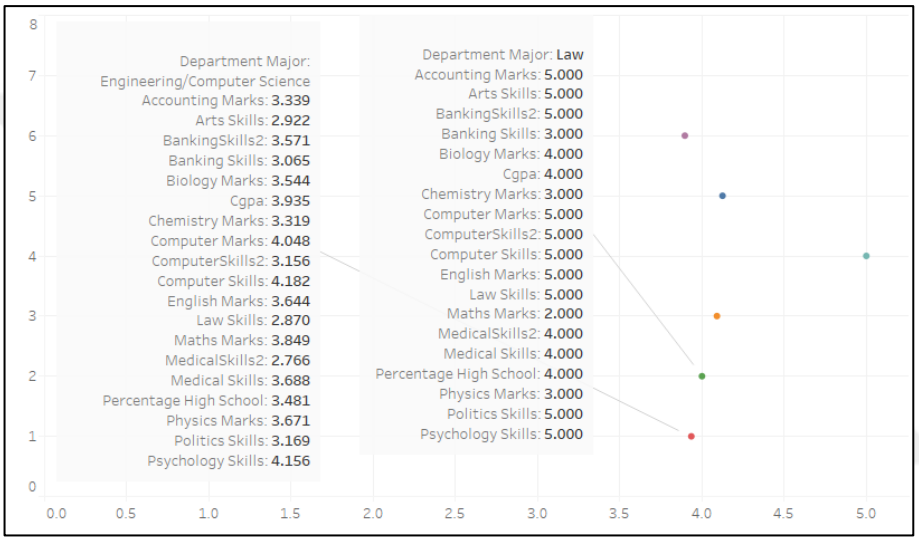
This skills part, that was filled by each student divided across the dataset as the following:

i. Strongly Disagree = 1
ii. Disagree = 2
iii. Neutral = 3
iv. Agree = 4
v. Strongly Agree = 5

3.12.3 Data Visualization

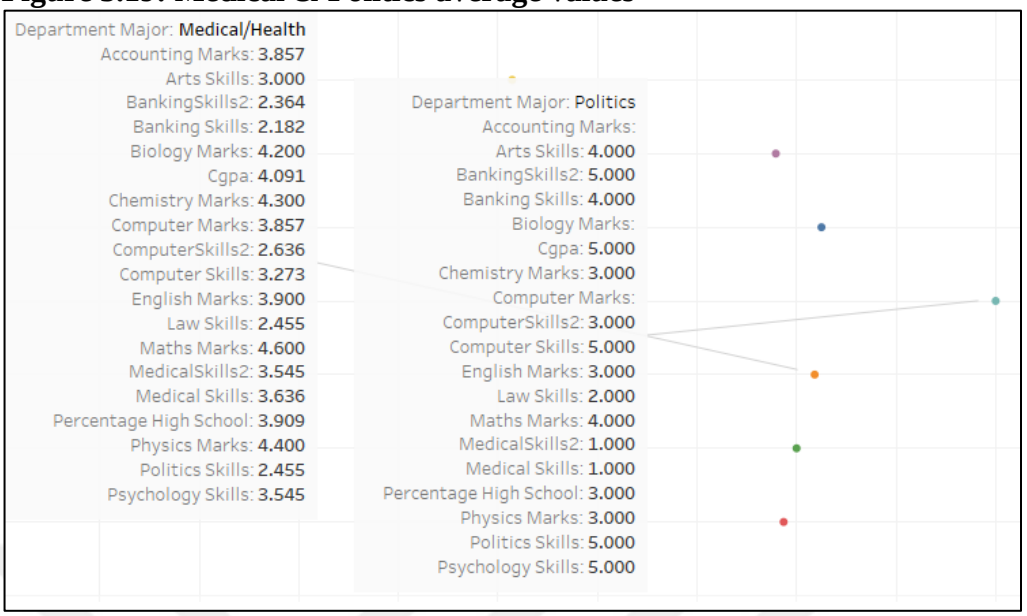
Figures 3.18-3.21 is a visualization of the university student's data that has been clustered according to their respective department in the university. After grouping them, the next step is to take out the average of each data within the cluster, thus displaying a rough idea of the information.

Figure 3.18: Engineering & Law average values



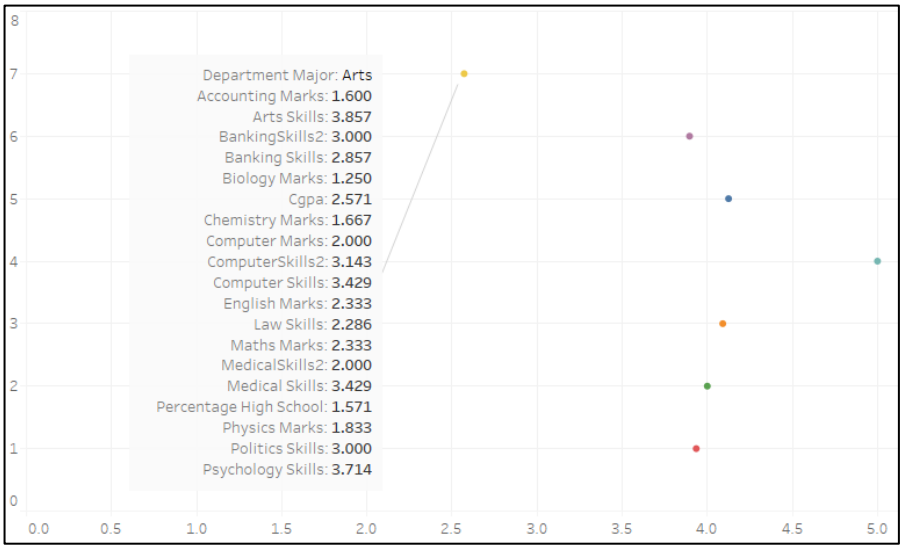
Considering the values that were assigned for the categorical variables, it is visible that data differs from the other clusters like the Department of Law and Engineering as shown in Figure 3.18

Figure 3.19: Medical & Politics average values



Similarly, as shown in Figure 3.19, the Medical & Politics departments consist of different average results indicating what are the characteristics of the average student.

Figure 3.20: Arts average values



Following up, the department of arts highlights the various characters of an average student as shown in Figure 3.20.

Figure 3.21: Business administration and psychology clusters

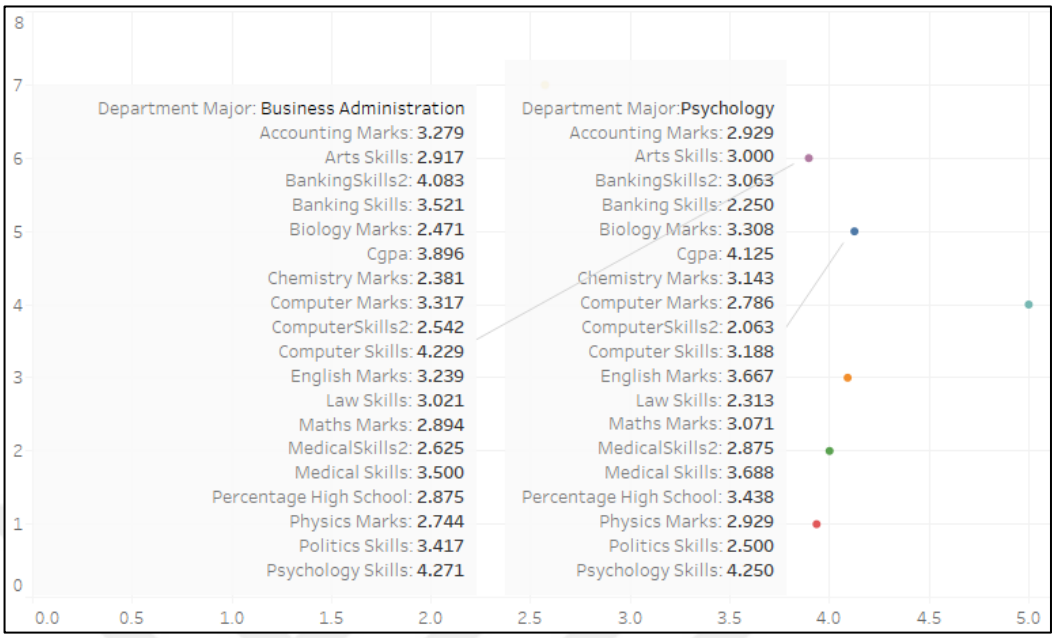


Figure 3.21 displays the average values of the business administration and psychology departments. Thus, this concludes the visualization of the data after having it pre-processed.

4. METHODOLOGY

In machine learning, classification alludes to a prescient demonstrating issue where a class name is anticipated for a given case of the information. A model will utilize the training dataset and will compute how to best guide instances of information to explicit class names. All things considered, the training dataset must be adequately illustrative of the issue and have numerous instances of each class name. There is a wide range of various classification calculations for modeling classification predictive modeling issues. There is nothing but a bad hypothesis on the most proficient method to plan calculations onto issue types; rather, it is largely suggested that an expert utilize controlled analyses and find which calculation and calculation setup brings about the best presentation for a given arrangement task. There is nothing but a bad hypothesis on the most proficient method to plan calculations onto issue types; rather, it is, by and large, suggested that an expert utilize controlled analyses and find which calculation and algorithm methods bring about the best presentation for a given classification task. Classification predictive demonstrating calculations are assessed depending on their outcomes. The accuracy of the classification is a well-known measurement used to assess the exhibition of a model depending on the anticipated class names. Classification precision isn't perfect however is a decent beginning stage for some, classification problems. After already assigning values to the categorical variables in the SPSS section, the same file is then used to apply the machine learning algorithms in the machine learning environments. The reason to use the pyreadstat library is that this library automatically reads the numerical values assigned to the original categorical variables, saving us the trouble from further assigning dummy variables.

Training data is used to ensure that the machine identifies patterns in the data, cross-validation data is used to ensure the algorithm used to train the machine is much more reliable and consistent, and test data is used to see how well the machine can predict new responses based on its training.

Figure 4.1 shows the test data being displayed after being imported whereas Figure 4.2 displays the training data after being imported.

Figure 4.1: The high school students' data after being imported

PercentageHighSchool	EnglishMarks	MathsMarks	PhysicsMarks	ChemistryMarks	BiologyMarks	AccountingMarks	ComputerMarks	ComputerSkills	ArtsSkills
2.0	2.0	2.0	2.0	2.0	NaN	NaN	2.0	1.0	2.0
2.0	2.0	3.0	2.0	3.0	2.0	3.0	2.0	1.0	2.0
5.0	5.0	5.0	5.0	5.0	5.0	5.0	3.0	5.0	1.0
3.0	2.0	5.0	2.0	1.0	3.0	5.0	4.0	1.0	2.0
5.0	5.0	4.0	4.0	3.0	NaN	NaN	3.0	4.0	1.0
...	...	...	...	...	...	...	...	...	...
4.0	4.0	3.0	NaN	NaN	NaN	4.0	4.0	4.0	2.0
4.0	4.0	3.0	NaN	3.0	4.0	NaN	NaN	3.0	2.0
3.0	2.0	2.0	2.0	1.0	NaN	NaN	2.0	4.0	1.0
5.0	5.0	5.0	NaN	5.0	5.0	NaN	NaN	4.0	2.0
3.0	4.0	3.0	2.0	2.0	NaN	NaN	NaN	5.0	3.0

Figure 4.2: The university students' data after being imported

PercentageHighSchool	EnglishMarks	MathsMarks	PhysicsMarks	ChemistryMarks	BiologyMarks	AccountingMarks	ComputerMarks	ComputerSkills	ArtsSkills
3.0	3.0	3.0	3.0	3.0	NaN	NaN	5.0	3.0	1.0
5.0	5.0	5.0	5.0	5.0	5.0	NaN	NaN	4.0	2.0
5.0	5.0	5.0	5.0	5.0	5.0	NaN	NaN	4.0	2.0
5.0	5.0	5.0	5.0	5.0	5.0	NaN	NaN	4.0	2.0
5.0	5.0	5.0	5.0	5.0	5.0	NaN	NaN	4.0	2.0
...	...	...	...	...	...	...	...	...	...
2.0	2.0	1.0	1.0	1.0	1.0	2.0	4.0	5.0	1.0
3.0	3.0	5.0	3.0	1.0	3.0	5.0	4.0	5.0	1.0
3.0	4.0	5.0	5.0	3.0	4.0	4.0	4.0	1.0	5.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4.0	5.0	4.0	4.0	4.0	4.0	4.0	4.0	1.0	1.0

In the dataset, many columns consisted of blank spaces which were automatically marked as NaN meaning it was null values.

Table 4.1 shows which columns consisted of NaN values and how many in total along with other columns in the test data i.e. High School Students

Table 4.1: Number of NaN values in high school student's dataset

English Marks	1
Math Marks	1
Physics Marks	7
Chemistry Marks	4
Biology Marks	24
Accounting Marks	29
Computer Marks	20

Precisely there were 86 blank spaces in the test data. Similarly, the same process is followed to identify the number of blank spaces in the training data. Table 4.2 displays the information of the University Students

Table 4.2: Number of NaN values in university student's

English Marks	2
Physics Marks	4
Chemistry Marks	6
Biology Marks	31
Accounting Marks	23
Computer Marks	19

The training data consists of 85 blank spaces that need to be processed. The next step is to fill these NaN values with 0 to further simplify the data. This enables the data to be in a single numeric data type, every string type variable is converted into a numeric one.

Figure 4.3: The high school students' data after NaN values being replaced

PercentageHighSchool	EnglishMarks	MathsMarks	PhysicsMarks	ChemistryMarks	BiologyMarks	AccountingMarks	ComputerMarks	ComputerSkills	ArtsSkills
2.0	2.0	2.0	2.0	2.0	0.0	0.0	2.0	1.0	2.0
2.0	2.0	3.0	2.0	3.0	2.0	3.0	2.0	1.0	2.0
5.0	5.0	5.0	5.0	5.0	5.0	5.0	3.0	5.0	1.0
3.0	2.0	5.0	2.0	1.0	3.0	5.0	4.0	1.0	2.0
5.0	5.0	4.0	4.0	3.0	0.0	0.0	3.0	4.0	1.0
...	...	...	...	...	...	...	...	...	...
4.0	4.0	3.0	0.0	0.0	0.0	4.0	4.0	4.0	2.0
4.0	4.0	3.0	0.0	3.0	4.0	0.0	0.0	3.0	2.0
3.0	2.0	2.0	2.0	1.0	0.0	0.0	2.0	4.0	1.0
5.0	5.0	5.0	0.0	5.0	5.0	0.0	0.0	4.0	2.0
3.0	4.0	3.0	2.0	2.0	0.0	0.0	0.0	5.0	3.0

Figure 4.3 shows the updated test data of high school students. All the NaN values have been replaced by 0

Figure 4.4: The university students' data after NaN values being replaced

PercentageHighSchool	EnglishMarks	MathsMarks	PhysicsMarks	ChemistryMarks	BiologyMarks	AccountingMarks	ComputerMarks	ComputerSkills	ArtsSkills
3.0	3.0	3.0	3.0	3.0	0.0	0.0	5.0	3.0	1.0
5.0	5.0	5.0	5.0	5.0	5.0	0.0	0.0	4.0	2.0
5.0	5.0	5.0	5.0	5.0	5.0	0.0	0.0	4.0	2.0
5.0	5.0	5.0	5.0	5.0	5.0	0.0	0.0	4.0	2.0
5.0	5.0	5.0	5.0	5.0	5.0	0.0	0.0	4.0	2.0
...	...	...	...	...	...	...	...	...	...
2.0	2.0	1.0	1.0	1.0	1.0	2.0	4.0	5.0	1.0
3.0	3.0	5.0	3.0	1.0	3.0	5.0	4.0	5.0	1.0
3.0	4.0	5.0	5.0	3.0	4.0	4.0	4.0	1.0	5.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4.0	5.0	4.0	4.0	4.0	4.0	4.0	4.0	1.0	1.0

Similarly, Figure 4.4 shows the updated training data of university students. Like the previous one, all the NaN values have been replaced with 0.0 in this also.

The next step after is to verify the data, That Is done by running through the entire dataset to check for columns that consist of numerical or categorical values that include numbers across both datasets. Table 4.3 displays all the columns from both datasets that contain numerical data.

Table 4.3: The columns that the datasets consist

High School Students Data (Test Data)	University Students Data (Training Data)
PercentageHighSchool	PercentageHighSchool
EnglishMarks	EnglishMarks
MathsMarks	MathsMarks
PhysicsMarks	PhysicsMarks
ChemistryMarks	ChemistryMarks
BiologyMarks	BiologyMarks
AccountingMarks	AccountingMarks
ComputerMarks	ComputerMarks
ComputerSkills	ComputerSkills
ArtsSkills	ArtsSkills
PsychologySkills	PsychologySkills
MedicalSkills	MedicalSkills

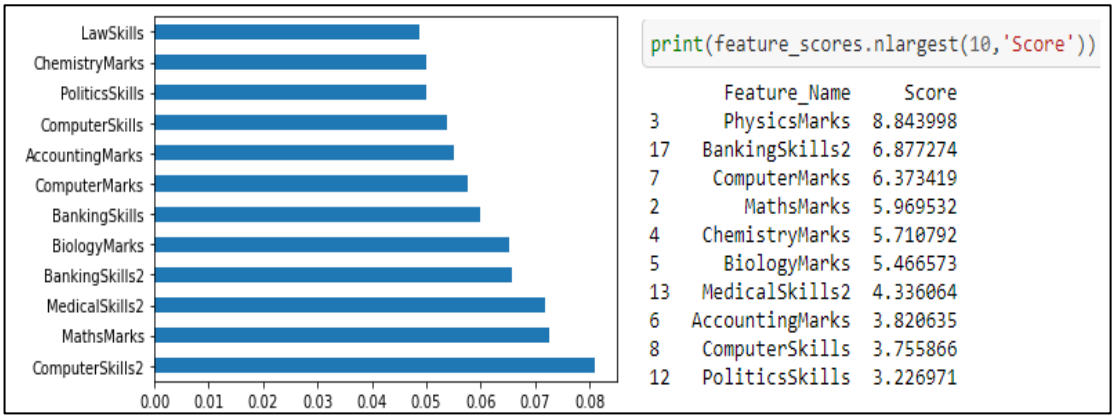
PoliticsSkills	PoliticsSkills
MedicalSkills2	MedicalSkills2
LawSkills	LawSkills
ComputerSkills2	ComputerSkills2
BankingSkills	BankingSkills
BankingSkills2	BankingSkills2
	CGPA
	UniversityDepartment

The CGPA and the University Department are the different columns as the university department is the column in which the prediction needs to be done. Therefore, it is unique.

4.1 FEATURE SELECTION

The most instrumental step that was needed to be done in our method was the feature selection procedure. Using the recursive feature elimination method, it was noteworthy to pinpoint the columns that had the potential to give us a good result by comparing the accuracies. Out of 18 columns, 12 columns were deemed worthy because they produced higher individual accuracies. Applying the classification algorithms under these columns produced fruitful results. The reason to use these columns is to get a higher accuracy level, after conducting many procedures it was noted that these specific columns escalated the accuracy level. It was to be noted that the fewer features are used then the higher the accuracy level climbs. Figure 4.5 depicts the mentioned information.

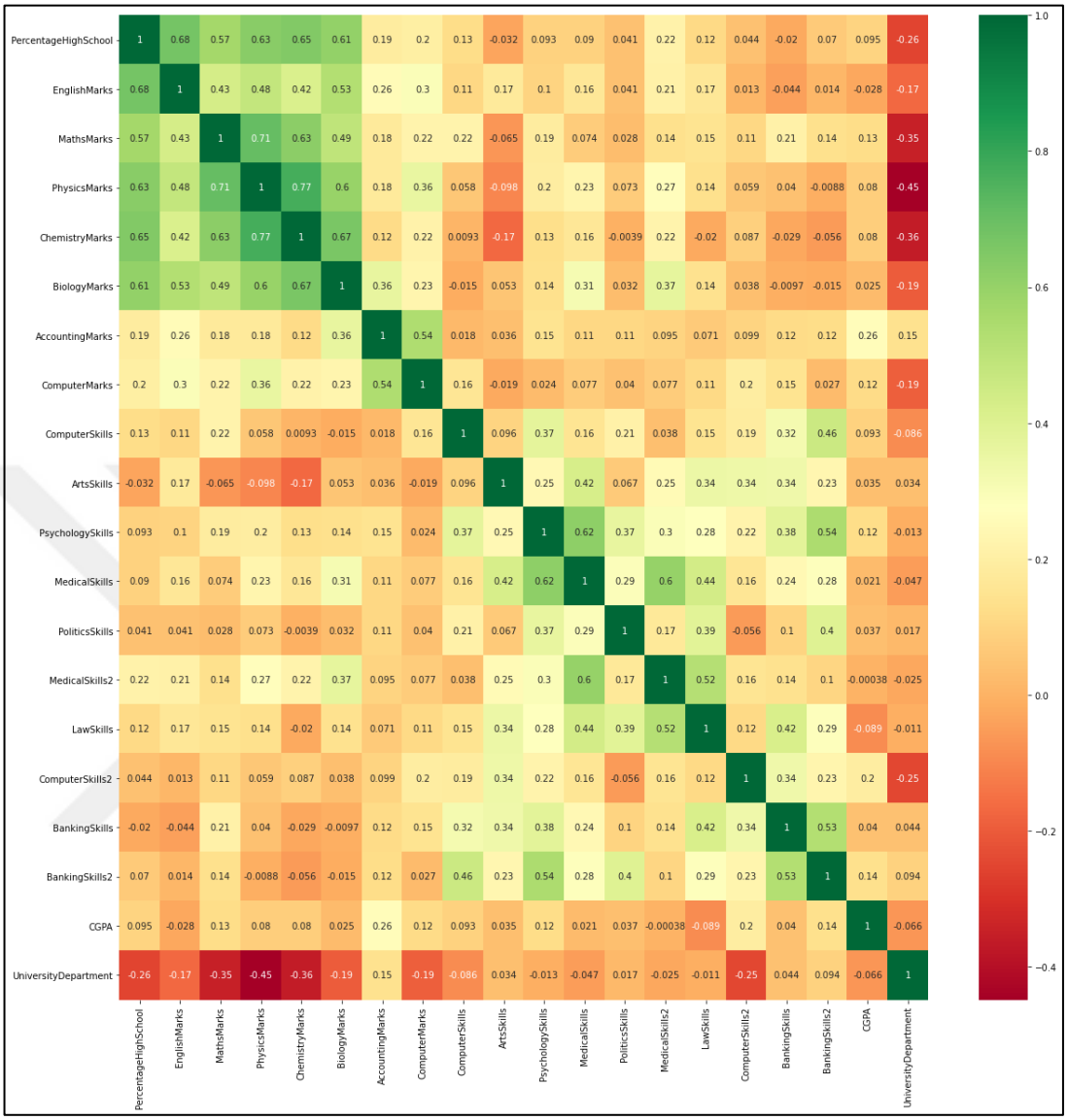
Figure 4.5: Best features accuracy



4.2 CORRELATION HEATMAP

The next step following this is to give a visualization of the correlations between the variables. The best to visualize it is to generate a heatmap. This gives a more simplified idea about which variables have a strong correlation with each other and which have a weak correlation among them.

Figure 4.6: The correlation heatmap



According to Figure 4.6, dark green suggests a strong positive correlation whereas, on the other hand, dark red suggests a strong negative correlation. However, only a couple of variables show a strong negative correlation whereas, on the contrary, a handful of variables show a positive correlation. The reason to perform this step is to find out what type of relationship do the variables have between them, The strong positive correlation dictates that the variables have a linear relationship among them and move forward in tandem, whereas the strong negative correlation means there's an opposite relation among them. This also outlines the effectiveness of the variables when using them for prediction.

4.3 SPLITTING DATA

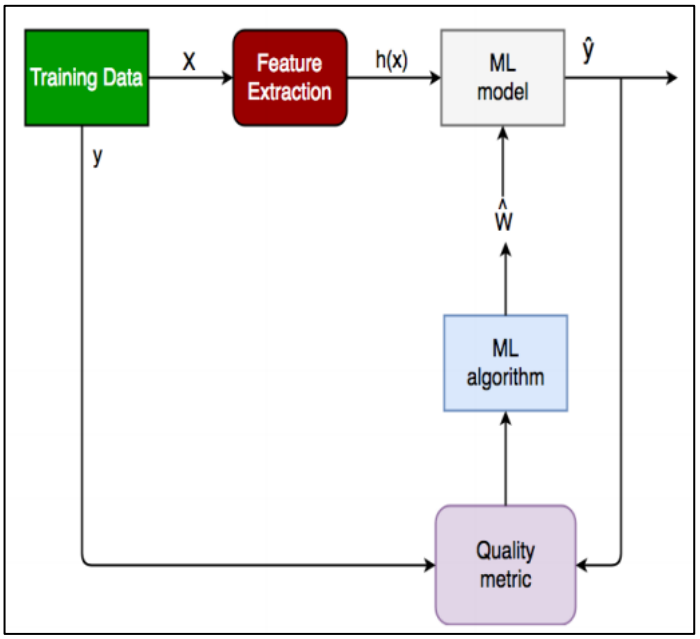
This entire dataset was isolated to two sets (70 percent) of the populace were considered as the set that was to labeled as training with 150 understudies, the leftover (30 percent) has been similarly separated to the approval set as 91 understudies as the test set.

4.4 CLASSIFICATION MODELS

In a machine learning setting, classification is a kind of supervised learning. Supervised learning implies that the information entered in the system is named beforehand, with the significant characteristics previously isolated into classifications heretofore. Figure 4.7 outlines the workflow structure of the machine learning technique.

This implies that the system realizes which parts of the information are significant, and there is additionally an objective or ground truth that the system can examine itself against.

Figure 4.7: Machine learning workflow diagram



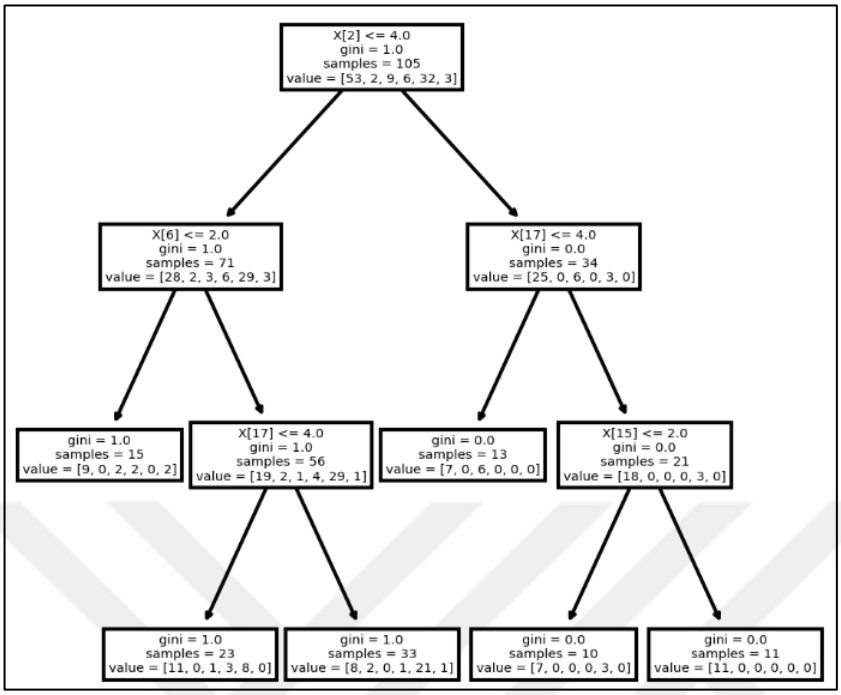
In other words, classification is the issue of distinguishing which of a lot of classes (sub-populaces), a groundbreaking perception has a place with, based on a preparation set of information containing perceptions and whose classifications enrollment is known.
On the dataset, the following classification algorithms are used.

4.4.1 Decision Trees

Tree-based learning calculations are viewed as truly outstanding and generally utilized administered learning techniques. Tree-based techniques enable prescient models with high precision, security, and simplicity of understanding. In contrast to straight models, they map nonlinear connections very well. They are versatile at tackling any sort of issue within reach.

Strategies like decision trees, random forests are being prominently utilized in a wide range of data science issues. Subsequently, for each analyst, it's critical to get familiar with these calculations and use them for modeling. A decision tree is a sort of regulated learning method that is generally utilized in characterization issues. It works for both categorical and consistent information yielding factors. In this strategy, the next step split the populace or test into at least two homogeneous sets (or sub-populaces) considering the most critical splitter/differentiator in input factors. Figure 4.8 shows the generated diagram after applying the decision tree model.

Figure 4.8: Decision Tree Model



The accuracy of the prediction of the decision tree that was computed after running it through the dataset using the Gini index and the same splitting method was around 62.2 percent. The Decision Tree helps to segregate students based on all principles. The most significant variable and its value are identified by the variables and determine the one that produces the best sets of a student decision tree. The Gini index says that if we randomly pick two objects from a population, they must be of the same class, and if the population is pure, the likelihood is 1 for this.

a. It operates with the Successor Loss categorical target variable.
b. It only conducts binary splits,
c. The higher the Gini value, the greater the homogeneity.

To have a reliable decision tree with high out-of-sample precision, the hyperparameters need to be carefully calibrated. Hyperparameters that were used and tuned in the decision tree classifier are as follows

i. Criterion - is one of the parameters. The function used to calculate the consistency of a split is this parameter and can be set as 'Gini' or 'entropy'. According to the

decision tree classifier that was used, the criterion of Gini was taken into consideration

ii. Gini – is a measure of how frequently, if it was randomly labeled according to the distribution of labels in the subset, a randomly selected variable from the set will be incorrectly labeled.

iii. Entropy - the measurement of the variability of the processed results. The higher the entropy, the more difficult it is to draw certain inferences from the data.

iv. max_depth: This determines what the scope of the decision tree should be. According to the decision that was used on this dataset, the maximum depth was up to 3.

v. max_leaf_nodes: It specifies what the highest leaf number in the decision tree should be.

vi. max_features: To get the best function from those randomly selected, the algorithm randomly selects these many features on each split.

vii. min_samples_leaf: The minimum number of samples that the leaf can produce is. In the decision tree classifier that was used on this dataset, the minimum number of sample leaves was tuned as 10.

viii. min_impurity_decrease: It is much easier to break the decrease in impurity so that the value of this parameter determines the least decrease in impurity that the algorithm can accept. The tree would not further break below this degree of impurity.

ix. class_weight: This makes the various groups different weights. If no value is given, then the weight of all groups is equal.

x. presort: If True value is supplied to this parameter, it accepts Ture or False value and orders the data before splitting.

4.4.2 Logistic regression

Logistic regression is a measurable model that in its essential structure utilizes a calculated capacity to show a paired ward variable, albeit a lot more mind-boggling augmentations exist. In regression analysis investigation, logistic regression is assessing the boundaries of a calculated model. Logistic regression gauges the connection between

the absolute ward variable and at least one free factor by assessing probabilities utilizing a calculated capacity, which is the combined conveyance capacity of strategic dispersion. Along these lines, it treats a similar arrangement of issues as logistic regression utilizing comparative methods, with the last utilizing an aggregate ordinary conveyance bend. Comparably, in the inactive variable translations of these two techniques, the first expects a standard calculated appropriation of mistakes and the second a standard ordinary circulation of blunders. Figure 4.9 depicts an example equation of logistic regression.

Figure 4.9: Logistic regression equation

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

The accuracy of the prediction of the multinomial Multinominal logistic regression that was computed after running it through the dataset using the same splitting method was around 71.1 percent. This is the multinomial linear regression because as the class column there are multiple values like arts, law, and more When the variable is continuous, the best regression analysis to be conducted is Multinominal logistic regression. Like all other regression analyses, Multinominal logistic regression is a statistical evaluation. Logistic regression is used to classify data and to explain the relationship between one dependent binary variable and one or more different nominal, ordinal, interval, or ratio-level variables. Y, from one or more response variables, X, Logistic regression (LR) is a statistical method relevant to linear regression, as LR finds an equation that specifies an outcome for a binary variable. However, unlike linear regression, the response variables may be categorical or continuous, because the model does not explicitly require continuous data. LR uses the ratio of log odds rather than percentages and an iterative method of cross-validation rather than a minimum square to fit the final model in predicting group membership.

4.4.2.1 Multinomial logistic regression

Multinomial logistic regression is a method of classification that generalizes logistic regression, i.e. with more than two possible isolated results, to multi-class concerns. That

is, it is a model that, given a set of independent variables, is used to predict the probabilities of the various possible outcomes of a categorically distributed dependent variable. The reason this is used because the target variable according to the dataset was although categorical, but it was of multiple classes which are multinomial logistic regression was used. Figure 4.10 depicts the results of the model

Figure 4.10: Logistic regression result

```
In [141]: accuracy_score(y_test,predictions)

Out[141]: 0.7111111111111111

In [142]: from sklearn.metrics import classification_report
          print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
1.0	0.70	0.95	0.81	20
3.0	0.33	0.50	0.40	2
4.0	0.00	0.00	0.00	1
5.0	0.50	0.67	0.57	3
6.0	0.91	0.67	0.77	15
7.0	0.00	0.00	0.00	4
accuracy			0.71	45
macro avg	0.41	0.46	0.42	45
weighted avg	0.66	0.71	0.67	45

5. FINDINGS

The prediction results were calculated and then it was concluded that the entire test data was predicted with an accuracy of 71.1 percent. The 91 students of high school that were of the test data consisted of most students recommending following a career in the "Engineering & Computer" field followed by "Business Administration". The remaining students were almost equally divided in "Medical & Health" and "Psychology" while a single student was recommended for the "Arts" section. Figure 5.1 shows the predicted results in the visual format of the number of students that are assigned to different majors according to their profiles and predictions using the multinominal logistic regression.

Figure 5.1: The number of students predicted for each university major

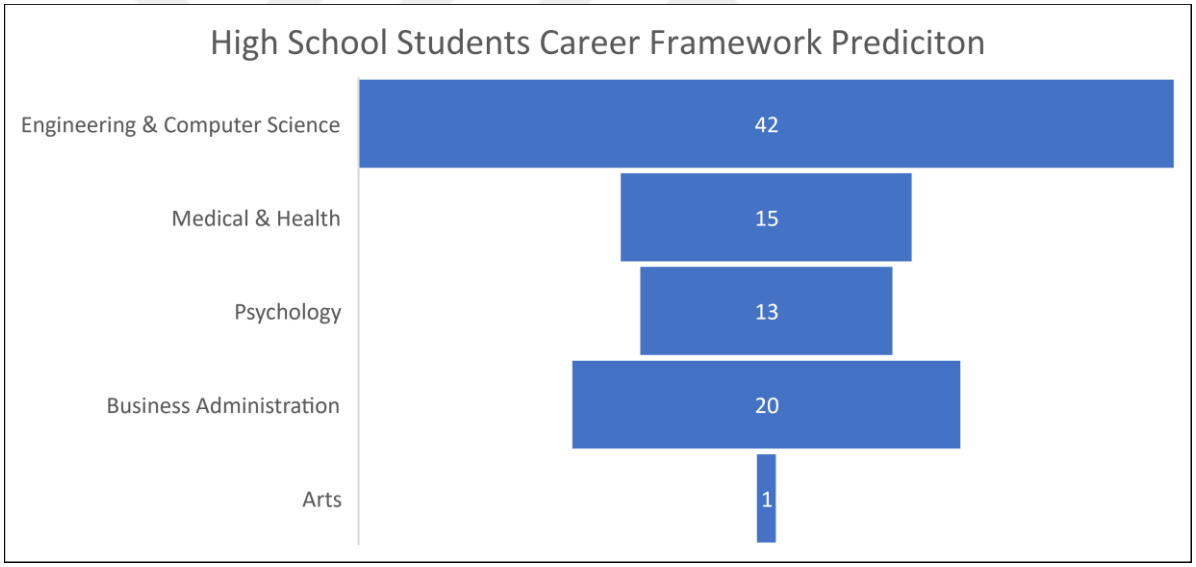
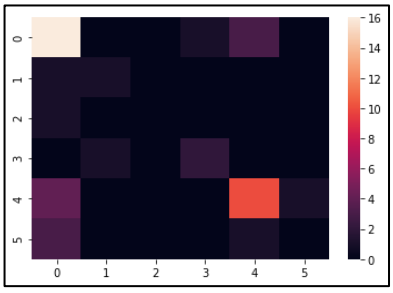


Figure 5.2 indicates the heatmap confusion matrix that gives visual information about the relationship between the predictions and the real values.

Figure 5.2: Confusion matrix heatmap



The information is prepared and tried with every one of the three algorithms and out of all, logistic regression gave more accuracy with 71.1 percent and followed the decision trees with 62.2 percent precision. As logistic regression gave the most noteworthy accuracy, all further information forecasts are picked to be followed with logistic regression. The foundation calculation being utilized is multinominal logistic regression and the new expectation is to continue adding to the dataset for additional more precision. According to Table 5.1, If a comparison is made of the results for both the algorithms, it depicts that the mean absolute error and the root mean square error have drastically decreased when looking at the accuracies.

Table 5.1: Values calculated of algorithms

Models	Decision Tree	Logistic regression
Accuracy Percentage	62.2 percent	71.1 percent
Precision	1.0 - 0.59 6.0 – 0.70	1.0 – 0.70 6.0 – 0.91
Recall	1.0 – 0.94 6.0 – 0.44	1.0 – 0.95 6.0 – 0.67
F1-score	1.0 – 0.72 6.0 – 0.54	1.0 – 0.81 6.0 – 0.77

The main thing that is to be noted was that the scores were high of 1.0 and 6.0 mainly because it's dependent on the data that was collected. Figure 5.3 indicates the evaluation metric of the classification models. Considering both the algorithms, it was a clear choice to go with the logistic regression algorithm as it showed an accuracy of above 70 percent.

These results were obtained after running the data through many procedures that involved different feature selections, different data splitting methods, and after extensively working on the data. In the end, the best possible results were recorded and used.

The ranges for the majority part of the datasets were between 1-5, whereas the university department had a range between 1-7.

Hence, the main reason behind Table 5.1 indicates the comparison between the evaluation metrics between the algorithms in which the records selected is 1.0 and 6.0 display a higher value than the rest due to its high density in the dataset than the rest of the values.

Figure 5.3: Evaluation metrics of algorithms

	precision	recall	f1-score	support
1.0	0.59	0.94	0.72	53
2.0	0.00	0.00	0.00	2
3.0	0.00	0.00	0.00	9
5.0	0.00	0.00	0.00	6
6.0	0.70	0.44	0.54	32
7.0	0.00	0.00	0.00	3
accuracy			0.61	105

Decision Tree Classifier

	precision	recall	f1-score	support
1.0	0.70	0.95	0.81	20
3.0	0.33	0.50	0.40	2
4.0	0.00	0.00	0.00	1
5.0	0.50	0.67	0.57	3
6.0	0.91	0.67	0.77	15
7.0	0.00	0.00	0.00	4
accuracy			0.71	45

Logistic Regression

6. DISCUSSION & CONCLUSION

Moving forward, the future study that can be done in this field is to break down the domains and help the students choose a more specific field i.e. choosing electrical engineering in the engineering department. This rundown a high extension for the understudies to choose for the more promising time to come with the explicit and exact investigation. As proficiency, exactness, and viability assume the imperative part during the time spent training framework, utilization of the Logistic relapse method gives us an ideal answer for this present reality understudy's instruction. In this proposition, it is necessary to have utilized the methodology of multinominal logistic regression to anticipate the professional choice for the passing out, understudies. The utilization of multinominal logistic regression has helped the understudies to accept the right fitting choice according to their advantage and abilities. The last objective is to give superior knowledge to plan a superior Education framework for understudies with a powerful result. This survey may reach out to bigger highlights to tackle complex choice information bases productively.

In the current scenario, the world is suffering from a pandemic known as COVID-19 which made access to student's profiles almost impossible because of strict lockdown and measures. This is one of the main reasons behind the accuracy level being on the lower side as it was related to the data and due to the exceptions, these days, it was the best data that was collected. If it were normal circumstances, then hopefully the accuracy level would have been much higher.

REFERENCES

Periodicals

A.O. Elfaki, K. A. (2015). Supporting students' learning-pathway choices by providing rule-based recommendation system,. *International Journal of Education and Information Technologies* .

AbuKhousa, E. &. (2014). Y. A Learning analytics approach to career readiness development in higher rducation. Tallinn, Estonia: *The 13th International Conference on Web-based Learning (ICWL 2014),.*

Adekola, B. ((2011)). Career Planning and Career Management as Correlates for Career Development and Job Satisfaction: *A Case Study of Nigerian Bank Employees. Australian Journal of Business and Management Research, 2.*

Ahmad Slim, G. L. (2014). Predicting student success based on prior performance. 2014 *IEEE Symposium on Computational Intelligence and Data Mining (CIDM).*

Albion, M. F. (n.d.). Factors influencing career decision making in adolescents and adults. *. J. Career Assess. 10(1), 91–126* .

Ali Daud, N. R. (2017). “Predicting Student Performance using Advanced Learning Analytics”,. *International World Wide Web Conference Committee (IW3C2).*

Alpaslan SAHIN, N. E. (2012). The Effects of High School Course Taking and SAT Scores on College Major Selection. *Sakarya University Journal of Education,.*

Arcidiacono. (2012). Ability sorting and returns to college major. *Journal of Econometrics.*

Ashish., D. a. (2015). "Clustering algorithms applied in educational data mining." . *International Journal of Information and Electronics Engineering, Volume 5, Issues.2, pp:112, March 2015.*

Athanasios S. Drigas, P. L. (2014). The Use of Big Data in Education . IJCSI *International Journal of Computer Science Issues, .*

Burns, R. P. (2017). Business research methods and statistics using SPSS. *Thousand Oaks, . CA: SAGE Publications Limited.*

Bydzovska, H. (2016). Course enrollment recommender system, in: T. Barnes, M.F. Min Chi (Eds.), . *Proceedings of the 9th International Conference on Educational Data*

Mining.

D, E. M. (2014). Comparative analysis of factors influencing career choices among senior secondary school students in rivers state Nigeria Arab . *J. Bus Manag. Rev. 4(4) pp 20–5.*

Dick, T. P. (2014). Factors and influences on high school students' career choices. *Journal for Research in Mathematics Education, 22(4), 281-292. .*

Eric Lichtenberger, C. G.-J. (2013). Predicting High School Students' Interest in Majoring in a STEM Field: *Insight into High School Students' Postsecondary Plans. Illinois Education Research Council.*

Ferry, N. M. (n.d.). Factors influencing career choices of adolescents and young adults in rural Pennsylvania. *Journal of Extension, 44(3), 1-6.*

Ferry, N. M. (n.d.). Factors influencing career choices of adolescents and young adults in rural Pennsylvania. . *Journal of Extension.*

G. Bordea, A. S. (2015). A review on predicting students performance using data mining techniques, *Procedia Computer Science.*

G. Gray, C. M. (2014). An application of classification models to predict learner progression in tertiary education, Advance Computing Conference . *IEEE International,.*

Gorad N, Z. I. (2017). Career counselling using data mining. *Int. J. Innov. Res. Comput. Commun. Eng 5(4) pp 0–5.*

Goyal, M. a. (2012). "Applications of data mining in higher education.". International journal of computer science, *Volume 9, Issues 2, pp: 113, March 2012.*

Gupta, A. &. (2014). Applying data mining techniques in job recommender system for considering candidate job preferences. In Advances in Computing, Communications and Informatics *(ICACCI, . International Conference on (pp. 1458-1465.*

K. Sripath Roy, ,. K. (2018). Student Career Prediction Using Advanced Machine Learning Techniques. *International Journal of Engineering & Technology.*

M C B Natividad, B. D. (2019). A fuzzy-based career recommender system for senior high school students in K to 12 education. *IOP Publishing.*

M, S. A. (2012). Factors influencing students' career choice and aspirations in South Africa . *J. Soc. Sci. 33(2) pp 169–78.*

M. Mayilvaganan, D. K. (2014). Comparison of classification techniques for predicting the performance of students academic environment, . *2014 International Conference on Communication and Network Technologies (ICCNT), 2014,.*

Maharani K, A. T. (2015). Comparison analysis of data mining methodology and student performance improvement influence factors in small data set . *Conf.Proc. Int. Conf. Sci. Inf. Technol. 15 pp 169–74.*

Malgwi, C. H. (n.d.). Influences on Students' Choice of College Major. *Journal of Education for Business, 80(5), 275-282.*

Mervat Adib Bamiah, S. N. (2018). BIG DATA TECHNOLOGY IN EDUCATION: ADVANTAGES, IMPLEMENTATIONS, AND CHALLENGES. *Journal of Engineering Science and Technology .*

Min Nie, L. Y. (2016). Forecasting Career Choice for College Students Based on Campus Big Data.

N. Jeeva, E. G. (2014). Application of data mining in educational database for predicting behavioural patterns of the students, . *International Journal of Computer Science and Information Technologies.*

P. Strecht, L. C.-M. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance . , *International Educational Data Mining Society,.*

P.M. Arsad, N. B. (2013). A neural network students' performance prediction model (NNSPPM), 2013 . *IEEE International Conference on Smart Instrumentation, Measurement and Applications.*

P.Veeramuthu. (2014). "Analysis of Student Result Using Clustering Techniques" . *International Journal of Computer Science and Information Technologies, Volume 5, Issues 4, pp: 5092- 5094, 2014.*

R, S. V. (2016). Analysis of students' performance evaluation using classification techniques . *Int. Conf. Comp. Tech. & Intell. Data Engi. p 16.*

Rajalakshmi Krishnamurth, M. G. (2018). Automatic Detection of Career Recommendation Using Fuzzy Approach. Noida: *Journal of Information Technology Research.*

Rao, K. P. (2016). "Predicting Learning Behavior of Students using Classification Techniques." . *International Journal of Computer Applications, Volume 139, Issues*

7, pp: 0975 – 8887, April 2016.

Ray, S. (2013). BIG DATA IN EDUCATION. GRAVITY Issue 20.

Razak T R, H. M. (2014). Career Path Recommendation System for UiTM Perlis Students Using Fuzzy Logic . *14 10.1109/ICIAS.2014.6869553.*

S, O. S. (2013). The factors determining the choice of career among secondary. *Int. J. Engineeering Sci. 2(6) pp 33–44.*

Saa, A. (2016). Educational data mining & students' performance prediction. (IJACSA). . *International Journal of Advanced Computer Science and Applications 7(5), 212–220 (2016).*

Shahiri A M, H. W. (2017). A proposed framework on hybrid feature selection techniques for handling high dimensional educational data . *AIP Conf. Proc. 1891020130.*

Simpson, J. C. (n.d.). Segregated differences by subject: Racial differences in the factors influencing academic major between European Americans, Asian Americans, and African, Hispanic, and Native Americans. *Journal of Higher Education, 72(1), 63-100. .*

T, A. C. (2016). Feature selection techniques to analyse student academic performance using Naïve Bayes student acadamic performance using Naïve. *Conf. Proc. Small Mediu Bus 2016 January 19 pp 345–50.*

T. Mishra, D. K. (2014). Mining students' data for prediction performance, . *2014 Fourth International Conference on Advanced Computing Communication Technologies.*

T.M. Christian, M. A. (2014). Exploration of classification using NBTree for predicting students' performance. *International Conference on Data and Software Engineering* .

V. Labatut, H. C. (2011). Accuracy measures for the comparison of classifiers, in: A.- D. Ali (Ed.),. *The 5th International Conference on Information Technology, AlZaytoonah University of Jordan, amman, Jordan, 2011, p. 1,5. .*

Verma, P. S. (2017). Student career path recommendation in engineering stream based on three dimensional model. Computer Applications in Engineering Education,. *Journal of Information Technology* .

Yadav, S. B. (n.d.). Data mining applications: a comparative study for predicting student's performance. *Int. J. Innov. Technol. Creative Eng.*

Yadav, S. K. (2012). Data mining applications: A comparative study for predicting

student's performance. Shahiri A M, H. W. (2017). A proposed framework on hybrid feature selection techniques for handling high dimensional educational data . *AIP Conf. Proc. 1891020130.*

Other Publications

A.-S. Hoffait, M. S. (2017). Early detection of university students with potential difficulties, Decision Support Systems .

Ankhtuya Ochirbat, T. K.-H.-H. (2017). Hybrid Occupation Recommendation for Adolescents on Interest, Profile, and Behavior.

Asanov, D. (2011). Algorithms and methods in recommender systems. . Berlin: Berlin Institute of Technology.

Atif, E. A. (2014). Big Learning Data Analytics Support for Engineering Career Readiness . Al-Ain: UAE University.

Baradwaj, B. P. (2011). Mining educational data to analyze students' performance. *(IJACSA) . Int. J. Adv. Comput. Sci. Appl. 2(6), 2011 (2011).*

Beggs, J. B. (2013). Distinguishing the Factors Influencing College Students' Choices of Major. College Student. *Journal, 42(2), 381-394.*

Bharat Patel, V. K. (2017). CaPaR: A Career Path Recommendation Framework. San Jose, CA: *IEEE Third International Conference on Big Data Computing Service and Applications.*

Bimrose, J. B. (2014). A systematic literature review of research into career-related interventions for higher education.

Borchert, M. (2012). CAREER CHOICE FACTORS OF HIGH SCHOOL STUDENTS . University of Wisconsin-Stou.

Borchert, M. (2010). Career choice factors of high school students. Unpublished Master of Science thesis.

Breiman, L. (2010). Random forests, Bagging predictors, Machine Learning

C. Márquez-Vera, A. C. (2016). Early dropout prediction using data mining: a case study with high school students, *Expert Systems* .

Chen, L. B. (2013). Matching skills of individuals and firms along the career path. Career Planning Beginning in Middle School. Educational Studies. 2013.

DAVID A. WALDMAN, T. K. (2004). Student Assessment Center Performance in the Prediction of Early Career Success.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management, Decision Sport Systems.

Dietz-Uhler, B. H. (2013). Using learning analytics to predict (and improve) student success: a faculty perspective. J. Interact. Online Learn.

E.B. Costa, B. F. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses, Computers in Human Behavior.

Edmonds, J. (2012). FACTORS INFLUENCING CHOICE OF COLLEGE MAJOR: WHAT REALLY MAKES A DIFFERENCE? College of Liberal Arts and Sciences.

Ephrem, B. B.-S. (2013). H.: Projection of Students' Exam Marks using Predictive Data Analytics.

F. Araque, C. R. (2009). Factors influencing university drop out rates, Computers & Education .

H. Guruler, A. I. (2010). A new student performance analysing system using knowledge discovery in higher educational databases.

J. Dougherty, R. K. (2011). Supervised and unsupervised discretization of continuous features, in: Machine Learning Proceedings.

J. Miao, W. H. (2014). High school graduation rates:alternative methods and implications, Education Policy Analysis Archives.

Janger, J. &. ((2013).). Career choices in academia . (working paper).

M.N.V.D. Berg, W. H. (2005). Student success in university education: a multimeasurement study of the impact of student and faculty factors on study progress, Higher Education .

Osmanbegovi, E. S. (n.d.). Data mining approach for predicting student performance. .

Peter Arcidiacono, V. J. (2012). Modeling college major choices using elicited measures of expectations and counterfactuals. Elsevier.

Porter, M. F. (2014). An algorithm for suffix stripping. Program. Career path recommendation system for UiTM Perlis students using fuzzy logic. In.

R.O. Aluko, O. A. (2016). Predicting the academic success of architecture students by pre-enrolment requirement: using machine-learning techniques, Construction Economics and Building 16 .

Robert W. Lent, A. M.-B. (2018). Social cognitive career theory and the prediction of interests and choice goals in the computing disciplines. Elsevier.

S. Huang, N. F. (2013). Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models, Decision Support Systems.

S.B. Aher, L. L. (2013). Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data, Knowledge-Based Systems .

Song, C. &. (n.d.). College attendance and choice of college majors among Asian American students. Social Science Quarterly. Blackwell Publishing Limited.

Subhalaxmi Panda, P. A. (2017). A Higher Education Predictive Model Using Data Mining Techniques.

Sudha, K. A. ((2013).). Career Guidance Counseling Needs of Graduate Students- A Study in India. Global research analysis, .

Tajul Rosli Razak, M. A. (2014). Career Path Recommendation System for UiTM Perlis Students Using Fuzzy Logic .

V.L. Miguéisa, A. F. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. Elsevier.

Yazici, S. &. (2010)). Students' Choice of College Major and their Perceived Fairness of the Procedure: Evidence from Turkey. Educational Research and Evalutation.

Yu Lou, R. R. (n.d.). A Machine Learning Approach for Future Career Planning.

Zun Hlaing Moe, T. S. (2018). Evaluation for Teacher's Ability and Forecasting Student's Career Based on Big Data. Myanmar: Myanmar Institute of Information Technology.