Accuracy Of Data Analysis On University Student's Major Subject And Job Sector Prediction Using KNN And Decision Tree Algorithm

By

Istear Ahmed
19101662
Tushar Karmakar
19101663
IffatSumaitaAnadi
17101353
MD.Shahriar Azad
17101181
Faiza Ibnat Sharif
18301287

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

September, 2021

Computer Science and Engineering
BRAC University

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I/We have acknowledged all main sources of help.

Student's Full Name & Signature:

Istear Ahmed

Istear Ahmed
19101662

Tushar Karmakar

Tushar Karmakar
19101663

Iffat Sumaita Anadi

Iffat Sumaita Anadi
17101353

MD.Shahriar Azad

MD.Shahriar Azad
17101181

Faiza Ibnat Sharif

Faiza Ibnat Sharif
18301287

Approval

The thesis titled “Accuracy of Data Analysis on University student's major subject and Job sector prediction using KNN and Decision Tree Algorithm” submitted by

1. Istear Ahmed (19101662)
2. Tushar Karmakar (19101663)
3. Iffat Sumaita Anadi (17101353)
4. MD.Shahriar Azad (17101181)
5. Faiza Ibnat Sharif (18301287)

of Summer, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of on September 28, 2021.

Examining Committee:

Hossain Arif

Supervisor:
(Member)

Hossain Arif
Assistant Professor
Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md Golam Robiul Alam
Associate Professor
Computer Science and Engineering
BRAC University

Departmental Head:
(Chair)

Sadia Hamid Kazi
Dean School of Data and Sciences
BRAC University

Abstract

The students are at a loss when it comes to choosing their major or field of work due to inefficient school system and decision making of the parents. This problem can be resolved to some extent by using classifiers. The accuracy of the decision the students will take can be gauged these using classifiers. These classifiers such as KNeighbor classifier and Decision tree classifier help to gauge the accuracy of the prediction. Decision tree classifier is more accurate in predicting the correct outcome. This use of classifier will help the students to take the right decision in case of their study and career field. Based on a student's extra-curricular activity, Olympiad and HSC grade it predicts the major subject of an undergraduate student. Again, based on an undergraduate student's major subject, CGPA and extra-curricular activity it predicts job of the student.

Acknowledgement

Our thesis has been completed without any egregious interruption by the grace of Great Almighty Allah. We also want to praise our supervisor Hossain Arif (Assistant Professor, Computer Science and Engineering, BRAC University) sir for his compassionate support and discreet advice. Without his clemency it was quite tough for us to complete our thesis. We would like to give gratitude to our beloved parents for their corroboration. With their support and prayer, we are now on the verge of our graduation.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Bangladesh and the major part of Southeast Asia are still stuck on the past when it comes to selecting the field of education. The parents usually choose the course of study for their child based on some arbitrary areas without looking into what their children want or what they'll be able to excel in. For this archaic method of major or field of study selection, students lose interest and are unable able to reach their full potential. The use of data science can assist the students in making the right choices when it comes to major and career. Educational Data Mining (EDM) can play a pivotal role in providing the students with a sense of direction according to their best ability. Educational data mining and analytics will transform the whole educational system, helping the students choose correct career paths according to their ability. As a result, the academic performance and interest in study will increase. Educational data mining isn't limited to choosing subjects that the students will study at university, but it'll also help provide an accurate decision for which job field to select. Our system will help gauge the students' choices regarding career paths and major selection.

1.2 Thesis Statement

Educational data mining is used to find out the accuracy of field of study in university of HSC students using their GPA, Extracurricular and Olympiad performance. Using the same methodology, the Job sector selection accuracy of the students can be selected by taking CGPA, Extracurricular and major subject. KN-neighbor classifier is used to gauge the accuracy of the decision as well as decision tree classifier is used to gauge the accuracy of the decision. By further scrutiny it was understood that decision tree classifier is more accurate.

1.3 Problem statement

Choosing the right subject for under graduation is the most important part of in everyone’s life. The best time to choose it when we are in high school but it is not that much easy. A student who is studying in high school always suffers under dilemma that what is best for him / her. Our research can be revolutionary for them who always suffer to choose the right subject for their under graduation. Anyone can easily find which is the best subject for him / her. This research can give us a solution for choosing a right subject for the students who are studying in High School but it remains some challenges. The challenges are:

1. The students who do multiple co-curriculum activities and good at academic result, it is hard to give them a proper result. For example: If a student is good at math and good at arts subject also, he can manage people in his or her co-curriculum activities then it may not work on the data of that student.

2. We will collect a huge amount of student’s data. So, protecting the data is another challenging part of this research.

3. This research is not that much common in our country and rely on us is not easy. For example: After analyzing a student we can get a result that he / she should be a computer engineer but the student like to be an architect. So, he / she is not relying on our research’s result and it might not helpful for this kind of student. It might be challenging for us to bring trust in the students.

4. In our country sports or other co-curricular activities is not taken seriously. Example- a student is good in sports but for some reason he/she cannot go furthermore in that activity after completing high school studies. So, this is also a concern.

5. Sometimes having a choice of desired subject most of the student relay on their parent’s decision of selecting the subject for higher studies.

1.4 Motivation

In this part of the world, parents still dictate the life choices for students when it comes down to the field of study and job. Students are genuinely in a dilemma on what to study and in which area they will excel. The same can be said for university students as they are confused about which job sector to go in. Our work is done to provide a base solution to these dilemmas. Helping the students to get to a concrete decision regarding their career using data science is the primary motive of this paper. This paper will act as the stepping stone for further research on educational data mining. We are testing the accuracy of the algorithms, so it is possible to work on this topic at further length in the future.

1.5 Contribution

Educational Data Mining is an emerging discipline. Being relatively new much work hasn't been done with this. The uniqueness of this paper is that it is exploring the novel topic of educational data mining in the context of Bangladesh. No other paper has explored the possibility of this in the context of Bangladesh and how it can impact the scenario of the educational landscape. This paper gauges the accuracy of major selection by HSC students as well as guides in job selection as well. Using data mining for job selection is also a new concept in our country. The proper selection of field of study will be a great motivating factor for the student as they'll be studying the thing which interests them. The same can be said for job selection. This will pioneer the way for the education system to move forward as everyone will be doing what is best to their ability. Educational data mining will be able to make our education system more polished and scientific, in a way changing the whole educational landscape for the greater good.

Chapter 2

Related Work

In the recent decade, numerous studies in the subject of educational data mining have been conducted. Various data sets gathered from educational institutions were classified by Dimi´c, Prokin, and K. Kuk using various classification algorithms.[1] used the Naive Bayes classifier and decision tree to predict student achievement based on previous data from the Moodle course regarding the first colloquium. The J48 decision tree classification method outperformed the Naive Bayes method on the investigated data set, according to the researchers.

Educational data mining entails data mining approaches that can be used in a variety of settings to find useful patterns in existing data. In the educational setting, the most common application of classification algorithms is the prediction of student performance on the final test. Dimic, Prokin, and K. Kuk investigated how two classifiers were used to create model predictions (Decision Tree and Naive Bayes). We utilized the GainRatio and InfoGain filters to estimate the appropriate properties. The approach of cross validation was used to estimate the classifiers.For elimination an unbalanced distribution values of class variables we were used Resample function and generated new prediction models.[1]

Dervisevic, Zunic, ,Eonko, & Buza, [2] analyzed data from a real-world production system called Edu720, which is designed for internal employee education and is used by a number of enterprises in Bosnia and Herzegovina and the surrounding region. The considered data set underwent a sophisticated data pretreatment process that included data cleaning and data transformation so that it could be used in a variety of classification tasks. The main objective of this study is to predict the performance of the education program that the company wants to implement for its employees. The usage of decision trees and KNN classifiers is examined in this paper [2]. The number of successfully categorized samples was used to measure the classifier's quality. There is also a comparison of applying classification methods to imbalanced and balanced data sets.

The number of class attribute occurrences in the dataset determines the dataset's balance. The use of classification methods has yielded better results when used with a balanced data set, rather than the one that is not. In order to generate a balanced data set, a SMOTE method was applied. The analysis of the data set partitioning method showed that data selection in the training data set and test data set affected the accuracy of the classification. Partitioning a data set that is based on data clustering has yielded much better results than partitioning data by random selection. Comparing the accuracy of the classifiers, it can be concluded that the KNN algorithm gave more accurate results in all cases than the decision tree.

Yadav and pal applied [3] different decision tree algorithms such as C4.5, ID3 and CART for predictions of engineering student performance in the final exam. Data was collected from VBS Purvanchal University. The number of pupils who are likely to fail, pass, or be promoted to the next year was a final result of those classification techniques. From prior year's student data, machine learning methods such as the C4.5 decision tree algorithm can create successful predictive models. The empirical results reveal that by applying predictive models to the records of entering new students, they may provide a brief but accurate prediction list for the student. C4.5 has a higher accuracy in predicting a student's final grade, with 67.77 percent of test instances correctly identified.The comparative analysis of the results stated by Jadav and pal shows that that the prediction has helped the weaker students to improve and brought out betterment in the result.

Ramalingam & Ilakkiya presented [4] a predictive analysis of the academic performance of the students in Kongu Engineering College based on the academic performance of the passed-out students. The performance of the students both in academic and in the placement is degrading every year which needs to be analyzed in order to improve the performance of the students for the upcoming batches. The placement performance of the students purely depends on academic performance and also on co-curricular performance. So, both were taken for consideration.

Students' participation in co-curricular and extra-curricular activities has an impact on their placement performance. For the forecast, data relating to academic outcomes, co-curricular activities, extracurricular engagement details, and other personal details must be evaluated. The statistics were gathered from a variety of sources, and the core 36 criteria were identified as having a significant impact on placement performance. The faculty assisted in manually filling in the missing details. Chi-square method is used for selecting the features that matches both the placed and non-placed students. Classification models were created to predict the placement performance of the students at the end of the last semester for each dataset.

Ramalingam & Ilakkiya applied [4] Machine Learning techniques such as Knn, decision tree, and others to predict student success in both academics and placement, as well as visualizing the findings.Predictive analysis aids in creating future predictions based on the data information accessible in the past. In the realm of education, data mining refers to the analysis of data using advanced technologies such as machine learning and statistics.It's used to extract meaning from massive data sets created by or connected to people's learning.

Machine learning gives you the deep perspectives of algorithms and statistical models which is used by the computer system to do a task effectively without using the pattern match. Data Mining and its analytics [4] helps in performing inspections, cleansing of data, transforming it into another form and also in making decisions based on the visualization support available in the tools currently in the market. Data visualization is the also another method which supports the data mining process to have a clear and understandable view of a particular process by extracting information from the data available from the past using various charts in the form of the graphs, charts and through various graphical representation.

This paper [5] k Nearest Neighbor (KNN) is widely used in document classification for dealing with the much more difficult problem such as large-scale or many of categories. It is a simple, effective and nonparametric classification method But KNN classifier may have a problem when training samples are uneven. The problem is that KNN classifier may decrease the precision of classification because of the uneven density of training data. To solve the problem, Lijuan, Linshuang, Xuebin & Qian Shi provided CLKNN, an improved KNN classification algorithm that uses clustering to improve training samples, and it is applied to the circumstance of uneven training sample distribution and a more evident border.

In comparison to classic classification algorithms, the KNN classification algorithm with divided clustering has a faster execution time and a higher accuracy rate. The experimental findings have demonstrated the benefits of the novel method in dealing with the uneven distribution of test samples., In Lijuan, Linshuang, Xuebin & Qian Shi’s experiment they have used only two types of data sets, they should take more data sets to experiment in future, how to choose the minimum number of samples and the values of the neighborhood radius also need to be studied further.

One of the most critical aspects of Customer Relationship Management is predicting churn (CRM). Researchers in a variety of industries, including business intelligence, marketing, and information technology, were inspired to examine the best strategies for delivering the finest services to clients in order to retain customers and preserve their pleasure. Many machine learning algorithms have been built with the goal of identifying potential churning customers and making the best selections at the correct time.

Hassonah, Rodan, Al-Tamimi, & Alsakran [6] conducted a comparison study of the performance towards churn prediction between two of the most powerful machine learning algorithms which are Decision Tree and K-Nearest Neighbor algorithms. Results were quite interesting showing a quite large dissimilarity in many areas between the two algorithms.

Statistical results were found to be in favor of the decision tree algorithm; accuracy, precision, recall, F-measure and the Lift measure were found to be better in DT than K-NN algorithm. The AUC was found to have nearly the same results for both algorithms which is caused by the low values of specificity for DT leading to the similarity of AUC values for both algorithms. However, results shown in the confusion matrix indicate that K-NN had more true positive rates than the DT algorithm telling us that the K-NN predicted better than the DT regarding real churning customers.[6]

Many service businesses employ data mining tools to help them make decisions. Educational institutions gradually began to employ business intelligence approaches to assess their current progress [7]. While using data mining techniques on academic data, educationalists will become aware of a variety of aspects that have an impact on academia. We were able to detect numerous patterns using data mining approaches, which helped institutions make strategic decisions to improve students' academic performance. Potential graduate students will have a dilemma on identifying the universities for their post graduate admissions and on the other hand an average graduate student would be uncertain on getting post graduate admission in a reputed university based on their academic scores.

Jeganathan, Parthasarathy, Lakshminarayanan, Ashok, P. M & Khan applied the classification techniques such as Logistic Regression, KNN Classification, Support Vector Classification, Naive Bayes Classification, Decision Tree Classification and Random Forest Classification on the given academic admission dataset. By comparing the accuracy and mean absolute error of each model, the Logistic Regression classifier outperformed others with an accuracy of 99% [7].

Academic institutions spend a lot of time and money each year trying to influence, forecast, and understand the decision-making choices of applicants who have been accepted. To predict student college commitment decisions, Basu K [8] et al used numerous supervised machine learning approaches on four years of data from 11,001 students admitted to a small liberal arts institution in California, each with 35 related attributes. By treating the question of whether a student offered admission will accept it as a binary classification problem, they implemented a number of different classifiers and then evaluated the performance of these algorithms using the metrics of accuracy, precision, recall, F-measure and area under the receiver operator curve.

Machine learning approaches were used by Basu K et al to estimate whether students would accept the admission offer [8]. This assists universities in identifying possible applicants who will accept the offer and those who will decline it. The author used many supervised learning approaches to find a Logistic Regression classifier with a high level of accuracy. By refining the parameters using the grid search method, the performance of the Logistic Regression classifier is increased from 77% to 78%. The significance of this research is that it demonstrates that many institutions could use machine learning algorithms to improve the accuracy of their estimates of entering class sizes, thus allowing more optimal allocation of resources and better control over net tuition revenue.

To admit a competent student, Surya Rebbapragada et al. [9] use a two-step approach. First, a model was developed utilizing data mining techniques to predict an applicant's quality based on academic performance data and application criteria. Second, revenue management tactics are employed to create an application's going-rate table.The output of the datamining technique is compared to the going-rate table, and a judgment on admission is taken.

The healthcare industry is dealing with billions of patients all over the world and producing massive data. The machine learning-based models are dissecting the multidimensional medical datasets and generating better insights. R. Jane Preetha Princy et al, [10] classified a cardiovascular dataset by using several state-of-the-art S upervised Machine Learning algorithms that are precisely used for disease prediction.

R. Jane Preetha Princy et al, analyzed the accuracy of supervised algorithms by using a cardiac prediction dataset [10]. By applying the dimensionality reduction technique Jane found that Decision tree provided the best performance with accuracy level of 73%.

Because dataset dimension influences algorithm performance, reducing dataset dimension reduces the capacity of Random Forest and KNN algorithms. The results show that the size of the dataset has a positive or negative impact on the algorithms. The High Correlation Filter and Principal Component Analysis will be used to reduce dimensionality at the next level. The CVD dataset will be used to test and create a better illness prediction model utilizing ensemble machine learning algorithms.

Due to the increasing volume of students applying for higher education, there may be various issues with college entrance and enrollment in big data universities. As a result, effective data mining algorithms are required for better data classification decision-making by students. Abdul Hamid M. Ragab et al study several difficulties related to admission and enrollment by connecting a large number of individuals applying for higher education [11]. They experimented with nine different classification algorithms in order to determine the best technique for classifying students' datasets. WEKA, a data mining tool, was utilized to classify the students who were qualified for admission. Student data from many areas such as medicine, engineering, computing, and other courses makes up the source data. The WEKA tool was used to examine six algorithms, and C4.5 was found to be the best performing algorithm for the dataset in terms of performance, accuracy, and lowest mistakes.

Ahmed, Tahid, Mitu, Kundu & Yeasmin, [12] explains how to use several data mining approaches to create a predictive model that can effectively examine and forecast student academic performance. This study also examines how well different classification algorithms perform on feature selection strategies and determines which of the various measure values used to student academic performance data sets yields the best accuracy result. On 800 student records obtained from the CSE, department of North Western University, Khulna, K-Nearest Neighbor, Nave Bayes, Bagging, Random Forest, and J48 classification algorithms were used to evolutionary algorithms, gain ratio, relief, and information gain feature selection methods.

The experimental outcome demonstrated that the qualitative model based on student performance greatly relies on the collection of the most related attributes of the record of the attribute used in the student dataset. Genetic algorithms with K-Nearest Neighbor classifier demonstrated the best accuracy measure compare to other methods.

Using Weka, Mythili M S and Shanavas A R devised a method based on classification algorithms for analyzing and estimating school student performance.On the academic data acquired from the student management system, they applied multiple categorization algorithms: Random Forest, J48, Multilayer perception, IBI, and decision table [13].

It is a primary concern to observe the student’s academic performance for high learning. . Imdad, Ahmad, Asif, & Ishtiaq proposed [14] a method for student’s results classification centered on the clustering of data and agreeing on their performance level. It is based on two traditional evolutionary algorithms. This system evaluates KNN and ANN using given data set for a student result classification by using multilayer perceptron. Data from educational boards of Pakistan is being used for the evaluation. The proposed system will be useful for academic planners for making an effective resolution.

This work can be extended by performing many artificial intelligences Algorithms like K-mean clustering, decision trees, and Naïve Bayes, etc., may also be applied to get a different result. Diverse datasets can be used with various methods for computing correctness and error rates. In future, may report hybrid classification models using KNN and ANN with other classification and AI techniques. Simulations can be done using different tools other than WEKA like MATLAB etc. to get better and more accurate results. Fuzzy logic is another important method based on content to forecast result. A fuzzy logic approach to the same problem can bring some new understandings into the problem. Evaluate data by using a different filter in WEKA.

Previous scholars always made an effort to make various formulations that were used to categorize and calcify hadith. At present, the process of categorization or classification is facilitated by the process of text mining technology. There are a variety of tools and approaches or algorithms that may be utilized in the study of text mining to help deliver the best outcomes in the process of extracting information from a text. The Decision Tree C4.5 and K-Nearest Neighbor algorithm are two examples. Based on this, [15] Awaludin, Gerhana, Maylawati, Darmalaksana, Arianti, Rahman & Musli wish to conduct research and complete this final project in order to compare the results of the text document classification process utilizing Decision Tree C4.5 and the K-Nearest Neighbor method for the classification of Imam At-Tirmidzi hadith. [15]

With Awaludin, Gerhana, Maylawati, Darmalaksana, Arianti, Rahman & Musli's research,[15] it is expected to be knowledgeable about the process of classifying text documents along with the performance of the two algorithms. Based on testing that has been done, the Decision Tree C4.5 algorithm produces an average accuracy value of 70.53% with an average processing time of 0.083 seconds. While the K-Nearest Neighbor algorithm produces an average accuracy value of 66.36% with an average processing time of 0.03 seconds.

Nicholas T. Young et al. employed machine learning to determine the percentage of admission granted in the physics department's doctoral program at a mid-western public research university in the United States [16]. Data on candidates' undergraduate GPAs and institutions, research interests, and GRE scores were acquired from a large, Midwestern public research university with a decentralized admissions procedure. We employed supervised machine learning methods to develop models that predicted who would be admitted to the PhD program because the acquired data was of varying scale. Nicholas found that using only the applicant's undergraduate GPA and physics GRE score, and he was able to predict with 75% accuracy who will be admitted to the program.

Chapter 3

Working Procedure

3.1 Methodology

Accuracy in research is a study attribute that indicates how closely sample parameters correspond to population characteristics. As a result, accuracy refers to how closely the measured value or findings correspond to the genuine or original values. One of the components of data quality is data accuracy. Accurate data allows for better decision-making. All who rely on the data benefit from the best data quality. Users will be able to generate better outputs if the data quality is excellent. This improves the efficiency of the firm and reduces the chance of negative results. It is used to determine whether the data values stored for an object are correct. A data value must be the correct value and must be represented in a consistent and unambiguous manner to be correct. We use KNN and Decision Tree Algorithm to predict the actual and accurate data.

3.2 Data collection procedure

This research requires a large amount of dataset of high school students and university students regarding their grades, extra-curriculum activity and university CGPA. As many students and educational institutes often do not want to disclose their data and these datasets also are not available in data-science related online communities like – Kaggle, so it is bit challenging to collect this vast amount of data for this research. So, the dataset which is needed for this research has to be collected through different sources like -Google forms and in person survey. The dataset required for this study needs to have the following attributes –

1. Extra curriculum activity.
2. Participation in International Olympiad.
3. Participation in National Olympiad
4. High school subject grades.
5. University department.
6. University CGPA.

3.3 Data cleaning procedure

The dataset which has been collected from different sources is not always ready for machine learning because all these collected data are in categorical value which is not friendly for machine learning. Moreover, there is always remain some extra and unrelated data that is not of no use for work. At first the remaining extra data which is not work related have to be cleaned properly. After that, in order to make these dataset machine learning friendly, it has to be encoded from categorical value to numerical value.

3.4 Working procedure of K-Nearest Neighbor (KNN)

K-Nearest Neighbor also known as KNN classifier. It is a supervised machine learning algorithm. A supervised machine learning algorithm uses labelled input data to develop a function after then when an unlabeled data given that gives a suitable output based on that function. KNN may be used to predict outcomes in both regression and classification. The KNN classifier method saves all existing data and classifies newer data sets based on their similarity.

In the first step for implementing KNN classifier a data set is needed. The data set is to be divided into two part – training data and test data. After that the nearest data point have to be chosen which is called K. It is very important to choose K perfectly. The value of K cannot be very bigger or very smaller.

In the next step the distance between the data points whose class will have to be predicted and all the training data points is then measured. Hamming, Manhattan, The Euclidean distance can be used in this case. The distance collection must be sorted in ascending order after it has been calculated.

The final process includes selecting K entries from the sorted collection. The labels of the chosen K entries then must be fetched. The mean of the K labels must be returned in the case of regression, and the mode of the K labels must be returned in the case of classification.

The accuracy of the KNN depends on the quality of data [16].

3.5 Working procedure of Decision Tree

The Decision tree algorithm is also a supervised machine learning algorithm. The objective of this algorithm is to construct a model that predicts the value of a target variable, and it solves the problem by using the tree representation, where the leaf node relates to a class label and properties which are represented on the inner node of the tree. Like KNN, it is also used for both classification and regression problem. In the decision tree algorithm for classification problem classification trees are used which are the tree models in which the target variable can take a discrete set of values and for regression problem regression tree models are used in which the target variable can take continuous values like real numbers. In statistics, data mining, and machine learning, a decision tree is one of the most used predictive modeling approaches.

In order to split the record, at first the best attribute in the dataset have to be chosen by attribute selection measure (ASM). Then a decision node has to be made by that attribute, and the dataset must be divided into small subsets. After then, the tree building will be started by redoing the process recursively for each child node from the root until all of the tuples correspond to the same attribute value or there are no more attributes.

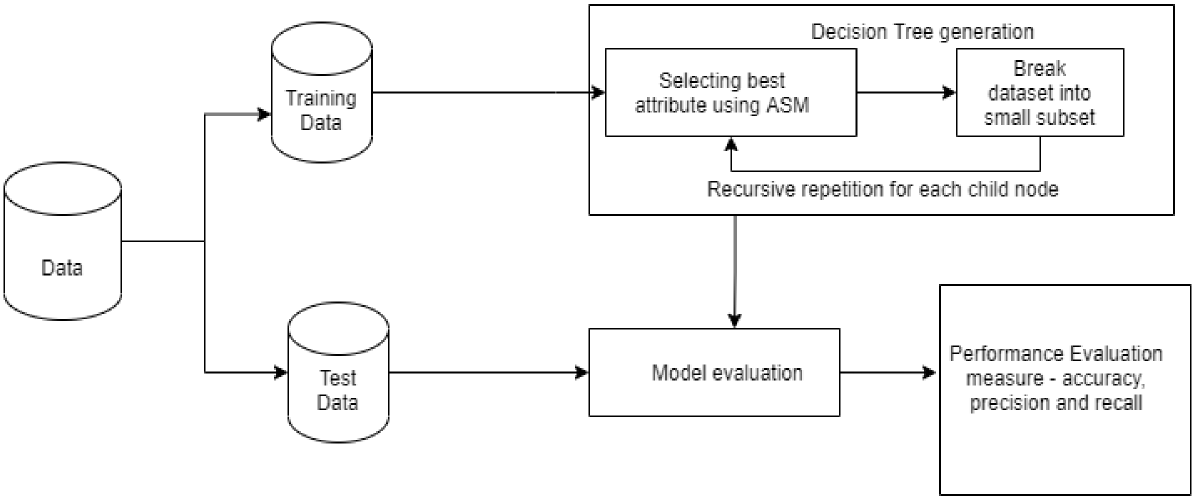

Data
Training Data
Test Data
Decision Tree generation
Selecting best attribute using ASM
Break dataset into small subset
Recursive repetition for each child node
Model evaluation
Performance Evaluation measure - accuracy, precision and recall

Figure 3.1: Decision Tree Algorithm Performing.

3.6 k-fold Cross-Validation procedure

k-fold Cross-Validation is widely used in machine learning to make comparisons and select a model for a specific predictive modeling problem because it is simple to understand, simple to execute, and outcomes in skill estimates that typically have a lower bias than other processes..The process of dividing the data set for training and testing is called folding. K -Fold basically means K ratio. If the data devided into three parts, One part will be used in testing and the other two parts in training. Then this folding will be called 3- fold cross validation.In this reasearch study K-fold Cross-Validation prpcedure has been used for getting better result.

Chapter 4

Workflow diagram

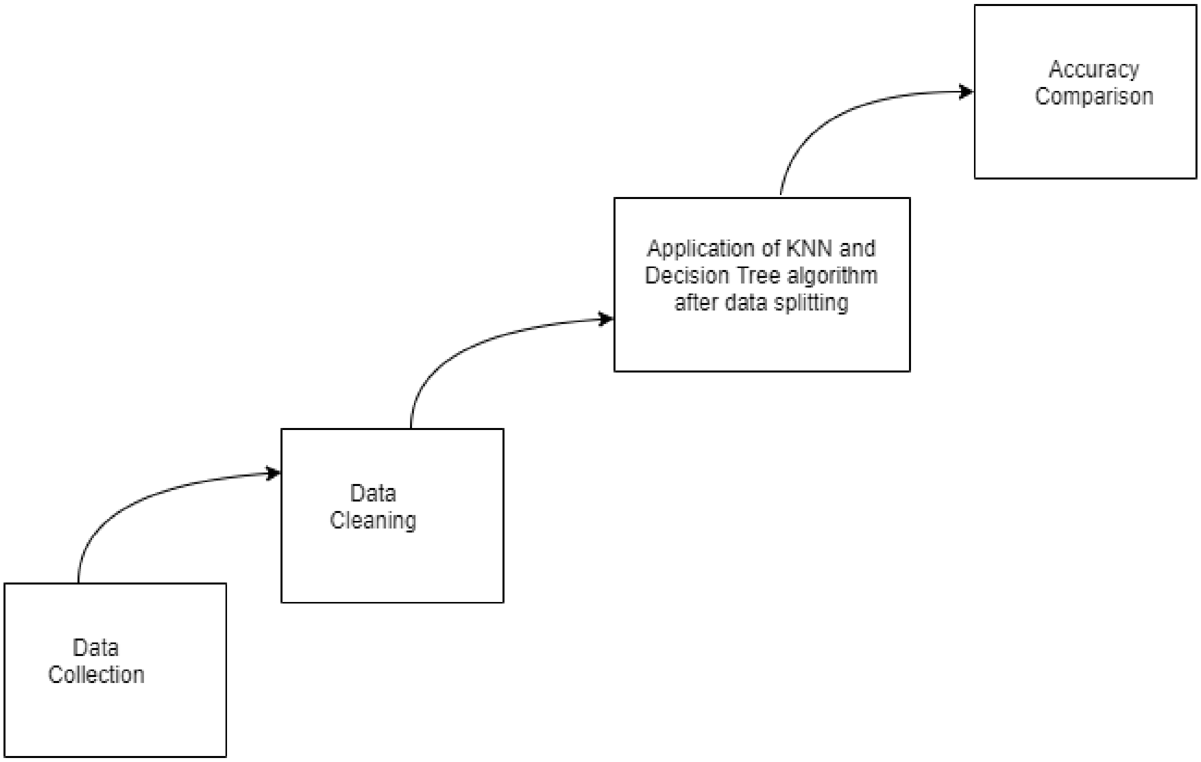

Accuracy Comparison
Application of KNN and Decision Tree algorithm after data splitting
Data Cleaning
Data Collection

Figure 4.1: Applied process of the research study.

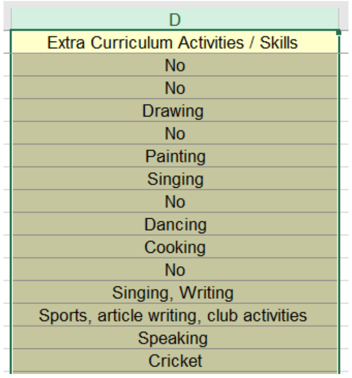
Chapter 5

Implementation

The entire implementation and experimentation were done in Python 3.8.5 environment. Pandas, Numpy, train_test_split, KNeighborsClassifier, DecisionTreeClassifier, classification_report, accuracy_score, StratifiedKFold library used in this project. Pandas is a widely used open-source Python library which work for Data Frame in data science, data analysis, and machine learning activities. Numpy is a Python tool for scientific computing that is widely used. Train_test_split library was used to test and train the data. KNeighborsClassifier and DecisionTreeClassifier were used to predict the accuracy. Classification_report is the visualizer which displays the precision. Accuracy_score function computes subset accuracy in multilabel classification, KFold was used for Cross Validation.

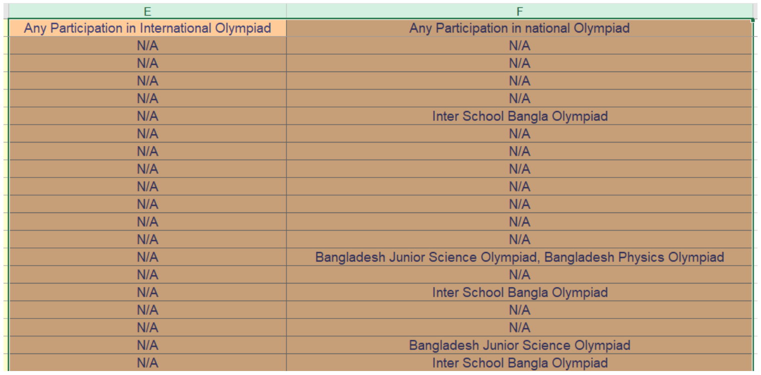
5.1 Data Collection of Student

Data collection is a very important part of Data analysis. The process of data collection divided into four categories. First of all, “Extra Curriculum of a student [Figure - 1.1.1]” Data was taken then the data of “International and National Olympiad participation [Figure - 1.1.2]” data was taken from the students. After that “Student Grades from high school [Figure - 1.1.3]” collected and then the collection of data was taken from the student who are currently studying in universities “University Major subject and CGPA [Figure - 1.1.4]”.



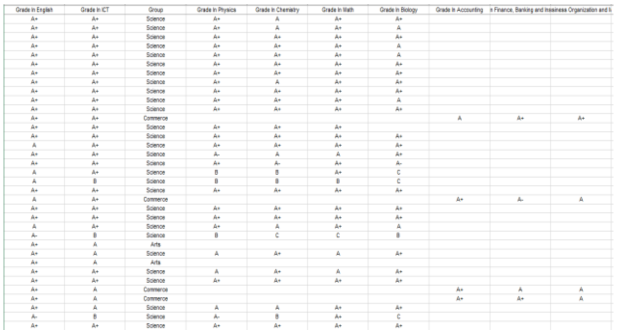
D
Extra Curriculum Activities / Skills
No
No
Drawing
No
Painting
Singing
No
Dancing
Cooking
No
Singing, Writing
Sports, article writing, club activities
Speaking
Cricket

Figure 5.1: ECA data



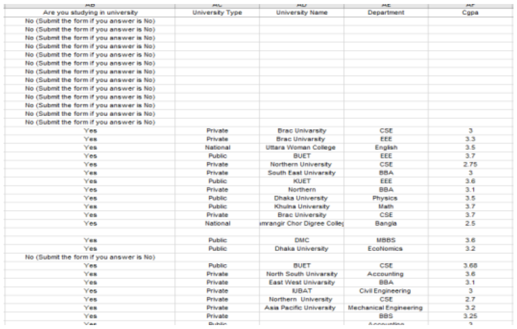
E	F
Any Participation in International Olympiad	Any Participation in national Olympiad
N/A	N/A
N/A	N/A
N/A	N/A
N/A	N/A
N/A	Inter School Bangla Olympiad
N/A	N/A
N/A	N/A
N/A	N/A
N/A	N/A
N/A	N/A
N/A	N/A
N/A	N/A
N/A	Bangladesh Junior Science Olympiad, Bangladesh Physics Olympiad
N/A	N/A
N/A	Inter School Bangla Olympiad
N/A	N/A
N/A	N/A
N/A	Bangladesh Junior Science Olympiad
N/A	Inter School Bangla Olympiad

Figure 5.2: International and National Olympiad data



Grade In English	Grade In ICT	Group	Grade In Physics	Grade In Chemistry	Grade In Math	Grade In Biology	Grade In Accounting	n Finance, Banking and In	usiness Organization and I
A+	A+	Science	A+	A	A+	A+			
A+	A+	Science	A+	A	A+	A			
A+	A+	Science	A+	A+	A+	A+			
A+	A+	Science	A+	A+	A+	A			
A+	A+	Science	A+	A+	A+	A			
A+	A+	Science	A+	A+	A+	A+			
A+	A+	Science	A+	A+	A+	A+			
A+	A+	Science	A+	A	A+	A+			
A+	A+	Science	A+	A+	A+	A+			
A+	A+	Science	A+	A+	A+	A			
A+	A+	Science	A+	A+	A+	A+			
A+	A+	Commerce					A	A+	A+
A+	A+	Science	A+	A+	A+				
A+	A+	Science	A+	A+	A+	A+			
A	A+	Science	A+	A+	A+	A+			
A+	A+	Science	A-	A	A	A+			
A+	A+	Science	A+	A-	A+	A-			
A	A+	Science	B	B	A+	C			
A	B	Science	B	B	B	C			
A+	A+	Science	A+	A+	A+	A+			
A	A+	Commerce					A+	A-	A
A+	A+	Science	A+	A+	A+	A+			
A+	A+	Science	A+	A+	A+	A+			
A	A+	Science	A+	A	A+	A			
A-	B	Science	B	C	C	B			
A+	A	Arts							
A+	A	Science	A	A+	A	A+			
A+	A	Arts							
A+	A+	Science	A	A+	A	A+			
A+	A+	Science	A+	A+	A+	A+			
A+	A	Commerce					A+	A	A
A+	A	Commerce					A+	A+	A
A+	A	Science	A	A	A+	A+			
A-	B	Science	A-	B	A+	C			
A+	A+	Science	A+	A+	A+	A+			

Figure 5.3: High School Grade

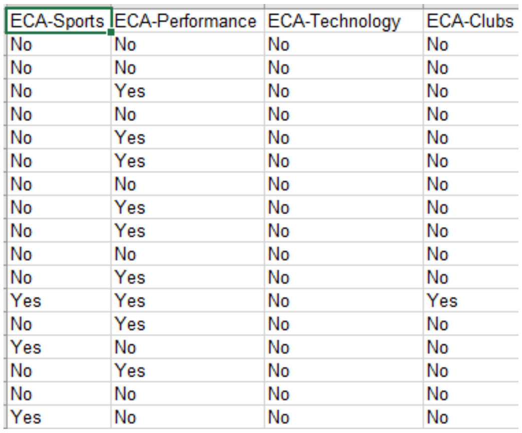


AB	AC	AD	AE	AF
Are you studying in university	University Type	University Name	Department	Cgpa
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
No (Submit the form if you answer is No)				
Yes	Private	Brac Univarsity	CSE	3
Yes	Private	Brac Univarsity	EEE	3.3
Yes	National	Uttara Woman College	English	3.5
Yes	Public	BUET	EEE	3.7
Yes	Private	Northern University	CSE	2.75
Yes	Private	South East Univarsity	BBA	3
Yes	Public	KUET	EEE	3.6
Yes	Private	Northern	BBA	3.1
Yes	Public	Dhaka University	Physics	3.5
Yes	Public	Khulna Univarsity	Math	3.7
Yes	Private	Brac University	CSE	3.7
Yes	National	ımrangir Chor Digree Colleg	Bangla	2.5
Yes	Public	DMC	MBBS	3.6
Yes	Public	Dhaka University	EcoNomics	3.2
No (Submit the form if you answer is No)				
Yes	Public	BUET	CSE	3.68
Yes	Private	North South Univarsity	Accounting	3.6
Yes	Private	East West Univarsity	BBA	3.1
Yes	Private	IUBAT	Civil Engineering	3
Yes	Private	Northern University	CSE	2.7
Yes	Private	Asia Pacific University	Mechanical Engineering	3.2
Yes	Private		BBS	3.25
Yes	Public		Accounting	3

Figure 5.4: University Data

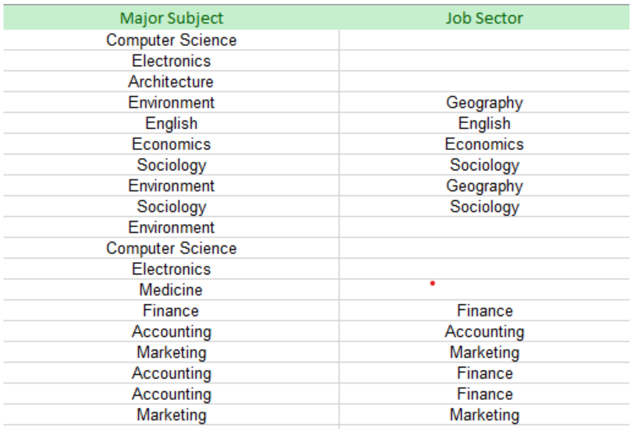
5.2 Data Pre-processing

The basic goal of Data Pre-processing is to find and remove errors and duplicate data so that a valid dataset may be created. This enhances the training data quality for analytics and allows for more precise decision-making. In this project, the data was collected from students via Google Forms. After collecting data, many inappropriate data was found. One of them was "Missing Data". If the missing data were dropped then there were a huge chance to lose a lot of data from the students. The missing data was labelled as 'No' to prevent the problem.



ECA-Sports	ECA-Performance	ECA-Technology	ECA-Clubs
No	No	No	No
No	No	No	No
No	Yes	No	No
No	No	No	No
No	Yes	No	No
No	Yes	No	No
No	No	No	No
No	Yes	No	No
No	Yes	No	No
No	No	No	No
No	Yes	No	No
Yes	Yes	No	Yes
No	Yes	No	No
Yes	No	No	No
No	Yes	No	No
No	No	No	No
Yes	No	No	No

Figure 5.5: ECA Divided and Labelled



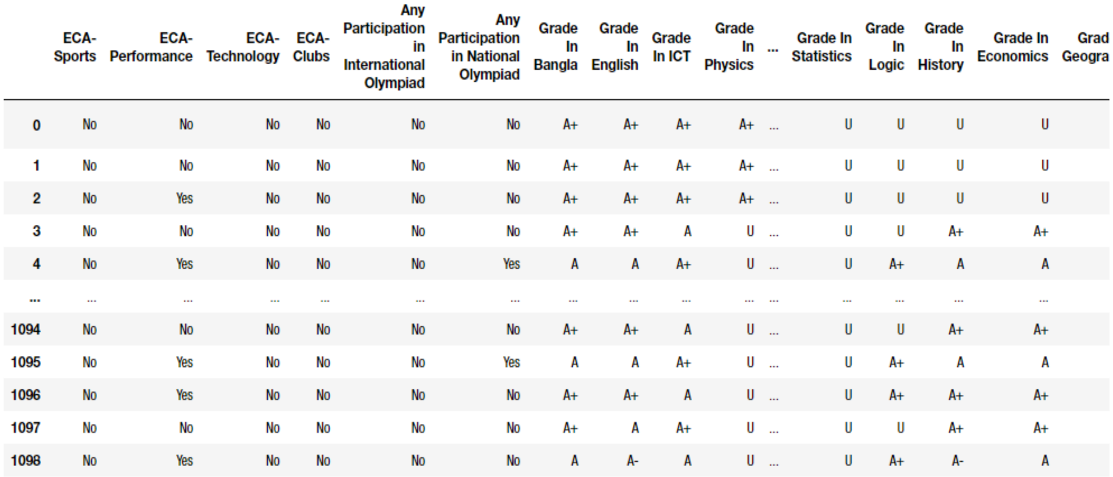
Major Subject	Job Sector
Computer Science	
Electronics	
Architecture	
Environment	Geography
English	English
Economics	Economics
Sociology	Sociology
Environment	Geography
Sociology	Sociology
Environment	
Computer Science	
Electronics	
Medicine	
Finance	Finance
Accounting	Accounting
Marketing	Marketing
Accounting	Finance
Accounting	Finance
Marketing	Marketing

Figure 5.6: Major Subject and Job sector prediction

Then another problem was arrived which was "Structural Errors". As the data were taken from human, they gave same data in different ways. For example, two people's major was Computer Science and Engineering. One of them wrote 'CSE' and other wrote 'Computer Science and Engineering'. This kind of problems was fixed manually. The data "Extra

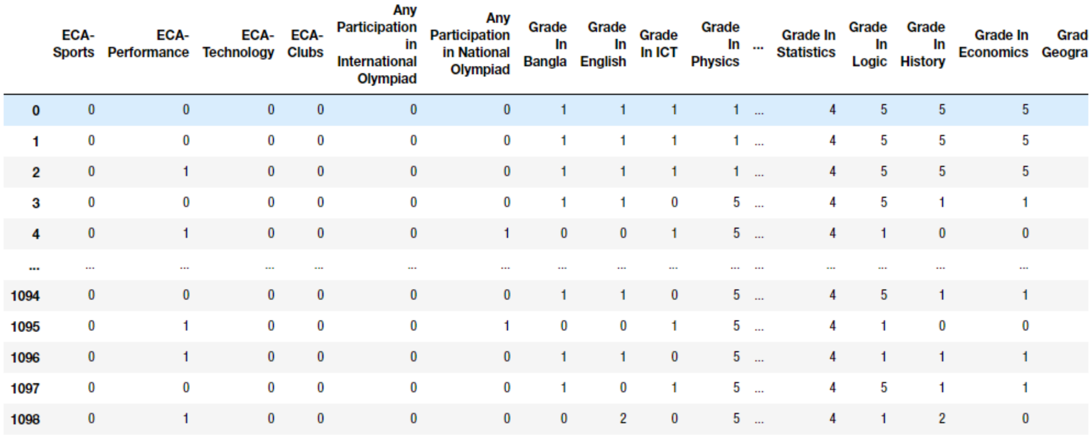
Curriculum of a student" was divided into four parts ECA-Sports, ECA-Performance, ECA-Technology and ECA-Clubs. Different kinds of student have different types of extra curriculum activity so that is why this section was divided into four parts. The main purpose of this project is to determine the accuracy of major subject prediction and job sector prediction of university students but these cannot be predicted automatically in machine learning. To train and test the data a major subject and job sector was labelled manually based on their ECA, High School Grade, University CGPA and Department.

After Cleaning the data, a function was created to encode the categorical values to numerical values. Machine learning algorithm do not work directly with categorical values so we created the function which will automatically encode into numerical data.



	ECA-Sports	ECA-Performance	ECA-Technology	ECA-Clubs	Any Participation in International Olympiad	Any Participation in National Olympiad	Grade In Bangla	Grade In English	Grade In ICT	Grade In Physics	...	Grade In Statistics	Grade In Logic	Grade In History	Grade In Economics	Grad Geogra
0	No	No	No	No	No	No	A+	A+	A+	A+	...	U	U	U	U	
1	No	No	No	No	No	No	A+	A+	A+	A+	...	U	U	U	U	
2	No	Yes	No	No	No	No	A+	A+	A+	A+	...	U	U	U	U	
3	No	No	No	No	No	No	A+	A+	A	U	...	U	U	A+	A+	
4	No	Yes	No	No	No	Yes	A	A	A+	U	...	U	A+	A	A	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1094	No	No	No	No	No	No	A+	A+	A	U	...	U	U	A+	A+	
1095	No	Yes	No	No	No	Yes	A	A	A+	U	...	U	A+	A	A	
1096	No	Yes	No	No	No	No	A+	A+	A	U	...	U	A+	A+	A+	
1097	No	No	No	No	No	No	A+	A	A+	U	...	U	U	A+	A+	
1098	No	Yes	No	No	No	No	A	A-	A	U	...	U	A+	A-	A	

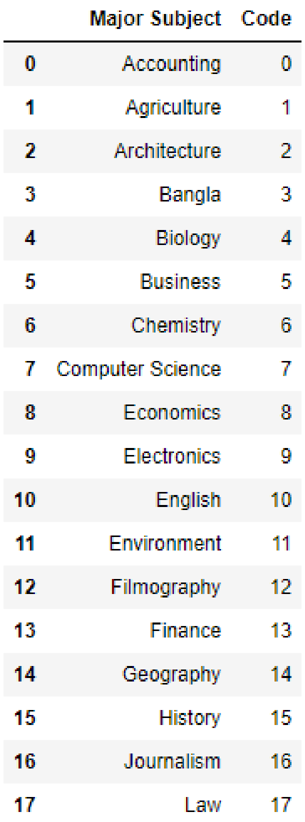
Figure 5.7: Categorical Values



	ECA-Sports	ECA-Performance	ECA-Technology	ECA-Clubs	Any Participation in International Olympiad	Any Participation in National Olympiad	Grade In Bangla	Grade In English	Grade In ICT	Grade In Physics	...	Grade In Statistics	Grade In Logic	Grade In History	Grade In Economics	Grad Geogra
0	0	0	0	0	0	0	1	1	1	1	...	4	5	5	5	
1	0	0	0	0	0	0	1	1	1	1	...	4	5	5	5	
2	0	1	0	0	0	0	1	1	1	1	...	4	5	5	5	
3	0	0	0	0	0	0	1	1	0	5	...	4	5	1	1	
4	0	1	0	0	0	1	0	0	1	5	...	4	1	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1094	0	0	0	0	0	0	1	1	0	5	...	4	5	1	1	
1095	0	1	0	0	0	1	0	0	1	5	...	4	1	0	0	
1096	0	1	0	0	0	0	1	1	0	5	...	4	1	1	1	
1097	0	0	0	0	0	0	1	0	1	5	...	4	5	1	1	
1098	0	1	0	0	0	0	0	2	0	5	...	4	1	2	0	

Figure 5.8: Numerical Values

Another function was created to label the Major Subject which was predicted based on the Grades, ECA and Participation of International and National Olympiad.



	Major Subject	Code
0	Accounting	0
1	Agriculture	1
2	Architecture	2
3	Bangla	3
4	Biology	4
5	Business	5
6	Chemistry	6
7	Computer Science	7
8	Economics	8
9	Electronics	9
10	English	10
11	Environment	11
12	Filmography	12
13	Finance	13
14	Geography	14
15	History	15
16	Journalism	16
17	Law	17

Figure 5.9: Major Subject Labelling

5.3 KNeighbors Classification Algorithm for Major Subject

Feature list and Label were used. To predict in machine learning feature list will contain those values which were used to predict the result ('ECA-Sports', 'ECA-Performance', 'ECA-Technology', 'ECA-Clubs', 'Any Participation in International Olympiad', 'Any Participation in National Olympiad', 'Grade In Bangla', 'Grade In English', 'Grade In ICT', 'Grade In Physics', 'Grade In Chemistry', 'Grade In Math', 'Grade In Biology', 'Grade In Accounting', ‘Grade In Finance', 'Grade In Management', 'Grade In Marketing', 'Grade In Statistics', ‘Grade In Logic', 'Grade In History', 'Grade In Economics', 'Grade In Geography', ‘Grade In Civics', 'Grade In Psychology'). Label represented the result. Feature encoded value and label encoded value were used for test and train the data set. Test and Train data were two most important part in this project. 20 percent encoded data was taken to test the data and rest 80 percent was trained so that they can tag themselves. A ravel was added so that the train data could store into one dimensional array. Then the KNeighbors was applied in the algorithm for finding the accuracy.

5.4 Decision Tree Classifier Algorithm for Major Subject

Decision Tree Classifier and KNeighbors Classification are quite similar algorithm. To compare with KNeighbors Classification and if there any chance to get better result on Decision Tree Classifier. Similar data and methods were used to run the classification. First of all, we test 20 percent data to train rest of 80 percent data which is as like as KNeighbors Classification. Also, a ravel was added to put the 80 percent data in one dimensional array.

5.5 K-fold Cross – Validation for Major Subject

Cross-validation is a resampling technique for evaluating machine learning models on a small sample of data. The process includes only one parameter, k, which specifies the number of groups into which a given data sample should be divided. As a result, the process is frequently referred to as k-fold cross-validation. When a precise value for k is specified, it can be substituted for k in the model's reference. Here k=3 for 3-fold cross-validation was done to get better accuracy.

5.6 KNeighbors Classification Algorithm for Job Sector

KNeighbors Classification Algorithm for Job Sector was also implemented as like as done before for Major Subject. Those who did not provided their CGPA marks were dropped because without it the process cannot be done properly. Then again, the data set was encoded Categorical to Numerical. The Job sector was labelled. This time the values of the feature were ('ECA-Sports', 'ECA-Performance', 'ECA-Technology', 'ECA-Clubs', 'Cgpa', 'Major Subject'). 20 percent of the data set was tested and 80 percent of data was trained similarly done for the Major Subjects.

5.7 Decision Tree Classifier Algorithm for Job Sector

Decision Tree also used on Job Sector for comparison and if there any chance to get better results. DecisionTree Classifier implemented as like as KNeighbors Classification where the data set trained and tested in a same way.

Chapter 6

Comparison

A decision tree is an Eagar Classification where KNN is a Lazy Classification. Decision tree is a supervised learning but in the other hand KNN is Unsupervised. Though KNN can only accepts numerical data, KNN is faster for large amount of data and has effectiveness on small data. On the other hand, Decision Tree accepts both categorical data and numerical data though the speed is slower for a large amount of data.

In this project we work with both Decision Tree and KNN or KNeighbors Classification. Though the result doesn't have huge differences, Decision Tree gave better results.

Classification	Major Subject Prediction Accuracy
KNeighbors	0.5636363636363636
Decision Tree	0.6227272727272727

Figure 6.1: Major Subject Prediction Accuracy Comparison

Here the accuracy value of KNeighbors Classification is less than Decision Tree Classification. So that a cross validation was done to get better result from KNeighbors classification. K fold cross – Validation was done 3 times to improve the result.

Classification	1st Time	2nd Time	3rd Time
KNeighbors	0.4986376021798365	0.5655737704918032	0.5218579234972678

Figure 6.2: K fold cross - Validation

From Figure – 2.2 we can see that the result still did not increase that much. The main reasons for this result are less data. We did not find a huge amount of data related to our criteria. So that we had to collect real life data and this was the reason why the accuracy is below 70 percent.

The value of Job Sector prediction accuracy was better than Major Subject prediction Accuracy.

Classification	Job Sector Prediction Accuracy
KNeighbors	0.7471910112359551
Decision Tree	0.8595505617977528

Figure 6.3: Job Sector Prediction Accuracy Comparison

From Figure – 2.3 we can see that both of the classification gave us good result but among them Decision Tree Classification gave us more accuracy. So, the Decision Tree Classification did better for Job Sector Prediction Accuracy.

Chapter 7

Result

The students suffer a lot after they complete their high school study because they cannot decide what major subject is suitable. Similarly, after completing the study from university they again suffer in dilemma about their job sector. So, we decided to do some research on these criteria. Our vision is to help the student to reduce their dilemma about these two sectors. When we started the procedure our first work to collect huge amount of data. Unfortunately, we choose an area where we did not find any kind of readymade data. For this reason, we had to collect real life data from the students. We faced another problem because of Covid-19 pandemic. Reaching to students were so difficult because of Covid-19. If there were no pandemic situation, it was easy to collect data from our university and also from our college but it did not possible to do. Also, many students did not take Google form fill up seriously so that they put many random data for this reason we had to drop many unusual data from our data set. When data cleaning of our data set started, we had to work a lot to clean it. Then choosing algorithm was a very vital part of our project. After doing some research we decided to go with KNeighbors Classification or KNN Classification because it delivers highly precise predictions, the KNN algorithm can compete with the most accurate models. The distance measure affects the accuracy of the predictions. As a result, the KNN method is appropriate for applications with significant domain knowledge. This understanding aids in the selection of an acceptable metric. The KNN algorithm is a sort of lazy learning in which the prediction calculation is postponed until after classification. Despite the fact that this approach has higher processing costs than other methods, it is still the best choice for applications where predictions aren't demanded frequently but accuracy is critical. After starting the work, we decided to go for a comparison with another popular algorithm which name is Decision Tree Classification. Decision trees resemble trees, and so appear to be a hierarchical model. It sorts all objects depending on attribute value to classify them. Based on attributes, each node in the tree represents the item. Branches represent the value of an object. Objects are categorized from the root node. The objects are then ordered based on the attribute value. A decision tree is a type of classifier that uses a tree structure to offer some rules. We choose this algorithm because KNeighbors is a lazy Classification where Decision tree is an Eager Classification. In this project these two classifications gave us a good result. Though we did not get good accuracy in Major Subject Prediction, the accuracy of Job Sector Prediction was quite good. If we got more data then Major Subject Prediction Accuracy definitely increase. These two classifications works very well but among them Decision Tree Classification gave us more accuracy on both Major subject and Job sector prediction accuracy. After getting prediction accuracy for the first time for Major Subject Prediction accuracy, we saw that decision tree gave us more accuracy then KNeighbors. Then we decided to use K-fold Cross Validation and thought that it the accuracy of KNeighbors will increase but did not get satisfactory values. The main problem was less data in data set. If we could collect more data then the value might have increased. So, Decision Tree Classification is better than KNeighbors classification regarding this project.

Chapter 8

Future Work

According to our proposal, only a few changes are required to get the system up and running as envisioned. As our research is for "Accuracy of Data Analysis on University student's major subject and Job Sector prediction". In future we want to research on High school students who are weak in terms of academic results and confused about which major subject they should study in university. We will continue to work on their data in the future, as well as data from persons who are successful in their careers but struggled in school. We will analyze and work on their data using other classification algorithms. On the whole, our target is to keep improving our research and the accuracy so that these research work can be helpful to the education policies and systems.

Chapter 9

Conclusion

To wrap up, our proposed model is to find the accuracy of university student's major subject and Job sector prediction using KNN and Decision Tree Algorithm. In our prediction Decision Tree Algorithm gives much good accuracy than KNN Algorithm. We face many challenges to collect student's data, as it is very much confidential. Also, many students hesitate to give their academic results. And our prediction level is in primary stage. If there was more data, than we can get better accuracy from KNN Algorithm. Our proposed model will be beneficial both for predicting suitable major subject at university after high school and job prediction after under graduation. The motto of this research work to ensure student's before and after graduation will be out of all kind of dilemma to choose right major subject and to choose right job sector. However, a few operations remain to be completed in order to complete the seal, which we expect to perform as part of our future work.

Bibliography

[1] G. Dimi ́c, D. Prokin, and K. Kuk, ‘’Primena Decision Trees i Naïve Bayes klasifikatora na skup podataka izdvojen iz Moodle kursa,”INFOTEH-JAHORINA, 2012.

[2] Dervisevic, O., Zunic, E., Eonko, D., & Buza, E. (2019). Application of KNN and Decision Tree Classification Algorithms in the Prediction of Education Success from the Edu720 Platform. 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech).

[3] S.K. Yadav, S. Pal, “Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification,” World of Computer Science and Information Technology Journal,2012.

[4] Ramalingam, M., & Ilakkiya, R. (2021). Data Mining Algorithms (KNN & DT) Based Predictive Analysis on Selected Candidates in Academic Performance. 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence).

[5] Lijuan Zhou, Linshuang Wang, Xuebin Ge, & Qian Shi. (2010). A clustering-Based KNN improved algorithm CLKNN for text classification. 2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010).

[6] Hassonah, M. A., Rodan, A., Al-Tamimi, A.-K., & Alsakran, J. (2019). Churn Prediction: A Comparative Study Using KNN and Decision Trees. 2019 Sixth HCT Information Technology Trends (ITT).

[7] Jeganathan, S., Parthasarathy, S., Lakshminarayanan, A. R., Ashok Kumar, P. M., & Khan, M. K. A. 2021). Predicting the Post Graduate Admissions using Classification Techniques. 2021 International Conference on Emerging Smart Computing and Informatics (ESCI).

[8] Nicholas T. Young, Marcos D. Caballero, “Using Machine Learning To Understand Physics Graduate School Admissions”, arXiv:1907.01570v2 [physics.ed-ph] 30 Sep 2019.

[9] Surya Rebbapragada, Amit Basu, and John Semple, “Data mining and revenue management methodologies in college admissions”Communications of the ACM,Vol 53, No 4, April 2010

[10] R. Jane Preetha Princy, Saravanan Parthasarathy,P. Subha Hency Jose, Arun Raj Lakshminarayanan, Selvaprabu Jeganathan. “Prediction of Cardiac Disease using Supervised Machine Learning Algorithms”, 2020 4th International Conference on Intelligent Computing and Control Systems ICICCS),2020.

[11] Abdul Hamid M. Ragab,Amin Y. Noaman,Abdullah S. Al-Ghamdi,Ayman I. Madbouly, "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining", Workshop on Interaction Design in Educational Environments, June 2014

[12] Ahmed, M. R., Tahid, S. T. I., Mitu, N. A., Kundu, P., & Yeasmin, S. (2020). A Comprehensive Analysis on Undergraduate Student Academic Performance using Feature Selection Techniques on Classification Algorithms. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

[13] M.S. Mythili1 and A.R.Mohamed Shanavas , "An analysis of students' Performance using classification algorithms ", IOSR-JCE, Volume 16, iss1, Jan. 2014.

[14] Imdad, U., Ahmad, W., Asif, M., & Ishtiaq, A. (2017). Classification of students results using KNN and ANN. 2017 13th International Conference on Emerging Technologies (ICET).

[15] Awaludin, G. N., Gerhana, Y. A., Maylawati, D. S., Darmalaksana, W., Arianti, N. D., Rahman, A., & Musli, M. (2020). Comparison of Decision Tree C4.5 Algorithm with K-Nearest Neighbor (KNN) Algorithm in Hadith Classification. 2020 6th International Conference on Computing Engineering and Design (ICCED).

[16] Ougiaroglou, S. and Evangelidis, G., 2015. Dealing with noisy data in the context of k-NN Classification. *Proceedings of the 7th Balkan Conference on Informatics Conference*.