Prediction Model for a Student to Find Best Department of Undergraduate Admission

BY

MD. SHOWAL HASAN
ID: 172-15-10139

AND

MD. AL-AMIN
ID: 172-15-9999

AND

MD. ABU TALEB
ID: 172-15-9929

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

Supervised By

Ms. Nazmun Nessa Moon
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
OCTOBER 2020

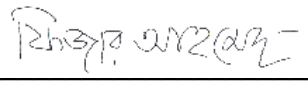
APPROVAL

This Project titled **"Prediction Model for a Student to Find Best Department of Undergraduate Admission"**, submitted by Md. Showal Hasan ID: 172-15-10139 and Md. Al-Amin, ID: 172-15-9999 and Md. Abu Taleb, ID: 172-15-9929 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on October 2020.

BOARD OF EXAMINERS

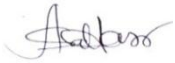
Dr. Syed Akhter Hossain
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



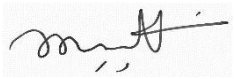
Dr. Fizar Ahmed
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Mr. Abdus Sattar
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



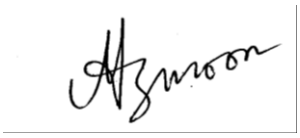
Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

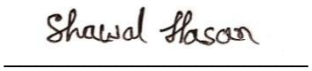
We hereby declare that, this project has been done by us under the supervision of **Ms. Nazmun Nessa Moon, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

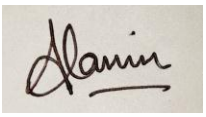


Ms. Nazmun Nessa Moon
Assistant Professor Department of CSE
Daffodil International University

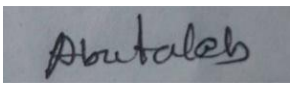
Submitted by:



Md. Showal Hasan
ID: 172-15-10139
Department of CSE
Daffodil International University



MD. Al-Amin
ID: 172-15-9999
Department of CSE
Daffodil International University



MD. Abu Taleb
ID: 172-15-9929
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We really grateful and wish our profound our indebtedness to **Ms. Nazmun Nessa Moon, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge and keen interest of our supervisor in the field of "Machine Learning" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to the Almighty Allah and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

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

ABSTRACT

Our research title is "Prediction model for a student to find best department of undergraduate admission". We have used SPSS statistical platform for storing data and we have use WEKA to find out best algorithm model for finding best outcome according to our data structure. We have selected KNN model for our project. Finally we have used python language to implement the project. In this study, originally implemented by using Bangladeshi students' & educated employed students' real data. The data has learned to a machine using python language which is helped us to find out departments' successor prediction in percent serially. In this system, it has focused on previous group & group courses' result first. Moreover it is focused on general courses which process the overall result & match with the learned result which priority is the best profession & success too. As we know Bangladeshi students are too much brilliant. So, we think if they get the opportunity to choose which department is best for him and then it will be helpful for their career.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Data mining is the process used by companies to analyze data from different outlook and outline to interpret it into useful information – information which can be used to increase revenue, reduce costs, or both. Data mining software is one of the numbers of analytical tools for analyzing range of database and detect pattern. It is one of the most used technologies and research topic to solve our problems about our study subject choice. Every year many students are admitted into universities for undergraduate degree in different subjects. But after passing the HSC examination, they became confuse about their undergraduate subject what they should choose.

1.2 Motivation

Students have a lot of worries when it comes to graduation. According to their results, it would be better to take any subject, they cannot make the right decision in which subject they can do well. This has a great impact on their education, exam results and future career. Some parents force to admit medical or engineering sectors but they were not enable to study in this subject .Then they became frustrate and made wrong decision in their life. As this situation we want to analysis their SSC & HSC result and predict which subject will be good for him or her.

1.3 Rationale of the Study

It is no doubt there are lots works at the Logistic regression function classifier. We can see many works with data, some of them can find expecting result what you want. But working with the K-Nearest Neighbor Classification is more efficient and reliable. It can show accurate output with less training data. This made me to be interested to work with this.

At the present time, we can see the use of AI everywhere and we can also see the crisis with the actual decision. So we decide to do something new to predict the best future goal subject to students.

1.4 Research Questions

- May we predict the best subject for student?
- How can we use it platform independently?
- How much will student get benefited?
- How much accuracy can we get?

1.5 Expected Output

Basically, we going to develop a research based project system which will help people to choose their career. Moreover who needs the best career background for him/her better future.

1.6 Report Layout

In **chapter 1,** all about this project is written here. The reason of choosing this project, how will we have completed this project, project motivation, expected outcome and so on is discussed briefly. In a word, chapter 1 is elaboration of introduction of this project.

In **chapter 2**, related works on this area which were studied are showed. Their findings and limitations are summarized and hence the scope and challenges of the research are also mentioned.

In **chapter 3:** research methodology will discusses Research Topic and Instrumentation, Data Collection Procedure, Statistical Analysis and Implementation Requirements

In **chapter 4:** experimental results and discussion Experimental Results and Descriptive Analysis

In **chapter 5:** presents a short conclusion. And list of references

CHAPTER 2
BACKGROUND

2.1 Introduction

In this chapter, we will discuss related research or project about data classifier or related to detect something with raw data. In the first section we will discuss about previous related work, then in the second section we will show the outcome or a summary of our study of the related work and then we will discuss about the benefits and challenges that we have faced to do this project.

2.2 Related Works

Data mining is the field of inventing fictions and dynamic effective information from huge dataset which is also known as "Knowledge Discovery in Databases" (KDD). Nowadays, Educational Data Mining (EDM) is still in its cradle [1]. Generally, EDM sector is totally fresh and outgoing in the field of learning and educational sector which could be used in other fields like gaming, sports, accounts, transportation and so on. Han and Kamber narrates in their paper that data mining software is one of the numbers of analytical tools for analyzing range of database and detect pattern which is accept the users to analyze data from different dosages, classify it and summarize the given relationships which are used during the data implement process [2]. M.Ramaswami and R.Bhaskaran had used CHAID prediction model that can be used to predict the outcome of the performance at higher secondary school education level which can be adjusted the interrelation between variables of the dataset [3]. The CHAID expectation model was developed with seven class indicator variable for understudy execution. They have used three methods to deal with the class unbalancing problem and all of them are shown the satisfactory results as well. They initially adjusted the datasets and utilized both cost-obtuse and delicate learning with SVM for the little datasets with Decision Tree for the bigger datasets. Arockiam et al. have developed FP Tree and K-means clustering technique for finding the similarity of programming skills between urban and rural students [4]. FP Tree mining has been applied to strainer the examples of the favorable dataset. K-implies grouping was utilized to decide the programming abilities of the understudies. The examination plainly demonstrates that the country and the metropolitan understudies contrast in

their programming aptitudes and found that enormous extents of metropolitan understudies were acceptable in programming ability contrasted with provincial understudies. It reveals that academicians give additional preparation to metropolitan understudies in the programming department. Cortez and Silva have attempted to predict failure percentage of two secondary school students from the Alentejo region of Portugal in the two prime courses (Mathematics and Portuguese) [5]. Basically they have used that by utilizing 29 predictive variables. They had used four popular algorithms for instance Neural Network (NN), Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM) that were applied on a dataset of 788 students, who had performed in 2006 examination. It was accounted for that DT and NN calculations have the prescient exactness of 93% and 91% for two-class dataset (pass/come up short) individually. A point was also reported that both Decision Tree (DT) and Neural Network (NN) algorithms had the predictive accuracy of 72% among four-class dataset; Galit had given a case study that uses student's data to analyze their learning behavior to be conscious students at risk before their final exams and to predict the results also [6].

V.Ramesh et al had tried to find out the key motives which are impact on the performance of students in their final examination [7]. They embraced review cum exploratory approach to create the information base. They have used the algorithms for implementations were Naïve Bayes, Multi-Layer Perception, KNN, Random forest, and Logistic. The got outcomes from speculation testing uncovers that kind of school isn't impact understudy execution yet parent's occupation assumes a significant part in anticipating grades.

2.3 Comparative Analysis and Summary

After researching some research papers and projects we decided to go with KNN(K-Nearest Neighbor)because,

- It works best with numerical data
- KNN has the best accuracy among other same classification algorithm with the accuracy of 90% or above with proper training and labeling
- It's easy to use and lots of resources to work with in further development
- It also works best in comparing data like float value data and analysis
- We can obtain good results and accuracy by using KNN properly and with

proper training.

As we have decided to use KNN (K-Nearest Neighbor) as our main classification model we will work with KNN layers and deep learning algorithms. We have used WEKA and Google cola in backend to implement the model. Our main goal was using an in house database and anaconda. But because of the wide use Colab and easy, fast implementation way we have decided to go with Google Colab. We have used Colab and Google own GPU in runtime to make the best outcome. As a result we had to use mount drive for database use using Google drive.

Our main goal is making a classifier of SSC & HSC result with subject wise and further comparing them in order to get the output. By using SSC & HSC subject wise result; we will be able to get good accuracy range in order to find the best comparison possible among their good output. As if we have a good range between good and bad at then we will be able to say how much bad or good at that subject is the best choice and is it edible or not.

2.4 Scope of the Problem

The major purpose of this research work is making e system that can predict the best subject for under graduate students and proved that KNN is the best algorithm model for data classifier. In future we will try to use my system in big data set.
We will make my system open source and freely available to all. So that anyone of Bangladesh can use this system to analysis previous result and can make a good subject for future goal. It will make their work easier. For our system normal people also can analysis data and get a good decision as well.

2.5 Challenges

1) Data Collection

Since there aren't available data in online, data collection had not that much difficult for this research project we thought. But when we have started collecting data locally it becomes too much difficult because most of the student did not like that we know about their result without any reason. For more accuracy we need more train data. But we was not able to collect huge amount of data.

2) Model selection

Model selection is a quiet tough task for any researcher. Because research project success depends on dataset and model selection. The right choice will lead you at your goal quickly and the wrong one will ruin it. We test different types of model with the test data and try to find the best one. We also try to work with Matlab. After trying everything we have choose a KNN algorithm because Google already did something great to make our work easy. When we have seen that python has built a library like pandas, numpy, seaborn etc for this type of data classifier work and we can work with Google CoLab when working with this kinds of data need an expensive type of GPU, but Google provides free virtual GPU for my work we have decided to work on this model.

3) Data labeling

Data labeling was the most important part of my work. Because it makes code faster to run and less time consuming for data as well. Labeling is one kind of surveying, the object of which is to be established or verify or measure the height of specified points relative to a datum. In this phase, we have renamed all of data as their name and also numbered them sequentially.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

In this session, we will describe about our research methodology and procedures. Moreover, tools have used for the research project, data collection, research topic, pre-processing, processing, statistical analysis and its implementation will be discussed in this session as well.

Fig 3.1: Methodology at a glance

We have done all of discussed steps which are shown above figure. Moreover, we can utilize our data and use online stored data in our project by some simple steps shown above fig 3.1.

3.2 Research Subject and Instrumentation

We think data is the heart of research. So, first of all we have to know about research data. Research data is information which has been organized, collected, observed, and generated to meet the original research findings. It is one of the most critical parts for

researchers to find out suitable data and compatible algorithm or model for their research work. They also have to study about related research papers and relevant topics. Then they need to make several decisions:

- Which type of data should be collected?
- How to ensure that collected data are real or okay?
- How should each data be organized?
- How should each data be labeled?
- How should each data be measure?
- How to analyze different types of data?

3.3 Data Collection Procedure

We have collected data from undergraduate students studying in different universities and job holders are servicing different companies. We have selected those types of organization to find real and raw data. Moreover, we have thought these types of data will provide more accurate outcome and these data will be easily labeled as well. We have used around 1000 academic and professional data. Further that, we have used about 300 data for testing purpose.

Data Pre-processing: Data pre-processing refers to a data mining technique which includes transformation of raw data into organized and efficient format. Usually raw data sets are not used or not able to operate operations and find out expected outcome. As we know, data pre-processing is very important and it is mandatory to go next data process stage. Moreover, it is considered that it is one of the most important parts of research. In this phase, we have filtered some of noisy and incomplete data and tried to discard those data. Furthermore, we had to cast floating value and convert them to our statistical format. Data had to convert into various level in our convenient stage.

Data organizing: Data organizing is a process of classifying and organizing data sets to make them more useful and efficient as it can also be applied to digital recodes. In this phase, we have divided data and store them into two data folder for

training and testing purpose. We also use here validation folder for check trained data validation.

Labeling Data: Labeling is one kind of surveying, the object of which is to be established or verify or measure the height of specified points relative to a datum. In this phase, we have renamed all of data as their name and also numbered them sequentially. We have shown some of data column name in the following fig 3.2

```
df.head()
```

	Department	Gender	Year	Current_CGPA	Subject_Choice	SSC_HSC_Group	SSC_Bangla	SSC_English	SSC_Mathematics	SSC_Physics	SSC_Chemistry	SSC_Biology	SSC_Highermath
0	1	1	0	3.50	1	1	4.0	3.0	4.0	4	5	4	5
1	2	1	1	3.25	2	1	5.0	4.0	4.0	4	5	4	5
2	3	1	2	4.00	3	1	3.5	5.0	5.0	3.5	3.5	5	5
3	5	2	4	2.75	0	1	4.0	4.0	5.0	4	4	4	4
4	6	2	0	3.00	0	1	5.0	3.0	3.0	5	5	3.5	4

Fig 3.2: Labeled table data in SPSS

Data Storing: Data storing is a process of recording or storing information in a structured way in a storage medium. In this step, we have stored all of data in Google drive because we have used online emulator which makes our task easier.

3.4 Statistical Analysis

The amount of my total individual data is more than 1000 that we have collected, but after preprocessing we have got Total 887 individual data. The accurate data amount are given in the following in table 3.1,

Table 3.1: Student related variables

Variable Name	Description	Domain
Gender	Student's Sex	{M,F}
Department	Student's running department	{CSE, EEE, BBA etc.}
Year	Running Semester	{1-12,Graduated}
Current_CGPA	Internal semester CGPA	{A+, A, B, C}
Subject_Choice	Who influence for subject choose	{Parents, Teacher, Others}
Group	SSC&HSC group	{Science, Arts, Commerce}
SSC_Subject GPA	Individual subject wise GPA for SSC	{A+, A, A-,B, C}
HSC_Subject GPA	Individual subject wise GPA for HSC	{A+, A, A-,B, C}
SSC_Result	Overall SSC Result	{A+, A, A-,B, C}

HSC_Result	Overall HSC Result	{A+, A, A-,B, C}
Satisfied	Are you satisfied with your subject	{Yes, No}
Higher Study	Willing to go abroad for higher study	{Yes, No}

3.5 Implementation Requirements

• **Python 3.8**

Python 3.8 is Python version. It is a high level programming language. Most of the researches use it to do their research. It is highly recommended programming language for AI based project and it is very popular among new generation programmers because it is very easy to learn and understand.

• **Google CoLab**

Google CoLab is free to use open source distributor of Python programming language. We can work here online by using our browser as like as Jupiter notebook, but the main benefit of this Google CoLab is, it provides us free online virtual GPU access.

• **SPSS**

SPSS refers to a software platform which offers advanced, improved and enrich statistical analysis, a rich library of machine learning algorithms, text analysis, open source platform, integration with big data and seamless deployment into applications which is developed by American tech giant company IBM (International Business Machine). Its user friendly interface, flexible and scalable features make SPSS more popular to users of skill levels from all perspectives which make perfect for projects of all categories and levels of complexity. That's why, we have use the platform as our data storing application which helps us to find out new opportunities, improve effectiveness, efficiency and minimize risks to develop the project.

• **WEKA**

WEKA is called as "Waikato Environment for Knowledge Analysis" which is developed at the University of Waikato, New Zealand. Basically, WEKA is an open

source software platform which is licensed by the GNU Company. It is platform independent software. Generally it is used to solve real life data mining problems. We have used the software to find out the suitable algorithm in different perspective with respect to our data for finding best fit outcome.

• **Hardware and Software Requirements**

- Operating System (Windows 7 or above)
- Web Browser(preferably chrome)
- Hard Drive or SSD (minimum 120 GB)
- Ram(more than 4 GB)

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

In my model implementation and code implementation We have collected the data first. Procedure given below,

- As we have worked for predicting subjects for student which is the best for them, that's why we have collected SSC & HSC result with subject wise result.
- For a larger part of our project we have collected data from our university's student.
- We also collect data from outside of our university to find out more data from students.
- Then we have finalized and normalized the data in order to perform our training.
- After labeling data was usable and good for further processes. After that we had to preprocess our data in below steps,
- First we have stored data in SPSS file.
- Then we had to convert dataset into csv & arff file for work with python & WEKA.

4.2 Model Summary

The preparation period of K-closest neighbor arrangement is a lot quicker contrasted with other grouping calculations. There is no compelling reason to prepare a model for speculation that is the reason KNN is known as the straightforward and occurrence based learning calculation. KNN can be valuable if there should be an occurrence of nonlinear information. Yield an incentive for the item is registered by the normal of k nearest neighbors esteem. The testing period of K-closest neighbor grouping is increasingly slow as far as time and memory. It requires huge memory for putting

away the whole preparing dataset for forecast. KNN requires scaling of information on the grounds that KNN utilizes the Euclidean separation between two information focuses to discover closest neighbors. Euclidean separation is touchy to extents. The features with high magnitudes will weigh more than features with low magnitudes. KNN also is not suitable for large dimensional data. For better outcomes, normalizing information on a similar scale is strongly suggested. For the most part, the standardization extend considered somewhere in the range of 0 and 1. KNN isn't appropriate for the enormous dimensional information. Here, we have shown the relationship between SSC and HSC GPA by using the model. In such cases, measurement needs to lessen to improve the presentation which is shown below in figure 4.1

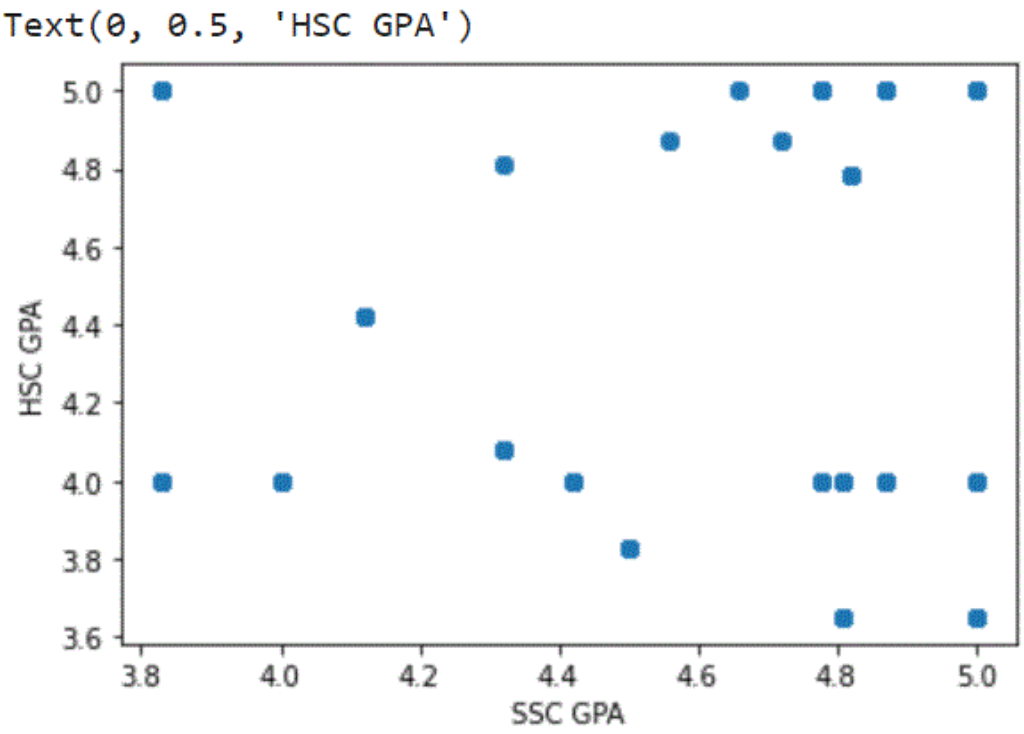


Fig 4.1: Data structure on KNN Model

We have implemented KNN (K nearest neighbor) model on our data, it’s providing satisfactory accurate data graph which is suggested by WEKA. Basically the K Nearest Neighbor (KNN) model is much simple and easier to implement and we have not need to build a new model as well. Generally it’s an algorithm which stores all of

recent percepts and cases that will be used in new problems based on similarity measurements.

4.3 Experimental Results & Analysis

During this project is being developed, we have tested and analyze our dataset with five popular classification algorithms. Algorithms are Multilayer Perception, Naïve Bayes, IBK, Random forest and Logistic. All of the numerical and statistical results have shown in table 4.1. We have also shown a comparison accuracy of all classifiers has completed and finally it has been decided that IBK (KNN- K Nearest Neighbor) algorithm model performs best with accuracy about 75%. The accuracy level of all the algorithms are provided below in table 4.2

Table 4.1 Statistical Analysis of Classifiers with Cross Validation

Name of Classification Algorithm	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
IBK	NQ	0.83	0.44	0.807	0.83	0.822	0.77
	Q	0.55	0.162	0.605	0.553	0.578	0.773
Logistic	NQ	0.76	0.596	0.741	0.762	0.751	0.648
	Q	0.40	0.238	0.432	0.404	0.418	0.648
Random forest	NQ	0.88	0.766	0.721	0.886	0.795	0.56
	Q	0.23	0.114	0.478	0.234	0.314	0.56
Naïve Bayes	NQ	0.81	0.574	0.761	0.819	0.789	0.713
	Q	0.42	0.181	0.513	0.426	0.465	0.713
Multilayer Perceptron	NQ	0.83	0.681	0.733	0.838	0.762	0.667
	Q	0.31	0.162	0.469	0.319	0.38	0.667

Table 4.2 Comparison of Classifiers on the basis of Correctly Classified Instances with Cross Validation

Mining Technique	Accuracy
IBK(KNN)	70%
Naïve Bayes	65.13%
Multi perceptron	68.42%
Logistic	69.73%
Random forest	67.76%

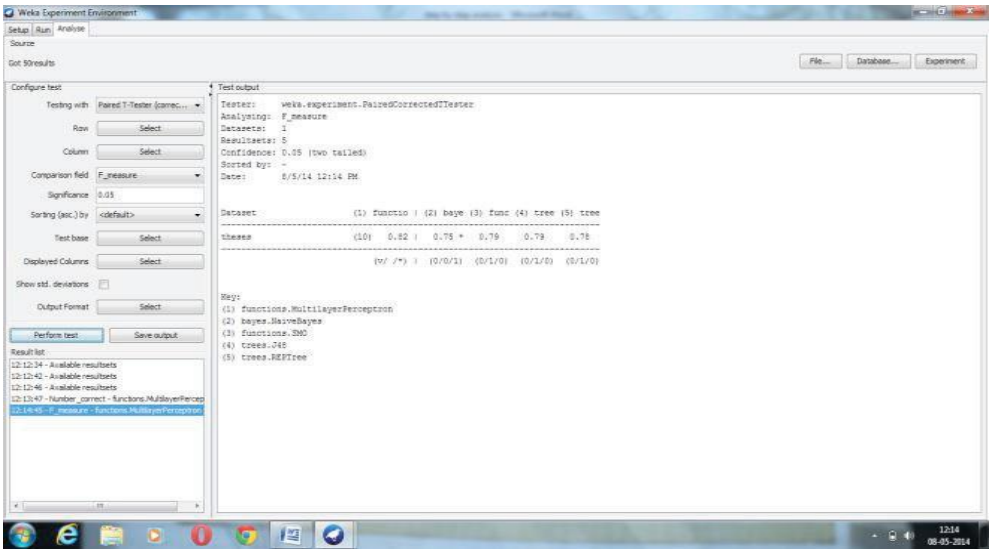


Fig 4.2: Comparison of Classifiers by using WEKA Experimenter

Here, we have shown comparison results of all classifiers by using WEKA Experimenter has been shown in Fig 4.2. In this case, we have also shown that IBK performs best among all other classifiers as well with F-Measure 82%.

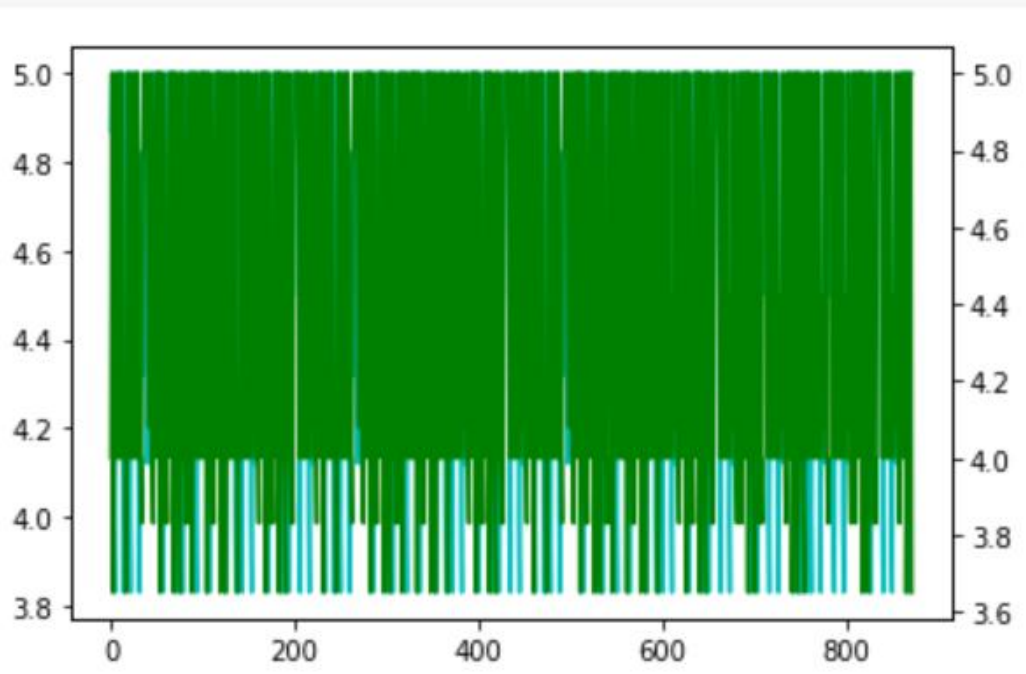


Fig 4.3: Validating & testing SSC & HSC result with whole dataset

In this figure 4.3, we have represented all of data in a form of students' secondary and higher secondary Grade point with respect to whole dataset.

If we divide major 3 portion of Higher Education Subject

1. Engineering (Civil, EEE, Mechanical, CSE, Architecture)
2. Science Subject (Physics, Chemistry, MBBS, Mathematics, Statistics)
3. General Subject (Bangla, English, Economics, Geography, Sociology)

Now which Student will be admitted from above departments?

Suppose one student has got GPA 5 SSC and HSC result and also got good result in Physics and Chemistry subject then this student can get Engineering or Science Subject

If any student can got GPA 4.30 SSC and HSC result and also got poor result in science subject like as physics and chemistry and math. Then this student get engineering subject or science subject this student fail there ambition or don't possible success in graduation.

If this student chooses Bangla or other general subject. We hope that thy will be success there ambition.

When ones student good result in Physic, Chemistry and Biology then they can get Medical science subject to complete higher Study.

Table 4.3 Priority based department choice based on result

	Physics	Chemistry	Biology	Math	Bangla	English	Admission Subject
Student-01	5	5	5	5	5	5	Engineering Subject
Student 02	5	4	5	5	4	5	Medical Subject
Student 03	5	5	4	5	4	5	Science Subject
Student 04	4	3	4	5	4	5	General Subject

In above table 4.3, we see that student grading and admission faculty.

4.4 Discussion

In this report, we have classification techniques used classification techniques for prediction in our dataset of 887 students. We have used that dataset to analyze student's previous result sheet and to predict future career background as well. A model has been developed based on some selected fields of data for this research. In this study, KNN (the K Nearest Neighbor) model performs best with 75% accuracy among all data mining classifiers. Therefore, KNN model has proved that it is potentially effective and efficient model algorithm according to our dataset. WEKA experimenter is also completed by comparing with all of classifiers by using that comparing platform. In that case, KNN model has also proved to be the best with F-measure of 82%. As a result, KNN model performance is comparatively much better than any other classifier algorithms as well. A performance model chart has been also plotted and shown.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

At present, career choosing is being one of the major problems for students of Bangladesh. As we know, literacy rate of Bangladesh is about 75% till now and the rate is increasing day by day. Maximum illiterate people of Bangladesh live in village. But nowadays people of village are being conscious rapidly and coping up with the rate very quickly. As a result, they are feeling confused to choose his/her career background. As we know, job place of Bangladesh is not growing rapidly with respect to number of educated student increasing every year. So, it is very important to balance rate of profession in a specific sector with job market as well. That will highly affect in our society. So, we think our research will play a role to face that unbalancing situation.

5.2 Impact on Environment

In this research, we think it will effect on students' mentally to study attentively. Moreover, it will be positive dimension for our educational environment. As we know, young generation is being addicted to different bad addiction day by day. In this circumstance, student will be able to find out what should to do right now and what will be effective to do for his/her future profession. We think it will be much effective for students and their guardians as well. Furthermore, company is much affective thing in the society and environment for young generation. So, if our research will affect 25% directly, then we can say it will affect half of negative impact in the environment.

5.3 Ethical Aspects

In our research, there are many ethical aspects like discipline for an attentive, sincere student and how to monitor a student by their guardians. Moreover, we have strongly recommended to guardians to monitor their students regularly. We believe that if students will follow our recommendation to finding their best career background, then students will be more conscious about their current activities as well.

5.4 Sustainability Plan

This project will be used for students and their guardians. In this project, student can input their result. No registration will be required. Project will prove a result or outcome, which profession or department will be best for the student. As we have developed the project with incremental model, so it will be updated easily. We will work furthermore with this project to improve better outcome day by day.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of the Study

The research project is developed for finding the best place for future life which subject is the best for students based on their SSC & HSC result and prove that KNN is the best classifier algorithm. For this research we have needed academic result data. So data sets are collected from different students. After that data pre-processing rules are maintained for all the data to make them compatible with the systems environment. Data sets are trained for data handling purpose. Here some of data sets are trained data sets and some are test datasets.

6.2 Conclusions

In this modern era of technology, data is not brought us new opportunities only, but also new challenges and complexities too. Sometimes we need to develop new method, new technology to handle new data and dataset. Collecting data from different students from different universities are very difficult and time consuming. After that our students better future we design this system for choosing their best subject choose by their previous result what they gained from SSC & HSC. So this simple system can solve all of those problems. This system may be a very easy and flexible, but more effective and efficient.

6.3 Recommendations

Some of flexible algorithms may be developed to identify objects from academic data. Because our future is AI based and data in machine learning is very important for AI sector. It makes our machine or technology more productive. So everyone should need to work with data classifier. It can be affected on the entire concept what you have learned from earlier cases as well. Moreover, we are recommending some of important issues to get more effective and efficient outcome.

- ✓ Students who are good at biology should be study in medical
- ✓ Students who are good at mathematics and physics should be study in

engineering

- ✓ Student should choose department based on their previous result

6.4 Implication for Further Study

As we know, there is no limitation of developing projects and applications. So, in future, we will try our best to improve following sectors:

- We will make more data to make this more accurate and efficient outcome
- We will use more labeling of data.
- In future we will develop a complete open source platform with huge amount of student data.

REFERENCES

[1] Zaïane, O. (2002), "Building a recommender agent for e-learning systems". Proceedings of the International Conference on Computers in Education,55–59

[2] Jiawei Han Michelin Kamber(2011), "Data Mining-Concepts and Techniques", Morgan Kaufmann Publishers.

[3] M.Ramaswami and R.Bhaskaran(2010), "A CHAID Based Performance Prediction Model in Educational Data Mining", International Journal of Computer Science Issues Vol. 7, Issue 1, pp 10-18.

[4] L.Arockiam, S.Charles,Arulkumar Association between Urban and Rural Students Programming Skills", International Journal on Computer Science and Engineering Vol. 02, No. 03, pp 687-690.

[5] P. Cortez, and A. Silva (2008), "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeira (Eds.), pp 5-12.

[6] Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme(2009), "Improving Academic Performance Prediction by Dealing with Class Imbalance", Ninth International Conference on Intelligent Systems Design and Applications,

[7] V.Ramesh, P.Parkavi, K.Ramar(2013), "Predicting student performance: A statistical and data mining approach", International journal of computer applications , Volume 63- no. 8, pp 35-39.

[8] https://www.datacamp.com. Last accessed on 23rd August 2020

APENDIX RESEARCH REFLECTION

During project activities, we have faced several problems. But two problems were major among them one is selecting best algorithm and another is data collection. Before working with KNN algorithm we have tried many ways to solve our problem and we failed to get best output. To collecting data we have faced too much obstacles. Because locally data collection is time consuming and uncomfortable for students. Country like Bangladesh, people didn't take it positively. As we can't collect the required data from online resources. After a long time and a lot of attempts and by hard work we have got success finally.

Shawal_report

ORIGINALITY REPORT

17%	17%	12%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	s3.amazonaws.com Internet Source	11%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	www.datacamp.com Internet Source	2%
4	Rosanna Costaguta. "chapter 14 Data Mining Applications in Computer-Supported Collaborative Learning", IGI Global, 2015 Publication	1%
5	doit.wp.txstate.edu Internet Source	<1%
6	digitalcommons.unl.edu Internet Source	<1%
7	Kaur, Parneet, Manpreet Singh, and Gurpreet Singh Josan. "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector", Procedia Computer Science, 2015.	<1%