Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 02, February-2021

Prediction of Finest Department Assortment on Undergraduate Admission Through Machine Learning Approach

Abdus Sattar
Department of Computer Science and Engineering,
Daffodil International University
Dhaka, Bangladesh

Rupak Bairagi
Department of Computer Science and Engineering,
Daffodil International University
Dhaka, Bangladesh

Shamsuzzaman Miah
Department of Computer Science and Engineering,
Daffodil International University
Dhaka, Bangladesh

Raihana Zannat
Department of Software Engineering,
Daffodil International University
Dhaka, Bangladesh

Ohidujjaman
Department of Computer Science and Engineering,
Daffodil International University
Dhaka, Bangladesh

***Abstract*— Students are feeling too much confusion to select the right department at undergraduate level after completing the higher secondary in Bangladesh. This study is emphasis on the perspective of students to choose a suitable department in a smart way. However, a system is developed where the students provide their prior data to find out which department is best for them in regarding the future profession. This article uses SPSS (Statistical Package for the Social Sciences) statistical platform for analyzing data and used WEKA tool to find out the best algorithm for finding the optimum outcome according to the data structure. The KNN (k-Nearest Neighbors) model has been selected for this study because it works well with numerical data. The KNN has the stunning accuracy among other classification algorithms with the accuracy of 90% or above with proper training and labeling. The major purpose of this research work is to making e-system that can predict the suitable department for under graduate students and also verified that KNN is the paramount algorithm model for the data classifier.**

Keywords— KNN Model , Machine learning, SPSS, WEKA,

I. INTRODUCTION

Data mining is the process used by companies to analyze data from different outlooks and outline to interpret in into beneficial information. This kind of beneficial information can be used to increase the revenue for any country or organization, to reduce to cost minimize or both can happen. Using data mining techniques, it is possible to analyze a large number of database and detect the pattern of the data [9]. It is one of the most used technologies and research topics to solve our problems with our study subject choice. Every year many students are admitted into universities for the undergraduate degree in different subjects. But after passing the HSC examination, they became confused about their undergraduate subject what they should choose. 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 to medical or engineering sectors but they were not enabled to study this subject .Then they became frustrate and made the wrong decision in their life. In this situation, they want to analyze their SSC and HSC result and predict which subject will be good for them. It is no doubt there are lots of works at the logistic regression function classifier. We can see many works with data, some of them can find expecting result what you want. The main objective of this study is to find out the successor function for the student's carrier. So that it will be possible to decide to do something new to predict the best future goal subject to students. The overall aim of this study is to determine the best subject for the students which one gets benefitted for the students. The following research questions will be answered by this study:

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

II. LITERATURE REVIEW

Data mining includes the field of inventing fictions and utilization of redefined data analysis process from a large number of the dataset which is also known as "Knowledge Discovery in Databases" (KDD). This also works for incorporates analysis and prediction from the database. Nowadays, emerging disciplined is Educational Data Mining (EDM) is to work for developing methods for educational settings. Generally, the EDM sector is totally fresh and outgoing in the field of learning and education sector which could be used in other fields like gaming, sports, accounts, transportation, and so on. By using these tools, accepts the user tools from different stages, then classifies these data, and summarizes them to make their relationship among them during the implementation process [1][2]. Another paper

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 02, February-2021

[3] had used CHAID prediction model tools used, which can predict the outcome and shows the performance of that outcome at a higher secondary school education level. This performance is work to influence many factors and created interrelation between variables and dataset. The CHAID expectation model was developed with seven class indicator variable for understudy execution [3]. Another tool developed FP tree and clustering technique which is used to find the similarity of programming skills between two different locations like, rural and urban area students. It reveals that academicians give additional preparation to metropolitan understudies in the programming department [4]. Cortez [5] tried to shows an educational economical display which works to find the child's interest in the school in Portugal. Basically, they have used that by utilizing 29 predictive variables. They had used four popular algorithms that were applied on a dataset of 788 students, who had performed in the 2006 examination. It was accounted for that DT and NN calculations have the prescient exactness of 93% and 91% for the two-class dataset (pass/come up short) individually. Nguyen [6] stated that two algorithm Decision Tree (DT) and Neural Network (NN) used had the predictive accuracy of 72% among four-class dataset and 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 [7] had tried to find out the key motives and tried to find the best performance during the final exam. They used three different algorithms and got outcomes from speculation testing uncovers that kind of school isn't impact understudy execution yet the parent's occupation assumes a significant part in anticipating grades[8]. After researching some research papers and projects decided to go with KNN (K-Nearest Neighbor) because it works best with numerical data. The KNN has the best accuracy among other same classification algorithms with the accuracy of 90% or above with proper training and labeling. It's easy to use and lots of resources to work within further development. It also works best in comparing data like float value data and analysis. It can obtain good results and accuracy by using KNN properly and with proper training. As it is decided to use KNN (K-Nearest Neighbor) as the main classification model, here will work with KNN layers and deep learning algorithms. It will use WEKA and Google cola in the backend to implement the model. The main goal was using an in house database and anaconda. But because of the wide use of Colab and the easy, fast implementation way, have decided to go with Google Colab. Here, used Colab and Google's own GPU in runtime to make the best outcome. As a result, had to use mount drive for database use using Google drive. Our main goal is making a classifier of SSC & HSC result subject wise and further comparing them in order to get the output. By using SSC & HSC subject wise result; it 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 it will be able to say how much bad or good at that subject is the best choice and is it edible or not. However, based on the literature review doesn't get any clear concept of how and which features can make the student's choice [10], [11]. Therefore, used a quantitative research approach in this study to find the research gap. The major purpose of this research work is to make a system that can predict the best subject for undergraduate students. 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. It will make my system open source and freely available to all. So that anyone of Bangladesh can use this system to analyze the previous result and can make a good subject for future goal. It will make their work easier and can analyze data and get a good decision as well.

III. RESEARCH METHODOLOGY

A. Research Methodology Steps

Research methodology, procedures and tools describe in this section. In this study used several steps like, e.g. data collection, pre-processing of the data, model selection, data organizing, data leveling, and statistical analysis. Later discussed the several tools which is used to implement will be discussed in this session as well. Moreover, can utilize data and use online stored data in this project by some simple steps shown in fig. 1.

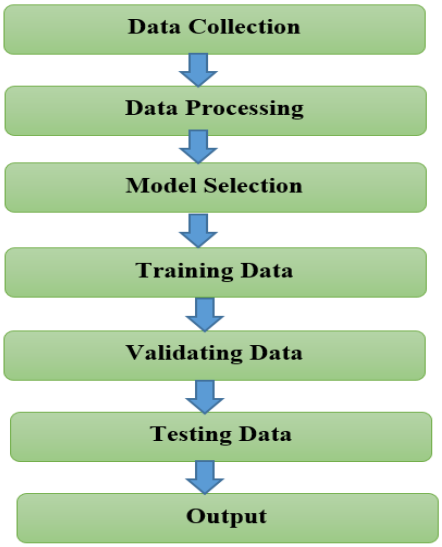

Fig. 1. Research Methodology

B. Research Subject and Instrumentation

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. To find the research subject and instrumentation needed to study about related research papers and relevant topics. It is also needed 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?

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 02, February-2021

- How should each data be measure?
- How to analyze different types of data?

C. Data Collection Procedure

In this study, collected data from undergraduate students those are studying in different universities and job holders are servicing different companies. And selected those types of organization to find real and raw data. Moreover, have thought these types of data will provide more accurate outcome and these data will be easily labeled as well. Around 873 academic and professional data collected for this study. Further that, have used about 300 data for testing purpose shown in table-I and table-II.

TABLE I: TRAIN STUDENT DATA (QUANTITY)

Department	Amount
Science	419
Arts	234
Commerce	358

TABLE II: TEST STUDENT DATA (QUANTITY)

Department	Amount
Science	388
Arts	213
Commerce	286

D. 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. However, it is known, 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, have filtered some of noisy and incomplete data and tried to discard those data. Furthermore, had to cast floating value and convert them to statistical format. Data had to convert into various level in the convenient stage.

E. 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, have divided data and store them into two data folder for training and testing purpose. And also use here validation folder for check trained data validation.

F. Labeling Data

In this phase, renamed all of data as their name and also numbered them sequentially. Here also shown some of data column name in the following table III.

TABLE III. LABELED TABLE DATA IN SPSS

	Department	Gender	Year	Current_CGPA	Subject_Choice	SSC_HSC_Group	SSC_Bangla	SSC_English	SSC_Mathematics	SSC_Physics	SS
1	1	1	0	3.50	1	1	4.00	3.00	4.00	4.00	
2	2	1	1	3.25	2	1	5.00	4.00	4.00	4.00	
3	3	1	2	4.00	3	1	3.50	5.00	5.00	3.50	
4	5	2	4	2.75	0	1	4.00	4.00	5.00	4.00	
5	6	2	0	3.00	0	1	5.00	3.00	3.00	5.00	
6	9	1	3	3.40	3	1	5.00	5.00	5.00	5.00	
7	10	2	4	2.50	1	1	4.00	4.00	4.00	5.00	
8	1	1	0	3.50	1	1	4.00	3.00	4.00	5.00	

G. Data Storing

Data storing is a process of recording or storing information in a structured way in a storage medium. In this step, stored all of data in Google drive which make the task easier to use.

H. Statistics Analysis

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

TABLE IV. VARIABLES NAMING AND DESCRIPTING

Name of the Variables	Variable Description	Domain
Gender	Student Gender	{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}

IV. RESULT AND ANALYSIS

A. Experimental Setup

In model implementation and code implementation, collected the data first. Procedure given below:

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 02, February-2021

B. Model Selection Summery

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.

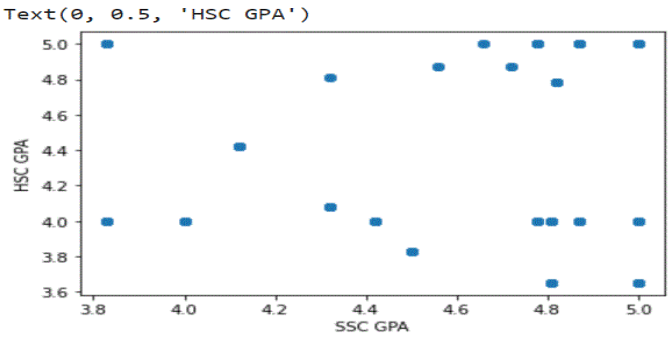

Fig. 2 .Data structure on KNN Model

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 is not appropriate for the enormous dimensional information. Here, 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 in fig. 2. Moreover, implemented KNN (K nearest neighbor) model on this data, it is 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 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.

C. Experimental Results and Analysis

During this project is being developed, have tested and analyze the 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 V. 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 90.05%. The accuracy level of all the algorithms are provided below in table VI.

TABLE V. VARIABLES NAMING AND DESCRIPTING

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 VI. COMPARISON OF CLASSIFIER TECHNIQUE WITH ACCURACY

Classifier Technique	Accuracy of the classifier
IBK(KNN)	90.05%
Naïve Bayes	65.13%
Multi Perceptron	68.42%
Logistic	69.73%
Random forest	67.76%

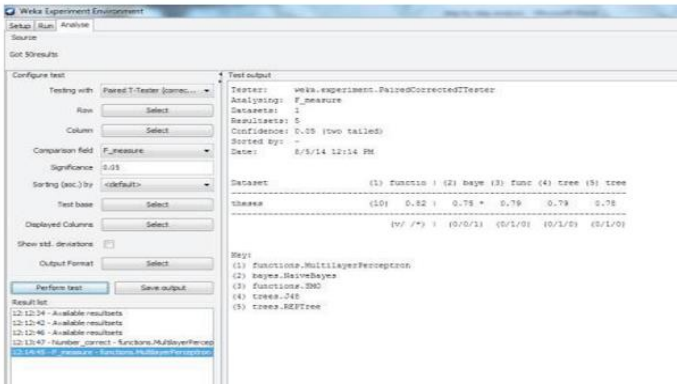

Fig. 3. Comparison of classifiers by using WEKA tool

Here, it is shown comparison results of all classifiers by using WEKA Experimenter has been shown in fig. 3. In this case, have also shown that IBK(KNN) algorithm shows performs best in this study among all other classifiers as well with F-Measure 82%. In the figure 4, shows the testing and validation accuracy of the SSC and HSC result from the whole data set. Moreover, the figure 5, shows the training and validation accuracy of the data set

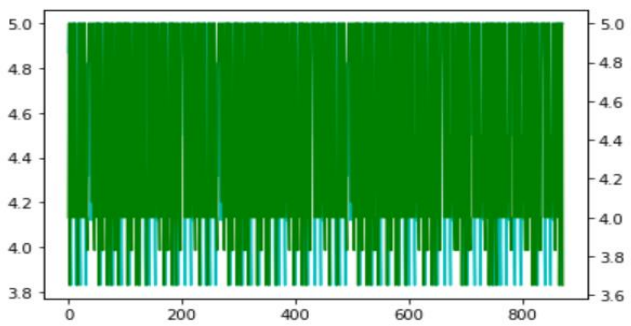

Fig. 4. Validating and testing for SSC and HSC result with whole dataset

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 02, February-2021

Fig. 5. Training and validating accuracy

In the fig. 4 and fig. 6, have represented all of data in a form of students' secondary and higher secondary Grade point with respect to whole dataset.

If it divide major 3 portion of Higher Education Subject

i. Engineering (Civil, EEE, Mechanical, CSE, Architecture).
ii. Science Subject (Physics, Chemistry, MBBS, Mathematics, Statistics).
iii. General Subject (Bangla, English, Economics, Geography, Sociology).

Now which Student will be admitted from above departments?

i. 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
ii. 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.
iii. If this student chooses Bangla or other general subject. Hope that thy will be success there ambition.
iv. When ones student good result in Physic, Chemistry and Biology then they can get Medical science subject to complete higher Study.

TABLE VII. 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

The table VII shows the students grading and admission in a faculty. For example, Student-01 received 5 grade from all the course like, Physics, Chemistry, Biology, Math, Bangla, English, then it he might possible to get the admission in Engineering Subject. Another example Student-4, they received 4 grade in Physics, 3 grade in Chemistry, 4 grade in Biology, 5 in Mathematics, 4 in Bangla, and 5 in English, it might possible to get their admission in general subject.

D. Discussion on Comparative Analysis

In this study uses classification techniques for prediction in the dataset of 887 students. Moreover it uses that dataset to analyze student's prior result sheet to predict future career as well. In this study, KNN (the K Nearest Neighbor) model performs best with 90.05% accuracy among all data mining classifiers. Therefore, KNN model has proved that it is potentially effective and efficient algorithm according to the dataset. The WEKA tool is also completed by comparing with all of classifiers by using that comparing platform. In that case, KNN model has also showed to be the best with measure of more than 90%.Hence the KNN model performance is comparatively much better than any other classifier algorithms as well.

V. CONCLUSION

It is always challenge to elicit paramount outcome in respect to prior existence. The matter of fact that the suitable department choosing is very complex phenomenon in respect to the targeted profession. This study completed this challenging jobs using WKEA tools by applying KNN methods and it is shown in the result discussion part the expected optimum outcome is more than 90 %. In the future it is anticipated to work with RNN (recurrent neural network) and deep learning to predict the suitable department in respect to the appropriate profession

REFERENCES

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

[2] Ardchir, S., Talhaoui, M. A., and Azzouazi, M., "Towards an adaptive learning framework for moocs", In MCETECH (2017).

[3] Bobadilla, J., Serradilla, F., and Hernando, A., "Collaborative filtering adapted to recommender systems of e-learning. Knowledge-Based Systems" Artificial Intelligence (AI) in Blended Learning, 22, 4 (2009), 261 – 265.

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

[5] 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.

[6] 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.

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

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

[9] 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

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

[11] Kalles D., Pierrakeas C., "Analyzing student performance in distance learning with genetic algorithms and decision trees", Hellenic Open University, Patras, Greece, 2004.

[12] Woodman, R., "Investigation of factors that influence student retention and success rate on Open University courses in the East Anglia region", M.Sc. Dissertation, Sheffield Hallam University, UK,2021.