"Analysis and manipulation of the data set obtained from the study of pre-primary education development using machine learning"

BY

Sushmita Laila Khan (10101008)

Nusrat Jahan Feroz (10101030)

SUPERVISOR

Zahidur Rahman

Signature:

Date:

Acknowledgement

We would humbly like to thank everyone who has helped in completion of this thesis work, for their advice, suggestion and help.

We cordially thank our supervisor Zahidur Rahman sir, our co-supervisor Zahingir Alom sir, for their endless support. We specially thank the department of Institute of educational development (IED) BRAC University for their support. This thesis would not have been possible without the data.

Finally we thank our beloved parents for their never-ending support, motivation and believe in us. We also would also like to specially thank our co-advisor Zahangir Alom sir, without whose guidance, assistance and encouragement this would not have been possible.

Abstract

A data set on ‘Preprimary education’ consisting of eleven hundred and fourteen data was collected from IED BRAC University. At the end of the study the students were tested out of eighteen and scored according to their performance .The entire data set was divided in three groups, and the results were classified in four parts. The first data set consisted of the socio economic attributes and the result obtained from the test, the second data set contained other likely attributes which could possibly have affected the results, like the education of parents, siblings, average education of student’s families etc. The third and final data set consisted of the marks they scored in each question and the net total mark obtained, to test their readiness. The results were classified into four groups called fully prepared, partially prepared, unprepared and needs help and they were categorized accordingly. An attribute relation file format, a format which is compatible with WEKA was made of each data set, before taking input. Once input was taken, these data sets were analyzed using several machine learning algorithms. The main goal of analyzing the data was to test whether or not the preprimary education prepares the students for primary education. The algorithms used in analyzing the data set were Super vector machine: Sequential Maximization Optimization, Multilayer perceptron: Back propagation algorithm, Naive Bayes algorithm, Random tree and random forest. The results obtained from each algorithm were compared and the algorithm performing the best was selected.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Machine learning being one of the most popular branches of artificial intelligence was used in this thesis to analyze and manipulate a set of data obtained on 'Preprimary education' from IED BRAC University. Several machine learning algorithms implemented in WEKA were used in the analysis of the data set.

1.1 Motivation

The idea of this thesis was driven by the desire of helping the underprivileged ones of this country. BRAC University IED department conducted the study 'Preprimary study' in September 2008. This study constituted of forty nine attributes, each attribute consisting of eleven hundred and sixteen data. After the statistical analysis of the data by SPSS, done by IED BRAC University, we sought them out and acquired this data set. To contribute to the deprived ones, understand the effect of the socio economic condition on their education, we examined how the preprimary education helped the students and whether this preprimary education contributed to the development of the students and helped them prepare for primary education and higher studies. Lack of motivation which is also a prime factor in the illiteracy rate of this country incentivized the direction of the analysis of the data set and the research of this study.

1.2 Literature Review

Analysis and manipulation of data using machine learning is very efficient and is gaining more and more popularity day by day. There are various types of data that has been analyzed using machine learning algorithm around the globe over the past years. However this data set is on 'pre- primary educational development'. A similar study has been conducted in India in 2010, but the data set was not analyzed using machine learning algorithms. But there were other studies which used machine learning algorithm to analyze the data and understand the pattern of the data set for example the paper on "Support vector machine classification and validation of cancer tissue samples using microarray expression", "Classification of cardiotocogram data using neural network based machine learning technique" and many more.

1.3 Research methodology

The aim of this research was to test how well the data set can be analyzed using various algorithms, the performance of each algorithm and thus selecting the best fitted algorithm. However, before the algorithms were applied the raw data set was analyzed and manipulated rigorously in order to achieve accuracy of the highest degree possible. The entire data set consists of 49 attributes, and each group comprised eleven hundred and sixteen data. The results obtained from the test were then scaled and classified into four groups to evaluate the students as per their results. The data set was divided into three groups, each group testing the effectiveness of a range of similar data. Group one consisted of eleven attributes. Of the eleven attributes ten of them were the questions from the questionnaire which was given to test the students. The eleventh attribute was the result obtained from the test. This data set was examined to assess to effect of preprimary education on student and study their readiness. The second data set, called the 'Factors affecting readiness's comprised of seven attributes. The first six attributes were the factors like education of father, mother, average qualification of family head etc. The seventh

factor once again was the result. In this data set the influence of the environment in which each child resided was inspected with respect to the result of each student. The third and final data set called the 'Effect of economy' had a total of eight attributes. The first seven attributes were the factors that gave an over view of the economic condition if the family like occupation of the family head, monthly/yearly income of the family, whether the student had a study table or not etc. The eighth and final attribute was the result of the students. The effects of the economic situation of on the result of the results were studied in this data set. However before each data set was analyzed using machine learning algorithms, they were preprocessed in a format compatible with WEKA called attribute relation file format. They were then taken input into WEKA, individually. Six different algorithms were applied on each data set and the results observed and compared. The algorithms used were the following:

- Naive Bayes algorithm
- Sequential minimal optimization
- Multilayer perceptron algorithm
- Decision Tree
- Random forest

These algorithms were applied on each data set, and the results of each data set were compared and the best one was selected. The time taken to analyze the data sets by each algorithm was compared and a bar chart for each data set was made to help choose the most efficient algorithm.

1.4 Outline

This thesis is comprised of four chapters, including this chapter. Chapter two consists of elaborate description of the data set, the questionnaire and methods of data acquisition. This chapter also involves the reasons and methods via which the data was preprocessed. Finally this chapter covers the algorithms used in the analysis of the preprocessed data and the grounds on which these algorithms were chosen. Chapter three talks about the results obtained from the analysis and then discusses about the results, exploring the why's and how's, and finally compares the obtained results. The last and final chapter of this paper, chapter four constitutes of the conclusion drawn as a result of this study and covers materials and ideas for future prospect of this study.

2.0 System Implementation

The data set has been collected from institute of educational development, BRAC University. The data set consists of 1111 raw data, which will be processed, analyzed, a result will be derived and a conclusion will be decided.

2.1 Data Collection

For this thesis work, we approached IED BRAC University for the data set. They provided with data from their study on 'Pre-Primary development'. This study was conducted in 2012. Before handing over the data certain terms and conditions were given and asked to maintain. The terms

and conditions dictated not to use the data or share the data in any purpose but the thesis work. The data set consists of 29 different attributes which are represented in codes and a questionnaire. Each attribute has property of its own and has a contribution to the study. The questionnaire reflects the effect of preprimary development. The correct answer of a question scored 2 marks, partially correct 1 mark and an incorrect answer is scored 0. These marks show exactly how much development has been achieved.

2.11 Attributes and their contribution

(i) Stratum

The attribute stratum is represented by the code STRID. The division of country into different parts are called stratum. They are Rural Dhaka, Rural Chittagong, Rural Khulna, Rural Rajshahi, Rural Barishal, Rural Sylhet, Rural Rangpur and urban Bangladesh.

(ii) District (Jela)

District is a type of administrative division in some countries managed by local government. The country is divided into eight parts. The environment of each district varies from each other. Depends on the varying results in different districts also imply how the geography and environment effects the pre-primary development.

(iii) Sub District (Upojela)

Sub district is the division of a large city into smaller parts for ease of identification. This study was conducted in 5 sub districts of every district. In each district they approached five schools possibly BRAC schools since this study has been conducted by BRAC. Each school was given a serial number, and the children of the schools were taught basics and then a questionnaire was handed out to each of the students. The result of their performance was recorded. This study was conducted in every sub districts. The results of the study in every sub district will be compared to see how people inhabiting in various sub districts perform to this test. The varying results in different sub districts also imply how the geography and environment effects the pre-primary development. The sub districts and the school were chosen randomly and in an unbiased fashion. Sub-district is represented by the code UPID.

(iv) School

In this study school is represented by the code SCID. Schools were randomly selected from the sub districts and then they were each given a serial. School can also have a major contribution in this study. It is obvious that different school will yield different results. However the reason behind his varying results is what matters. After the test has been conducted and the results been obtained, they are compared. The first comparison is between the results of the schools from the same district and then the results are compared with the results from different district. School is the first step on education. How a school nourishes and teaches their student will definitely be reflected in the results of this study. This is because if students are taught well they will perform well. Hence the role of pre-primary development on children is significant as they respond as how the school teaches to.

(v) Name of the Child

Name of student is represented by the code STID. Needless to say name will definitely not have any effects what so ever on the development, growth or the performance of the student. Names are given for identification purposes. However it is possible for two different students to have the same there. Therefore to avoid confusion and problem, each student has been assigned a 'number' and throughout the study they have been referred to as the numbers. This simplifies the process and adds ease while conducting the study.

(vi) Age of the Child

The attribute age of the child is represented by the code STAG and recorded in months. For example in the dataset the age of student 1 is 84months which means the student is seven years old. Unlike the name age has had a significant effect on the performance of the students. The maturity and the development of brain depend on the age, in general cases. For example a nine year old will definitely be matured than a seven year old. The ability to concentrate and have a grasp on what is being taught depends on brain development and thus the age. Once something is taught, how well a student perceives it and understands the situation is also effected by his or her age. Therefore it can be easily said that the relationship between age and performance is directly proportional. The table shows how actually age effects pre-primary development:

Age	2002		2003		2004		2005		2006		2007
(Year)	PP	NPP	PP	NPP	PP	NPP	PP	NPP	PP	NPP	PP
4-5	25.9	20.4	21.8	18.9	18.9	18.3	26.4	25.9	27.0	27.0	10.6
6	45.1	49.1	52.6	45.7	51.0	41.1	52.3	49.1	54.2	51.2	63.6
7	24.9	23.9	17.1	22.6	20.3	28.6	15.4	16.2	15.6	14.3	20.8
8+	4.0	6.6	8.5	12.8	9.8	12.0	5.9	8.9	3.3	7.4	5.0

Note: PP = having pre-primary orientation, NPP = without pre-primary orientation

Table: 1

(vii) Sex of the Child

STSX code is used to represent the attribute sex of the children. Code for boys is 1 and code for girls is 2. It is not likely that sex will influence their performance; however in countries like Bangladesh it might have an effect. As parents are not willing to let their daughters study more often than not results in them not being supportive and disapproving of what they do. Thereby the girls will get demoralized and she might even lose interest and give up. Therefore sex can actually have a remarkable effect. Statistical analysis result of performance of different sex:

Issues	Boys	Girls	Both	Level of significance
Average number of correctly answering items	20	19	19.4	$p<0.001$
% of correctly answering items	80.0	76.0	77.8	$p<0.001$

Table: 2

(viii) Did the child attend pre-primary school?

This attributes code is PBST. This attribute has seven sub attributes. They are:

- Attended a particular school: Represented by code 1. The school the child has been to will affect the pre-primary development since the grooming from school matters
- Attended a particular NGO (Non-Government Institution): Represented by code 2. This attribute examines whether or not going to a NGO school or a government school has any impact on pre-primary development.
- Did not complete the particular school: Represented by code 3. This sub attribute can either mean the student did not finish his/her pre-primary education, or the student has changed that school and went to a new one. In this case how changing school can effect pre-primary development can be studied and seen if there is any at all. In case the student hasn't finished school this effect can also be seen in their test scores.
- Did not complete the particular NGO School: Represented by code 4. This sub attribute has portrayed the difference in results between students who has not finished or changed a normal school and those who has finished or changed a NGO school.
- Attended a private school: Represented by code 5. From the results of this sub attribute the differences in results of students studying in private schools can be compared with others. The results of the conclusion will give a clear understanding which school prepares their student best and where the pre-primary growth of the children is maximized.
- Did not complete the particular private school: Represented by code 6. This sub attribute shows the consequence of not completing a particular school or changing schools. If this value is compared to those of the children who dropped out from government schools or NGO the impact of different school on the children can be seen once again.
- Did not study anywhere: Represented by code 7. This can either mean the student has never studied ever in his/her and knows absolutely nothing or the student has been taught at home. However if they are even taught at home this sub attribute and it results will give an understanding of how schooling effect children's pre-primary development and its importance.

Grade	No. of Items	PP orientation					
		PP			Non-PP		
		Boys	Girls	Sig.	Boys	Girls	Sig.
Class I	8	6.5	6.5	ns	5.8	6.3	P<.01
Class II	11	6.0	5.6	ns	5.7	5.9	ns
Class III	10	5.8	5.8	ns	5.5	6.0	ns
Class IV	10	5.0	5.0	ns	4.9	4.9	ns
Class V	11	7.6	7.6	ns	7.4	7.1	ns

Table: 3

The above table shows the difference of results due to preprimary education. From the table we have seen that students who has been to preprimary education yielded a better result.

(ix) The number of the child's house

The attribute child's house number is represented by the code STHN. Every child's house has a specific number. This is important for counting the number of child.

(x) Name of the head of the family

The name of the head of the family for every child is saved in the data table. This attribute is important to identify child's guardian.

(xi) House head's occupation code

This field is represented by the code HHOC. From occupation code we can know the house head's occupation. And from that we can know the economic condition of the family.

(xii) Amount of land of each household

In this study amount of land of each household is represented by LOHH. From this data we can calculate the approximate affluences of each household.

(xiii) The earner's total income throughout the year

This field is represented by the code HHMI. The total income of the earning member of the family is recorded. This is an important contribution in this study. This attribute shows how money effects the education, in this case the pre-primary education. From the results obtained, the effect of economy of pre-primary development and the contrast of how economy actually effects can be seen and understood.

(xiv) The same source of income

This attribute is represented as HHMSI. Each source of income is denoted by a number and there are sixty eight different income source. For example the number 8 stands for doctor, number 33 for barber and so on. The result of the test from each different family is analyzed and observed and compared to see how the results of different families with different source of income vary.

(xv) Highest level of education of the earner

The code HHED has been used to represent this attribute in the data table and the level of education is defined by numbers. For example number 17 means the person has completed his/hers master's degree, number 99 means the person has never been admitted anywhere, number 88 means the persons educational qualification is not known and so on. The level of education and income in most cases are proportional. But in this case educational level of the earner, his/her income and hence the effect on pre-primary development will be tested.

(xvi) Education of the Father

This attribute is represented using the code STFED. Fathers from different family have different education level. The educational level is recorded, thus the kind of job they do and the amount the earn and its impact on the student's pre-primary development

(xvii) Education of the Mother

STMED is used as a code for this attribute. The mother's educational level is inquired and recorded. Usually in families from sub districts the females of the family do not contribute in the overall income as they don't work. However if some does work, or if they are the head of the family is the main source of income their job and the amount of money they earn is recorded.

This attribute is examined and assessed to make out how the mother's educational level influences the child's pre-primary development. This can also be analyzed if the mother is working and earning and thus increasing the overall income how that results in the pre-primary development.

(xviii) The child's siblings are studying or not

This attribute is represented by EBSUC. Sibling's educational level is inquired and recorded. This can analyzed if the child's brothers and sisters are educated. In which circumstances the child's are growing up.

(xix) What is the total number of family member in that Household?

Total number of members in a family will have direct effect on the economic condition of the family. If there are more members the economy will not be high and if there are less members the economy will be relatively higher. Thus if economy is stronger, the families will have liberty to spend on the children's education. This will usually mean that the pre-primary development of the student will be better because he/she has been sent to good school and has been provided with sufficient resources for studying.

(xx) Does this Household have newspaper delivered regularly?

DNHH is used as a code in this case. Having newspaper delivered regularly reflects their affluences. Not a single family of the first 50 gets newspaper delivered. This data set is secondary and is mainly for test and control purpose. Nevertheless this data will be evaluated to see if electricity has any or very minimal effect or no effect at all.

(xxi) Does this Household have electricity Supply?

EHH is used as a code. Answer yes is represented by 1 and no by 2. From the data set it is seen that 3/50 families has electricity only. Although this is not a very important attribute and neither will it be used to conduct this study, however this data set is secondary and is mainly for test and control purpose. Nevertheless this data will be evaluated to see if electricity has any or very minimal effect or no effect at all. Basically having electrical supply shows the affluences of the family in concern.

(xxii) Does this Household have Radio?

RHH is used as a code. Answer yes is represented by 1 and no by 2. From the data set it is seen that 1/50 family has radio only. Radio provides a mean of communicating with the outer world, and also a source of entertainment. Although this is not a very important attribute and neither will it be used to conduct this study, however this data set is secondary and is mainly for test and control purpose. Nevertheless this data will be evaluated to see if having radio has any or very minimal effect or no effect at all.. Basically having a radio shows the affluences of the family in concern.

(xxiii) Does this Household have Television?

TVHH is used as a code. Answer yes is represented by 1 and no by 2. From the data set it is seen that 4/50 families has radio only. Television provides a mean of communicating with the outer

world, helps in gaining knowledge if used correctly, and is also a source of entertainment. Although this is not a very important attribute and neither will it be used to conduct this study, however this data set is secondary and is mainly for test and control purpose. Nevertheless this data will be evaluated to see if having Television has any or very minimal effect or no effect at all. Basically having televisions shows the affluences of the family in concern.

(xxiv) Does this Household have Mobile phones?

MPHH is used as a code. Answer yes is represented by 1 and no by 2. From the data set it is seen that 29/50 families has radio only. Mobile phone provides a mean of communicating with the outer world, and is sometimes a source of entertainment. Although this is not a very important attribute and neither will it be used to conduct this study, however this data set is secondary and is mainly for test and control purpose. Nevertheless this data will be evaluated to see if having mobile phones has any or very minimal effect or no effect at all. . Basically having mobile phones shows the affluences of the family in concern. However mobile phones have become really cheap and having mobile phone cannot determine the wealth of a family.

(xxv) Does the child's have reading table?

STRT is used as a code for enquiring if the child has a reading table. Answer yes is represented by 1 and no by 2. If the child has a reading table then he/she can study more attentively.

(xxvi) Did the child need any expenditure for pre-primary school?

CPED is used as a code. It is for if the child need any expenditure for pre-primary school. Answer yes is represented by 1 and no by 2.

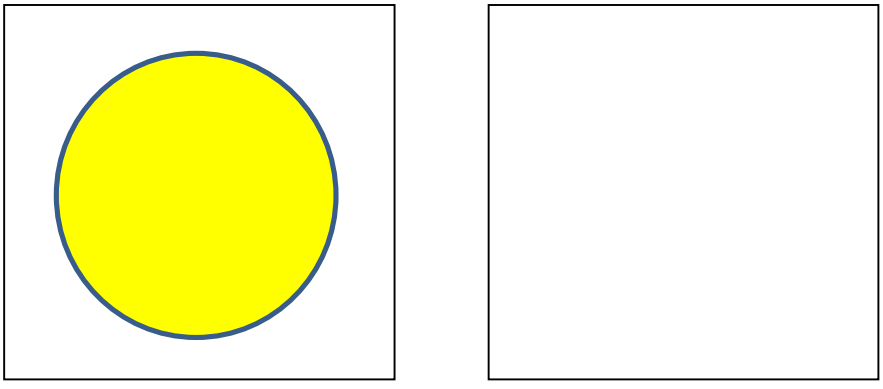
(xxvii) How much did each cost?

For admission fee, the code is CPAD. For tuition fee, the code is CTF. For Books, the code CFB. For other expenses, the code is COM. For house tutors, the code is CHT. For tiffin, the code is CFT. These attribute is used to know about the expenditure that are caused by different purpose.

2.12 Questionnaire, its Importance and Effects

The questionnaire had 10 questions and marks allocated for answering each question. For answering correctly the student will score 2 marks, partially correct answer will yield in 1 mark and an incorrect will yield zero mark.

(i) Question 1: Draw a similar circle in box2. (Question represented by code cc)



If the students managed to follow the instruction and understand it correctly and draw a perfect square like the one in box 1 he/she has been scored 2 and otherwise 1. If the student didn't manage to draw the circle at all he/she had been awarded zero. This question tested their ability on understanding shapes. From the data set it has been seen that 18/50 scored 2 marks, only three scored 0 and the rest scored 1.

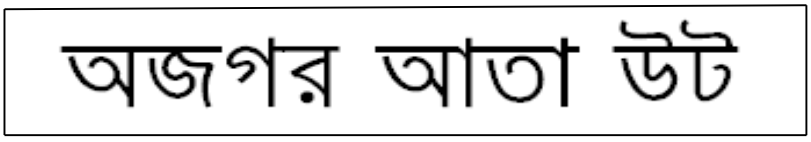
(ii) Bengali alphabets (Shoroborno) were written in the following box and the students were asked to fill in the missing alphabets. (Question represented by code WAB)

This question was set to try to figure the students' knowledge of bangle alphabets and to test whether or not they knew it thoroughly enough to fill in missing spaces and understand which alphabet comes after what. From the data set, 37/50 scored a full mark of two out of 2 in this question. Only 4/50 scored zero. And the remaining scored one.

অ	আ		ঈ	
উ		ঋ	এ	ঐ

(iii) Draw a circle to identify "আ". (Question represented by code IALW)

This question tests whether the students can identify the shoroborno "আ" in words. None of the first 50 student was able to identify the desired alphabet correctly. In fact not just the first 50 but no student from the data set of 1111 data managed to answer this question to perfection. However 43/50 students managed to answer it partially and hence scored 1.



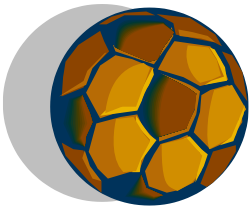
(iv) Draw a circle around 'আম' from the sentence below to identify it. (Question represented by IAWS)

This question tests whether the students can identify the word 'আম' in the sentence. None of the first 50 student was able to identify the desired alphabet correctly. In fact not just the first 50 but no student from the data set of 1111 data managed to answer this question to perfection. However 46/50 students managed to answer it partially and hence scored 1.

(v) Write your name in the box below. (Question represented by code WYN)

This question tests whether the students can write their name correctly. 30/50 students managed to answer it to perfection and score a perfect two.

(vi) Identify the diagram below and write its name in the next box. (Question represented by the code WNOP)

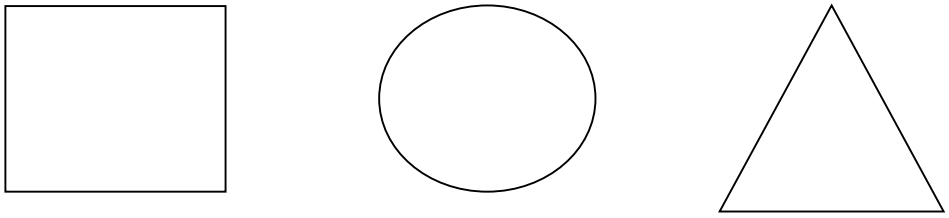


This question tests whether the students can identify and write the name of the figure correctly. 34/50 students managed to identify the figure and spell it correctly hence scoring 2 marks

(vii) Tick the taller person. (Question represented by ITO)

None of the students managed to answer this question to perfection. However, 46/50 answered this question partially hence scoring one.

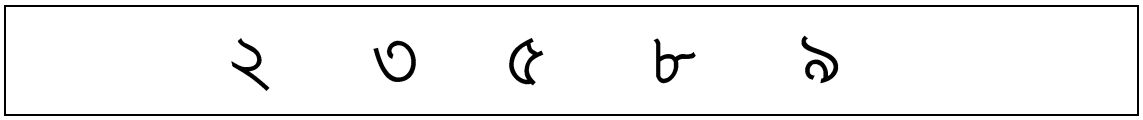
(viii) Identify the triangle and tick the correct answer. (Question represented by code ITGS)



None of the students managed to answer this question to perfection. However 45/50 answered this question partially hence scoring one. Partially correctly implies not being able to tick the answer properly but somehow explaining they know which figure is the triangle.

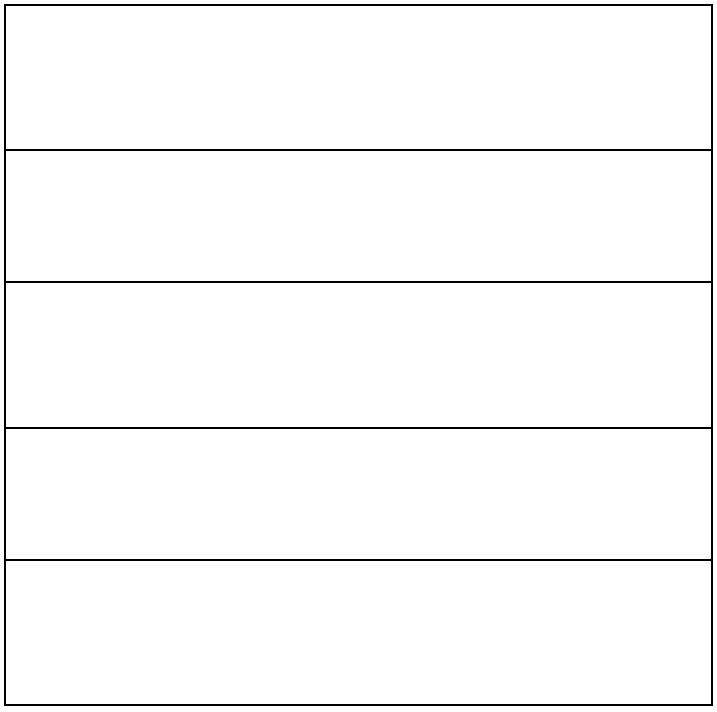
(ix) Which of the following number is a bangle eight? Circle around the eight. (Question represented by code IANGN)

None of the students managed to answer this question to perfection. However 48/50 answered this question partially hence scoring one. Partially correctly implies not being able to tick the answer properly but somehow explaining they know which number are eight.



(x) Write one to five in Bangla in the boxes bellows. (Question represented by code WN)

46/50 answered this question to perfection being able to write one to five in bangle correctly.



Institution of Educational Development

Social and Economic Condition of the Family

Serial Number	Question	Answer/Code/ID	
1.	Stratum	STRID	
2.	District		
3.	Sub-District	UPID	
4.	School	SCID	
5.	Name of the child	STID	
6.	Age of the child	STAG	
7.	Sex of the child Code: STSX	Boys	1
		Girls	2
8.	Did the child attend pre-primary school? Code: PBST	Attend a particular school	1
		Attend a particular NGO	2
		Did not complete school	3
		Did not complete NGO	4
		Attended a private school	5
		Did not complete the particular private school	6
		Did not study anywhere	7
9.	Child's house number	STHN	
10.	Name of the head of the house		
11.	Profession of the house head	HHOC	
12.	Amount of land owned by the house head	LOHH	
13.	Total income of all earners of a family throughout the year	HHMI	
14.	Source of income of a house	HHMSI	
15.	The economic condition of the house in the last one year Code: ECONC	All time shortage	1
		Sometimes shortage	2
		Equal	3
		Sometimes	4
		All time	5
16.	Highest level of education of the head	HHED	
17.	Education of the father	STFED	
18.	Education of the mother	STMED	
19.	Average education level of the house	HAED	
20.	Does the sibling's attend school? Code: EBSUC	Yes	1
		No	2
21.	Total number of family member's	HTNM	
22.	Does the household have newspaper supply? Code: DNHH	Yes	1
		No	2
23.	Does the household have electricity supply? Code: EHH	Yes	1
		No	2
24.	Does the household have radio? Code: RHH	Yes	1
		No	2

25.	Does the household have television? Code: TVHH	Yes	1
		No	2
26.	Does the household have mobile phones? Code: MPHH	Yes	1
		No	2
27.	Does the child have reading table? Code: STRT	Yes	1
		No	2
28.	Did the child need any expenditure for pre-primary school? Code: CPED	Yes	1
		No	2
29.	How much did each cost?	Admission Fee	CPAD
		Tuition Fee	CTF
		Book	CFB
		Other Expenses	Com
		House Tutor	CHT
		Tiffin	CFT

2.2 Preprocessing of data set

Institute of educational development BRAC University has conducted the investigation to explore how preprimary education helps school readiness of children. From their reports the result obtained stated that preprimary education did help in the readiness of the students. This investigation will serve as the test case in this study. The table below shows how preprimary education influences in better performance of the student.

Grade	No. of Items	PP orientation					
		PP			Non-PP		
		Boys	Girls	Sig.	Boys	Girls	Sig.
Class I	8	6.5	6.5	ns	5.8	6.3	P<.01
Class II	11	6.0	5.6	ns	5.7	5.9	ns
Class III	10	5.8	5.8	ns	5.5	6.0	ns
Class IV	10	5.0	5.0	ns	4.9	4.9	ns
Class V	11	7.6	7.6	ns	7.4	7.1	ns

Table: 4 This table clearly indicates that students who had preprimary education yields better result.

2.21 Classification of dataset into groups and evaluation of students

In order to input the data into WEKA, and to apply machine learning algorithms on them the data has to be in a specific format, known as the Attribute relation file format (ARFF). The marks obtained from the test that were obtained and scaled accordingly. The students were examined on a scale of eighteen and then classified accordingly. Questions one, two, five and six were scored out of two, questions three, four, seven and nine were scored out of one and question ten was scored out of 5. The marks obtained by each of the students were dependent of the following factors:

- Question 1:
  - ✓ Complete and correct: 2
  - ✓ Incomplete and partially correct: 1
  - ✓ Incorrect: 0
- Question 2:
  - ✓ Complete and correct:2
  - ✓ Incomplete and partially correct:1
  - ✓ Incorrect: 0
- Question 3:
  - ✓ Correct: 1
  - ✓ Incorrect: 0
- Question 4 :
  - ✓ Correct: 1
  - ✓ Incorrect: 0
- Question 5:
  - ✓ Complete and correct:2

    - ✓ Incomplete and partially correct: 1
    - ✓ Incorrect: 0
- Question 6:
    - ✓ Complete and correct: 2
    - ✓ Incomplete and partially correct: 1
    - ✓ Incorrect: 0
- Question 7:
    - ✓ Correct: 1
    - ✓ Incorrect: 0
- Question 8:
    - ✓ Correct: 1
    - ✓ Incorrect : 0
- Question 9:
    - ✓ Correct : 1
    - ✓ Incorrect: 0
- Question 10:
    - ✓ Successfully wrote one to five: 5
    - ✓ Missed one number:4
    - ✓ Missed two numbers:3
    - ✓ Missed three numbers:2
    - ✓ Missed four numbers:1
    - ✓ Did not manage at all: 0

Once the whole process of correcting and hence giving marks were complete, the students were evaluated according to the marks that they obtained. Students were divided into four different groups depending on their marks, as per following:

- Above 15: Fully prepared(FP)
- Above 10: Partially prepared(PP)
- Above 5: Needs Help(NH)
- Below 5: Unprepared(UP)

After completion of categorizing the students according to the marks that they obtained, the data set was divided into three groups. Group one tested the readiness of the students on the basis of their test results. Group one consisted of eleven attributes, of which the first ten was the questions of the test and the eleventh attribute, the result of the test.

Group two examined how the external factors like parents educational back ground, siblings education etc. affected the readiness of the students for this test. Group two consisted of the seven following attributes:

- Maximum qualification of the student's family head(HHED)
- Student's father qualification(STFED)
- Student's mother qualification(STMED)
- Average qualification of family members(For families with seven plus members)(HAED)
- Did the student's sibling/s attend class

- Total number of family members

Group three examined how the socio economic factors of a family affected the result of the students. This group comprised of eight attributes, each reflecting their effects on the student's results. The attributes are:

- Total monthly/yearly income of the family
- Family head's source of income
- What was the financial condition of the family in the last one year
    - ✓ Always broke
    - ✓ Broke during some parts of the year
    - ✓ Not broke but not well off either
    - ✓ Sometimes solvent
    - ✓ Always solvent
- Does this family have electricity supply?
- Does this family have television
- Does the student have a study table
- Did the student's preprimary education require any expense

Once the data set was divided into three groups, the students classified as per their result, and then these groups were further analyzed for input into WEKA before the machine learning algorithms could be applied.

2.22 Attribute relation file format

After classification of the data in three different groups, to examine their effect on the readiness of the students, each group of data was preprocessed, before they were taken as inputs in WEKA. The raw data's were preprocessed into a format called the attribute relation file format, which is a compatible input format for WEKA.

The attribute relation file format has two distinct sections. The first section is the **HEADER,** followed by the data section. The header segment consists of the name of the dataset, the @relation line and the information of the attributes, the @attribute line.

ARFF Header section

- **The @relation declaration**

The name of the relation is defined as the first line in the ARFF file. The format is:

@relation<relation name>

- **The @attribute declaration**

Attribute declarations takes place in the form of ordered sequence of **@attribute** statements. Each attribute in the data set has its own **@attribute** statement which uniquely defines the name of that attribute and its data type. The order the attributes are declared indicates the column position in the data section of the file. For example, if an attribute is the third one declared then

Weka expects that all that attributes values will be found in the third comma delimited column. The format for @attribute statement is:

@attribute <attribute-name> <data-type>,

Where the attribute name must start with an alphabetic character and if spaces are to be included, then it has to be enclosed by a quotation.

The data types used in this thesis, which are also supported by WEKA are the following:

- ✓ Numeric data type: numeric attributes can be real or integer numbers
- ✓ Boolean data type: Boolean attributes are either true or false depending on the values.

For example @attribute education_of_mother real

@attribute average_education_of_family_sevenplusmembers_haed numeric

@attrirbute did_the_sibling_study_in_higher_class_ebsuc {TRUE,FALSE}

ARFF Data section

The @data is a single line denoting the start of the data segment in the file. The format is:

@data

The instance data

Each instance is represented on a single line, with carriage returns which denotes the end of an instance. Attribute values for each instance are separated by commas. It is mandatory that the values appear in the order the attributes were declared in the header section. Missing values are represented by a question mark. The Boolean values are case sensitive, and it has to be ensured that if the Boolean values are declared in upper case in the header section, they remain in upper case in the @data section as well.

Example of an ARFF file format

HEADER
@relation weather.symbolic
@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

DATA SECTION:
@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes

rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no

2.23 ARFF format of data set

- **Data set one: Socio economic factors affecting readiness**

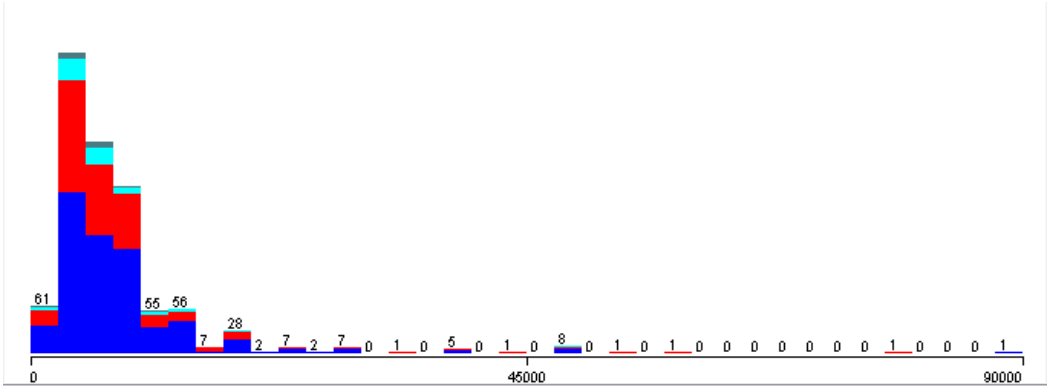


Figure 2.23a

- **Data set two: Factors affecting readiness**

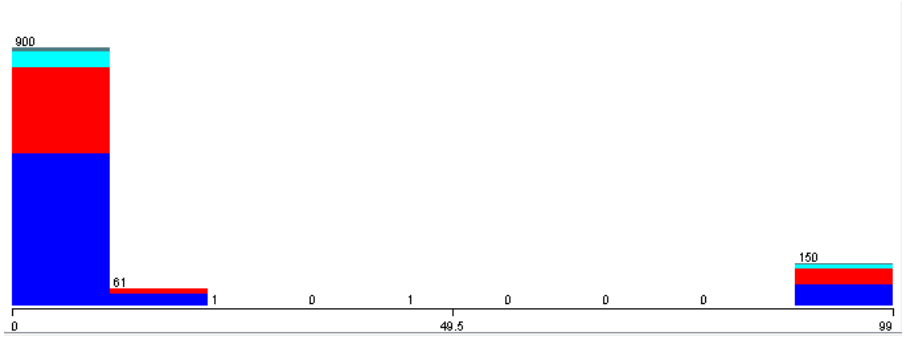


Figure 2.23b

- **Data set three: Readiness of students**

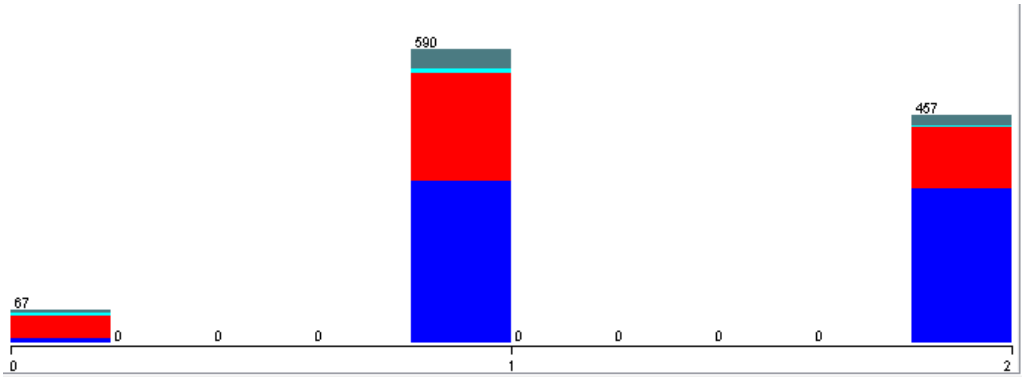


Figure 2.23c

2.3 Algorithm details

The data set was made to undergo rigorous preprocessing. Once the preprocessing was complete, machine learning algorithms were applied on the data set. The algorithms support vector machine: sequential minimization optimization, Multilayer perceptron algorithm, Naïve Bayes algorithm, random tree and random forest were used in the analysis.

2.31 Super vector machine: Sequential Minimization Optimization

A Support Vector Machine (SVM) is a differential classifier previously defined by a separating hyper plane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples. This raises the question of as to how this hyper plane is drawn and how is it made optimal. Assuming there was a given set of 2D data, as shown in the graph below; we attempted to classify the training data set with the hyper plane.

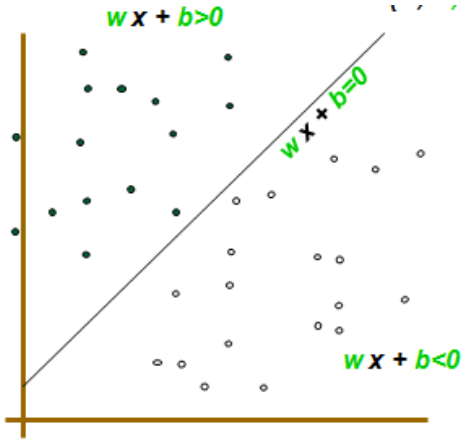


Figure 2.31a

The ‘w’ in the equation denotes the gradient of the line while ‘x’ represents each training data set. The point/s closest to the hyper plane are called super vectors. However separating just once with one line is not sufficient since there might have been certain misclassification. So the gradient, ‘w’ of this line was adjusted several times till there were minimal mistakes as shown in the figures below

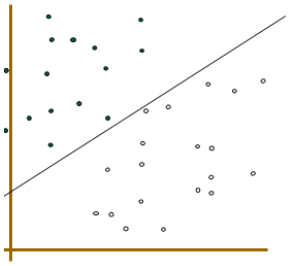


Figure 2.31b

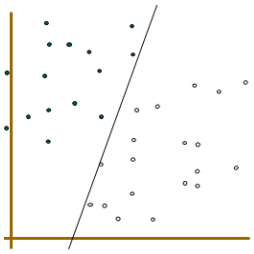


Figure 2.31c

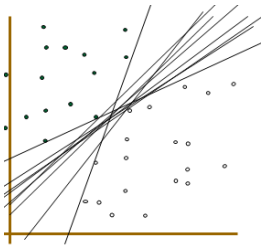


Figure 2.31d

As we can see, there are multiple hyper planes which attempted to classify the data accurately. It was necessary to determine which line classified the training data set to nearest perfection. If a line passes too close to the points then it will be too noise sensitive. So the optimal hyper plane will be the one which will be as far possible from the training data set, but cannot misclassify. The operation of the SVM algorithm is based on finding the hyper plane that gives the largest minimum distance to the training examples. This distance is called the margin**.** According to intuition and PAC theory, maximizing the margin is good. It also implies that only the support vectors are good and it works well empirically.

Maximum margin formalization

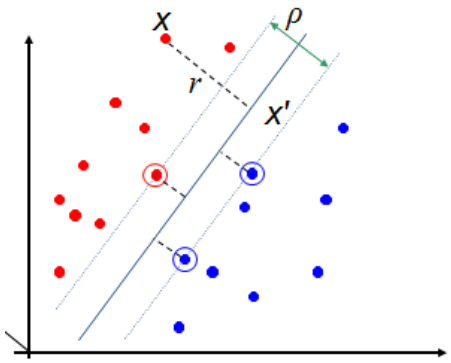


Figure 2.31e

- The distance from the example to the separator is $r = y\dfrac{\mathbf{w}^T\mathbf{x}+b}{\|\mathbf{w}\|}$ (equation 2.31a)
- Margin ρ of the separator is the width of separation between support vectors of classes.

It was assumed that all data is at least at distance 1 from the hyper plane. Then the following two constraints were followed for a training set $\{(\mathbf{x_i}, y_i)\}$:

$\mathbf{w^T x_i} + b \geq 1$ if $y_i = 1$ (equation 2.31b)

$\mathbf{w^T x_i} + b \leq -1$ if $y_i = -1$ (equation 2.31c)

For support vectors each inequality becomes equality. As each examples distance from hyperplane is $r = y\frac{\mathbf{w}^T\mathbf{x}+b}{\|\mathbf{w}\|}$,(equation 2.31d) the margin is,

$$\rho = \frac{2}{\|\mathbf{w}\|}$$ (equation 2.31e)

Finally, the problem of maximizing M is equivalent to the problem of minimizing a function subject to some constraints. The constraints model the requirement for the hyper plane to classify correctly all the training examples $\mathbf{x_i}$.

Minimize: $\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$, and for all $\{(\mathbf{x_i}, yi)\}$: $y_i(\mathbf{w^T x_i} + b) \geq 1$ (equation 2.31f)

Solving the optimization problem

This data set was analyzed using **Sequential Minimal Optimization**, as the data was classified into four groups. This called for an explanation as to how to solve the optimization problem. SMO is a simple algorithm which quickly solves the SVM quadratic problem without even adding an extra matrix. SMO decomposes the overall quadratic problem into quadratic sub problems. SMO chooses to optimize the smallest possible optimization problem at every step. The smallest possible optimization for a standard SVM QP requires two Lagrange multipliers (α_i) since they must always obey a linear equality constant. At every step, SMO chooses two multipliers which jointly optimizes, finds the maximum values for these multipliers and then updates the SVM to reflect new optimal values. These values are saved and used for reprocessing till the optimal hyperplane is found.

The advantages of using SMO:

- Solving for two Lagrange multipliers can be done analytically. Thus an entire inner iteration due to quadratic problem can be avoided.

- SMO does not require extra matrix storage. Thus very large SVM training program can be stored in personal computers or work station.

There are three components to SMO:

Analytic method to solve for the two Lagrange multipliers.

- Heuristic to choose which multipliers to optimize.
- A method for computing the threshold

In this paper we have discussed about the Lagrange multiplier approach in order to solve the optimization problem.

The above equations (2.31f) are optimizing a quadratic function subject to *linear* constraints. The solution involves constructing a dual problem where a Lagrange multiplier α_i is associated with every constraint in the primary problem. For example:

Find $\alpha_1 \ldots \alpha_N$ such that

$\mathbf{Q}(\boldsymbol{\alpha}) = \Sigma\alpha_i - \frac{1}{2}\Sigma\Sigma\alpha_i\alpha_j y_i y_j \mathbf{x_i}^T\mathbf{x_j}$ is maximized and

(1) $\Sigma\alpha_i y_i = 0$ (equation 2.31g)

(2) $\alpha_i \geq 0$ for all α_i (equation 2.31h)

The solution to this optimization problem has the following form:

$\mathbf{w} = \Sigma\alpha_i y_i \mathbf{x_i}$ (equation 2.31i) $\quad b = y_k - \mathbf{w}^T\mathbf{x_k}$ for any $\mathbf{x_k}$ such that $\alpha_k \neq 0$ (equation 2.31j)

Each classifying non-zero α_i indicates that the corresponding $\mathbf{x_i}$ is a support vector. Then the classifying function will have the following form:

$f(\mathbf{x}) = \Sigma\alpha_i y_i \mathbf{x_i}^T\mathbf{x} + b$ (equation 2.31k)

Solving the optimization problem involved computing the inner products $\mathbf{x_i}^T\mathbf{x_j}$ between all pairs of training points.

In this thesis, we have used Linear SVM for analysis. However for data sets where the data cannot be linearly classified, non linear SVM's are used. Another way to classify to data with

noise is to use soft margin-hard margin. Although it is not very effective with data sets with extreme noise level. When the data set is too noisy and requires non linear SVM to separate them a kernel is used. Kernel is a mapping done to the training data to improve its resemblance to a linearly separable set of data. This mapping consists of increasing the dimensionality of the data and is done efficiently using a kernel function. There are several types of kernel but the most widely used ones are the polynomial kernel and Gaussian kernel. Since this thesis was concerned with linear SVM only, no kernel or in oher words linear kernel was used.

2.32 Multilayer Perceptron

The multilayer perceptron (MLP) is a simple model of biological neural networks and is based on the principle of a feed-forward-flow of information. The network is planned in a hierarchical manner. Multilayer perceptron is very popular in both areas of application and theoretical research.

The concept for MLP was inspired from the neuron networks of human brain and how they work. The interconnected neurons of the human brain transmit information through each other and through their synapses. Neural network is made up of artificial neurons, although they have little to no similarity as to how the human brain works. Feedforward is the most widely used and efficient way of connecting the neurons.

MLP was mainly used for classification. The MLP consists of different layers where the information flows only from one layer to the next layer. Layers between the input and output layer are called hidden layers.

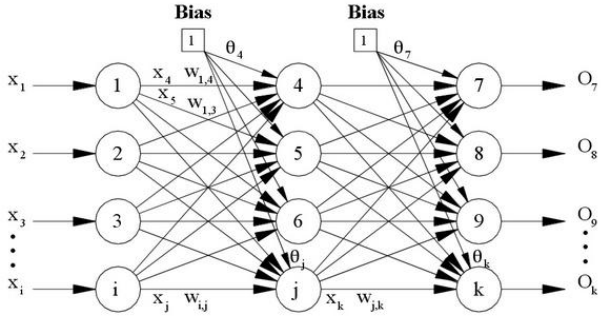


Figure 2.32a

There are typically three layers, but the number of layers varies depending upon the task. The first layer draws linear boundary, the second layer combines these boundaries and the third layer can generate arbitrarily complex boundaries. These layers have the following properties:

- Each layer has no connection between themselves

- The input and output layers are not directly connected
- Fully connected within a layer
- The number of output units need not equal number of input unit
- Number of hidden units per layers can be more or less than input or output unit

Each neuron in the given layers is assigned a specific weight, and the outputs are achieved depending on the weight. However the problem arises in which weights are to be altered, by how much and in which direction. The solution to this credit assignment problem which is also used in training the MLP is called the back propagation algorithm.

Back propagation algorithm

The operations of the Back propagation neural networks can be divided into two steps: **feed forward** and **back propagation**. In the feed forward step, an input pattern is applied to the input layer and it effectively propagates, layer by layer, through the network until an output is produced. The network's actual output value is then compared to the expected output, and an error signal is computed for each of the output nodes. Since all the hidden nodes have, to some degree, contributed to the errors evident in the output layer, the output error signals are transmitted backwards from the output layer to each node in the hidden layer that immediately contributed to the output layer. This process is then repeated, layer by layer, until each node in the network has received an error signal that describes its relative contribution to the overall error. Once the error signal for each node has been determined, the errors are then used by the nodes to update the values for each connection weights until the network converges to a state that allows all the training patterns to be encoded. The Back propagation algorithm looks for the minimum value of the **error function** in weight space using a technique called the delta rule or **gradient descent**. The weights that minimize the error function are then considered to be a solution to the learning problem. The back propagation networks come up theories as how to classify the samples. These are then tested against the correct outputs to see how accurate the guesses of the network are. There are also issues regarding generalizing the neural networks. The data set can be under trained or over trained depending on the data types. Under training occurs when neural network is not complex enough to detect the problem, as shown in the illustration below:

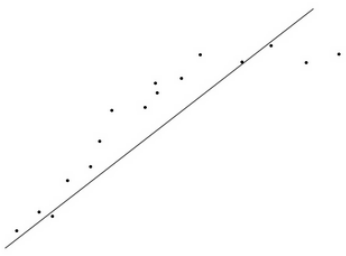


Figure 2.32b Under fitting data

Over training occurs when the network is too complex, resulting in predictions are far beyond the range of training data, as shown in the illustration below:

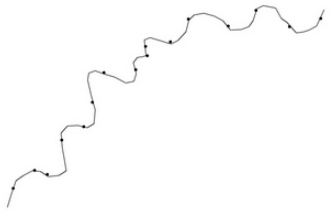


Figure 2.32c Over fitting data

The aim is create aneural network with satisfactorily good number of hidden nodes that will produce a good result, like the one in the illustration below:

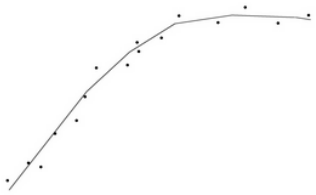


Figure 2.32c Good fitting of data

The back propagation algorithm has the following steps:

Feed forward

When a set of training data is given to the input layer, the weighted sum of the input to the j^{th} layer is denoted by:

$$Net_j = \sum w_{i,j} x_j + \theta_j$$ Equation 2.321

The above equation is used to calculate the total input to the neuron. The θ_j term is the weighted value from a **bias** node that always has an output value of 1. The bias node is used to overcome the problems associated with situations where the values of an input pattern are zero. If any input pattern has zero values, the neural network could not be trained without a bias node. To decide whether a neuron should fire, the "Net" term, also known as the action potential, is passed onto an appropriate activation function. The resulting value from the activation function determines the neuron's output, and becomes the input value for the neurons in the next layer connected to it. Since one of the requirements for the Back propagation algorithm is that the activation function is differentiable; a typical activation function used is the Sigmoid equation

$$O_j = x_k = \frac{1}{1 + e^{-Net_j}}$$ Equation 2.322

The equations 2.321 and 2.322 are also used to calculate the output value for the nodes in the output layer.

Error Calculations and Weight Adjustments:

Output layer

Assuming that activation value of the output node, k, is O_k, and the expected target output for node k is t_k, the difference between the actual output and the expected output is represented by:

$$\Delta_k = t_k - O_k \quad \text{Equation 2.323}$$

The error signal for node k in the output layer can be calculated using the following equations:

$$\delta_k = \Delta_k O_k (1 - O_k) \text{ or } \quad \delta_k = (t_k - O_k) O_k (1 - O_k) \quad \text{Equation 2.324}$$

Where the $O_k(1-O_k)$ term is the derivative of the Sigmoid function.

With the delta rule, the change in the weight connecting input node j and output node k is proportional to the error at node k multiplied by the activation of node j.

The formulas used to modify the weight, $w_{j,k}$, between the output node, k, and the node, j is:

$$\Delta w_{j,k} = l_r \delta_k x_k \quad \textbf{(5)}$$

$$w_{j,k} = w_{j,k} + \Delta w_{j,k} \quad \textbf{(6)}$$

where $\Delta w_{j,k}$ is the change in the weight between nodes j and k, l_r is the **learning rate**. The learning rate is a relatively small constant that indicates the relative change in weights. If the learning rate is too low, the network will learn very slowly, and if the learning rate is too high, the network may oscillate around minimum point, overshooting the lowest point with each weight adjustment, but never actually reaching it. Usually the learning rate is very small, with 0.01 not an uncommon number. Some modifications to the Back propagation algorithm allow the learning rate to decrease from a large value during the learning process. This has many advantages. As learning develops, the learning rate decreases as it approaches the optimal point. Slowing the learning at optimal point helps the network to come near a solution and reduce overshooting. If, however, the learning process initiates close to the optimal point, the system may initially oscillate, but this effect is reduced with time as the learning rate decreases.

To improve the process of updating the weights, a modification to equation (5) is made:

$$\Delta w_{j,k}^{n} = l_r \delta_k x_k + \Delta w_{j,k}^{(n-1)} \mu \quad \textbf{(7)}$$

Hidden layer

The error signal for node j in the hidden layer can be calculated as

$$\delta_k = (t_k - O_k) O_k \sum (w_{j,k} \delta_k)$$

Where the Sum term adds the weighted error signal for all nodes, k, in the output layer.

As before, the formula to adjust the weight, $w_{i,j}$, between the input node, i, and the node, j is:

$$\Delta w_{i,j}^{n} = l_r \delta_j x_j + \Delta w_{i,j}^{(n-1)} \mu$$

$$w_{i,j} = w_{i,j} + \Delta w_{i,j}$$

Global error

Back propagation is derived by assuming that it is desirable to minimize the error on the output nodes over all the patterns presented to the neural network. The following equation is used to calculate the **error function**, E, for all patterns

$$E = \tfrac{1}{2} \sum (\sum (t_k - O_k)^2)$$

Ideally, the error function should have a value of zero when the neural network has been correctly trained. This, however, is numerically unrealistic.

2.33 Naïve Bayes

Naïve Bayes classification is based on the Bayesian theorem. Naïve Bayes works best when the range of input is large. Despite the over simplified assumptions of naïve Bayes, it does work well in complex situations. The advantage of using naïve Bayes is that it performs well even with a small amount of data. Naïve Bayes classifier is a simple probabilistic classifier. The efficiency naïve Bayes classifier is mainly because it is less computationally intensive and requires a small amount of data. This classifier also has an added advantage in terms of CPU and memory consumption. In cases it performs almost as well as other complex algorithms.

Naïve Bayes trains the data set pretty fast and can also classify them fast enough. It is also indifferent to irrelevant data, so can minimize if not completely deduct all the errors which can potentially occur due to the irrelevant data. However naïve Bayes tends to assume independence

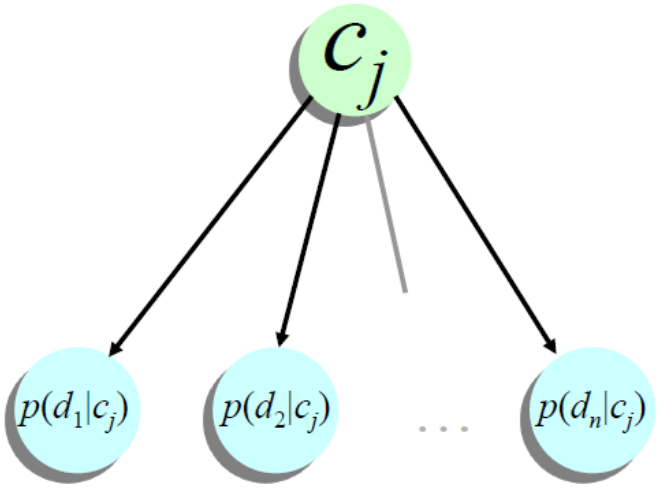
of data, hence causing some error.

Naïve Bayes, a branch of Bayesian is used when a data set has more than one feature and the Bayesian theory can no longer handle it. Naïve Bayes handles this issue by considering all the variable independent of each other by estimating

$$p(d|cj) = p(d1|cj) * p(d2|cj) * ....* p(dn|cj)$$ Equation 2.331

where, $p(d|cj)$ means the probability of class c_j generating instance d equals the equation, $p(d1|cj)$ is probability of class c_j generating the observed value for feature one and $p(d2|cj)$ is the probability of class c_j generating the observed for feature two and so on.

A naïve Bayesian network is represented with the graph below:



The direction of the arrows represents how each class causes certain features, with a certain probability. All the probabilities can be found by scanning the database once and thus are stored in a table.

Figure: 2.33a

Example

Let us assume we have the following feature, and will determine each person's sex using them.

Name	Over 170cm	Eye	Hair length	Sex
Drew	No	Blue	Short	Male
Claudia	Yes	Brown	Long	Female
Drew	No	Blue	Long	Female
Drew	No	Blue	Long	Female
Alberto	Yes	Brown	Short	Male
Karin	No	Blue	Long	Female
Nina	Yes	Brown	Short	Female
Sergio	Yes	Blue	Long	Male

Table: 5

Using equation 2.331, we will get the following results:

$P(\text{officer Drew}|C_{j)} = p(\text{over\_170=yes}|C_j) * p(\text{eyes=blue}|C_j) * p(\text{hair=short}|C_j)$

$P(\text{officer Drew}|\text{female}) = p(2/5) * p(3/5) * p(1/5) = 0.048$

$P(\text{officer Drew}|\text{male}) = p(2/3) * p(2/3) * p(2/3) = 0.2963$

So the probability officer Drew being a female is more than the being male. Therefore Officer Drew is a female. When the results were matched with the database, it was seen tha Officer Drew indeed was a female.

2.34 Decision Tree

A schematic tree-shaped diagram used to determine a course of action or show a statistical probability. Each branch of the decision tree represents a possible decision or occurrence. The tree structure shows how one choice leads to the next, and the use of branches indicates that each option is mutually exclusive.

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown below. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

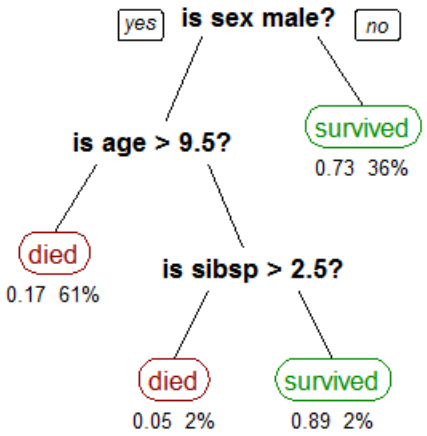


Figure: 2.34a

A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of *top-down induction of decision trees* (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data, but it is not the only strategy. In fact, some approaches have been developed recently allowing tree induction to be performed in a bottom-up fashion.

In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

Data comes in records of the form:

$$(\mathbf{x}, Y) = (x_1, x_2, x_3, ..., x_k, Y)$$

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector **x** is composed of the input variables, x_1, x_2, x_3 etc., that are used for that task.

Decision tree advantages

Amongst other data mining methods, decision trees have various advantages:

- **Simple to understand and interpret:** People are able to understand decision tree models after a brief explanation.
- **Requires little data preparation:** Other techniques often require data normalization, dummy variables need to be created and blank values to be removed.
- **Able to handle both numerical and categorical data:** Other techniques are usually specialized in analyzing datasets that have only one type of variable. (For example, relation rules can be used only with nominal variables while neural networks can be used only with numerical variables.)
- **Uses a white box model:** If a given situation is observable in a model the explanation for the condition is easily explained by Boolean logic. (An example of a black box model is an artificial neural network since the explanation for the results is difficult to understand.)
- **Possible to validate a model using statistical tests:** That makes it possible to account for the reliability of the model.
- **Robust:** Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.
- **Performs well with large datasets:** Large amounts of data can be analyzed using standard computing resources in reasonable time.

2.35 Random Forest

The random forest starts with a standard machine learning technique called a "decision tree" which, in ensemble terms, corresponds to our weak learner. In a decision tree, an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets. For details see here, from which the figure below is taken.

In this example, the tree advises us, based upon weather conditions, whether to play ball. For example, if the outlook is sunny and the humidity is less than or equal to 70, then it's probably OK to play.

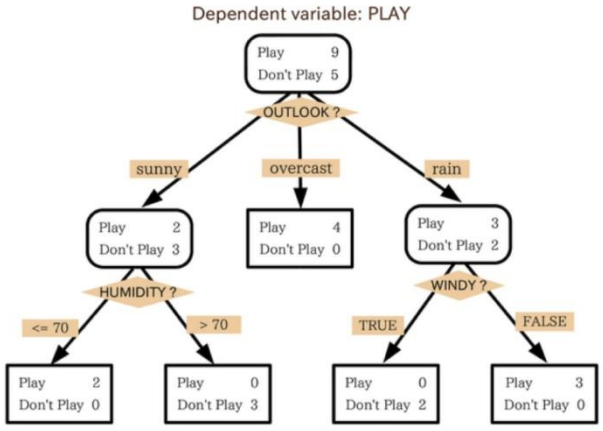


Figure: 2.35a

The random forest (see figure below) takes this notion to the next level by combining trees with the notion of an ensemble. Thus, in ensemble terms, the trees are weak learners and the random forest is a strong learner.

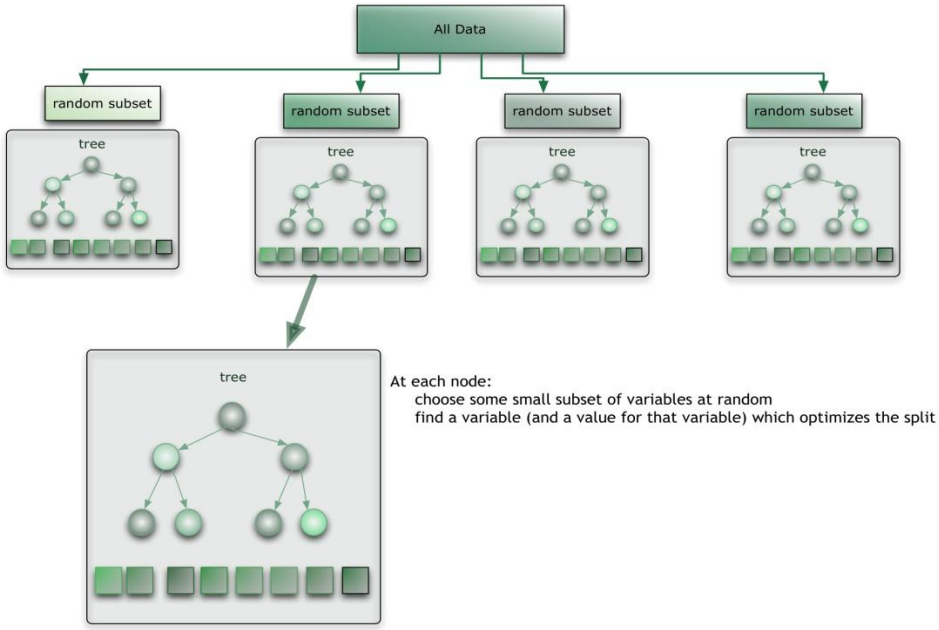


Figure: 2.35b

Here is how such a system is trained; for some number of trees *T*:

1. Sample *N* cases at random with replacement to create a subset of the data (see top layer of figure above). The subset should be about 66% of the total set.
2. At each node:
    1. For some number *m* (see below), *m* predictor variables are selected at random from all the predictor variables.
    2. The predictor variable that provides the best split, according to some objective function, is used to do a binary split on that node.
    3. At the next node, choose another *m* variables at random from all predictor variables and do the same.

Depending upon the value of *m*, there are three slightly different systems:

- Random splitter selection: $m = 1$
- Breiman's bagger: m = total number of predictor variables
- Random forest: m << number of predictor variables. Brieman suggests three possible values for m: $\frac{1}{2}\sqrt{m}$, $\sqrt{m}$, and $2\sqrt{m}$

Running a Random Forest: When a new input is entered into the system, it is run down all of the trees. The result may either be an average or weighted average of all of the terminal nodes that are reached, or, in the case of categorical variables, a voting majority.

Note that:

- With a large number of predictors, the eligible predictor set will be quite different from node to node.
- The greater the inter-tree correlation, the greater the random forest error rate, so one pressure on the model is to have the trees as uncorrelated as possible.
- As *m* goes down, both inter-tree correlation and the strength of individual trees go down. So some optimal value of *m* must be discovered.

Strengths and weaknesses*:* Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data. Random Forest weaknesses are that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy. Of course, the best test of any algorithm is how well it works upon our own data set.

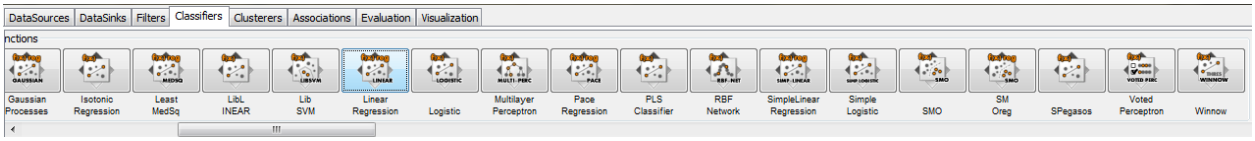
3.0 System Implementation

3.1 System Setup: WEKA

WEKA is a collection of machine learning algorithms. The algorithm can be directly applied to the data set or called from the JAVA code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. WEKA's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes. WEKA's main user interface is the *Explorer*, but the same functionality can be obtained through component-based knowledge flow. There is also the **Experimenter,** which allows the systematic comparison of the predictive performance of WEKA's machine learning algorithms on a collection of datasets.

The **Explorer** interface features several panels providing access to the main components of the

workbench.

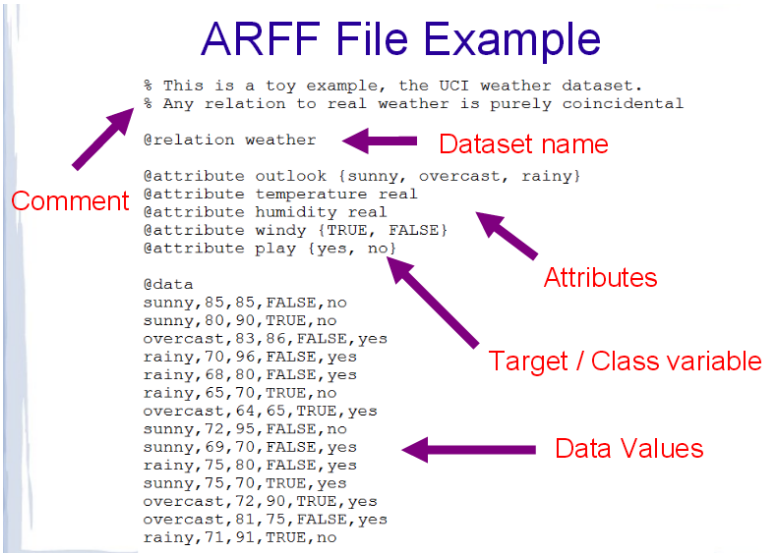


The **Cluster** pane gives access to the clustering techniques in WEKA, for example, the simple k-means algorithm.

The **Classify** panel allows the user to apply classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive models, and to visualize erroneous predictions.

Attribute Relation File Format(ARFF) is the default file type for data analysis in weka but data can also be imported from various formats.

•ARFF (Attribute Relation File Format) has two sections:
–the Header information defines attribute name, type and relations.
–the Data section lists the data records.



3.2 Result and discussion

The data sets were preprocessed and taken input into WEKA, and then algorithms were applied on it. Each algorithm yielded different result, although there were certain similarities. WEKA produces output in a certain pattern. The output comprises of confusion matrix, true and false positive values, precision and recall values and weighted average for each of these values. In case of tree algorithm, WEKA also generates a tree.

Confusion matrix: A confusion matrix consist information about the actual and predicted

classifications done by a classification system.

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	Positive	**c**	**d**

Table: 6

This is an illustration of what a two classifier confusion matrix typically looks like, where a, b, c and d stands for the following:

- 'a' is the number of correct prediction that an instance is negative
- 'b' is the number of incorrect prediction that an instance is positive
- 'c' is the number of incorrect prediction that an instance is negative
- 'd' is the number of correct prediction that an instance is positive

True positive or Recall: It is the proportion of the positive cases that were correctly identified, using the equation below:

$$TP = \frac{d}{c+d}$$

False positive rate: It is the proportion of the negative classes that were incorrectly classified as positive, using the equation below:

$$FP = \frac{b}{a+b}$$

True negative rate: It is the proportion of negatives cases that were classified correctly using the equation below:

$$TN = \frac{a}{a+b}$$

False negative rate: It is the proportion of positive cases that were incorrectly classified using the equation below:

$$FN = \frac{c}{c+d}$$

Precision: It is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$P = \frac{d}{b+d}$$

Tree: Tree indicates how the classifier uses the attributes to make a decision.

```
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9
```

The illustration above is what a tree generated in WEKA typically looks like. The leaf nodes indicate which class an instance will be assigned to should that node be reached. The numbers in brackets after the leaf nodes indicate the number of instances assigned to that node, followed by how many of those instances is incorrectly classified as a result. With other classifiers some other output will be given that indicates how the decisions are made.

3.21 Super Vector Machine: Sequential Minimization Optimization

This algorithm was applied on three data set, each of them yielding different output depending on the attributes which were effecting the results.

Data set one: Exam preparing students for readiness

After applying this algorithm on this data set, almost 98% was classified correctly and only two percent were misclassified, proving that the exam helped to prepare the students for further education. However despite the very good outcome certain classes were misclassified. In the confusion matrix, eleven instances were misclassified in row a, eight instances misclassified in row b, none were misclassified in row d and four in row d.

```
=== Summary ===

Correctly Classified Instances        1090               97.8456 %
Incorrectly Classified Instances        24                2.1544 %
Kappa statistic                          0.96
Mean absolute error                      0.2518
Root mean squared error                  0.3147
Relative absolute error                 92.5306 %
Root relative squared error             85.3658 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      75      %
Total Number of Instances             1114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.023    0.983      0.998    0.991      0.978    0.988     0.983     FP
                 0.971    0.012    0.976      0.971    0.974      0.960    0.980     0.959     PP
                 0.778    0.000    1.000      0.778    0.875      0.880    0.997     0.861     UP
                 0.884    0.004    0.938      0.884    0.910      0.905    0.991     0.870     NH
Weighted Avg.    0.978    0.018    0.978      0.978    0.978      0.966    0.986     0.966

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 643   1   0   0 |   a = FP
  11 372   0   0 |   b = PP
   0   0  14   4 |   c = UP
   0   8   0  61 |   d = NH
```

Data set two: Student's family's education affecting readiness

After applying this algorithm on this data set, only 58% was classified correctly and forty seven percent were misclassified, proving the fact that the family's education does not affect the readiness and hence the result of students. There was certain misclassification in this data set, as shown in the confusion matrix. In the row 'a' of the confusion matrix, 376 instances of class 'b' were classified into class a, 69 instances of class 'c' were misclassified in class 'a', and twenty instances of class 'd' were misclassified in class 'a'. The time taken to test this training data set was 0.03 seconds.

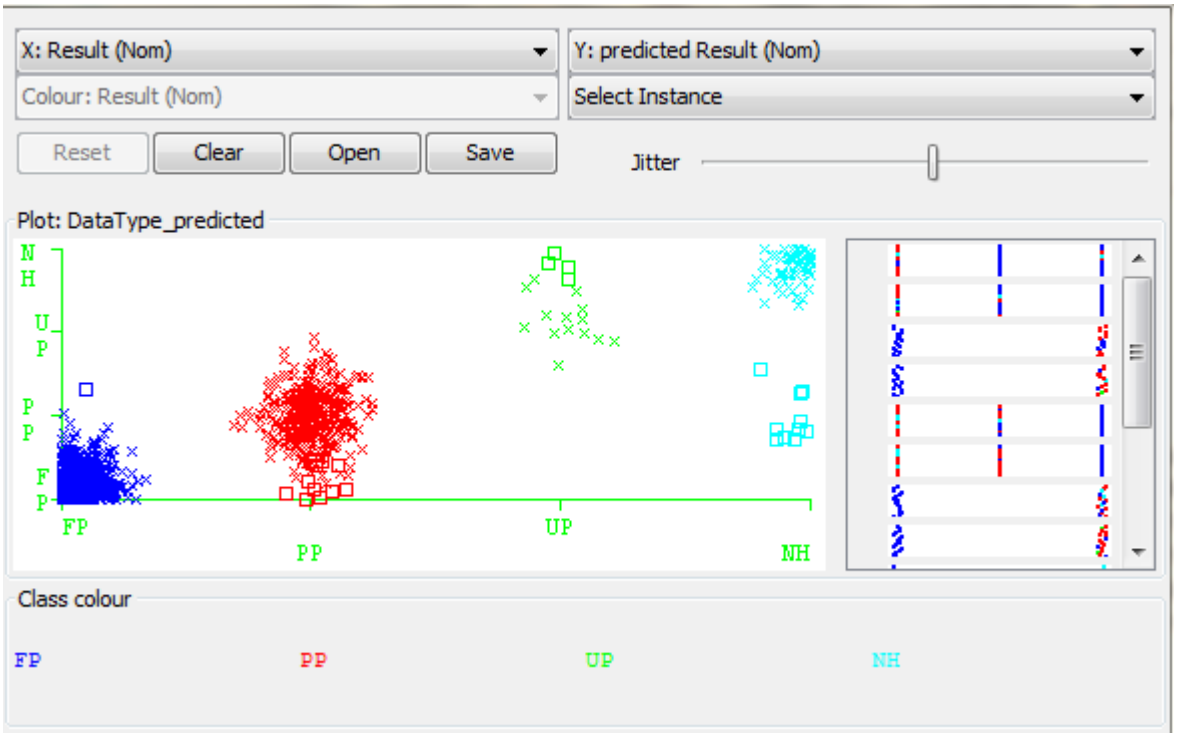
```
Time taken to test model on training data: 0.03 seconds

=== Summary ===

Correctly Classified Instances         648               58.221  %
Incorrectly Classified Instances       465               41.779  %
Kappa statistic                          0
Mean absolute error                      0.293
Root mean squared error                  0.3744
Relative absolute error                107.815  %
Root relative squared error            101.6497 %
Coverage of cases (0.95 level)          98.2031 %
Mean rel. region size (0.95 level)      75      %
Total Number of Instances             1113

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.582      1.000    0.736      0.000    0.500     0.582     FP
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.338     PP
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.062     NH
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.018     UP
Weighted Avg.    0.582    0.582    0.339      0.582    0.428      0.000    0.500     0.457

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 648   0   0   0 |   a = FP
 376   0   0   0 |   b = PP
  69   0   0   0 |   c = NH
  20   0   0   0 |   d = UP
```

D**ata set three: Socio economic factors affecting readiness and result**

After applying this algorithm on this data set, only 58% was classified correctly and forty one percent were misclassified, proving the fact that the family's economic condition barely affect the readiness and hence the result of students. There was certain misclassification in this data set, as shown in the confusion matrix. In the row 'a' of the confusion matrix, 376 instances of class 'b' were classified into class a, 69 instances of class 'c' were misclassified in class 'a', and twenty instances of class 'd' were misclassified in class 'a'. The time taken to test this training data set was 0.03 seconds.

```
Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         649               58.2585 %
Incorrectly Classified Instances       465               41.7415 %
Kappa statistic                          0
Mean absolute error                      0.2929
Root mean squared error                  0.3744
Relative absolute error                107.8453 %
Root relative squared error            101.6568 %
Coverage of cases (0.95 level)          98.2047 %
Mean rel. region size (0.95 level)      75      %
Total Number of Instances             1114

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
               1.000    1.000    0.583      1.000    0.736      0.000    0.500     0.583     FP
               0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.338     PP
               0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.062     NH
               0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.018     UP
Weighted Avg.  0.583    0.583    0.339      0.583    0.429      0.000    0.500     0.457

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 649   0   0   0 |   a = FP
 376   0   0   0 |   b = PP
  69   0   0   0 |   c = NH
  20   0   0   0 |   d = UP
```

The classification of the data set, as a result of applying the respective algorithm.



3.22 Multilayer Perceptron: Back propagation Algorithm

Back propagation algorithm was applied on the three respective preprocessed data set, and the results obtained from applying this algorithm is shown below.

Data set one: Exam preparing students for readiness

After applying this algorithm on this data set, almost 99% was classified correctly and only 0.1percent were misclassified, proving that the exam helped to prepare the students for further education. However despite the very good outcome certain classes were misclassified. In the confusion matrix, only one instance was misclassified in row a, one instance misclassified in row b, none were misclassified in row c and four in row d.

```
Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances        1112               99.8205 %
Incorrectly Classified Instances         2                0.1795 %
Kappa statistic                          0.9967
Mean absolute error                      0.0035
Root mean squared error                  0.0295
Relative absolute error                  1.2714 %
Root relative squared error              8.0019 %
Coverage of cases (0.95 level)          99.9102 %
Mean rel. region size (0.95 level)      25.1795 %
Total Number of Instances             1114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.002    0.998      0.998    0.998      0.996    1.000     1.000     FP
                 0.997    0.001    0.997      0.997    0.997      0.996    1.000     1.000     PP
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     UP
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     NH
Weighted Avg.    0.998    0.002    0.998      0.998    0.998      0.997    1.000     1.000

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 643   1   0   0 |   a = FP
   1 382   0   0 |   b = PP
   0   0  18   0 |   c = UP
   0   0   0  69 |   d = NH
```

Data set two: Student's family's education affecting readiness

After applying this algorithm on this data set, only 58% was classified correctly and forty one percent were misclassified, proving the fact that the family's education does not affect the readiness and hence the result of students. There was certain misclassification in this data set, as shown in the confusion matrix. In the row 'a' of the confusion matrix, 373 instances of class 'b' were classified into class a, 69 instances of class 'c' were misclassified in class 'a', and twenty

instances of class ‘d’ were misclassified in class ‘a’. The time taken to test this training data set was 0.03 seconds.

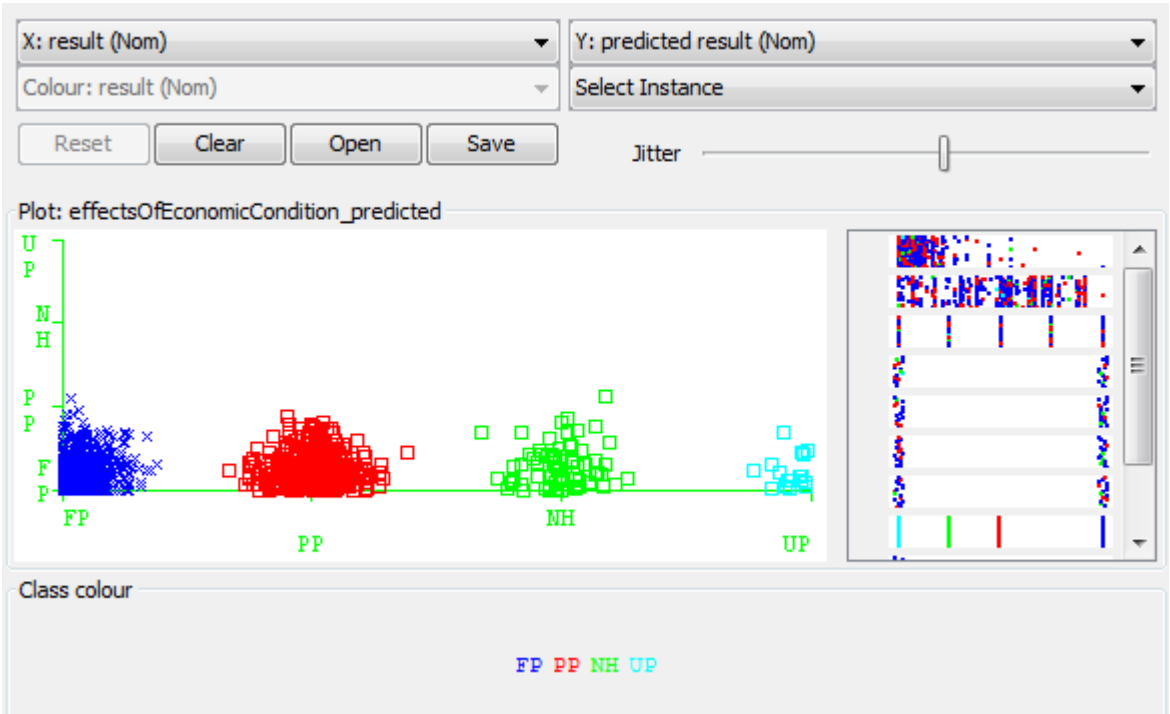
```
Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         651               58.4906 %
Incorrectly Classified Instances       462               41.5094 %
Kappa statistic                          0.008
Mean absolute error                      0.2545
Root mean squared error                  0.3689
Relative absolute error                 93.6496 %
Root relative squared error            100.1435 %
Coverage of cases (0.95 level)          97.4843 %
Mean rel. region size (0.95 level)      65.7233 %
Total Number of Instances             1113

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.994    0.584      1.000    0.737      0.061    0.601     0.667     FP
                 0.008    0.000    1.000      0.008    0.016      0.073    0.552     0.398     PP
                 0.000    0.000    0.000      0.000    0.000      0.000    0.686     0.109     NH
                 0.000    0.000    0.000      0.000    0.000      0.000    0.661     0.029     UP
Weighted Avg.    0.585    0.578    0.678      0.585    0.435      0.060    0.591     0.530

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 648   0   0   0 |   a = FP
 373   3   0   0 |   b = PP
  69   0   0   0 |   c = NH
  20   0   0   0 |   d = UP
```

Data set three: Socio economic factors affecting readiness and result

After applying this algorithm on this data set, only %60 was classified correctly and thirty nine percent were misclassified, proving the fact that the family’s economic condition had some sort of affect on the readiness and hence the result of students. There was certain misclassification in this data set, as shown in the confusion matrix. In the row ‘a’ of the confusion matrix, 337 instances of class ‘b’ were classified into class a, 64 instances of class ‘c’ were misclassified in class ‘a’, and 18 instances of class ‘d’ were misclassified in class ‘a’. In row ‘b’ the 16 instances of class ‘a’ were misclassified, five instances of class ‘c’ and 2 instances of class ‘d’ .The time taken to test this training data set was 0.01 seconds.

```
Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         672               60.3232 %
Incorrectly Classified Instances       442               39.6768 %
Kappa statistic                          0.0795
Mean absolute error                      0.2507
Root mean squared error                  0.3569
Relative absolute error                 92.2825 %
Root relative squared error             96.9161 %
Coverage of cases (0.95 level)          96.7684 %
Mean rel. region size (0.95 level)      61.5799 %
Total Number of Instances             1114

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.975    0.901    0.602      0.975    0.744      0.160    0.645     0.702     FP
               0.104    0.031    0.629      0.104    0.178      0.150    0.603     0.460     PP
               0.000    0.000    0.000      0.000    0.000      0.000    0.674     0.130     NH
               0.000    0.000    0.000      0.000    0.000      0.000    0.686     0.054     UP
Weighted Avg.  0.603    0.535    0.563      0.603    0.494      0.144    0.633     0.573

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 633  16   0   0 |   a = FP
 337  39   0   0 |   b = PP
  64   5   0   0 |   c = NH
  18   2   0   0 |   d = UP
```

The classification of the data set, as a result of applying the respective algorithm.



3.23 Naïve Bayes

This algorithm was applied on three data set, each of them yielding different output depending on the attributes which were effecting the results.

Data set one: Exam preparing students for readiness

After applying this algorithm on this data set, almost 85% was classified correctly and fifteen percent were misclassified, proving that the exam helped to prepare the students for further education. However despite the very good outcome certain classes were misclassified. In the confusion matrix, one hundred and thirteen instances were misclassified in row a, thirty nine instances misclassified in row b, none were misclassified in row c and eleven in row d.

```
Time taken to test model on training data: 0.23 seconds

=== Summary ===

Correctly Classified Instances         951               85.368  %
Incorrectly Classified Instances       163               14.632  %
Kappa statistic                          0.7214
Mean absolute error                      0.0803
Root mean squared error                  0.2218
Relative absolute error                 29.495  %
Root relative squared error             60.1661 %
Coverage of cases (0.95 level)          98.0251 %
Mean rel. region size (0.95 level)      33.281  %
Total Number of Instances             1114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.958    0.240    0.845      0.958    0.898      0.746    0.977     0.986     FP
                 0.676    0.053    0.869      0.676    0.761      0.668    0.947     0.857     PP
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     UP
                 0.826    0.011    0.838      0.826    0.832      0.821    0.991     0.815     NH
Weighted Avg.    0.854    0.158    0.855      0.854    0.848      0.728    0.968     0.931

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 617  27   0   0 |   a = FP
 113 259   0  11 |   b = PP
   0   0  18   0 |   c = UP
   0  12   0  57 |   d = NH
```

Data set two: Student's family's education affecting readiness

After applying this algorithm on this data set, only 56% was classified correctly and forty six percent were misclassified, proving the fact that the family's education does not affect the readiness and hence the result of students. There was certain misclassification in this data set, as shown in the confusion matrix. In the row 'a' of the confusion matrix, 302 instances of class 'b'

were classified into class a, 69 instances of class ‘c’ were misclassified in class ‘a’ and ‘b’ , and twenty instances of class ‘d’ were misclassified in class ‘a’ and ‘b’. The time taken to test this training data set was 0.07 seconds.

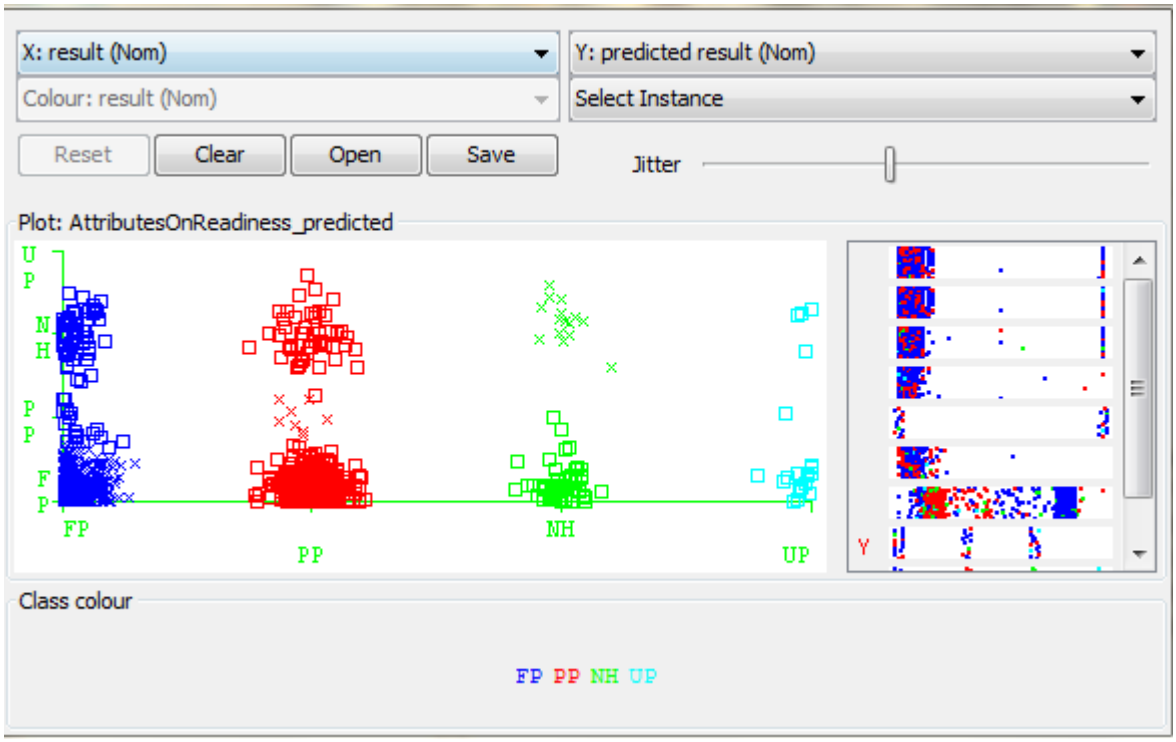
```
Time taken to test model on training data: 0.07 seconds

=== Summary ===

Correctly Classified Instances         626               56.1939 %
Incorrectly Classified Instances       488               43.8061 %
Kappa statistic                          0.0511
Mean absolute error                      0.2699
Root mean squared error                  0.3682
Relative absolute error                 99.3757 %
Root relative squared error             99.9854 %
Coverage of cases (0.95 level)          98.474  %
Mean rel. region size (0.95 level)      73.8555 %
Total Number of Instances             1114

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.851     0.776     0.605       0.851   0.707      0.095    0.604     0.672     FP
               0.197     0.172     0.368       0.197   0.256      0.030    0.561     0.375     PP
               0.000     0.000     0.000       0.000   0.000      0.000    0.679     0.117     NH
               0.000     0.000     0.000       0.000   0.000      0.000    0.694     0.039     UP
Weighted Avg.  0.562     0.510     0.476       0.562   0.498      0.066    0.596     0.526

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 552  97   0   0 |   a = FP
 302  74   0   0 |   b = PP
  44  25   0   0 |   c = NH
  15   5   0   0 |   d = UP
```

Data set three: Socio economic factors affecting readiness and result

After applying this algorithm on this data set, only 53% was classified correctly and forty seven percent were misclassified, proving the fact that the family’s economic condition barely affect the readiness and hence the result of students. There was certain misclassification in this data set, as shown in the confusion matrix. In the row ‘a’ of the confusion matrix, 316 instances of class ‘b’ were classified into class a, 52 instances of class ‘c’ were misclassified in class ‘a’, and fifteen instances of class ‘d’ were misclassified in class ‘a’. In the row ‘b’ of the confusion matrix, 20 instances of class ‘a’ were classified into class b, 2 instances of class ‘c’ were misclassified in class ‘b’, and 1 instances of class ‘d’ were misclassified in class ‘b’. In the row ‘c’ of the confusion matrix, 60 instances of class ‘a’ were classified into class c, 50 instances of class ‘b’ were misclassified in class ‘c’, and four instances of class ‘d’ were misclassified in class ‘c’. The time taken to test this training data set was 0.01 seconds.

```
Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         593               53.2794 %
Incorrectly Classified Instances       520               46.7206 %
Kappa statistic                          0.0373
Mean absolute error                      0.2639
Root mean squared error                  0.4124
Relative absolute error                 97.097  %
Root relative squared error            111.9621 %
Coverage of cases (0.95 level)          94.7889 %
Mean rel. region size (0.95 level)      69.9461 %
Total Number of Instances             1113

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.877    0.824    0.597      0.877    0.710      0.074    0.590     0.642     FP
                 0.027    0.031    0.303      0.027    0.049      -0.013   0.505     0.342     PP
                 0.217    0.109    0.116      0.217    0.152      0.082    0.673     0.104     NH
                 0.000    0.000    0.000      0.000    0.000      0.000    0.718     0.047     UP
Weighted Avg.    0.533    0.497    0.457      0.533    0.440      0.044    0.568     0.497

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 568  20  60   0 |   a = FP
 316  10  50   0 |   b = PP
  52   2  15   0 |   c = NH
  15   1   4   0 |   d = UP
```

The classification of the data set, as a result of applying the respective algorithm.



3.24 Decision Tree

Decision Tree algorithm was applied on the three respective preprocessed data set, and the results obtained from applying this algorithm is shown below.

Data set one: Exam preparing students for readiness

After applying this algorithm on this data set, almost 99% was classified correctly and only 0.1percent were misclassified, proving that the exam helped to prepare the students for further education. However despite the very good outcome certain classes were misclassified. In the confusion matrix, only one instance was misclassified in row a, one instance misclassified in row b, none were misclassified in row c and d. The time taken to test this training data set was 0.18 seconds.

```
Time taken to test model on training data: 0.18 seconds

=== Summary ===

Correctly Classified Instances        1112               99.8205 %
Incorrectly Classified Instances         2                0.1795 %
Kappa statistic                          0.9967
Mean absolute error                      0.0016
Root mean squared error                  0.0279
Relative absolute error                  0.5702 %
Root relative squared error              7.5564 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      25.4264 %
Total Number of Instances             1114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.002    0.998      0.998    0.998      0.996    1.000     1.000     FP
                 0.997    0.001    0.997      0.997    0.997      0.996    1.000     1.000     PP
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     UP
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     NH
Weighted Avg.    0.998    0.002    0.998      0.998    0.998      0.997    1.000     1.000

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 643   1   0   0 |   a = FP
   1 382   0   0 |   b = PP
   0   0  18   0 |   c = UP
   0   0   0  69 |   d = NH
```

Data set two: Student's family's education affecting readiness

After applying this algorithm on this data set, only 93% was classified correctly and only seven percent were misclassified, proving the fact that the family's education affect the readiness and hence the result of students. There was certain misclassification in this data set, as shown in the confusion matrix. In the row 'a' of the confusion matrix, 44 instances of class 'b' were classified into class a, 10 instances of class 'c' were misclassified in class 'a', and three instances of class 'd' were misclassified in class 'a'. In the row 'b' of the confusion matrix, 8 instances of class 'a' were classified into class 'b', 6 instances of class 'c' were misclassified in class 'b', and four instances of class 'd' were misclassified in class 'a' and 'b'. The time taken to test this training data set was 0.16 seconds.

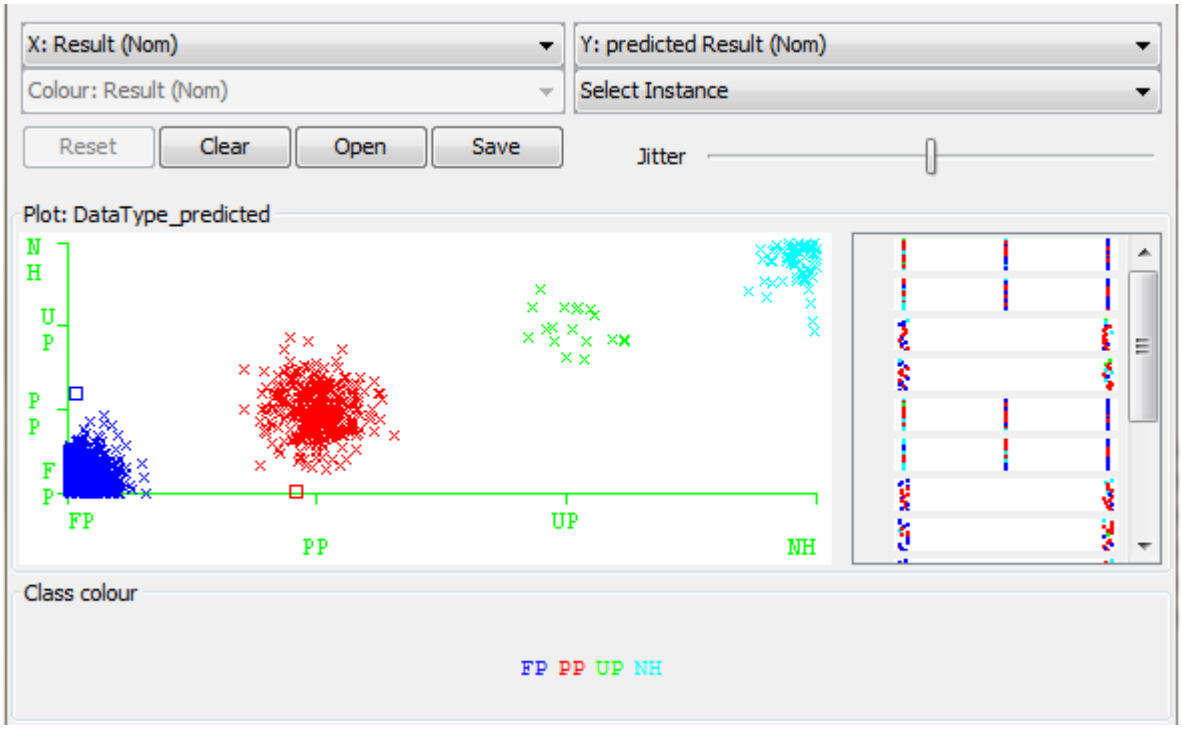
```
Time taken to test model on training data: 0.16 seconds

=== Summary ===

Correctly Classified Instances        1042               93.5368 %
Incorrectly Classified Instances        72                6.4632 %
Kappa statistic                          0.8773
Mean absolute error                      0.036
Root mean squared error                  0.1342
Relative absolute error                 13.2576 %
Root relative squared error             36.4358 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      29.219  %
Total Number of Instances             1114

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.988     0.123     0.918       0.988   0.952      0.882    0.993     0.993     FP
               0.883     0.020     0.957       0.883   0.918      0.881    0.993     0.982     PP
               0.768     0.000     1.000       0.768   0.869      0.870    0.998     0.959     NH
               0.800     0.000     1.000       0.800   0.889      0.893    1.000     0.967     UP
Weighted Avg.  0.935     0.078     0.938       0.935   0.934      0.881    0.994     0.987

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 641   8   0   0 |   a = FP
  44 332   0   0 |   b = PP
  10   6  53   0 |   c = NH
   3   1   0  16 |   d = UP
```

Data set three: Socio economic factors affecting readiness and result:

After applying this algorithm on this data set, only %60 was classified correctly and thirty nine percent were misclassified, proving the fact that the family's economic condition had some sort of affect on the readiness and hence the result of students. There was certain misclassification in this data set, as shown in the confusion matrix. In the row 'a' of the confusion matrix, 38 instances of class 'b' were classified into class a, 10 instances of class 'c' were misclassified in class 'a', and 1 instances of class 'd' were misclassified in class 'a'. In row 'b' the 8 instances of class 'a' were misclassified, five instances of class 'c' and 2 instances of class 'd' .The time taken to test this training data set was 0.04 seconds.

```
Time taken to test model on training data: 0.04 seconds

=== Summary ===

Correctly Classified Instances        1049               94.2498 %
Incorrectly Classified Instances        64                5.7502 %
Kappa statistic                          0.8913
Mean absolute error                      0.0339
Root mean squared error                  0.1302
Relative absolute error                 12.4792 %
Root relative squared error             35.35   %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      29.4025 %
Total Number of Instances             1113

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.988    0.105    0.929      0.988    0.957      0.896    0.994     0.995     FP
                 0.899    0.020    0.958      0.899    0.927      0.893    0.994     0.986     PP
                 0.783    0.000    1.000      0.783    0.878      0.878    0.998     0.963     NH
                 0.850    0.000    1.000      0.850    0.919      0.921    1.000     0.974     UP
Weighted Avg.    0.942    0.068    0.944      0.942    0.942      0.894    0.995     0.990

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 640   8   0   0 |   a = FP
  38 338   0   0 |   b = PP
  10   5  54   0 |   c = NH
   1   2   0  17 |   d = UP
```

The classification of the data set, as a result of applying the respective algorithm.



3.25 Random Forest

Decision Tree algorithm was applied on the three respective preprocessed data set, and the results obtained from applying this algorithm is shown below.

Data set one: Exam preparing students for readiness

After applying this algorithm on this data set, almost 99% was classified correctly and only 0.1 percent was misclassified, proving that the exam helped to prepare the students for further education. However despite the very good outcome certain classes were misclassified. In the confusion matrix, only one instance was misclassified in row a, one instance misclassified in row b, none were misclassified in row c and d. The time taken to test this training data set was 0.05 seconds.

```
Time taken to test model on training data: 0.05 seconds

=== Summary ===

Correctly Classified Instances        1111               99.7307 %
Incorrectly Classified Instances         3                0.2693 %
Kappa statistic                          0.995
Mean absolute error                      0.0145
Root mean squared error                  0.0623
Relative absolute error                  5.3169 %
Root relative squared error             16.8888 %
Coverage of cases (0.95 level)          99.9102 %
Mean rel. region size (0.95 level)      28.7478 %
Total Number of Instances             1114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.002    0.998      0.998    0.998      0.996    1.000     1.000     FP
                 0.997    0.003    0.995      0.997    0.996      0.994    1.000     0.999     PP
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     UP
                 0.986    0.000    1.000      0.986    0.993      0.992    1.000     1.000     NH
Weighted Avg.    0.997    0.002    0.997      0.997    0.997      0.995    1.000     1.000

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 643   1   0   0 |   a = FP
   1 382   0   0 |   b = PP
   0   0  18   0 |   c = UP
   0   1   0  68 |   d = NH
```

Data set two: Student's family's education affecting readiness

After applying this algorithm on this data set, only 90% was classified correctly and only ten percent were misclassified, proving the fact that the family's education affect the readiness and hence the result of students. There was certain misclassification in this data set, as shown in the confusion matrix. In the row 'a' of the confusion matrix, 51 instances of class 'b' were classified into class a, 13 instances of class 'c' were misclassified in class 'a', and 1 instances of class 'd' were misclassified in class 'a'. In the row 'b' of the confusion matrix, 23 instances of class 'a' were classified into class 'b', 9 instances of class 'c' were misclassified in class 'b', and one instances of class 'd' were misclassified in class 'a' and 'b'. In the row 'c' of the confusion matrix, 1 instances of class 'a' were classified into class 'c' and 1 instances of class 'b' were misclassified in class 'c'. In the row 'd', 3 instances of class 'a' were classified into class 'd'. The time taken to test this training data set was 0.02 seconds.

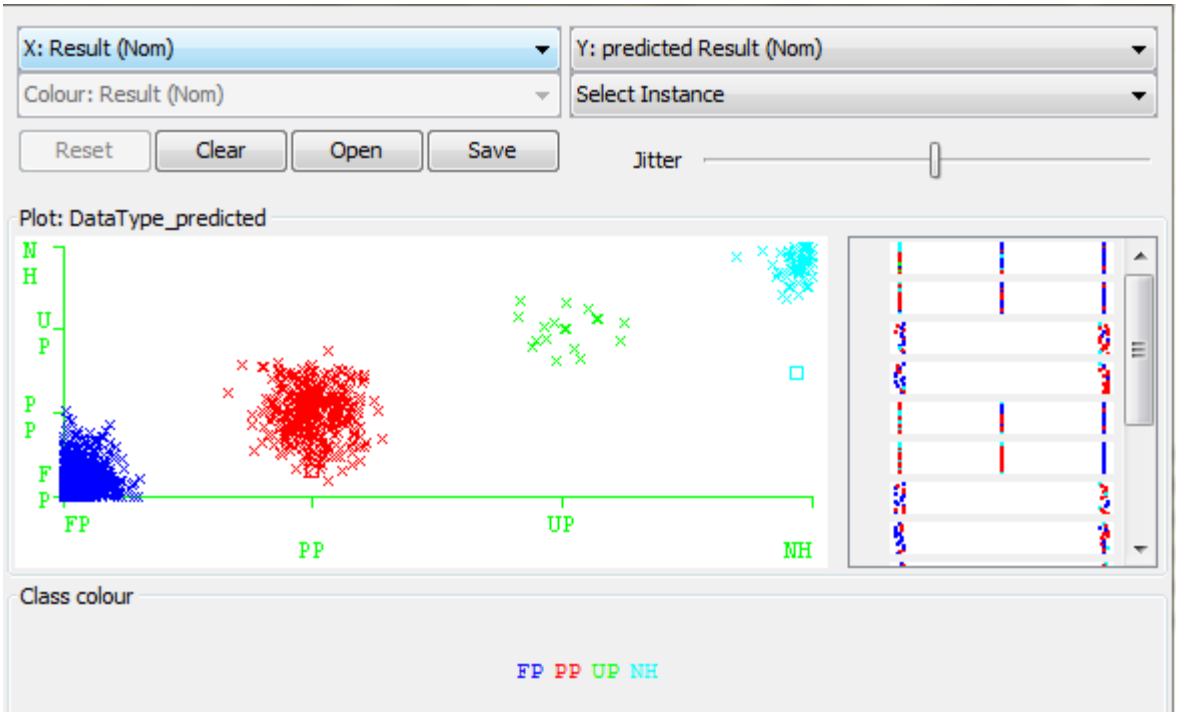
```
Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances        1011               90.754  %
Incorrectly Classified Instances       103                9.246  %
Kappa statistic                          0.8254
Mean absolute error                      0.1157
Root mean squared error                  0.2048
Relative absolute error                 42.6016 %
Root relative squared error             55.6096 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      50.0449 %
Total Number of Instances             1114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.958    0.140    0.905      0.958    0.931      0.830    0.977     0.984     FP
                 0.862    0.045    0.908      0.862    0.884      0.828    0.978     0.959     PP
                 0.681    0.002    0.959      0.681    0.797      0.798    0.994     0.921     NH
                 0.900    0.003    0.857      0.900    0.878      0.876    0.999     0.937     UP
Weighted Avg.    0.908    0.097    0.909      0.908    0.906      0.828    0.979     0.971

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 622  23   1   3 |   a = FP
  51 324   1   0 |   b = PP
  13   9  47   0 |   c = NH
   1   1   0  18 |   d = UP
```

Data set three: Socio economic factors affecting readiness and result

After applying this algorithm on this data set, 92 percent was classified correctly and only 8 percent were misclassified, proving the fact that the family's economic condition had a huge affect on the readiness and hence the result of students. There was certain misclassification in this data set, as shown in the confusion matrix. In the row 'a' of the confusion matrix, 43 instances of class 'b' were classified into class a, 12 instances of class 'c' were misclassified in class 'a', 6 instance of class 'b' and 1 instances of class 'd' were misclassified in class 'a'. In row 'b' 16 instances of class 'a' were misclassified, six instances of class 'c' and 1 instances of class 'd' were misclassified. In the row 'c' of the confusion matrix, 2 instances of class 'a', 4 instance of class 'b' were misclassified. In the row 'd', 1 instances of class 'b' were classified into class 'd'. The time taken to test this training data set was 0.02 seconds.

```
Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances        1027               92.2731 %
Incorrectly Classified Instances        86                7.7269 %
Kappa statistic                          0.8545
Mean absolute error                      0.1122
Root mean squared error                  0.1995
Relative absolute error                 41.284  %
Root relative squared error             54.1666 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      49.8428 %
Total Number of Instances             1113

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.972    0.120    0.918      0.972    0.945      0.864    0.981     0.985     FP
                 0.872    0.031    0.934      0.872    0.902      0.856    0.979     0.963     PP
                 0.739    0.006    0.895      0.739    0.810      0.802    0.992     0.923     NH
                 0.900    0.001    0.947      0.900    0.923      0.922    0.999     0.952     UP
Weighted Avg.    0.923    0.081    0.923      0.923    0.922      0.859    0.981     0.973

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 630  16   2   0 |   a = FP
  43 328   4   1 |   b = PP
  12   6  51   0 |   c = NH
   1   1   0  18 |   d = UP
```

The classification of the data set, as a result of applying the respective algorithm.



3.3 Comparative studies

The performance of the data sets on each data set were studied, listed, compared, and illustrated using bar chart for better understanding. The performance of the algorithms was compared on percentage correctness of the results and time to test the data sets.

3. 31 Data set one: Socio economic factors affecting readiness

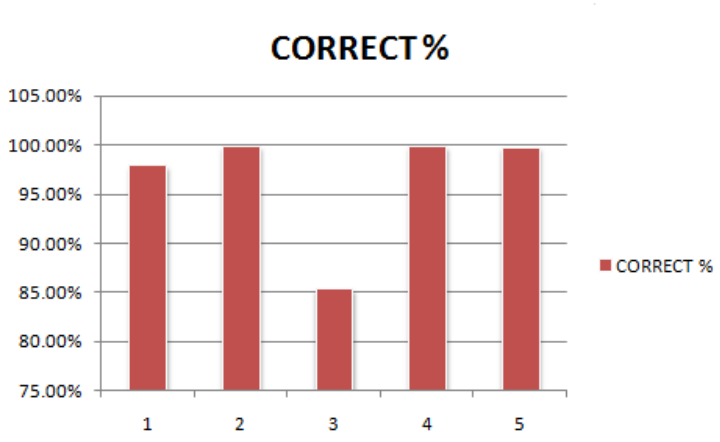


Figure 3.31a

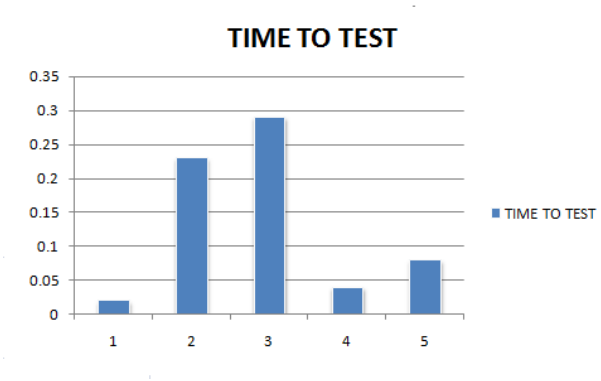


Figure 3.31b

The y-axis in figure 1 represents the percentage of correctness and the y-axis in figure 2 represents the time required in second, and the x-axis in both the graph represents the algorithms.

1. SVM:SMO
2. Multilayer perceptron: Back propagation algorithm
3. Naïve Bayes
4. Random tree
5. Random forest

From both the bar charts, we see that different algorithms perform differently, in terms of both correctness and time. From the bar charts we can see that the performance of SVM: SMO is relatively good and time required is also very less. Whereas in both MLP and naïve Bayes, the performance and the time required is not much favorable in comparison to SVM: SMO.

3. 32 Data set 2: Factors affecting readiness

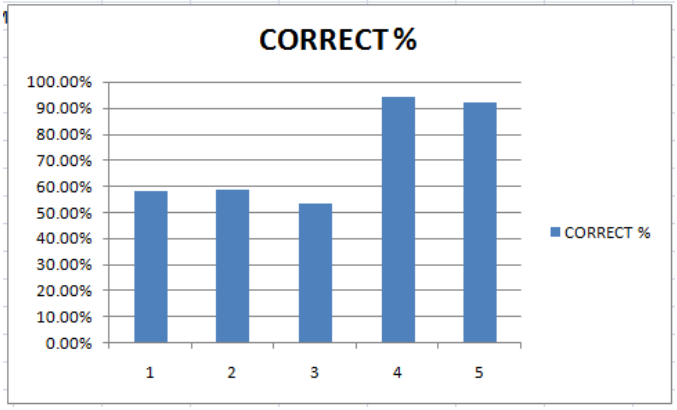


Figure 3.32a

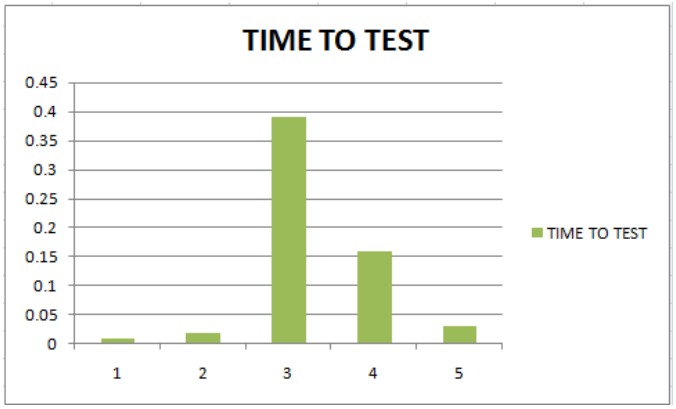


Figure 3.32b

The y-axis in figure 3 represents the percentage of correctness and the y-axis in figure 4 represents the time required in second, and the x-axis in both the graph represents the algorithms.

1. SVM:SMO
2. Multilayer perceptron: Backpropagation algorithm
3. Naïve bayes
4. Random tree
5. Random forest

Although the performance of both SVM:SMO and MLP are almost same, the time taken for MLP is slightly more than that of SVM:SMO. Naïve bayes on the other hand takes more time and performs poorly n comparison to both the other algorithms.

3. 33 Data set 3: Readiness of students

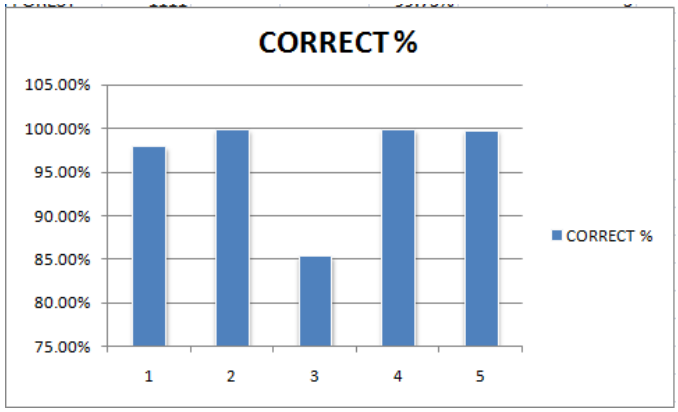


Figure 3.33a

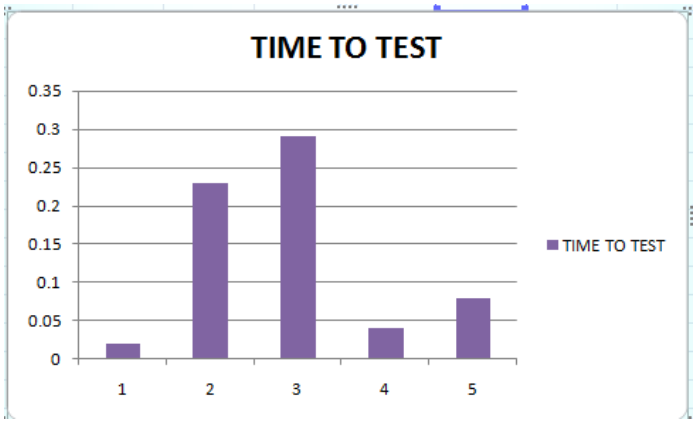


figure 3.33b

The y-axis in figure 3 represents the percentage of correctness and the y-axis in figure 4 represents the time required in second, and the x-axis in both the graph represents the algorithms.

1. SVM:SMO
2. Multilayer perceptron: Back propagation algorithm
3. Naïve Bayes
4. Random tree
5. Random forest

Regardless of both SVM: SMO and MLP performing exceptionally well, SVM here is the better choice since MLP takes more than ten times than SVM to finish testing.

Comparing the algorithms both in terms of correctness and time, we have seen that in every data set SVM has outperformed all other algorithms, thus it being the best option to choose to carry on with this analysis.

4.0 Conclusion and future works

In this paper,we have studied the nature of the data and tested how it affected on performance of the students thus determining whether or nor they are prepared for primary studies as a result of their pre primary education. however the result of the test taken from the students suggests that around ninety percent of the students were prepared after pre primary education.however the result of this study also suggests that external factors like the socio economic condition of the family of the student and the educational back ground has minimal effect on the preparation of the students.five algorithms were applied on the three respective data sets and their outcome analyzed and compared.it was seen that SVM: SMO performed the best result of all classifying and analyzing the data set to near perfection in the minimum possible time. this data set will further be analyzed in the future with respect to geographical location, and geography effecting their result. IED also plans to conduct another study and collect data about the teachers which will once again be analyzed using artificial intelligence to examine whether teaching skill of the teacher effects the readiness and hence the result of the students.

References

1. Christos Stergiou and DimitriosSiganos.NEURAL NETWORKS.
http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html
2. Jason Weston. Support Vector Machine (and Statistical Learning Theory)
Tutorial.http://www.cs.columbia.edu/~kathy/cs4701/documents/jason_svm_tutorial.pdf
3. Thorsten Joachims. SVMlight, Support Vector Machine. (14.08.2008)
http://svmlight.joachims.org/
4. Thorsten Joachims, Learning to Classify Text Using Support Vector Machines.Dissertation, Kluwer, 2002. [B&N] [Amazon] [Kluwer]
5. . T. Joachims, Optimizing Search Engines Using Clickthrough Data, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2002. Online [Postscript] [PDF]
6. An introduction to neural computing. Aleksander, I. and Morton, H. 2nd edition
Neural Networks at Pacific Northwest National Laboratory
7. R. Klinkenberg and T. Joachims, Detecting Concept Drift with Support Vector Machines. Proceedings of the Seventeenth International Conference on Machine Learning (ICML), Morgan Kaufmann, 2000.Online [Postscript (gz)] [PDF (gz)]
8. Wu, Shih-Hung, Support vector Machine
9. Choochart Haruechaiyasak, (2008), **A Tutorial on Naive Bayes Classification**
10. Mirja Mohammad Shahjamal, Samir RanjanNath. (2008) An Evaluation Of BRAC Preprimary
Education Programme. BRAC Research Report, September 2008.
11. NelloCristianini. Support Vector And Kernel Machines.
12. R. Klinkenberg and T. Joachims, Detecting Concept Drift with Support Vector Machines. Proceedings of the Seventeenth International Conference on Machine Learning (ICML), Morgan Kaufmann, 2000.Online [Postscript (gz)] [PDF (gz)]
13. IED BRAC University
14. Jamal M. Nazzal, Ibrahim M. El-Emary, and Salam A. Najim(2008), Multilayer Perceptron Neural Network (MLPs) For Analyzing the Properties of Jordan Oil Shale
15.Eric Meisner(November 22,2003) ,Naive Bayes Classifier example
Stephan Trenn, Multilayer perceptrons: Approximation order and necessary number of hidden Units

16.Eamonn Keogh, Naïve Bayes Classifier