Predicting Subject Area Interest of a Student Using Naive Bayes

By

Shakib Hossain
16141005

Under Supervision of

Moin Mostakim

BRAC
UNIVERSITY

Inspiring Excellence

Department of Computer Science and Engineering
BRAC University

A dissertation submitted to BRAC University in accordance with the requirements of the degree of Bachelors of Science in Computer Science in the Faculty of Computer Science and Engineering.

April 2016

ABSTRACT

Choosing a subject is a difficult decision for a young person on the whole. Some are pre determined but according to statistics most are confused when selecting a subject. This crucial decision is usually made right after finishing high school. This research refers to the possibility of applying a machine learning algorithm called Naive Bayes in order to predict the subject area interest of a student.

DEDICATION AND ACKNOWLEDGEMENTS

Firstly, I am thankful to Almighty Allah that my thesis have been completed. Secondly, I would like to thank my thesis supervisor Mr. Moin Mostakim. He guided me quite well towards completion of my work. And finally I would like to mention my parents. Without their constant support it may not have been possible. With their kind support and prayer I am on the verge of graduation.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signature of Author:

SIGNED: .... DATE: ....
Shakib Hossain, 16141005

Signature of Supervisor:

SIGNED: .... DATE: ....
Moin Mostakim
Department of Computer Science and Engineering
BRAC University

Table of Contents

LIST OF TABLES

TABLE **Page**

List of Figures

Figure **Page**

Chapter 1

Introduction

Machine learning can be defined as a set of processes which can autonomously capture patterns in data. Using these patterns, prediction on future data can be carried out[1]. For instance, lets consider a simple weather prediction model with attributes temperature, humidity and windy. Here, we would have to determine whether outcome is rain or no rain. A set of training data with attribute values and corresponding outcomes would be provided to the model. It would learn from the given data and later it would predict the outcome for new attribute values.

In recent times, an increase in research interest in educational data mining has been observed. Data mining in education is a new field which deals with developing methods for knowledge discovery from data acquired from academic environments[2].

1.1 Motivation

Everyone has to make choices at different stages in their life. Some of the most crucial relate to their education, in particular what subject to take for higher-level studies. For most young people such choices take place at the age of 17 or 18. In general, they are expected to make decisions on higher or further education program or their chosen area of employment right about that age. Here, the question is how or on the basis of what aspect do most young people make these decisions. Most students get confused during enrolling themselves to universities based on these factors. Also, some people find courses and curriculum unsuitable for them after enrollment in a particular department.

Eventually they have to switch or start new, causing loss of money and time. Students face problems taking decisions regarding which subject to take or which university to study or whom to take advice from. Here, a system with education counseling abilities can be of great help. An online student assessment system, analyzing the capability and interest of the students and helping them determine their subject of interest can be built here.

1.2 Goals

I plan to prepare a model which will predict the subject area interest of a student. My main objective is to come up with a system which can be used to solve this crucial problem. I plan to use Naive Bayes algorithm to design the predictive model.

CHAPTER 2

LITERATURE REVIEW

2.1 Related Work

Naive Bayes has been previously used to solve crucial diagnostic problems like Heart disease prediction and Breast cancer detection [1]. And it has proven to be very efficient in developing such diagnostic systems [2]. But it has not been used in educational systems to solve problems as such. Different machine learning algorithms have been applied on different educational problem sets like performance prediction of students, prediction for military career choice, medical diagnosis like heart disease prediction and breast cancer prediction[13].

Machine learning is sometimes conflated with data mining and in [3], it states that application of data mining in higher education system would benefit all the participants in the educational process. Therefore, I have chosen machine learning as a tool to solve this problem.

2.2 Introduction to Machine Learning

Machine learning is broadly divided into three categories-

- Supervised or predictive learning
- Unsupervised or descriptive learning
- Reinforcement learning

2.2.1 Supervised Learning

The most widely applied form of machine learning is supervised learning. It can be defined as a method where the output variable is given. All necessary information to make predictions are provided. The overall goal is to learn a general rule that maps inputs to outputs[1].

2.2.2 Unsupervised Learning

In unsupervised learning, the outcome variable is unknown. In other words, the data is not labeled, so the program will have to extract values and information from the data given. Unsupervised learning algorithms can be used for reduction, clustering and visualization, but in general it is not used for prediction. The main concern here is to find interesting structure in the data which is also sometimes referred to as knowledge discovery[1].

Chapter 3

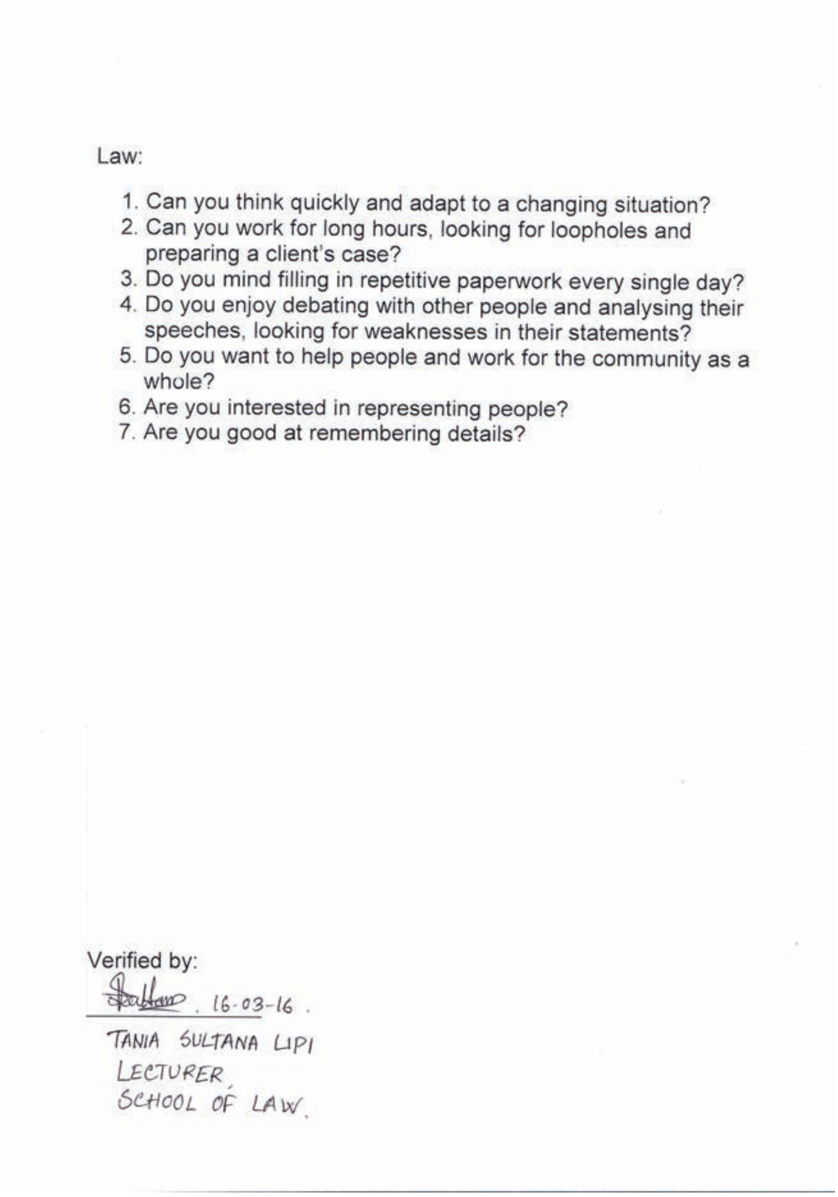
Methodology

3.1 Data Collection

Data collection was the toughest part of the research. During the first phase of my research time, I looked for ways to come around the problem. I searched for existing data but was unable to find any. With the help of my supervisor and few other faculties I came around the problem and soon was back in track of my research work.

At first, I selected the subjects that I would like to focus on. I went for four subjects which are Business, Computer Science, Law and Electrical Engineering. Why these subjects? BRAC University was the major data collection hub of my research. These four subjects have plenty of students who can help me acquire sufficient amount of data for analysis. I made questionnaires for each subject and got it verified by respective faculty members of BRAC University. Later I proceeded to carrying out surveys. Data accumulated from these surveys was to be used for training and testing purposes which I will be explaining in details in the next chapter. For each subject, there were seven questions and the number of questions was selected in such a way that would bring a balance between prediction accuracy and user satisfaction. By user satisfaction it is meant there is a limit to the number of questions that can be asked to a student. Answering say 20 or 30 questions per subject would surely discourage most to continue further. I made the questionnaires into pre-filled links and conducted surveys in BRAC University via different means of social media. Data collection period lasted for about four months.

The questions for each subject were gathered from different sources.I mainly followed [14] while making questionnaires. During this time, I consulted with few faculty members of BRAC University and later they verified the questions. The questionnaires became the basis of my analysis. The verified questions are as follows.



Law:

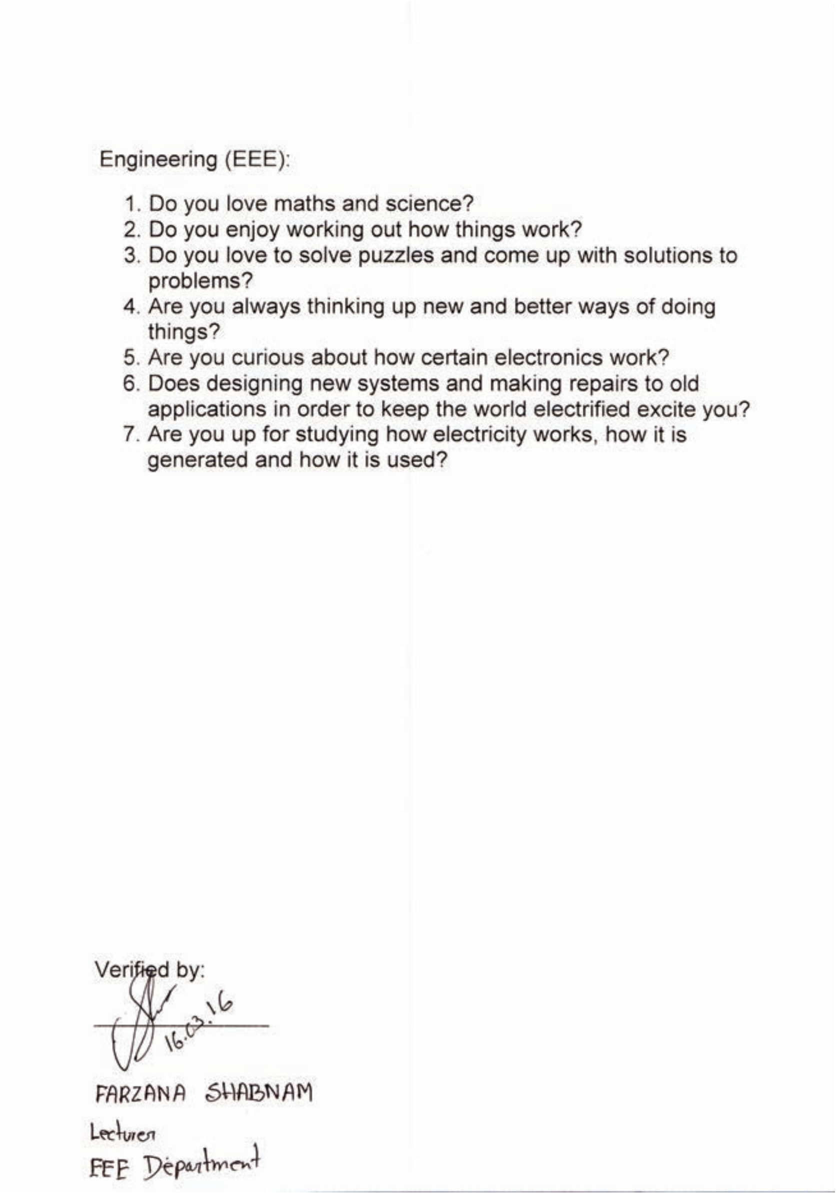
1. Can you think quickly and adapt to a changing situation?
2. Can you work for long hours, looking for loopholes and preparing a client's case?
3. Do you mind filling in repetitive paperwork every single day?
4. Do you enjoy debating with other people and analysing their speeches, looking for weaknesses in their statements?
5. Do you want to help people and work for the community as a whole?
6. Are you interested in representing people?
7. Are you good at remembering details?

Verified by:

. 16-03-16 .

TANIA SULTANA LIPI
LECTURER
SCHOOL OF LAW.

FIGURE 3.1. Questions for Law



Engineering (EEE):

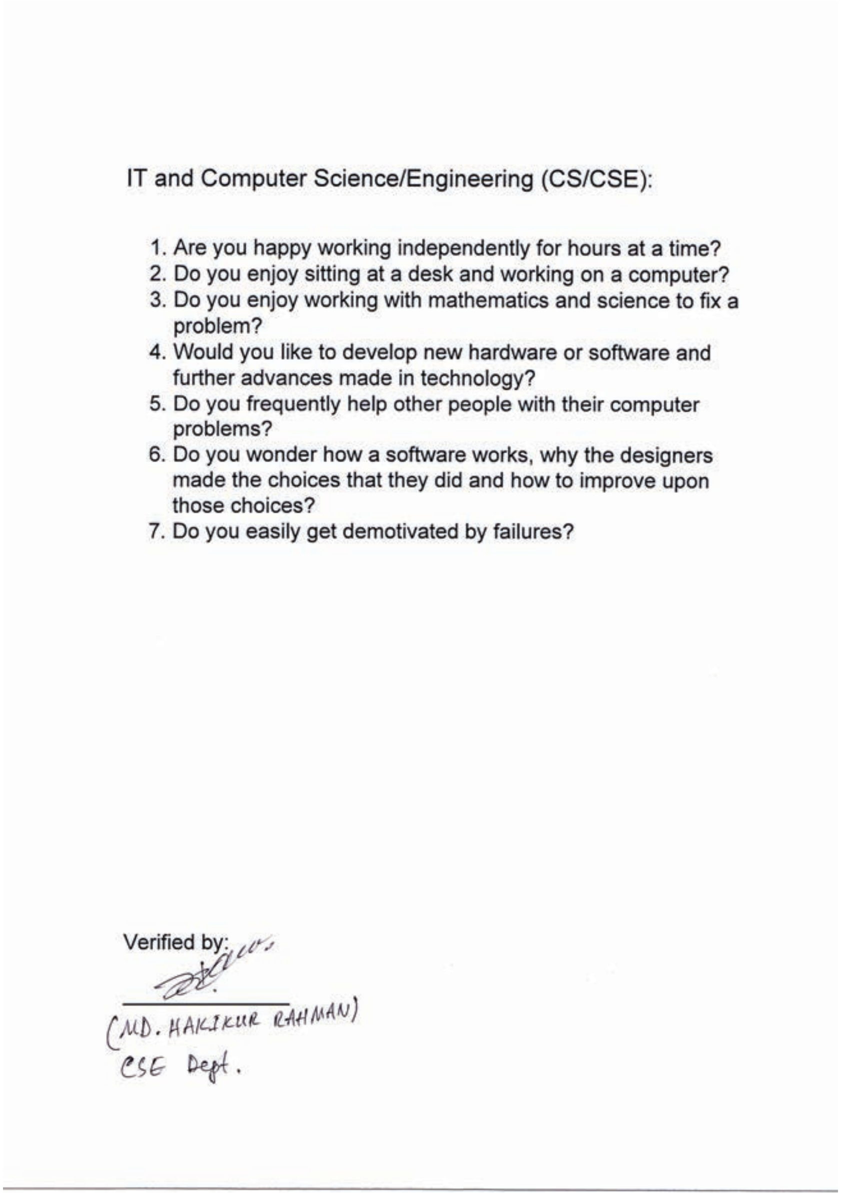
1. Do you love maths and science?
2. Do you enjoy working out how things work?
3. Do you love to solve puzzles and come up with solutions to problems?
4. Are you always thinking up new and better ways of doing things?
5. Are you curious about how certain electronics work?
6. Does designing new systems and making repairs to old applications in order to keep the world electrified excite you?
7. Are you up for studying how electricity works, how it is generated and how it is used?

Verified by:

16.03.16

FARZANA SHABNAM
Lecturer
EEE Department

FIGURE 3.2. Questions for Electrical Engineering



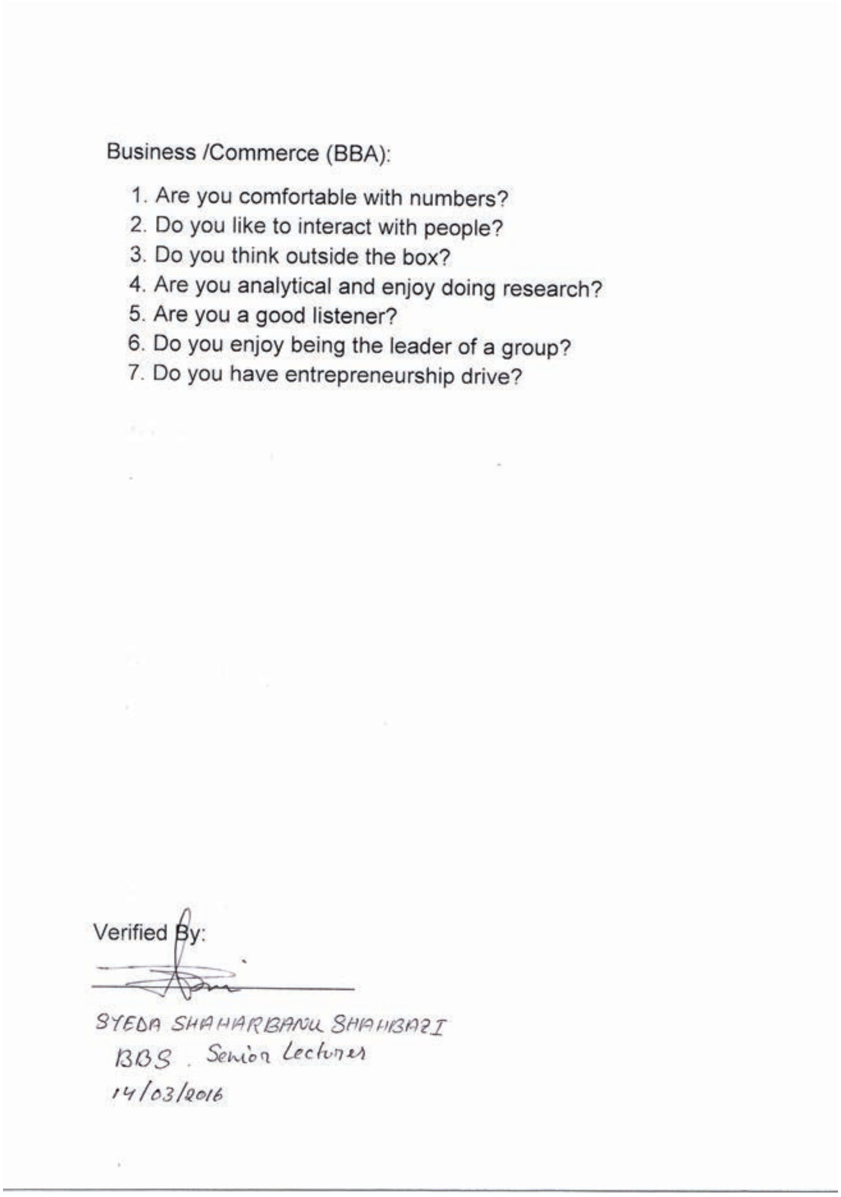
IT and Computer Science/Engineering (CS/CSE):

1. Are you happy working independently for hours at a time?
2. Do you enjoy sitting at a desk and working on a computer?
3. Do you enjoy working with mathematics and science to fix a problem?
4. Would you like to develop new hardware or software and further advances made in technology?
5. Do you frequently help other people with their computer problems?
6. Do you wonder how a software works, why the designers made the choices that they did and how to improve upon those choices?
7. Do you easily get demotivated by failures?

Verified by:

(MD. HAKIKUR RAHMAN)
CSE Dept.

FIGURE 3.3. Questions for Computer Science



Business /Commerce (BBA):

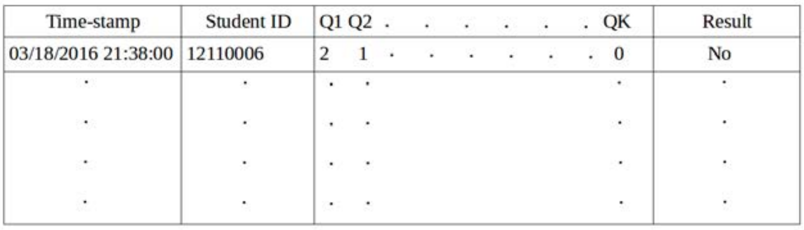
1. Are you comfortable with numbers?
2. Do you like to interact with people?
3. Do you think outside the box?
4. Are you analytical and enjoy doing research?
5. Are you a good listener?
6. Do you enjoy being the leader of a group?
7. Do you have entrepreneurship drive?

Verified By:

SYEDA SHAHARBANU SHAHBAZI
BBS . Senior Lecturer
14/03/2016

FIGURE 3.4. Questions for Business

The diagram below shows structure of the data collected from survey.



Time-stamp	Student ID	Q1 Q2 . . . . QK	Result
03/18/2016 21:38:00	12110006	2 1 0	No
.	.	. . .	.
.	.	. . .	.
.	.	. . .	.
.	.	. . .	.

FIGURE 3.5. Data Instances from Survey

3.2 Data Description

The data for all the subjects are similar in terms of rows and columns. The columns are the attributes Q1, Q2, Q3, Q4, Q5, Q6, Q7 and Result. And the rows contain responses to different attribute scenarios.

Attributes	Attribute description	Possible Values
Q1	Are you comfortable with numbers?	0,1,2
Q2	Do you like to interact with people?	0,1,2
Q3	Do you think outside the box?	0,1,2
Q4	Are you analytical and enjoy doing research?	0,1,2
Q5	Are you a good listener?	0,1,2
Q6	Do you enjoy being the leader of a group?	0,1,2
Q7	Do you have entrepreneurship drive?	0,1,2
Result	Outcome based on the above seven attributes	Yes,No

Table 3.1: Attributes for Business; where 2=Yes, 0=No, 1=To Some Extent.

Attributes	Attribute description	Possible Values
Q1	Are you happy working independently for hours at a time?	0,1,2
Q2	Do you enjoy sitting at a desk and working on a computer?	0,1,2
Q3	Do you enjoy working with mathematics and science to fix a problem?	0,1,2
Q4	Would you like to develop new hardware or software and further advances made in technology?	0,1,2
Q5	Do you frequently help other people with their computer problems?	0,1,2
Q6	Do you wonder how a software works, why the designers made the choices that they did and how to improve upon those choices?	0,1,2
Q7	Do you easily get demotivated by failures?	0,1,2
Result	Outcome based on the above seven attributes	Yes,No

Table 3.2: Attributes for Computer Science; where 2=Yes, 0=No, 1=To Some Extent.

Attributes	Attribute description	Possible Values
Q1	Do you love maths and science?	0,1,2
Q2	Do you enjoy working out how things work?	0,1,2
Q3	Do you love to solve puzzles and come up with solutions to problems?	0,1,2
Q4	Are you always thinking up new and better ways of doing things?	0,1,2
Q5	Are you curious about how certain electronics work?	0,1,2
Q6	Does designing new systems and making repairs to old applications in order to keep the world electrified excite you?	0,1,2
Q7	Are you up for studying how electricity works, how it is generated and how it is used?	0,1,2
Result	Outcome based on the above seven attributes	Yes,No

Table 3.3: Attributes for Electrical Engineering; where 2=Yes, 0=No, 1=To Some Extent.

Attributes	Attribute description	Possible Values
Q1	Can you think quickly and adapt to a changing situation?	0,1,2
Q2	Can you work for long hours, looking for loopholes and preparing a client‚Äôs case?	0,1,2
Q3	Do you mind filling in repetitive paperwork every single day?	0,1,2
Q4	Do you enjoy debating with other people and analyzing their speeches, looking for weaknesses in their statements?	0,1,2
Q5	Do you want to help people and work for the community as a whole?	0,1,2
Q6	Are you interested in representing people?	0,1,2
Q7	Are you good at remembering details?	0,1,2
Result	Outcome based on the above seven attributes	Yes,No

Table 3.4: Attributes for Law; where 2=Yes, 0=No, 1=To Some Extent.

3.3 Applying Naive Bayes

3.3.1 Introduction to Naive Bayes Classifier

The Naive Bayes classifier is a simple classifier which uses probability to make predictions. It is mainly built on Bayes theorem. The presumptions it makes are strongly naive but still it has been proven to perform quite well in many real world applications[8]. The classifier is also referred to as Idiot Bayes, Naive Bayes or Simple Bayes[9]. A more descriptive term for the underlying probability model would be independent feature model. In simple terms, a Naive Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature. For instance, an object may be considered to be a ball if it is round, bouncy, about 4 inches to 8 inches in diameter. Even if these features depend on each other or upon the presence of other features, a Naive Bayes classifier considers all of these properties to independently contribute to the probability that the object is a ball.

Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood;

in other words, one can work with the Naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

In spite of their naive design and apparently over-simplified assumptions, Naive Bayes classifiers have worked brilliantly in solving many complex problems. It is one of the most effective and efficient learning algorithms for data mining and machine learning[10].

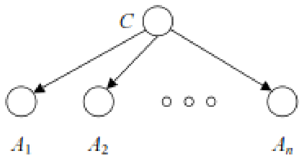
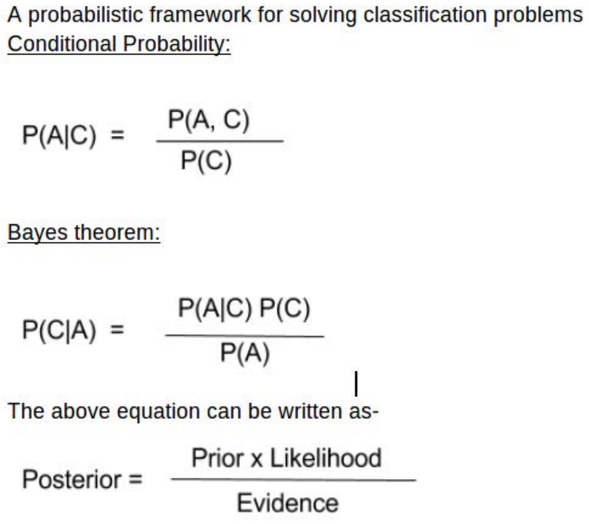


FIGURE 3.6. Structure of Naive Bayes

3.3.2 Probabilistic Model

The discussion so far has derived the independent feature model, that is, the naive Bayes probability model. The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as MAP decision rule.



A probabilistic framework for solving classification problems

Conditional Probability:

$$P(A|C) = \frac{P(A, C)}{P(C)}$$

Bayes theorem:

$$P(C|A) = \frac{P(A|C)\,P(C)}{P(A)}$$

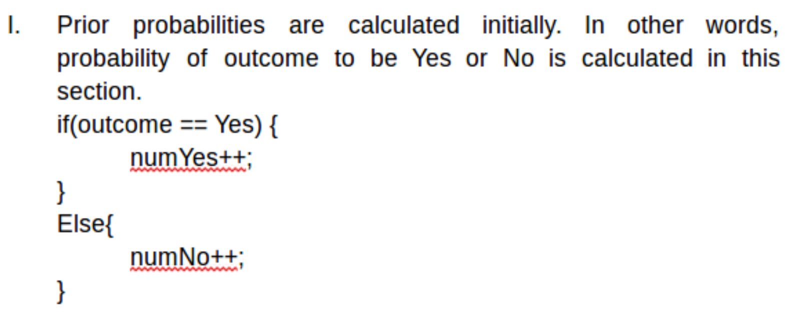
The above equation can be written as-

$$\text{Posterior} = \frac{\text{Prior x Likelihood}}{\text{Evidence}}$$

3.3.3 Algorithm

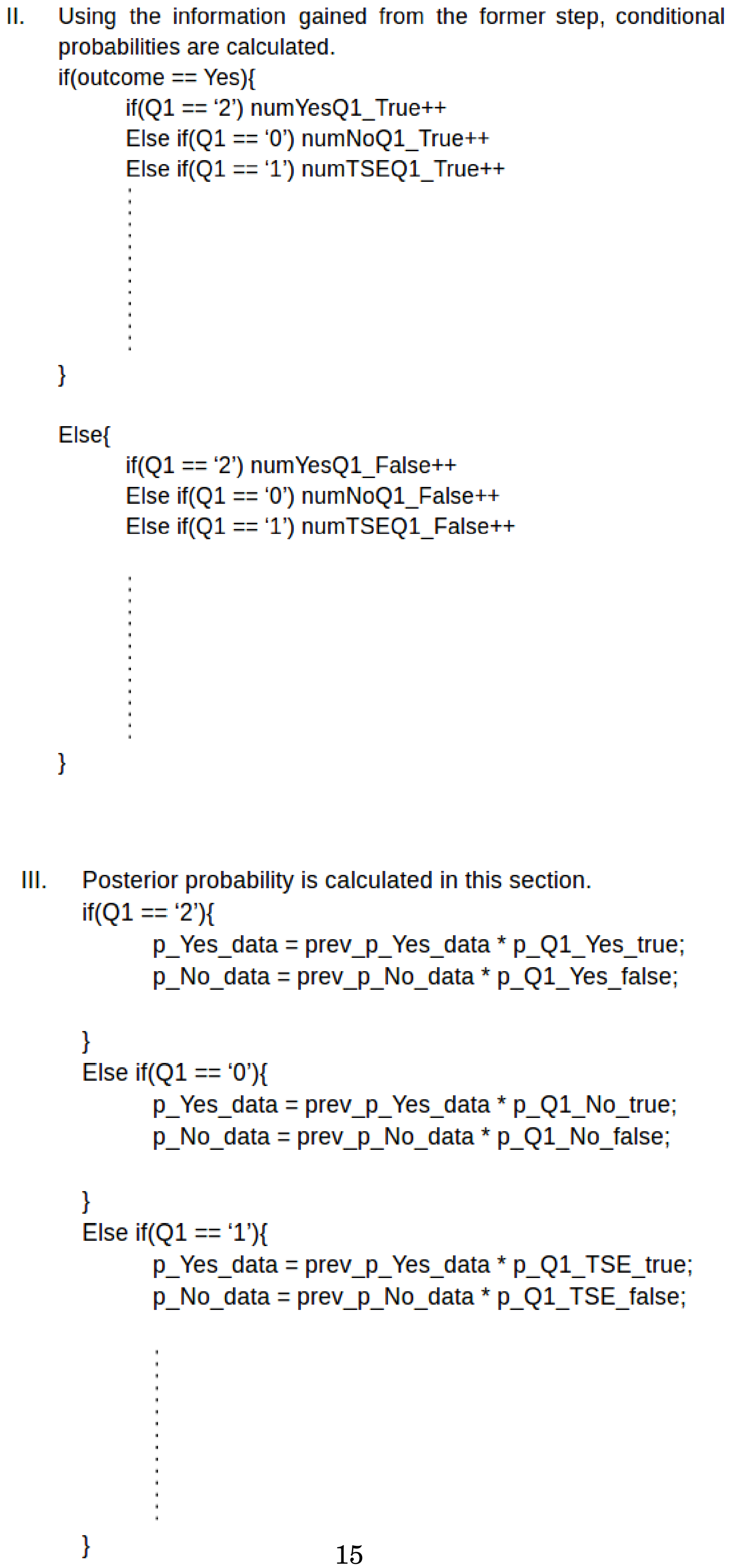
In our proposed model, Naive Bayes assumes that all variables are mutually independent. The attribute variables are Q1, Q2, Q3, Q4, Q5, Q6, Q7 and the result is the outcome variable. Our model takes the seven attributes as input and firstly it calculates the prior probabilities of outcome variables, then the conditional probabilities of the attribute variables are calculated. Lastly, using the previous calculation, the posterior probabilities of the outcome variables are calculated. Based on the posterior probabilities, prediction of whether to study a particular subject or not to study a particular subject is made.

The steps are as follows-



I. Prior probabilities are calculated initially. In other words, probability of outcome to be Yes or No is calculated in this section.

```
if(outcome == Yes) {
        numYes++;
}
Else{
        numNo++;
}
```



II. Using the information gained from the former step, conditional probabilities are calculated.

```
if(outcome == Yes){
        if(Q1 == ‘2’) numYesQ1_True++
        Else if(Q1 == ‘0’) numNoQ1_True++
        Else if(Q1 == ‘1’) numTSEQ1_True++
        ⋮
}

Else{
        if(Q1 == ‘2’) numYesQ1_False++
        Else if(Q1 == ‘0’) numNoQ1_False++
        Else if(Q1 == ‘1’) numTSEQ1_False++

        ⋮
}
```

III. Posterior probability is calculated in this section.

```
if(Q1 == ‘2’){
        p_Yes_data = prev_p_Yes_data * p_Q1_Yes_true;
        p_No_data = prev_p_No_data * p_Q1_Yes_false;

}
Else if(Q1 == ‘0’){
        p_Yes_data = prev_p_Yes_data * p_Q1_No_true;
        p_No_data = prev_p_No_data * p_Q1_No_false;

}
Else if(Q1 == ‘1’){
        p_Yes_data = prev_p_Yes_data * p_Q1_TSE_true;
        p_No_data = prev_p_No_data * p_Q1_TSE_false;

        ⋮
}
```

3.3.4 Solving Zero Frequency Problem

If one of the conditional probabilities is zero, then the entire expression becomes zero. To solve this problem, quite a few techniques can be used. My first approach was to not consider the ones with zero probabilities. I excluded them during calculation.

Another way to solve this problem is to use Laplace Smoothing. It is also referred to as add-one smoothing. This technique adds one to every combination of category and categorical variable. This helps since it prevents knocking out an entire class just because of one variable. Since we add one to each cell, the proportions are essentially the same. Usually, the more data there is, the smaller the impact the added one will have on a Naive Bayes model[11].

3.4 Comparison with Logistic Regression

3.4.1 Introduction to Logistic Regression

Functions of the form P(Y|X) where Y is discrete valued and X can be either discrete or continuous variables can be determined by Logistic Regression. It supposes a parametric form for the distribution P(Y|X), then directly deduces its parameters from the training data[5]. Logistic regression is somewhat similar to linear regression but it uses a binomial response variable. A model designed with logistic regression will represent the likelihood of an outcome with respect to individual features[4]. As stated in [7], logistic regression should be used when an outcome variable takes only two values such as Yes/No or 0/1. Our proposed model has two outcome values which are Yes and No. Therefore, we have chosen logistic regression for comparison purposes.

Logistic Regression estimates class probabilities directly[6]:

$$\Pr[1 \mid a_1, a_2, \ldots, a_k] = 1/(1 + \exp(-w_0 - w_1 a_1 - \ldots - w_k a_k))$$

Here, it chooses weights to maximize the log-likelihood[6]:

$$\sum_{i=1}^{n}(1-x^{(i)})\log(1-\Pr[1 \mid a_1^{(1)}, a_2^{(2)}, \ldots, a_k^{(k)}])+x^{(i)}\log(\Pr[1 \mid a_1^{(1)}, a_2^{(2)}, \ldots, a_k^{(k)}])$$

3.4.2 Relationship between Naive Bayes and Logistic Regression

Naive Bayes is sometimes referred to as a generative classifier and Logistic Regression is sometimes referred to as a discriminative classifier. The former directly estimates for P(Y) and P(X|Y), whereas the latter directly estimates parameters of P(Y|X). A variation of Naive Bayes classifier called Gaussian Naive Bayes classifier tends to be similar to Logistic Regression when the number of training data reaches infinity. In cases where training data is available, Logistic Regression outperforms GNB. But, GNB outperforms Logistic Regression when the training data is scarce[5]. Overall, Logistic Regression is a linear classifier over X. provided the Naive Bayes assumptions hold, the classifiers produced by GNB and Logistic Regression are identical when the training data is infinite. On the other hand, if these presumptions do not hold, the Naive Bayes will perform less accurately than Logistic Regression. In other words, Naive Bayes is a learning algorithm with greater bias, but lower variance, than Logistic Regression. If this bias is appropriate given the actual data, Naive Bayes will be preferred. Otherwise, Logistic Regression will be preferred[5].

CHAPTER 4

Experiment and Results

4.1 Training phase

Our model was trained with BBA, CS, EEE and Law data sets. The number of training data overall was not in abundance. This mainly lead to the selection of Naive Bayes algorithm as it is known to perform well when training data is limited.

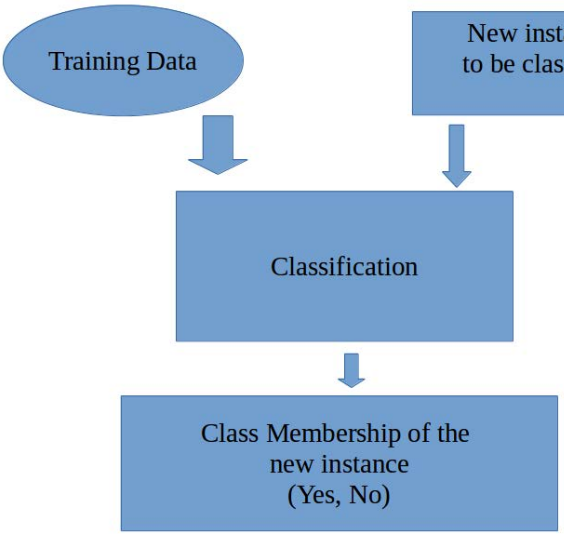

Training Data
New instance to be classified
Classification
Class Membership of the new instance (Yes, No)

FIGURE 4.1. Classification

The figure below shows the amount of training data for each data set.

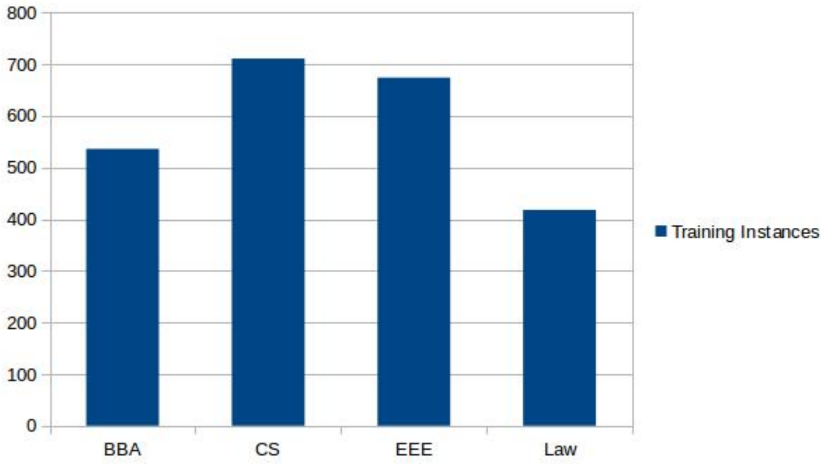


FIGURE 4.2. Training data

The algorithm discussed in the previous chapter deals with the zero frequency problem in two ways. Below is the comparison of the two approaches.As the two approaches result similar accuracy we can choose any one of them to proceed. However, we select the Laplace Smoothing technique as it has been previously used in many scenarios as such.

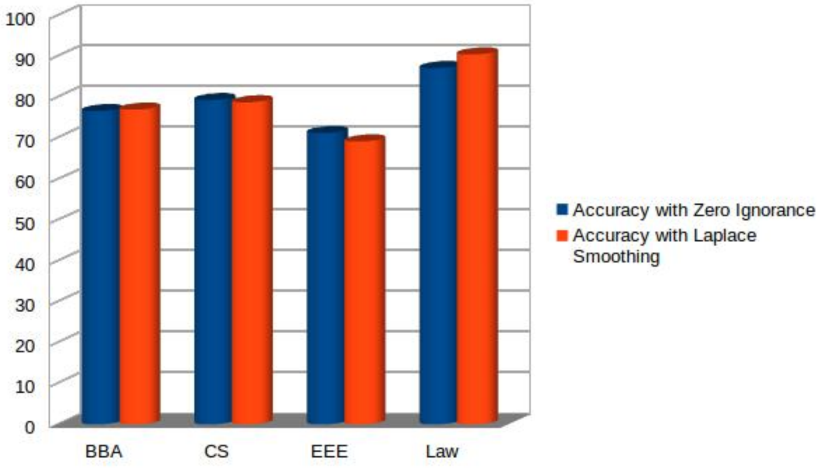
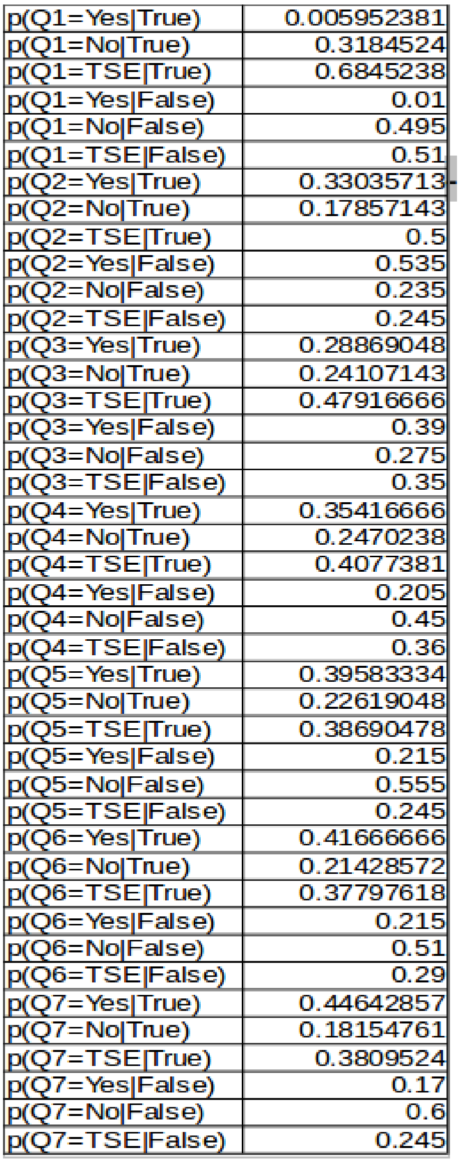
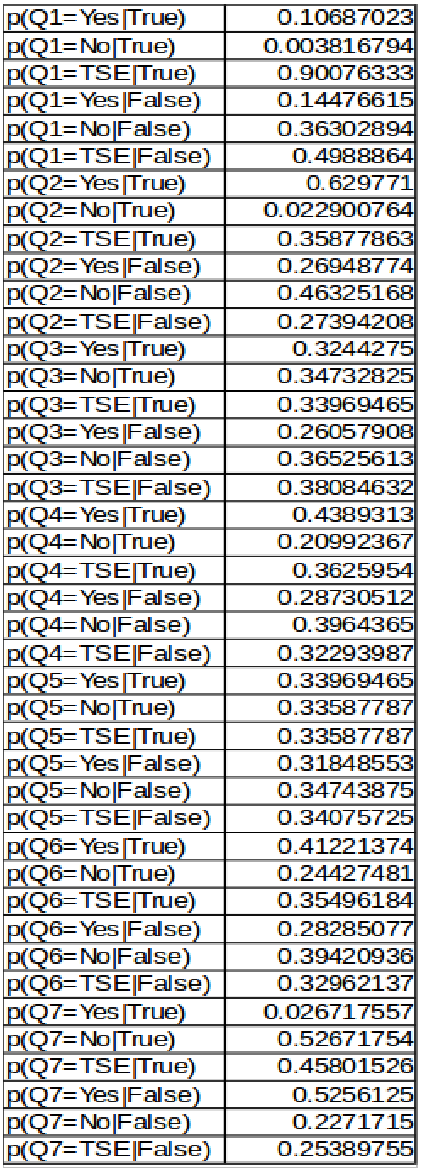


FIGURE 4.3. Comparison of the two approaches to solve zero frequency problem



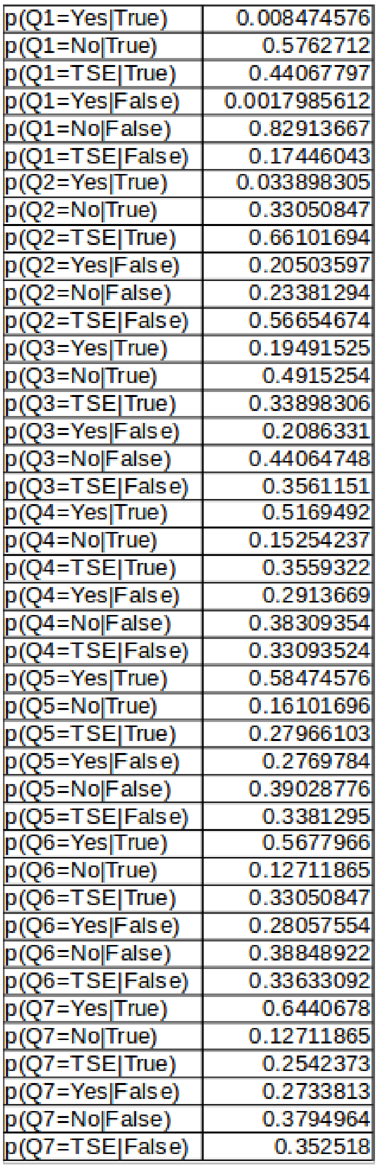
p(Q1=Yes|True)	0.005952381
p(Q1=No|True)	0.3184524
p(Q1=TSE|True)	0.6845238
p(Q1=Yes|False)	0.01
p(Q1=No|False)	0.495
p(Q1=TSE|False)	0.51
p(Q2=Yes|True)	0.33035713
p(Q2=No|True)	0.17857143
p(Q2=TSE|True)	0.5
p(Q2=Yes|False)	0.535
p(Q2=No|False)	0.235
p(Q2=TSE|False)	0.245
p(Q3=Yes|True)	0.28869048
p(Q3=No|True)	0.24107143
p(Q3=TSE|True)	0.47916666
p(Q3=Yes|False)	0.39
p(Q3=No|False)	0.275
p(Q3=TSE|False)	0.35
p(Q4=Yes|True)	0.35416666
p(Q4=No|True)	0.2470238
p(Q4=TSE|True)	0.4077381
p(Q4=Yes|False)	0.205
p(Q4=No|False)	0.45
p(Q4=TSE|False)	0.36
p(Q5=Yes|True)	0.39583334
p(Q5=No|True)	0.22619048
p(Q5=TSE|True)	0.38690478
p(Q5=Yes|False)	0.215
p(Q5=No|False)	0.555
p(Q5=TSE|False)	0.245
p(Q6=Yes|True)	0.41666666
p(Q6=No|True)	0.21428572
p(Q6=TSE|True)	0.37797618
p(Q6=Yes|False)	0.215
p(Q6=No|False)	0.51
p(Q6=TSE|False)	0.29
p(Q7=Yes|True)	0.44642857
p(Q7=No|True)	0.18154761
p(Q7=TSE|True)	0.3809524
p(Q7=Yes|False)	0.17
p(Q7=No|False)	0.6
p(Q7=TSE|False)	0.245

FIGURE 4.4. Probability distribution for BBA



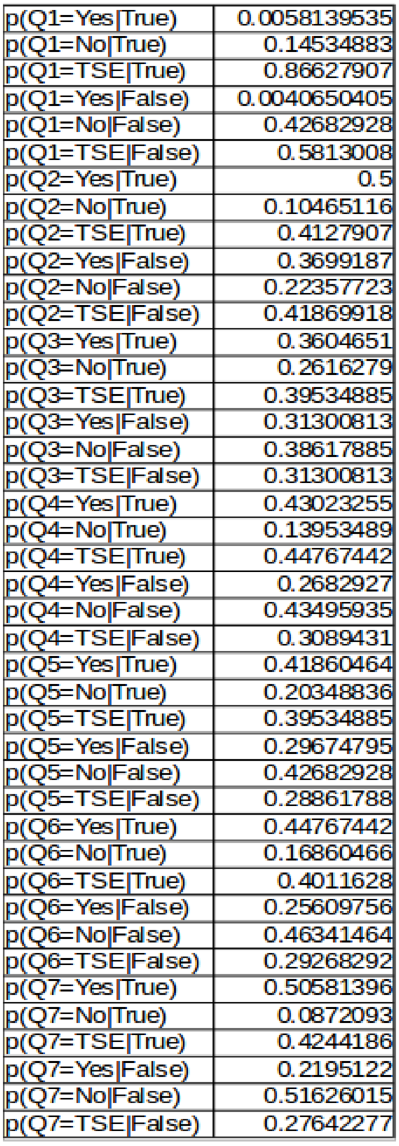
p(Q1=Yes|True)	0.10687023
p(Q1=No|True)	0.003816794
p(Q1=TSE|True)	0.90076333
p(Q1=Yes|False)	0.14476615
p(Q1=No|False)	0.36302894
p(Q1=TSE|False)	0.4988864
p(Q2=Yes|True)	0.629771
p(Q2=No|True)	0.022900764
p(Q2=TSE|True)	0.35877863
p(Q2=Yes|False)	0.26948774
p(Q2=No|False)	0.46325168
p(Q2=TSE|False)	0.27394208
p(Q3=Yes|True)	0.3244275
p(Q3=No|True)	0.34732825
p(Q3=TSE|True)	0.33969465
p(Q3=Yes|False)	0.26057908
p(Q3=No|False)	0.36525613
p(Q3=TSE|False)	0.38084632
p(Q4=Yes|True)	0.4389313
p(Q4=No|True)	0.20992367
p(Q4=TSE|True)	0.3625954
p(Q4=Yes|False)	0.28730512
p(Q4=No|False)	0.3964365
p(Q4=TSE|False)	0.32293987
p(Q5=Yes|True)	0.33969465
p(Q5=No|True)	0.33587787
p(Q5=TSE|True)	0.33587787
p(Q5=Yes|False)	0.31848553
p(Q5=No|False)	0.34743875
p(Q5=TSE|False)	0.34075725
p(Q6=Yes|True)	0.41221374
p(Q6=No|True)	0.24427481
p(Q6=TSE|True)	0.35496184
p(Q6=Yes|False)	0.28285077
p(Q6=No|False)	0.39420936
p(Q6=TSE|False)	0.32962137
p(Q7=Yes|True)	0.026717557
p(Q7=No|True)	0.52671754
p(Q7=TSE|True)	0.45801526
p(Q7=Yes|False)	0.5256125
p(Q7=No|False)	0.2271715
p(Q7=TSE|False)	0.25389755

FIGURE 4.5. Probability distribution for CS



p(Q1=Yes|True)	0.008474576
p(Q1=No|True)	0.5762712
p(Q1=TSE|True)	0.44067797
p(Q1=Yes|False)	0.0017985612
p(Q1=No|False)	0.82913667
p(Q1=TSE|False)	0.17446043
p(Q2=Yes|True)	0.033898305
p(Q2=No|True)	0.33050847
p(Q2=TSE|True)	0.66101694
p(Q2=Yes|False)	0.20503597
p(Q2=No|False)	0.23381294
p(Q2=TSE|False)	0.56654674
p(Q3=Yes|True)	0.19491525
p(Q3=No|True)	0.4915254
p(Q3=TSE|True)	0.33898306
p(Q3=Yes|False)	0.2086331
p(Q3=No|False)	0.44064748
p(Q3=TSE|False)	0.3561151
p(Q4=Yes|True)	0.5169492
p(Q4=No|True)	0.15254237
p(Q4=TSE|True)	0.3559322
p(Q4=Yes|False)	0.2913669
p(Q4=No|False)	0.38309354
p(Q4=TSE|False)	0.33093524
p(Q5=Yes|True)	0.58474576
p(Q5=No|True)	0.16101696
p(Q5=TSE|True)	0.27966103
p(Q5=Yes|False)	0.2769784
p(Q5=No|False)	0.39028776
p(Q5=TSE|False)	0.3381295
p(Q6=Yes|True)	0.5677966
p(Q6=No|True)	0.12711865
p(Q6=TSE|True)	0.33050847
p(Q6=Yes|False)	0.28057554
p(Q6=No|False)	0.38848922
p(Q6=TSE|False)	0.33633092
p(Q7=Yes|True)	0.6440678
p(Q7=No|True)	0.12711865
p(Q7=TSE|True)	0.2542373
p(Q7=Yes|False)	0.2733813
p(Q7=No|False)	0.3794964
p(Q7=TSE|False)	0.352518

FIGURE 4.6. Probability distribution for EEE



p(Q1=Yes|True)	0.0058139535
p(Q1=No|True)	0.14534883
p(Q1=TSE|True)	0.86627907
p(Q1=Yes|False)	0.0040650405
p(Q1=No|False)	0.42682928
p(Q1=TSE|False)	0.5813008
p(Q2=Yes|True)	0.5
p(Q2=No|True)	0.10465116
p(Q2=TSE|True)	0.4127907
p(Q2=Yes|False)	0.3699187
p(Q2=No|False)	0.22357723
p(Q2=TSE|False)	0.41869918
p(Q3=Yes|True)	0.3604651
p(Q3=No|True)	0.2616279
p(Q3=TSE|True)	0.39534885
p(Q3=Yes|False)	0.31300813
p(Q3=No|False)	0.38617885
p(Q3=TSE|False)	0.31300813
p(Q4=Yes|True)	0.43023255
p(Q4=No|True)	0.13953489
p(Q4=TSE|True)	0.44767442
p(Q4=Yes|False)	0.2682927
p(Q4=No|False)	0.43495935
p(Q4=TSE|False)	0.3089431
p(Q5=Yes|True)	0.41860464
p(Q5=No|True)	0.20348836
p(Q5=TSE|True)	0.39534885
p(Q5=Yes|False)	0.29674795
p(Q5=No|False)	0.42682928
p(Q5=TSE|False)	0.28861788
p(Q6=Yes|True)	0.44767442
p(Q6=No|True)	0.16860466
p(Q6=TSE|True)	0.4011628
p(Q6=Yes|False)	0.25609756
p(Q6=No|False)	0.46341464
p(Q6=TSE|False)	0.29268292
p(Q7=Yes|True)	0.50581396
p(Q7=No|True)	0.0872093
p(Q7=TSE|True)	0.4244186
p(Q7=Yes|False)	0.2195122
p(Q7=No|False)	0.51626015
p(Q7=TSE|False)	0.27642277

FIGURE 4.7. Probability distribution for Law

4.2 Testing phase

Testing of this kind of model can be done in various ways. Some of the ways are cross-validation, percentage spit and using a test set. For our model, we have used individual test sets for the data sets. The test instances for each data set are as follows.

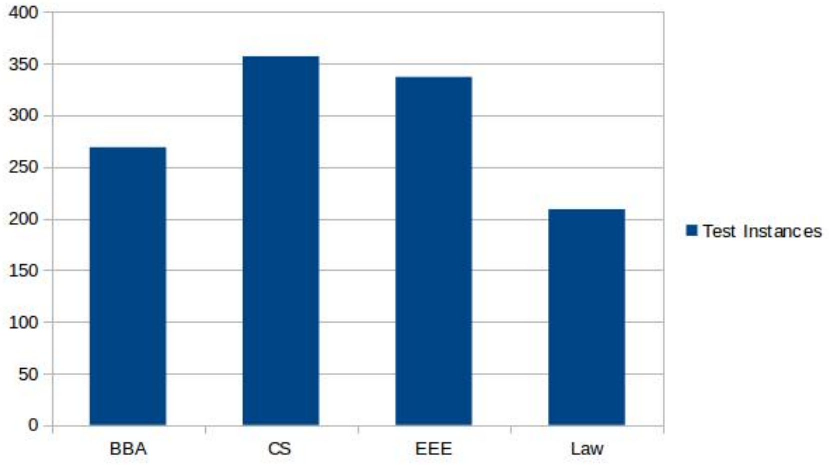


FIGURE 4.8. Test Data

4.3 Experimental Result

The observations and results are discussed in detail in this section. The figure below shows the classified instances.

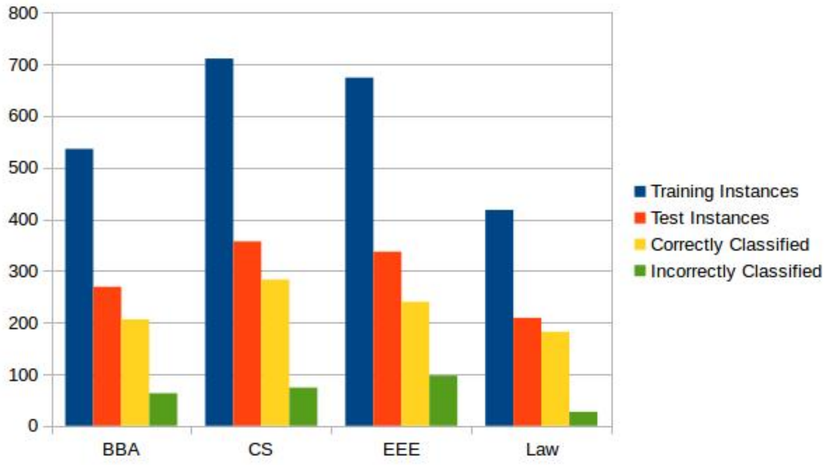


FIGURE 4.9. Classified Data

The following are the confusion matrices for the four data sets.

	Class = Yes	Class = No
Class = Yes	103	12
Class = No	51	103

Table 4.1: Confusion matrix for BBA.

	Class = Yes	Class = No
Class = Yes	67	59
Class = No	15	216

Table 4.2: Confusion matrix for CS.

	Class = Yes	Class = No
Class = Yes	17	94
Class = No	3	223

Table 4.3: Confusion matrix for EEE.

	Class = Yes	Class = No
Class = Yes	61	26
Class = No	1	121

Table 4.4: Confusion matrix for Law.

The prediction accuracy is displayed in the diagram below.

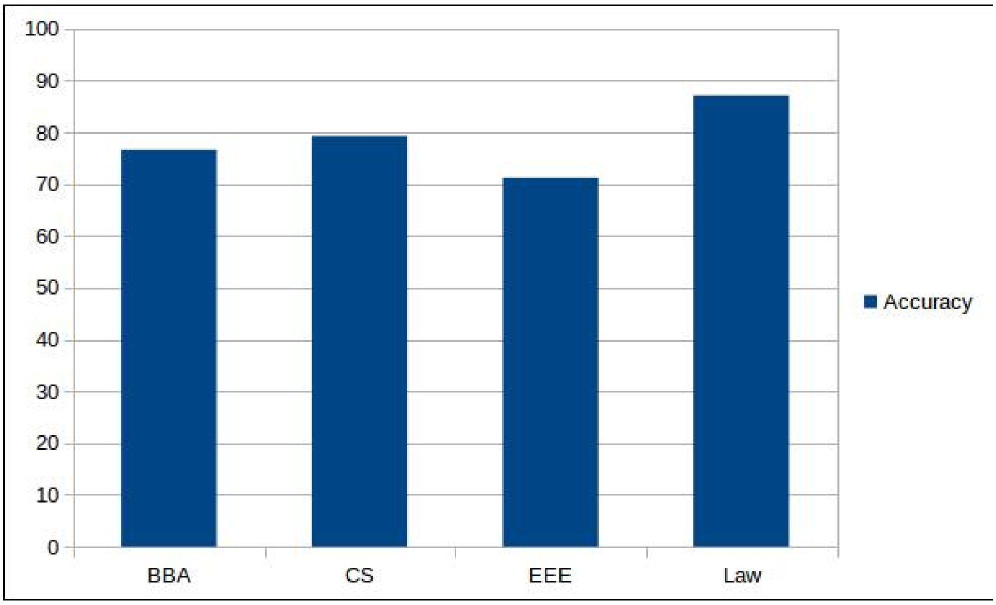


FIGURE 4.10. Accuracy with Naive Bayes

Our proposed Naive Bayes model is compared with Weka's Logistic Regression.

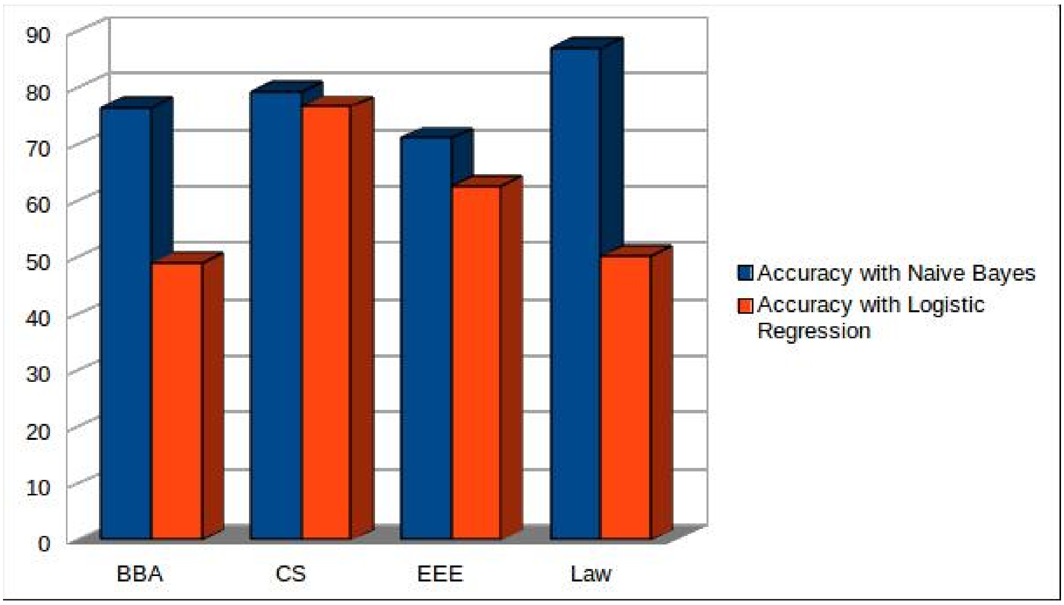


FIGURE 4.11. Accuracy Comparison with Logistic Regression

CHAPTER 5

FUTURE WORK

5.1 Holland Codes

The Holland Code is a theory developed by the psychologist John L. Holland. The theory states that Holland Occupational Themes or RIASEC is a model of careers and vocational choice based on personality types. The RIASEC model is broken down as follows. The R stands for Realistic or Doers, I equals Investigative or thinkers, A for Artistic or Creators, S for Social or Helpers, E for Enterprising or Persuaders and lastly C for Conventional or Organizers. Most of us do not fall under one Holland Code but under a combination of few.

Holland Codes by Subject for BRAC University-

Computer Science - IRE
Economics - ICE
Anthropology - IRE
Electrical and Electronic Engineering - RES
English and Humanities - SAI
Mathematics - ICA
Physics - IR
Business Administration - ECS

5.2 Project Proposal

With sufficient data and human expertise, a Subject Prediction System specific to an institution can be built in the near future. A system can be made where initially a student is expected to select Holland Codes according to his/her personality. List of subjects will be filtered according to specific Holland Codes. Then, the student will need to answer subject specific questions in random order from the filtered subject list. For example, for Computer Science, the student is required to answer seven questions. Based on the answers, probability of Yes and probability of No will be calculated by Naive Bayes algorithm. Lastly, after answering for all the subjects only the ones with probability of Yes will be displayed.

Chapter 6

Conclusion

In this paper, Naive Bayes Algorithm was used to make a probabilistic model. The resulting model was compared with Logistic Regression. The observations indicate that Naive Bayes outperforms Logistic Regression. Therefore, it can be concluded that the proposed system successfully classifies the given data into respective categories. Now, it can be determined whether a student is interested in a particular subject or not. Here, a base model has been proposed. More work can be done here. This model can be used in the back end of a system which would make predictions by asking questions. The answers received would be analyzed and a decision can be made based on that.

BIBLIOGRAPHY

[1] "Heart Disease Prediction System using Naive Bayes", INTERNATIONAL JOURNAL OF ENHANCED RESEARCH IN SCIENCE TECHNOLOGY & ENGINEERING, vol. 2, no. 3, 2013.

[2] "Naive Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer", International Journal of Computer Applications, vol. 92, no. 10, 2014.

[3] "DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE", Journal of Economics and Business, vol., no. 1, 2012.

[4] S. Sperandei, "Understanding logistic regression analysis", Biochemia Medica, pp. 12-18, 2014.

[5] T. Mitchell, GENERATIVE AND DISCRIMINATIVE CLASSIFIERS: NAIVE BAYES AND LOGISTIC REGRESSION. 2016.

[6] I. Witten, "Data Mining with Weka", University of Waikato, New Zealand.

[7] "When to Use Logistic Regression | LogisticRegressionAnalysis.com", Logisticregressionanalysis.com,2016.[Online].Available:http://logisticregressionanalysis.com/33-when-to-use-logistic-regression/. [Accessed: 01- Apr- 2016].

[8] V. Vryniotis and V. Vryniotis, "Machine Learning Tutorial: The Naive Bayes Text Classifier | Datumbox", Blog.datumbox.com, 2015. [Online]. Available: http://blog.datumbox.com/machine-learning-tutorial-the-naive-bayes-text-classifier/. [Accessed: 18- Apr- 2016].

[9] E. Keogh, "Naïve Bayes Classifier".

[10] H. Zhang, "The Optimality of Naive Bayes".

[11] "Naive Bayes Classification Simple Explanation–Learn By Marketing". Learnbymarketing.com. N.p., 2016. Web. 18 Apr. 2016.

[12] "UNF - Career Services - Holland Codes and Majors at UNF", Unf.edu. [Online]. Available:
https://www.unf.edu/careerservices/Holland_Codes_and_Majors_at_UNF.aspx.
[Accessed: 02- Feb- 2016].

[13] I. IONIi$, "DATA MINING FOR PREDICTING THE MILITARY CAREER CHOICE", 2015.

[14] "Studying IT and Computer Science/Engineering | Study My Way", Studymyway.com.Online].Available:http://www.studymyway.com/study-options/studying-it-and-computer-scienceengineering. [Accessed: 03- Apr- 2016].