Prediction of Potential Future IT Personnel in Bangladesh using Machine Learning Classifier

Md. Hasnat Parvez[1], Most. Moriom Khatun[2*], Sayed Mohsin Reza[3], Md. Mahfujur Rahman[4], Md. Fazlul Karim Patwary[5]

[1]Lecturer, Department of Computer Science and Engineering, Gono University, Savar, Dhaka, **BANGLADESH**
[2]Sher-e-Bangla Agricultural University, Dhaka, **BANGLADESH**
[3,4]Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, **BANGLADESH**
[5]Professor, Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, **BANGLADESH**

*Corresponding Contact:
Email: moriom.rima@gmail.com

ABSTRACT

Bangladesh is one of the most promising developing countries in IT sector, where people from several disciplines and experiences are involved in this sector. However, no direct analysis in this sector is published yet, which covers the proper guideline for predicting future IT personnel. Hence this is not a simple solution, training data from real IT sector are needed and trained several classifiers for detecting perfect results. Machine learning algorithms can be used for predicting future potential IT personnel. In this paper, four different classifiers named as Naive Bayes, J48, Bagging and Random Forest in five different folds are experimented for that prediction. Results are pointed out that Random Forest performs better accuracy than other experimented classifier for future IT personnel prediction. It is mentioned that the standard accuracy measurement process named as Precision, Recall, F-Measure, ROC Area etc. are used for evaluating the results.

Key Words: Future IT Personnel, IT in Developing Country, Machine Learning Classifier

INTRODUCTION

Machine learning algorithms are extremely used for the extraction of valuable information which is obtainable in data warehouses and other database sources. Prediction of potential future IT personnel in developing country such as Bangladesh is very much needed for job arena and administrative level. In the arena of IT job sector of a developing country, it is difficult know how many people currently involved with IT jobs, because of their unavailable of proper database. It is also challenging for future IT graduated students to choose IT based job or Non-IT based jobs. That’s why the prediction of potential IT personnel for job arena and administrative level are considered in this study.

The machine learning algorithms could be applied to find the final result of potential IT personnel is one of the most inspiring work and a difficult task. The machine learning procedures have become a well-known research instrument for potential future IT personnel to identify and interaction among large number of variable quantities. Different procedures of machine learning algorithms use different purpose of uses. From statistics, artificial intelligence and data warehouses, it is very easy to design methods and processes to classify the data for the use of real world applications (Qiang, 2007). From this problem we can use machine learning algorithms for prediction. In this research we used Naive Bayes, J48, Bagging and Random Forest classification algorithms. These machine learning algorithms are frequently used for predicting future IT personnel (Gaganjot, 2014)

Knowledge discovery with machine learning classifier is the procedure of finding previously unknown and potentially interesting patterns and relations in large databases (Gaganjot, 2014). This research work examines the various classification algorithms compared using WEKA tool environment and results are discussed. This would classify the IT or Non-IT graduated dataset which will predict future IT personnel. Several types of classification algorithms were selected and the dataset was applied with these algorithms. This work also creates simple strategy for the researcher or programmer to select the input parameter for classification creation of potential future IT personnel data in administration domain (Asma, 2011).

RQ: Prediction of Potential Future IT Personnel in Bangladesh

> Prediction of potential future IT personnel in Bangladesh is one of the major issues. Advance prediction will help the administration in planning the creation of job sector. Future prediction and decision can be made based on the knowledge discovery through classification.

RQ: Finding the Best Classifier for that prediction.

> The classifiers used in this research work consist of Naive Bayes, J48, Bagging and Random Forest. We used IT or Non-IT graduated dataset as input for every classifier. Then machine learning tool WEKA process this dataset. Finally we processed result for every classifier and select the most appropriate classifier.

This research presents prediction of potential future IT personnel in Bangladesh. In this research first of all we collect data from IT or Non-IT graduates, Software firm or various multinational companies. We create a well-defined questionnaire for the purpose of data collection from IT or Non-IT graduates, Software firm or various multinational companies. We collected information via mail, personal interview or over telephone. After processing the data, selected attributes are processed in SPSS. In attributes selection, chi-square and graph box are used. Then WEKA machine learning tool is used for classification. Naive Bayes, J48, Bagging and Random Forest Classifiers with 5, 10, 15, 20 and 25 fold are considered for classification. Accuracy matrix for each cross validation fold has been analyzed for measurement. Finally best classifier from the value of accuracy matrix and the percentage of correctly classified instances are selected.

IT JOB MARKET STUDY OF BANGLADESH

Bangladesh, as a country with a vast population, is full of potentials. Information Technology is the key to utilize those potentials. From 2009, Bangladesh has been experiencing numerous groundbreaking developments in IT sector. (Gaganjot, 2014) According to the BASIS 2012 survey the ICT industry has consistently grown in recent years at 20 to 30 percent per annum. Over 800 registered ICT companies generated total revenues of approximately $250 million. More than 75 percent of companies are involved in customized application development and maintenance, 50 percent are dedicated to IT enabled services, and 45 percent offer E-commerce/Web services. The survey also shows that 60 percent of companies solely focus on the domestic market.

In addition to the registered workforce, thousands of independent freelancers offer their services at online market places and 5.500 students annually graduate from ICT courses at more than 80 public and private universities. According to the ITC Exporter Directory there are over 10,000 ICT freelancers active in Bangladesh in 2012. According to BASIS the average monthly compensation for an ICT employee is approximately US$ 200 per month. According to BASIS estimations within the next five years over 150,000 Software and IT professionals will be working in the Bangladeshi IT industry. 65 percent of BASIS' member companies have between 10 and 30 employees and it is estimated that not more than 20 companies have over 100 employees. There are around 100 software houses, 35 data entry centers, thousands of formal and informal IT training centers and numerous computer shops (M Alam, 2012). Approximately 30,000 professionals, majority IT and other graduates, are employed in the industry. Today there are more than 320 software firms and IT services firms (Robert, 2010). Bangladesh now has 500,000 freelancers which makes the second height exporting sectors in fiscal year 2015-2016 (Gaganjot, 2014).

LITERATURE STUDY

In extracting actionable knowledge from decision trees novel post processing technique was used to extract actionable knowledge from decision tree. Customer relationship management was used as a case. Two cases were considered in this paper. One was unlimited resources cases and another one was limited resources cases. In both cases target is to maximize the profit. This paper described finding optimal solution for the limited resource problems and designing a greedy heuristic algorithm to solve it efficiently (Zengyou, 2005).

In post processing decision trees to extract actionable knowledge proposed a method to find out the best action rules. The best k-action rules are selected on the basis of maximizing the profit of moving from one decision to another. A new algorithm is presented that proposes action to change customer status from an undesired status to desired one (Feixiang, 2012). In data mining for actionable knowledge: a survey authors mine the actionable information from the perspective of data mining tasks and algorithms. The tasks such as clustering, association, outlier's detection etc are described along with the actionable techniques (Zengyou, 2005).

In actionable knowledge, mining from improved post processing decision trees, authors have presented an innovative algorithm applying decision trees to maximize the profit function under resource constraints. In this paper they take any decision tree as input and mine the best actions to be chosen in order to maximize the expected net profit of all the customers (M Alam, 2012).

In decision tree, post processing for extraction of actionable knowledge, new algorithms are discovered that suggest actions to change customers from an undesired status to a desired one while maximizing the expected net profit. These algorithms can discover cost effective actions to transform customers from undesirable classes to desirable ones (Rahul, 2012).

In data mining for direct marketing: problems and solutions authors discuss methods of coping with some problems based on their experience on direct marketing projects using data mining. During data mining several specific problems may arise. The class distribution is extremely imbalanced; the predictive accuracy is no longer suitable for evaluating learning methods and the number of examples can be too large (Charles, 1998).

In decision tree discovery for the diagnosis of type II diabetes of the patients is calculated by using the decision tree in two segments: data pre-processing in which the attributes are identified and second is diabetes prediction model constructed with the help of using the decision tree method. Both the segments are executed using WEKA data mining tool (Qiang, 2007).

In predicting disease by using data mining based on healthcare information system hypertension has been predicted by generating J48 and Naive Bayesian classifiers in WEKA. A slight improvement of ensembles five J48 classifier was seen over pure Naive Bayesian and J48 in sensitivity, accuracy and F-measure (Feixiang, 2012). In data mining for the diagnosis of type 2 diabetes the diabetes disease analysis is performed by using the artificial meta plasticity on multilayer perceptron. The outcomes conquered by artificial meta plasticity on multilayer perception were compared with Bayesian classifier, decision tree using same database. On the basis of standard deviation decision tree performed the best classification (Alexis, 2012).

In medical data mining with extended WEKA J48, Random Forest, Naive Bayes etc. algorithms are used for disease diagnosis as they led to good accuracy. In experimental comparison of classifiers for breast cancer diagnosis classification algorithms Naive Bayes, decision tree (J48), Sequential Minimal Optimization (SMO), Instance Based for K-Nearest neighbor (IBK) and Multi-Layer Perception are compared by using matrix and classification accuracy. Three different breast cancer databases have been used and classification accuracy is presented on the bases of 10-fold cross validation method. In machine learning methods for the detection of rwis sensor malfunctions has been predicted by generating J48, Naive Bayesian and Bayesnet classifiers. When classification algorithms are run using WEKA, the output is presented as a confusion matrix and statistical results such as the classification error, root mean squared errors and the percentage of correctly classified instances are given by WEKA (Aditya, 2006).

In this paper Naive Bayes, J48, Bagging and Random Forest classification are used to predict future IT personnel in Bangladesh. Output is presented as accuracy matrix and statistical results such as the classification error, root mean squared errors, the percentage of correctly and incorrectly classified instances are given by WEKA. The classification accuracy is presented on the bases of 5-fold, 10-fold, 15-fold, 20-fold and 25-fold cross validation method (Aditya, 2006).

METHODOLOGY

Classification is the process of finding a set of models that describe and distinguish data classes and concepts for the purpose of being able to use the model to predict the class whose label is unknown (Varun, 2011). We compared Naive Bayes, J48, Bagging and Random Forest classifiers. We only considered qualitative and quantitative attributes and removed instances with missing values from the datasets (Pat, 1994). First of all we describe data preprocessing function. Here described how to processed data using SPSS and WEKA tool. Then we described process of steps. Then we described classification algorithms are Naive Bayes, J48, Bagging and Random Forest classifiers. Then we describe evaluation matrix which are TP Rate, FP Rate, Precision, Recall, F-Measure and ROC Area (Pat, 1994).

LAYER 1: ATTRIBUTE SELECTION

We analyzed nearly 181 instances of IT or Non-IT graduates data. In this data set there are 95 instances of IT graduates and 86 instances of Non-IT graduates. Our predictor dataset contained of 57 attributes. In this attributes there are name, designation, duration of job, education, family, family income, social resource, ability and motivational factors, personal preference and interest etc. information. We make a well-designed question for data collection from graduates (Aditya, 2006). We selected graduates for data collection who already graduated from Information Technology, Jahangirnagar University, who admitted into PMIT in IT, JU and who are already involved with various software firm or multinational companies. We collected information via mail, personal interview or over telephone.

Chi-square and Graph box selection strategies are used for selecting attributes (Aditya, 2006). If any attributes chi-square asymp sig(2-Sided) value is less than .05, then those attributes will significant for classification, Otherwise those attributes will insignificant for classification. In this graph bar there is a mean value of one attribute with class value attribute. If the bar chart is level for those attributes then there have not any significant information. That means those attributes are insignificant for classification. MSExcel, Weka and Statistical Package for the Social Sciences (SPSS) have been used for Explorative Data Analysis (EDA) and for classification.

Layer 2: Attribute Classification

Algorithms that classify a given instance into a set of separate categories are called classification algorithms. These algorithms work on a training set to arise up with a model or a set of procedures that classify a given input into one of a set of discrete output values. Most classification algorithms can take inputs in any form discrete or continuous although some of the classification algorithms require all of the inputs also to be discrete. The output is always in the form of a discrete value (Aditya, 2006). To be capable to apply classification algorithms on our IT or Non-IT graduates example we need to convert the output attribute into classes. This is usually pre- pared by discrete selection, which is the procedure of separating a continuous variable into classes. We describe in detail the classification algorithms that have been used in this thesis in the sub-sections below (Aditya, 2006).

Naive Bayes Classifier

The Naive Bayes classifier does not carry out a widespread search through a space of probable descriptions in compare to many induction methods. There are no choices about how to partition the data which way to move in a weight space (Pat, 1994). In probability theory, Bayes theorem relates the conditional and marginal probabilities of two random events. Let x = (x^1, x^2, . . . , x^d) be a d-dimensional instance which has no class label, and our goal is to build a classifier to predict its unknown class label based on Bayes theorem. Let C = C^1, C^2, . . . , C^k be the set of the class labels. P(C_k)is the prior probability of C_k (k = 1, 2, ..., K)that are inferred before new evidence; P(x | C_k) be the conditional probability of seeing the evidence x if the hypothesis C_k is true.

A technique for constructing such classifiers to employ Bayes theorem to obtain:

$$P(Ck\,|\,x) = \frac{P(x|C_K)P(C_K)}{\sum_{K\prime} P(x|C_{K\prime})P(C_{K\prime})} \qquad (4.1)$$

A Naive Bayes classifier assumes that the value of a particular feature of a class is unrelated to the value of any other feature so that,

$$P(x\,|\,Ck) = \prod_{J=1}^{d} P(X^j|C_K) \qquad (4.2)$$

J48 Classifier

J48 which is observes the normalized information gain or difference in entropy that outcome for splitting the data from selecting an attribute (Youvrajsinh , 2014). Maximum standardized information achievement of features is used to make the result and then the algorithm repeats on the smaller subsets of component. While construction a tree, J48 can handle both distinct and constant attributes, missing attribute values and attributes of training data which in different costs (Youvrajsinh , 2014). In J48 classifier measure of the data disorder is called 'Entropy'. The Entropy of $\vec{y}$ is calculated by:

$$Entropy(\vec{y}) = -\sum_{j=1}^{n} \frac{|y_i|}{|\vec{y}|} log\left(\frac{|y_i|}{|\vec{y}|}\right) \qquad (4.3)$$

$$Entropy(j|\vec{y}) = \frac{|y_j|}{|\vec{y}\,|} \, log\,(\frac{|y_j|}{|\vec{y}\,|}) \qquad (4.4)$$

$$Gain(\overrightarrow{y}, j) = Entropy(\overrightarrow{y} - Entropy(j|\vec{y})) \qquad (4.5)$$

The objective is to maximize the Gain (Gaganjot, 2014). The elementary idea is to distribute the data into range constructed on the attribute values for that element that are originate in the training sample. J48 permits classification via either decision trees or rules generated from them.

Random Forest Classifier

In the field of data mining random forests is a new entry and is planned to produce correct estimates that do not over fit the data (Anantha, 2006). In random forest we cannot examine the individual trees separately that's why it seems more of a 'black box' approach. It offers some metrics that help in explanation. The subsequent tables can be used to compare relative significance among predictor variables (Anantha, 2006). A random forest multi-way classifier contains of an amount of trees, with every tree grown-up using particular method of randomization (Anna, 2007). A Random Forest classification model is a collection of classification tree predictors.

$$\{h(\bar{x}, \theta_k), k=1,2,\ldots,T\} \qquad (4.6)$$

Where θ_k independent identically distributed random vectors which each cast a vote for a class for a given input vector $\bar{x}$. Each of the classification tree models are grown fully without pruning as to keep bias at a minimum. The Gini measure of impurity is used to determine the variable selected to make the nodal split. The Gin impurity measure at node t is defined as (Gene, 2010):

$$g(t) = \sum_{j \neq i} p(j|t)p(i|t) \qquad (4.7)$$

Bagging Classifier

Bagging is a technique for improving outcomes of machine learning classification algorithms (Kristína, 2006). Bagging might be useful to construct a better classifier on training sample groups with misleaders (Kristína, 2006). When a bootstrap resample is drawn some of the data is omit-ted from the sample, but other data are simulated to convey the sample to full size.

In case of classification into two possible classes, a classification algorithm creates a classifier H : D → {-1,1} on the base of a training set of example descriptions D. The bagging method creates a sequence of classifiers H_m, m=1,...,M in respect to modifications of the training set.

$$H(d_i) = sign(\sum_{m=1}^{M} \alpha_m \, H_m \, (d_i)) \qquad (4.8)$$

The meaning of the above given formula can be interpreted as a voting procedure. An example d_i is classified to the class for which the majority of particular classifiers vote. Parameters α_k, m=1,...,M are determined in such way that more precise classifiers have stronger influence on the final prediction than less precise classifiers. The precision of base classifiers H_m can be only a little bit higher than the precision of a random classification.

LAYER 3: FINDING BEST CLASSIFIER

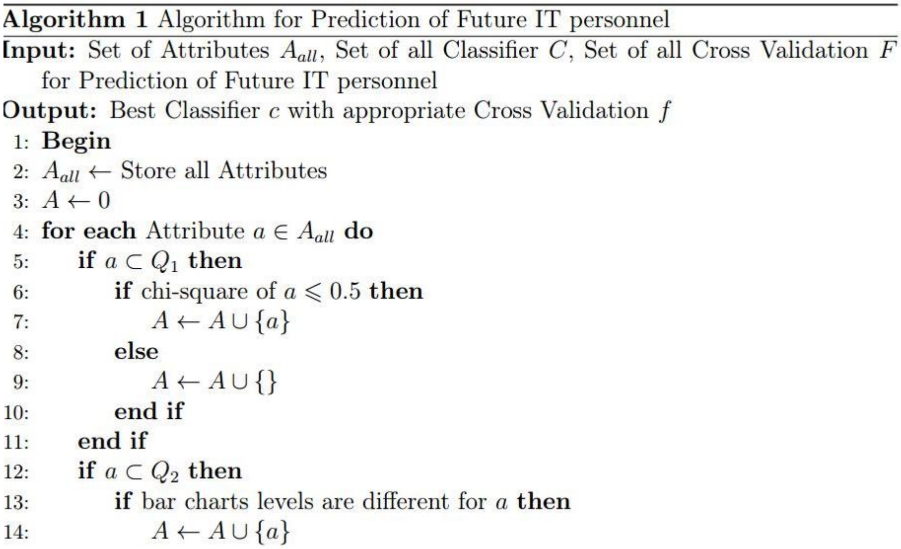
Classification is the procedure of finding a set of models that describe and differentiate data classes. This research has been experimented on four classifiers named as Naive Bayes, J48, Bagging and Random Forest. From the above four classifiers, we will find out best classifiers for this research. Accuracy matrix for every classifier with separate fold systems is used for performance comparison. After calculating accuracy matrix for every classifier with separate fold systems we calculate maximum, average and minimum value of accuracy matrix. TP

Rate, FP Rate, Precision, Recall, F-Measure and ROC Area are used as accuracy matrix calculations, which are described below.

Accuracy is basically a proportion of ((no. of correctly classified instances) / (total no. of instances))*100). True positive rate measures the proportion of positives that are correctly identified. It is basically the proportion of true positives to the sum of true positives and false negatives. It is basically the proportion of false positives to the sum of false positives and true negatives. In an ideal world we want this value to be zero. In information recovery positive predictive value is called precision. It is considered as number of correctly classified instances belongs to X divided by number of instances classified as belonging to class X; that is, it is the percentage of true positives out of all positive results. Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents. F-measure is a method of combining recall and precision scores into a single measure of performance. ROC Area is a graphical approach for displaying the tradeoff between true positive rate (TPR) and false positive rate (FPR) of a classifier. The entire prediction process including attribute selection, classification, and classifier prediction is detailed in Algorithm 1, where Table 1 presents the symbols used in that algorithm.

Table 1: Symbols used for Prediction of Future IT personnel (Algorithm 1)

Symbol	**Description**
Q1	Set of Quantitative Attributes
Q2	Set of Qualitative Attributes
A	Selected Attributes for Prediction of Future IT personnel
Aall	All Attributes for Prediction of Future IT personnel
a	Single Attributes for Prediction of Future IT personnel
C	Set of all Classifier (NB, J48, Bagging, and Random Forest)
F	Set of all Cross Validation (5, 10, 15, 20, and 25 Folds)
cij	Single *ith* Classifier with *jth* Fold
Accuracyij	Accuracy for *ith* Classifier with *jth* Fold



Algorithm 1 Algorithm for Prediction of Future IT personnel

Input: Set of Attributes A_{all}, Set of all Classifier C, Set of all Cross Validation F for Prediction of Future IT personnel

Output: Best Classifier c with appropriate Cross Validation f

```
Begin
$A_{all} \leftarrow$ Store all Attributes
$A \leftarrow 0$
for each Attribute $a \in A_{all}$ do
    if $a \subset Q_1$ then
        if chi-square of $a \leqslant 0.5$ then
            $A \leftarrow A \cup \{a\}$
        else
            $A \leftarrow A \cup \{\}$
        end if
    end if
    if $a \subset Q_2$ then
        if bar charts levels are different for $a$ then
            $A \leftarrow A \cup \{a\}$
        else
            $A \leftarrow A \cup \{\}$
        end if
    end if
end for
for each Classifier $c_i \in C$ do
    for each Cross Validation $f_j \in F$ do
        $Accuracy_{ij} \leftarrow$ calculated accuracy of $c_i$ with $f_j$ fold
    end for
end for
Sort $Accuracy$ list with descending order
Select top value of $Accuracy_{ij}$ list
Return $i^{th}$ Classifier with $j^{th}$ Fold for best prediction
End
```

CASE STUDY RESULT ANALYSIS AND DISCUSSION

This chapter presents the experimental setup and the results of Bangladesh Case Study, which describes the prediction of potential future IT personnel using various machine learning classifier methods with different execution folds.

Dataset Preparation

Dataset contains 181 instances of 57 attributes, which were collected from real life IT personnel inside the country named Bangladesh. Using Chi-square and Graph box selection process 23 most important attributes are selected for classification processing. The selected attribute list with their type and probable answer are listed in Table 2.

Table 2: Selected Attributes for Processing

Attribute	Type	Possible Answer
Education back	Qualitative	IT or Non-IT etc.
Designation	Quantitative	Sr. Officer, Officer, Jr. Officer, Ass. Officer
Duration of Current Position (months)	Quantitative	Various number of months
Duration of Total Job (months)	Quantitative	Various number of months
SSC (CGPA)	Quantitative	CGPA: 5.00, 4.94, 4.75, 4.50 etc.
Family Member (FQ1)	Quantitative	Four, Five, Six, Seven etc.
Father's Educational Background (FQ2)	Qualitative	Did not attend school, Primary school, High school etc.
Mother's Educational Background (FQ3)	Qualitative	Did not attend school, Primary school, High school etc.
Average Education of Parent's AvgEdu)	Quantitative	Various numbers
Parents Yearly Income (FIQ5)	Quantitative	Various numbers
Career in IT Field (SRQ1)	Qualitative	Friend, Family Member, Work, Teacher etc.
Most Encouraging Sources for IT Careers (SRQ3-1)	Qualitative	Family, Teacher, Friends, Faculty, Employer etc.
Second Most Encouraging Sources for IT Careers (SRQ3-2)	Qualitative	Family, Teacher, Friends, Faculty, Employer etc.
Third Most Encouraging Sources for IT Careers (SRQ3-3)	Qualitative	Family, Teacher, Friends, Faculty, Employer etc.
Most Essential Factors for Success in IT Field (AQ1-B)	Quantitative	Not essential, Essential, Very Essential etc.
Second Most Essential Factors for Success in IT Field (AQ1-C)	Quantitative	Not essential, Essential, Very Essential etc.
Third Most Essential Factors for Success in IT Field (AQ1-D)	Quantitative	Not essential, Essential, Very Essential etc.

Prepare Careering IT Field (PQ4-1)	Qualitative	Obtain information early in the educational process, Adequate preparation in the basic skills etc.
Most Influential part about to pursuit of a career (PQ6-1)	Qualitative	Graduates in subject matter courses, Interest in subject matter related to field etc.
Second Most Influential part to pursuit of a career (PQ6-2)	Qualitative	Study habits, Curriculum, Earning Potential etc.
Discouraged Factors That Student Have Never Considered a Career in IT/CSE (PQ7-1)	Qualitative	Never interest, Don't know much about it, Discouraged by friends etc.
Statements That True for Student's (PQ8)	Qualitative	I took IT/CSE classes in university, I enjoyed IT/CSE classes in university etc.
Greatest Potential Barrier to ursuit of an IT/CSE Career (PQ9)	Qualitative	Gender related stereotypes, Race-related stereotypes etc.

Remarks from the above Table 2 are listed below.

- Duration of Current Position in months like six, twelve, fifteen, twenty, thirty etc.
- Duration of total job in months like six, twelve, fifteen, twenty, thirty etc.
- Average Education of Parent's possible answer is various numbers. Various numbers means twenty two, twenty six, thirty three etc. We multiplied two by highest school level completed his parents and divided by two this result. Then we get this various numbers.
- Parent's yearly income means that father and mothers average income in every year.
- Not essential means one, essential means two and very essential means six.
- In designation attributes possible answers are Sr. Officer, Officer, Jr. Officer, and Assistant Officer. Value of Sr. Officer = 1, Officer = 2, Jr. Officer = 3 and Assistant Officer = 4.

EXPERIMENTAL RESULTS AND DISCUSSION

This section describes the experiments we performed using the classification algorithms to predict potential future IT personnel. The experiments were performed using the default options provided by WEKA for the respective algorithms. We employ only classification algorithms because potential future IT personnel type is reported by data set in the form of class values and classification algorithms are used to classify the given instance into class values taken by the output attribute. Experimental results of predefined five folds valued as 5, 10, 15, 20 and 25 are presented in Table 3.

Table 3 : Experimental Results of Different Classifier in Various Folds (5, 10, 15, 20, 25)

	Fold	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Naive Bayes (NB)	NB 5F	0.619	0.393	0.617	0.619	0.617	0.711
	NB 10F	0.657	0.36	0.657	0.657	0.653	0.714
	NB 15F	0.641	0.376	0.639	0.641	0.637	0.729
	NB 20F	0.652	0.365	0.651	0.652	0.648	0.715
	NB 25F	0.674	0.339	0.673	0.674	0.672	0.731
J48	J48 5F	0.586	0.421	0.585	0.586	0.585	0.63
	J48 10F	0.619	0.384	0.62	0.619	0.619	0.631
	J48 15F	0.575	0.436	0.573	0.575	0.573	0.619
	J48 20F	0.58	0.424	0.581	0.58	0.58	0.648
	J48 25F	0.657	0.349	0.657	0.657	0.657	0.695
Bagging (BG)	BG 5F	0.635	0.373	0.634	0.635	0.635	0.686
	BG 10F	0.685	0.315	0.687	0.685	0.686	0.763
	BG 15F	0.696	0.315	0.695	0.696	0.695	0.709
	BG 20F	0.635	0.368	0.636	0.635	0.636	0.695
	BG 25F	0.674	0.324	0.677	0.674	0.675	0.719

Random Forest (RF)	RF 5F	0.674	0.33	0.674	0.674	0.674	0.74
	RF 10F	0.674	0.328	0.675	0.674	0.674	0.735
	RF 15F	0.674	0.335	0.673	0.674	0.673	0.753
	RF 20F	0.696	0.305	0.697	0.696	0.696	0.756
	RF 25F	0.652	0.357	0.657	0.652	0.651	0.762

The TP Rate, FP Rate, Precision, Recall, F-Measure, and ROC Area values (maximum, minimum and average) of different experimented classifiers are listed in Table 4 and the graphical representation on those results are pointed out at Figure 1. Both of that table focus that the Random Forest (RF) classifier value performs better than any other classifier for detecting perfect potential IT graduate in Bangladesh Case Study.

Table 4: Maximum, Average and Minimum Results of Different Classifiers

Classifier	Measurement	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Naive Bayes	Maximum	0.674	0.696	0.673	0.674	0.672	0.731
	Average	0.6486	0.3666	0.6474	0.6486	0.6454	0.72
	Minimum	0.619	0.635	0.617	0.619	0.617	0.711
J48	Maximum	0.657	0.436	0.657	0.657	0.657	0.695
	Average	0.6034	0.4028	0.6032	0.6034	0.6028	0.6446
	Minimum	0.575	0.349	0.573	0.575	0.573	0.619
Bagging	Maximum	0.696	0.373	0.695	0.696	0.695	0.763
	Average	0.665	0.339	0.6658	0.665	0.6654	0.7144
	Minimum	0.635	0.315	0.634	0.635	0.635	0.686
Random Forest	Maximum	0.696	0.357	0.697	0.696	0.696	0.762
	Average	0.674	0.331	0.6752	0.674	0.6736	0.7492
	Minimum	0.652	0.305	0.657	0.652	0.651	0.735

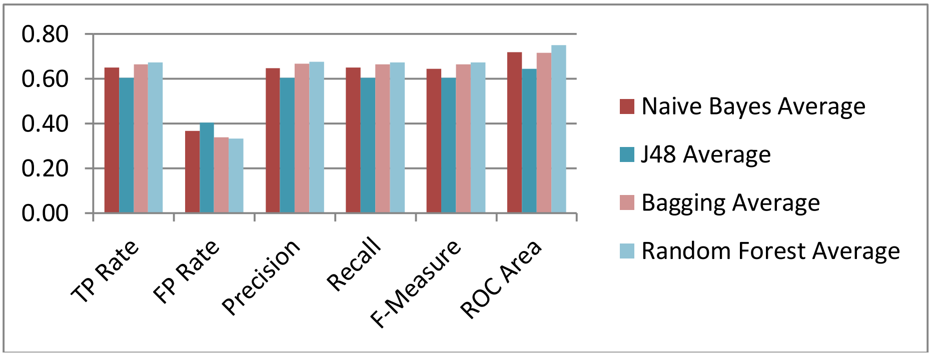


Figure 1 : Maximum, Average and Minimum Results of Different Classifiers

The prediction of future IT personnel type is a tough task. Here we see the average value of evaluation matrix.

Naive Bayes = TP Rate + FP Rate + Precision + Recall + F-Measure + ROC Area
$= 0.6486 + 0.3666 + 0.6474 + 0.6486 + 0.6454 + 0.72 \approx 3.6766$
J48 = TP Rate + FP Rate + Precision + Recall + F-Measure + ROC Area
$= 0.6034 + 0.4028 + 0.6032 + 0.6034 + 0.6028 + 0.6446 \approx 3.4602$
Bagging = TP Rate + FP Rate + Precision + Recall + F-Measure + ROC Area
$= 0.665 + 0.339 + 0.6658 + 0.665 + 0.6654 + 0.7144 \approx 3.7146$
Random Forest = TP Rate + FP Rate + Precision + Recall + F-Measure + ROC Area
$= 0.674 + 0.331 + 0.6752 + 0.674 + 0.6736 + 0.7492 \approx 3.777$

Random Forest classifies instances correctly 67.40332 percent and classifies instances incorrectly 32.59668 percent. From the below Figure 2, it has been shown that the average value of random forest classifier is higher than other classifiers.

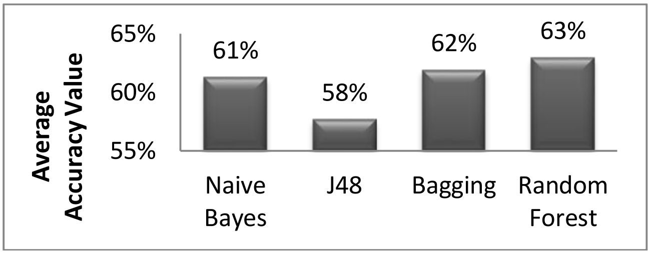


Figure 2: Average Accuracy Value of Different Classifiers

Thread to Validity

- The number of learning data in this research may not reflect the real statistics however we used random sampling to mitigate this inconsistency threats.
- Another threat to the data collection is that our chosen learning data attribute did not always fit the research goal. Based on the experience from the initial learning data and empirical analysis we will update the form further to ensure consistent collection.
- This learning data was collected from several IT industries and IT or Non-IT graduated students. Hence this prediction may not perform perfectly outside the country domain. (Robert , 2010).

Conclusion and Future Work

This research work has proposed an innovative approach for efficiently predicting the future IT personnel from IT or Non-IT graduated dataset in Bangladesh Case Study. It has come up with the information of graduates with and without having future IT personnel. SPSS has been used for attribute selection. The data mining tool WEKA has been used for experimental result. Four classification algorithms were used for this research, named as Naive Bayes, J48, Bagging and Random Forest. To find out better classifier, four classification algorithms with 5 fold, 10 fold, 15 fold, 20 fold and 25 fold cross validation are considered. From these cross validation folds, accuracy matrix and correctly classified average measurements such as Precision, Recall, F-Measure, ROC Area etc. are selected. It has been proved that the Random Forest classifies instances correctly 67.40332 percent, which is better than other classifiers. It has been proved that the random forest classifier achieve accuracy 3.777 which is better than other classifier. That means Random Forest is the best selection for predicting future IT graduate in this dataset of Bangladesh Case Study. Different feature selection processes may be used in future direction of this research track.

References

Aditya Polumetla. Machine learning methods for the detection of RWIS sensor malfunctions. PhD thesis, Citeseer, 2006.

Alexis Marcano-Cedeno and Diego Andina. Data mining for the diagnosis of type 2 diabetes. In World Automation Congress (WAC), 2012, pages 1–6. IEEE, 2012.

Anantha M Prasad, Louis R Iverson, and Andy Liaw. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems, 9(2):181–199, 2006.

Anna Bosch, Andrew Zisserman, and Xavier Munoz. Image classification using random forests and ferns. In 2007 IEEE 11th International Conference on Computer Vision, pages 1–8. IEEE, 2007.

Asma A Al Jarullah. Decision tree discovery for the diagnosis of type ii diabetes. In Innovations in Information Technology (IIT), 2011 International Conference on, pages 303–307. IEEE, 2011.

Charles X Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In KDD, volume 98, pages 73–79, 1998.

Feixiang Huang, Shengyong Wang, and Chien-Chung Chan. Predicting disease by using data mining based on healthcare information system. In Granular Computing (GrC), 2012 IEEE International Conference on, pages 191–194. IEEE, 2012.

Gaganjot Kaur and Amit Chhabra. Improved j48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 98(22), 2014

Gene M Ko, SA Reddy, Sunil Kumar, Barbara A Bailey, and Rajni Garg. A random forest model for the analysis of chemical descriptors for the elucidation of hiv1 protease protein–ligand interactions. Applied Computational Science and Engineering Student Support (ACSESS), San Diego State University, USA, 2010.

Kristína Machová, Frantisek Barcak, and Peter Bednár. A bagging method using decision trees in the role of base classifiers. Acta Polytechnica Hungarica, 3(2):121–132, 2006.

M Alam and SA Alam. Actionable knowledge mining from improved post processing decision trees. In International Conference on Computing and Control Engineering (ICCCE 2012), Chennai, pages 1–8, 2012.

Pat Langley and Stephanie Sage. Induction of selective bayesian classifiers. In Proceedings of the Tenth int. conference on Uncertainty in artificial intelligence, pages 399–406. Morgan Kaufmann Publishers Inc., 1994.

Qiang Yang, Jie Yin, Charles Ling, and Rong Pan. Extracting actionable knowledge from decision trees. IEEE Transactions on Knowledge and data Engineering, 19(1): 43–56, 2007.

Rahul A Patil, Prashant G Ahire, Pramod D Patil, and Avinash L Golande. Decision tree post processing for extraction of actionable knowledge. Int. Journal of Engineering and Innovative Technology, 2(1):152–55, 2012.

Robert Feldt and Ana Magazinius. Validity threats in empirical software engineering research-an initial survey. In SEKE, pages 374–379, 2010.

Varun Kumar and Nisha Rathee. Knowledge discovery from database using an integration of clustering and classification. International Journal of Advanced Computer Science and Applications, 2(3):29–33, 2011.

Youvrajsinh Chauhan and Jignesh Vania. J48 classifier approach to detect characteristic of bt cotton base on soil micro nutrient. International Journal of Computer Trends and Technology (IJCTT), vol 5. 2014

Zengyou He, Xiaofei Xu, and Shengchun Deng. Data mining for actionable knowledge: A survey. arXiv preprint cs/0501079, 2005.

--0--