Prediction analysis on short data based social media communication

[1]Pranali jyotiram jadhav, [2]Prof. Priya chandran
[1]Student, [2]Professor
BVIMIT CBD Navi mumbai

Abstract— Predictive analysis use to detect the relationships and patterns in data in order to predict the future by analyzing the past and taking better preventive decisions. It is used to detect orientations and patterns in data by databases. In this case, the data is gathered from young people survey. In this research, the comparison and classification between the original data and short data are used to predict the mindset of young people in future, for that some Machine learning techniques are commonly used to predict this situation. This research work mainly focused on finding the best classification algorithm based on different evaluation criteria like performance accuracy and root mean square error. We have performed a comparative study of the performance of machine learning algorithms Regression method, Bayes Classifier. The results show that Bayes Classifier is giving minimum root mean square error value compared to Regression Method.

Index Terms— orientations and patterns, predictive, short data, original data, regression method, Bayes classifier.

Introduction

Predictive analysis is defined as technologies and methods that allow us to detect orientations and patterns in data, by developing graphs and models, identify the future outcomes based on past data. Predictive models and graphs use know results to develop a model by database that can be used to predict values for different or new data. In this paper we are going to talk about prediction which is based on young people survey of original data and short data. In this paper we thoroughly go through the concept, comparison and classification between the original data and short data , all possible future opportunities of prediction on young people survey.

for that some Machine learning techniques are commonly used to predict this situation. This research work mainly focused on finding the best classification algorithm based on different evaluation criteria like performance accuracy and root mean square error. We have performed a comparative study of the performance of machine learning algorithms Regression method, Bayes Classifier applied on WEKA tool.

Literature Review

Several data prediction techniques are being used by researchers on different young people predictions. Large volume of data is available on young people for this study. Many different studies have been done on young people to check mindset in early stage. Researchers have implemented various data mining techniques to check mindset of young people.

Sharma & Shukla (2016) have argued that young people engage in social media because this dynamic and busy world does not allow people to physically stay in touch also they are using short data instead of original data while they chatting. Ajayi (1995) further argues that, for many people, social media use is a way of dealing with a society where spending time with humans is less valued than time because of short data used with technology. Others have supported the view that for students, university life can be a stressful owing to the demanding school work and exams (Tandon, Ferrucci & Duffy, 2015) so social media use serves to reduce tension because of this all data sources. It is expected that college students would be heavy social media users because students are far away from home, are free from parental home supervision (Arnett, 2007) all are using short data not a original data.

Research Methodology

In this section, we discuss some methods used in prediction with social media communication using short data. This section describes the proposed methodology for data mining from CKD dataset. The dataset has been used for the prediction of mindset of people. This dataset contains original data and short data are used in this comparative analysis. Classification techniques are applied to all features and selected features. In order to carry out experiments and implementations, WEKA is used as the data mining tool to classify the accuracy on the basis of datasets by applying different algorithmic approaches. In this work, we have used different machine learning algorithms Regression method and Bayes classifier to predict the mindset of people through classification algorithms

We have some collected data from the young people survey which is shown in below table no 1 Original and Short Data for communication. The data includes young people communication information which has original data and short data for communication. In this paper we used that collected data to check mindset of the youngster.

No	Original Data	Short Data
1	I enjoy listening to music.	Music
2	I prefer.	Slow songs or fast songs

3	Dance, Disco, Funk	Dance
4	Folk music	Folk
5	Country	Country
6	Classical	Classical music
7	Musicals	Musical
8	Pop	Pop
9	Rock	Rock
10	Metal, Hard rock	Metal or Hardrock
11	Punk	Punk
12	Hip hop, Rap	Hiphop, Rap
13	Reggae, Ska	Reggae, Ska
14	Swing, Jazz	Swing, Jazz
15	Rock n Roll	Rock n roll
16	Alternative music	Alternative
17	Latin	Latino
18	Techno, Trance	Techno, Trance
19	Opera	Opera
20	I really enjoy watching movies.	Movies
21	Public speaking	Fear of public speaking
22	Smoking habits	Smoking
23	Drinking	Alcohol
24	I live a very healthy lifestyle.	Healthy eating
25	I take notice of what goes on around me.	Daily events
26	I try to do tasks as soon as possible and not leave them until last minute.	Prioritising workload
27	I always make a list so I don't forget anything.	Writing notes
28	I often study or work even in my spare time.	Workaholism
29	I look at things from all different angles before I go ahead.	Thinking ahead
30	I believe that bad people will suffer one day and good people rewarded.	Final judgement
31	I am reliable at work and always complete all tasks given to me.	Reliability
32	I always keep my promises.	Keeping promises
33	I can fall for someone very quickly and then completely lose interest.	Loss of interest
34	I would rather have lots of friends than lots of money.	Friends versus money
35	I always try to be the funniest one.	Funniness
36	I can be two faced sometimes.	Fake
37	I damaged things in the past when angry.	Criminal damage
38	I take my time to make decisions.	Decision making
39	I always try to vote in elections.	Elections
40	I often think about and regret the decisions I make.	Self-criticism
41	I can tell if people listen to me or not when I talk to them.	Judgment calls
42	I am a hypochondriac.	Hypochondria

Table No. 1 Original and Short Data for Communication

Regression method

Regression methods analyze relationship between the dependent variable, prediction result, and one or more independent variables, such as the social communicational characteristics. Regression model could be linear and non-linear. But the linear model seems to describe the relation best [48]. Thus most times, we use the linear regression models, rather than non-linear ones, such as exponential, logarithmic, and polynomial models. In linear regression model, the variables could be the raw or transformed data. For example, between the orignal and short data usage on some consoles, the correlation is based upon the logarithmically transformed data [52]. Currently, this is the simplest and most used method.

Bayes classifier:

Bayes classifier is a probabilistic classifier using Bayes' theorem. Based upon the 14 priori probability of the prediction event, Bayes classifier uses the Bayesian formula to calculate its posterior probability, that the object belongs to the result classes, and then select the class with the largest posterior probability, as the event is most likely to have that result. If the prediction result is discrete, the Bayes classifier can be applied directly. Otherwise, the prediction result must be discretized first [43]. This classifier has an assumption that the predictors must be conditionally independent. There is no solid evidence always that the discussed metrics satisfy this assumption.

Results and Discussions

The input dataset were classified using different methods of data mining. Performance of various algorithms was studied. These methods include Regression and Bayes classifier. Classification is a data mining algorithm which finds out the output of a new

data instance. In this paper the experimental study is conducted on various classification algorithms and best algorithm is identified for mindset of people.

The sensitivity and specificity is calculated by following formulas:

$$Sensitivity\ (TPR) = \frac{TP}{P} = \frac{TP}{TP + FN} \qquad Specificity\ (TNR) = \frac{TN}{N} = \frac{TN}{FP + TN}$$

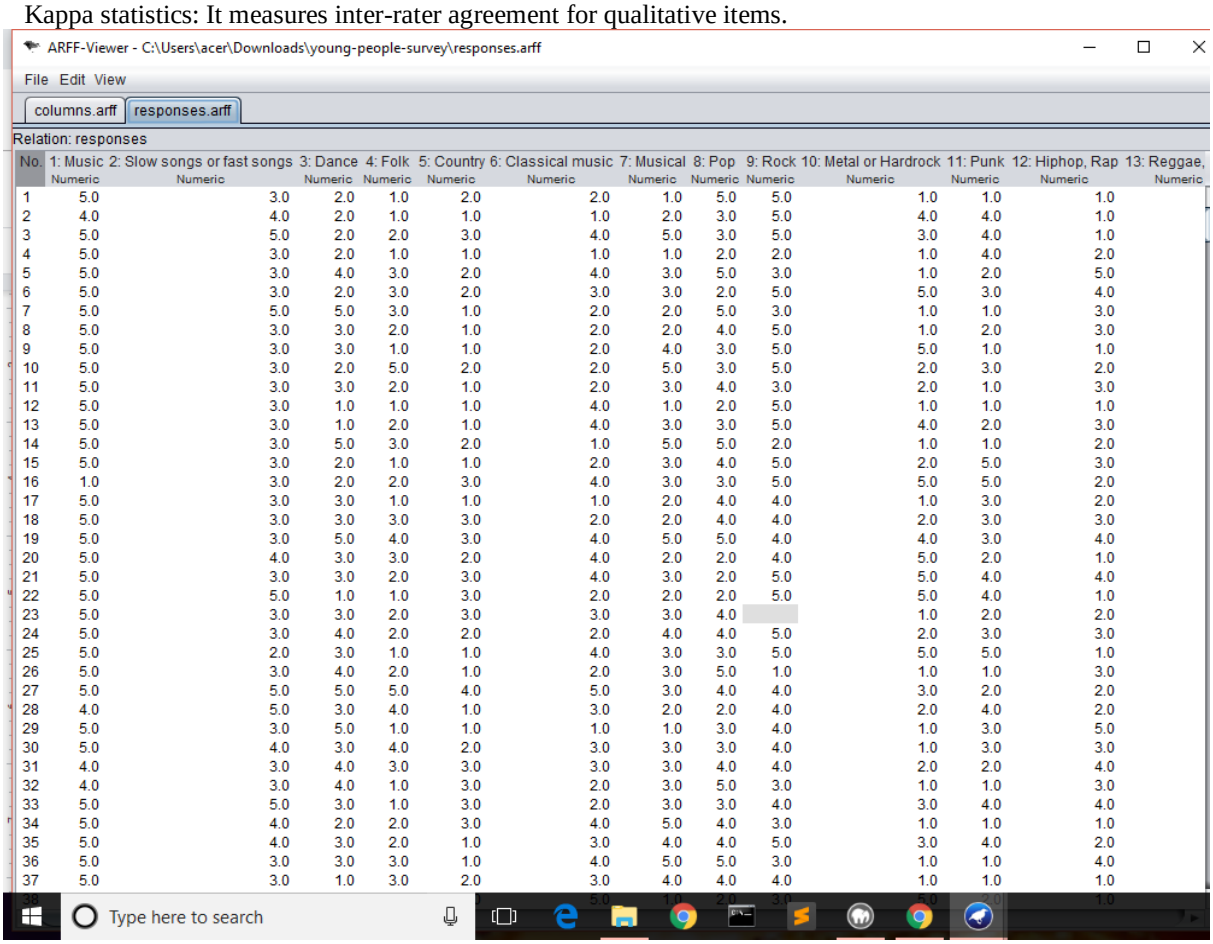
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

The number of real positive cases in the data is denoted by P. The number of real negative cases in the data is denoted by N.

A. Dataset Analysis

For prediction of mindset of young people we have used young people dataset for prediction and classification [7]. The dataset used for our experiment contain 150 attributes and 1010 instances. In order to obtain better accuracy 10 fold cross validation was performed. For each classification we selected training and testing sample randomly from the base set to train the model and then test it in order to estimate the classification and accuracy measure for each classifier. The thrust classifications and accuracy used by are:

- Correctly Classified Accuracy: It shows the accuracy percentage of test that is correctly classified.
- Incorrectly Classified Accuracy: It shows the accuracy percentage of test that is incorrectly classified.
- Mean Absolute Error: It shows the number of errors to analyze algorithm classification accuracy.
- Kappa statistics: It measures inter-rater agreement for qualitative items.



No.	1: Music	2: Slow songs or fast songs	3: Dance	4: Folk	5: Country	6: Classical music	7: Musical	8: Pop	9: Rock	10: Metal or Hardrock	11: Punk	12: Hiphop, Rap	13: Reggae,
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	5.0	3.0	2.0	1.0	2.0	2.0	1.0	5.0	5.0	1.0	1.0	1.0	
2	4.0	4.0	2.0	1.0	1.0	1.0	2.0	3.0	5.0	4.0	4.0	1.0	
3	5.0	5.0	2.0	2.0	3.0	4.0	5.0	3.0	5.0	3.0	4.0	1.0	
4	5.0	3.0	2.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	4.0	2.0	
5	5.0	3.0	4.0	3.0	2.0	4.0	3.0	5.0	3.0	1.0	2.0	5.0	
6	5.0	3.0	2.0	3.0	2.0	3.0	3.0	2.0	5.0	5.0	3.0	4.0	
7	5.0	5.0	5.0	3.0	1.0	2.0	2.0	5.0	3.0	1.0	1.0	3.0	
8	5.0	3.0	3.0	2.0	1.0	2.0	2.0	4.0	5.0	1.0	2.0	3.0	
9	5.0	3.0	3.0	1.0	1.0	2.0	4.0	3.0	5.0	5.0	1.0	1.0	
10	5.0	3.0	2.0	5.0	2.0	2.0	5.0	3.0	5.0	2.0	3.0	2.0	
11	5.0	3.0	3.0	2.0	1.0	2.0	3.0	4.0	3.0	2.0	1.0	3.0	
12	5.0	3.0	1.0	1.0	1.0	4.0	1.0	2.0	5.0	1.0	1.0	1.0	
13	5.0	3.0	1.0	2.0	1.0	4.0	3.0	3.0	5.0	4.0	2.0	3.0	
14	5.0	3.0	5.0	3.0	2.0	1.0	5.0	5.0	2.0	1.0	1.0	2.0	
15	5.0	3.0	2.0	1.0	1.0	2.0	3.0	4.0	5.0	2.0	5.0	3.0	
16	1.0	3.0	2.0	2.0	3.0	4.0	3.0	3.0	5.0	5.0	5.0	2.0	
17	5.0	3.0	3.0	1.0	1.0	1.0	2.0	4.0	4.0	1.0	3.0	2.0	
18	5.0	3.0	3.0	3.0	3.0	2.0	2.0	4.0	4.0	2.0	3.0	3.0	
19	5.0	3.0	5.0	4.0	3.0	4.0	5.0	5.0	4.0	4.0	3.0	4.0	
20	5.0	4.0	3.0	3.0	2.0	4.0	2.0	2.0	4.0	5.0	2.0	1.0	
21	5.0	3.0	3.0	2.0	3.0	4.0	3.0	2.0	5.0	5.0	4.0	4.0	
22	5.0	5.0	1.0	1.0	3.0	2.0	2.0	2.0	5.0	5.0	4.0	1.0	
23	5.0	3.0	3.0	2.0	3.0	3.0	3.0	4.0		1.0	2.0	2.0	
24	5.0	3.0	4.0	2.0	2.0	2.0	4.0	4.0	5.0	2.0	3.0	3.0	
25	5.0	2.0	3.0	1.0	1.0	4.0	3.0	3.0	5.0	5.0	5.0	1.0	
26	5.0	3.0	4.0	2.0	1.0	2.0	3.0	5.0	1.0	1.0	1.0	3.0	
27	5.0	5.0	5.0	5.0	4.0	5.0	3.0	4.0	4.0	3.0	2.0	2.0	
28	4.0	5.0	3.0	4.0	1.0	3.0	2.0	2.0	4.0	2.0	4.0	2.0	
29	5.0	3.0	5.0	1.0	1.0	1.0	1.0	3.0	4.0	1.0	3.0	5.0	
30	5.0	4.0	3.0	4.0	2.0	3.0	3.0	3.0	4.0	1.0	3.0	3.0	
31	4.0	3.0	4.0	3.0	3.0	3.0	3.0	4.0	4.0	2.0	2.0	4.0	
32	4.0	3.0	4.0	1.0	3.0	2.0	3.0	5.0	3.0	1.0	1.0	3.0	
33	5.0	5.0	3.0	1.0	3.0	2.0	3.0	3.0	4.0	3.0	4.0	4.0	
34	5.0	4.0	2.0	2.0	3.0	4.0	5.0	4.0	3.0	1.0	1.0	1.0	
35	5.0	4.0	3.0	2.0	1.0	3.0	4.0	4.0	5.0	3.0	4.0	2.0	
36	5.0	3.0	3.0	3.0	1.0	4.0	5.0	5.0	3.0	1.0	1.0	4.0	
37	5.0	3.0	1.0	3.0	2.0	3.0	4.0	4.0	4.0	1.0	1.0	1.0	
38	5.0					5.0	1.0	2.0	3.0	5.0		1.0	

Table 3.1 shows the description of the attributes of young people.The below Figure 3.1 shows Normalized Dataset.Then one by one the classification algorithms Bayes classifier and regression are applied on the filtered dataset attributes and their

distributions is shown in figure 3.2

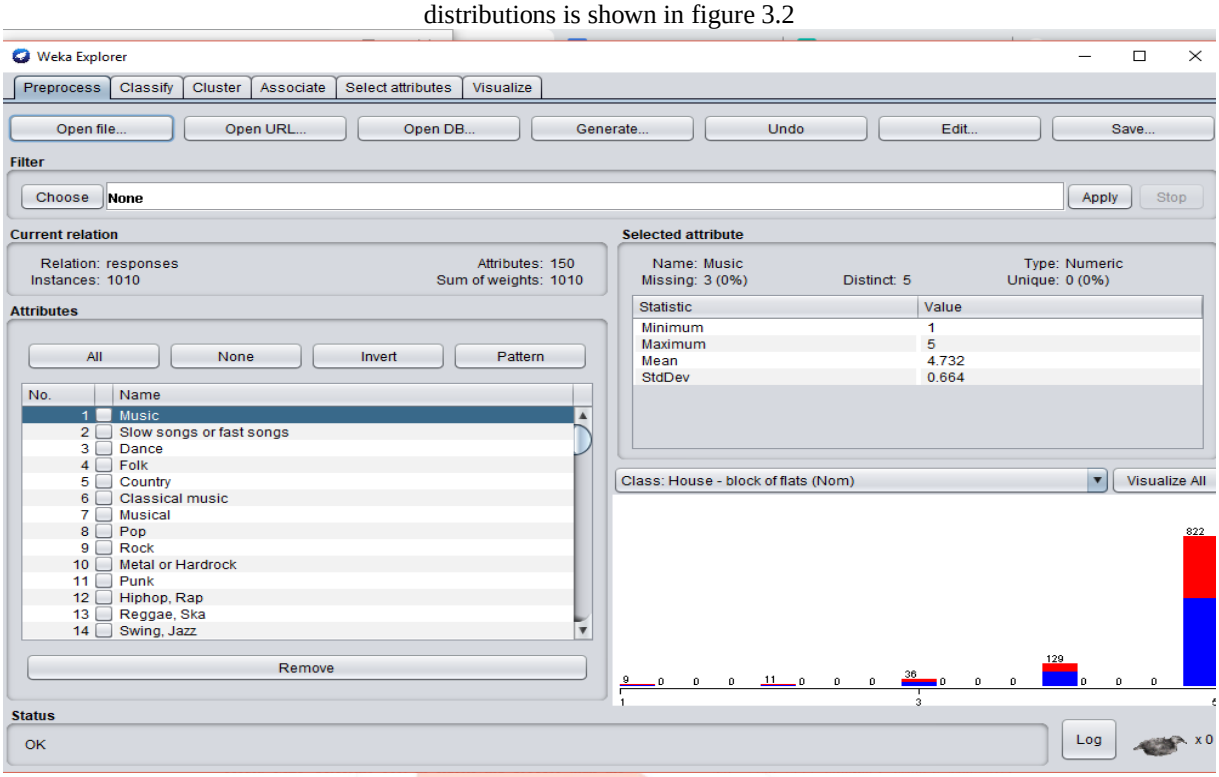


Figure 3.1: Normalized Dataset

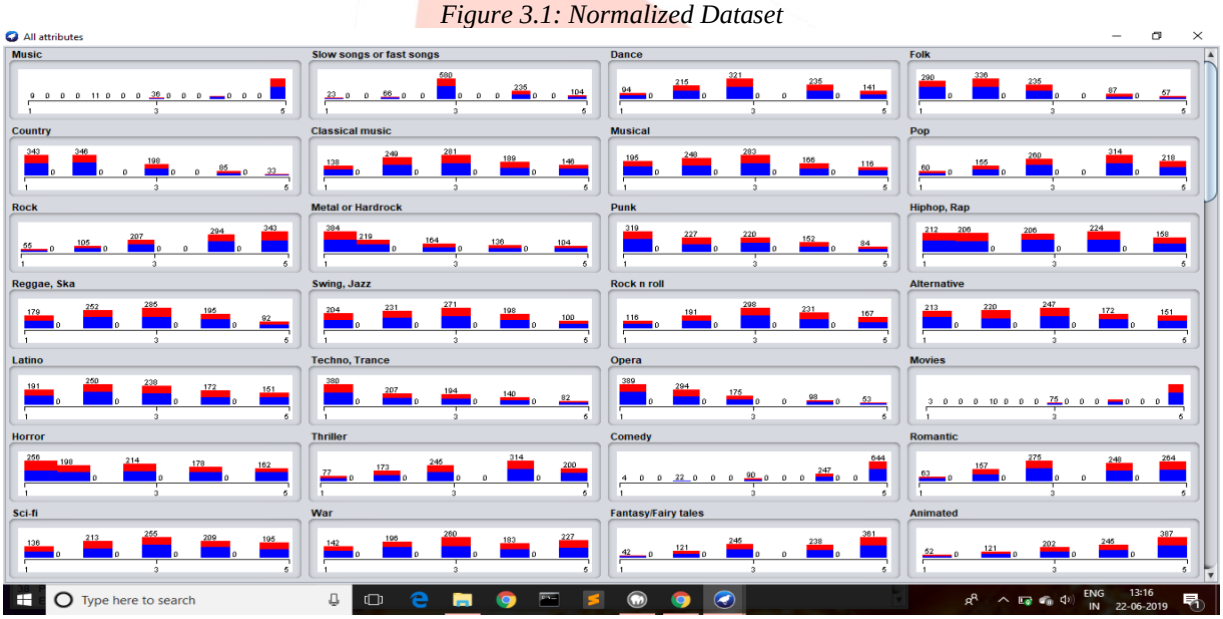


Figure 3.2: Attribute Distribution

Confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice versa).The matrix is represented in the form of nxn, where n is the number of classes. The accuracy of each classification algorithms can be calculated from that.

In our experiment we have two classes, and therefore we have a 2x2 confusion matrix, which is shown in table 3.2. The confusion matrix obtained for bayes classifier and regression are given in table 3.3, 3.4 and 3.5 respectively.

Table 3.2: Confusion Matrix

	a (CKD)	**b (NOT CKD)**
a (CKD)	TP	FN
b (NOT CKD)	FP	TN

Table 3.3: bayes Confusion Matrix

	a (CKD)	**b (NOT CKD)**
a (CKD)	461	134
b (NOT CKD)	159	252

Table 3.4: bayes Confusion Matrix

	a (CKD)	b (NOT CKD)
a (CKD)	400	130
b (NOT CKD)	159	200

B. Predictive Performance of classifier

Evaluation of performance is compared using predictive Accuracy, Mean absolute error, Root mean squared error and Receiver Operating Characteristic (ROC) Area and Kappa statistics.

Table 3.6: Predictive Performance of classifier

Evaluation Criteria	**Classifier**	
	Bayes	**Regression**
Timing to Build Model (sec)	0.4	0.07
Correctly classified Instance	751	690
Incorrectly Classified Instance	225	125
Predictive Accuracy	99.75	98.75
Kappa statistics	0.4693	0.2123
Mean Absolute error	0.3076	0.3056
Root mean square	0.4418	0.4112
Relative absolute error	63.6356	63.63
Root relative squared error	89.8804	81.8804

Table 3.6 depicts the performance of each algorithm based on different evaluation criteria. Here the prediction accuracy is 99.75, 98.75, for Bayes and regression. Bayes having the minimum root mean square error with a value 0. 3076.

C. Perform Analysis

In figure 3.3 and 3.4, the performance analysis is identified with the help of a graph. The fig 4.4 shows the correctly classified instances, the predictive accuracy for these three techniques are 97.75, 98.75 for Bayes and regression.

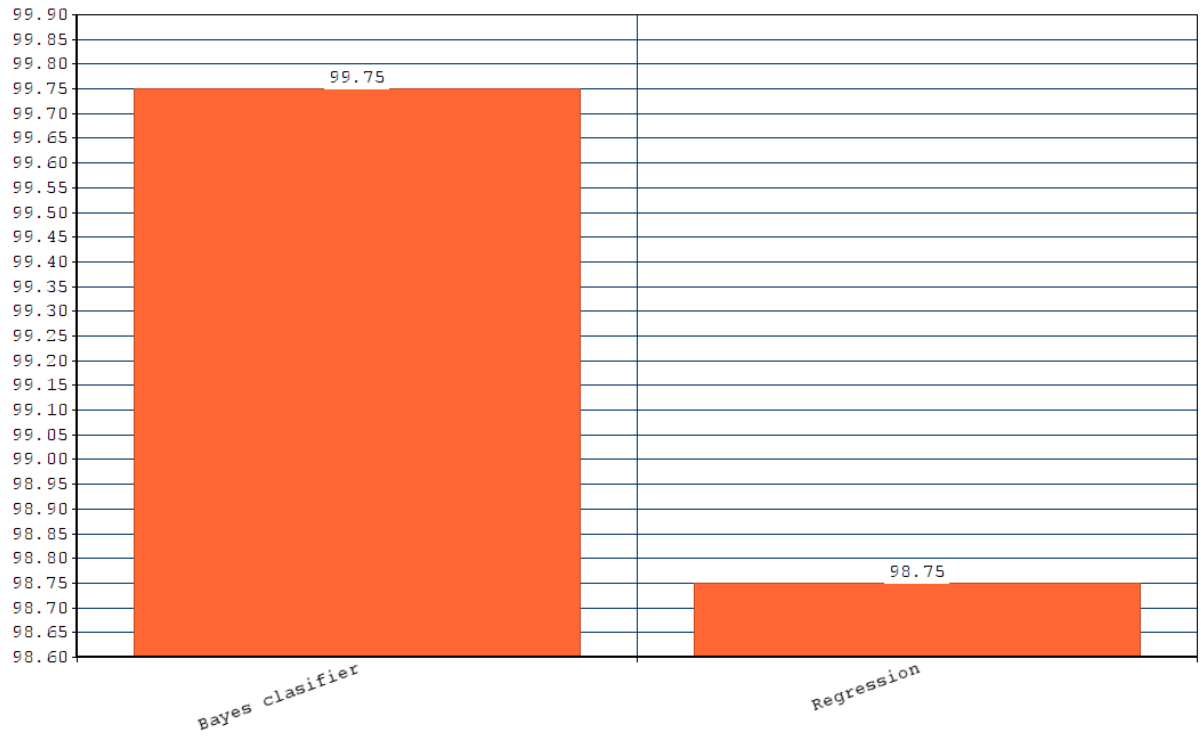


Figure 3.3: Correctly Classified Instances

CONCLUSION

In this research, we first demonstrated that it is possible to predict mindset of young people. Combining original data with short data of young people it is estimated that in future people will only use short data not an original data. We proposed a set of important consoles based on survey youngers are uses laptops, computers and mobiles, etc., and found some interesting associations between that consoles. Experiment results show that our predictive model is highly correlated with the ground truth. In the future, larger datasets will be explored to prove the effectiveness of this survey. Different types of short data will be use to enhance our predictive model. Also, recognizing that original data will become another important direction in our future research.