**Research Article**

**Betul Districts Primary School Performance Prediction Model Using Data Mining**

**Manmohan Singh¹, Anjali Sant²**

*Corresponding author: Manmohan Singh

1Research Scholar, Mewar University, Chittorgarh, Rajasthan
2Prof. and HOD, Dept of Mathematics, BIST, Bhopal

**A b s t r a c t**

As this academic performance is influenced by many factors, it is essential to develop predictive data mining model for students' performance so as to identify the slow learners and study the influence of the dominant factors on their academic performance. In the present investigation, a survey cum experimental methodology was adopted to generate a database and it was constructed from a primary. While the primary data was collected from the regular students and irregular student the secondary data was gathered from the school in class 3, 4 and 5 a total of 1000 datasets of the 2014 year from five different schools in three different districts of BETUL state Madhya Pradesh were collected. The raw data was preprocessed in terms of filling up missing values, transforming values in one form into another and relevant attribute/ variable selection. As a result, we had 700 student records, which were used for primary school prediction model construction. A set of prediction rules were extracted from primary school prediction model and the efficiency of the generated student prediction model was found. The accuracy of the present model was compared with other model and it has been found to be satisfactory.

**Keywords:** student performance, Decision Tree, Data Mining, WEKA

**Introduction**

In the present investigation Betul district in M.P is chosen for the study because district is comprises of Tribal as well as nontribal blocks. There is vast difference in culture of tribal and nontribal blocks in a district. The present study reveals that the drop out at primary level especially among socially backward and economically poor people is due to their peculiar type of culture, which hinders educational processes in tribal society of tribal blocks. The government has opened some more schools and EGS under SSA norms in remote, rural and tribal villages in district so that education can be the teachers remain usually absent from their schools for a long period and the situation becomes deteriorated and suffer student.

Generation of a data source of predictive variables,
Identification of highly influencing predictive Variables on the academic performance of higher secondary students
Construction of a primary school prediction model in data mining on the basis of identified predictive variables.

**Previous Studies**

A number of reviews pertaining to not only the diverse factors like personal, socio-economic, psychological and other environmental variables that influence the performance of students but also the models that have been used for the performance prediction are available in the literature and a few specific studies are listed below for reference.

Walters and Soyibo [1] conducted a study to determine Jamaican high school students' (population n=305) level of performance on five integrated science process skills with performance linked to gender, grade level, school location, school type, student type, and socio-economic background (SEB). The results revealed that there was a positive significant relationship between academic performance of the student and the nature of the school.

Khan [2] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses.

It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Hijazi and Naqvi [3] conducted as study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis that was stated as “Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' family income, students' mother's age and mother's education are significantly related with student performance” was framed. By means of simple linear regression
Kristjansson, Sighusdottir and Allegrante [7] made a study to estimate the relationship between health behaviors, body mass index (BMI), self-esteem and the academic achievement of adolescents. The authors analyzed survey data related to 6,346 adolescents in Iceland and it was found that the factors like lower BMI, physical activity, and good dietary habits were well associated with higher academic achievement.

Moriana et al. [8] studied the possible influence of extra-curricular activities like study-related (tutoring or private classes, computers) and/or sports-related (indoor and outdoor games) on the academic performance of the secondary school students in Spain. A total number of 222 students from 12 different schools were the samples and they were categorized into two groups as a function of student activities (both sports and academic) outside the school day. Analysis of variance (ANOVA) was used to verify the effect of extracurricular actives on the academic performance and it was observed that group involved in activities outside the school yielded better academic performance.

Bray [9], in his study on private tutoring and its implications, observed that the percentage of students receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Sri Lanka. It was also observed that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socio-economic conditions.

Modeling of student performance at various levels is discussed in [4], Ma, Liu, Wong, Yu, and Lee [4] applied a data mining technique based on association rules to find weak tertiary school students (n=264) of Singapore for remedial classes. Three scoring measures namely Scoring Based on Associations (SBA-score), C4.5-score and NB-score for evaluating the prediction in connection with the selection of the students for remedial classes were used with the input variables like sex, region and school performance over the past years. It was found that the predictive accuracy of SBA-score methodology was 20% higher than that of C4.5 score, NB-score methods and traditional method.

Kotsiantis, et al. [5] applied five classification algorithms namely Decision Trees, Perceptron-based Learning, Bayesian Nets, Instance-Based Learning and Rule-learning to predict the performance of computer science students from distance learning stream of Hellenic Open University, Greece. A total of 365 student records comprising several demographic variables like sex, age, marital status were used. In addition, the performance attribute namely mark in a given assignment was used as input to a binary (pass/fail) classifier. Filter based variable selection technique was used to select highly influencing variables and all the above five classification models were constructed. It was noticed that the Naive-Bayes algorithm.

Al-Radaideh, et al. [10] applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. They used 12 predictive variables and a 4-class response variable for the model construction. Three different classification methods namely ID3, C4.5, and the Naive Bayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models with the predictive accuracy of 38.33% for four-class response variable.

Cortez and Silva [6] attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school students from the Alentejo region of Portugal by utilizing 29 predictive variables. Four data mining algorithms such as Decision Tree (DT), Random Forest (RF), Neural Network (NN) and Support Vector Machine (SVM) were applied on a data set of 788 students, who appeared in 2006 examination. It was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (pass/fail) respectively. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset.

From these specific studies, we observe that the student performance could depend on diversified factors such as demographic, academic, psychological, socio-economic and other environmental factors. It was learnt that the predictive accuracy of constructed models ranged between 38.33 to 93%. The variation in the predictive accuracy could be correlated with the nature of student data set and utilization of number of records, predictive variables and class values of response variable.

**Methodology**

All the needed information was available in the school head master or self by student and its parents in form of through Questionnaire. Information not to be collected from the District Education office or the using web side and district project office, specially designed questionnaire were used in collecting information and data of student. Accuracy and Efficiency.

Classifier accuracy: predicting class label to be.

Predictor accuracy: guessing value of predicted attributes and its field.

Speed or Time.

Time to construct the model or training time.

Time to use the model by using classification or prediction time.

Robustness: handling noise and missing values of data.

Scalability: efficiency in disk-resident databases of student.

Interpretability: Understanding and insight provided by the model of real.

Survey cum experimental methodology was adopted to generate a database for the present study. The basis of this database was from the primary and secondary sources. While the primary data was collected from the regular students, the secondary data was gathered from the school office.

**Data Source**

A detailed questionnaire was prepared with the reference, assistance and guidance from the (i) Review of related literature,
(ii) Teachers of diversified schools, (iii) Parents of primary school students, (iv) Educational experts in school, colleges and Universities. The designed four page A4 size close-ended questionnaire was used for the collection of student details. Most of the information of the variables was collected directly from the students through this questionnaire. Based on this information, a few derived variables were generated. While some of the information for the variables was extracted from the school records, the mark details were collected from the education office. All the predictor and response variables which were derived from the questionnaire are given in Table 1 for reference.

### Table 1: Data Source

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Attribute Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>School_Code</td>
<td>School dice code</td>
</tr>
<tr>
<td>2</td>
<td>Name_Place</td>
<td>Place of school</td>
</tr>
<tr>
<td>3</td>
<td>Name_Block</td>
<td>Name of block (three blocks of Batul District)</td>
</tr>
<tr>
<td>4</td>
<td>District</td>
<td>Batul (M.P.)</td>
</tr>
<tr>
<td>5</td>
<td>Scholar_Number</td>
<td>Student scholar number</td>
</tr>
<tr>
<td>6</td>
<td>Name_Student</td>
<td>Name of student</td>
</tr>
<tr>
<td>7</td>
<td>Student_Father_Name</td>
<td>Student father name</td>
</tr>
<tr>
<td>8</td>
<td>Student_Mother_Name</td>
<td>Mother’s name</td>
</tr>
<tr>
<td>9</td>
<td>Age</td>
<td>Age of student (06-10 years)</td>
</tr>
<tr>
<td>10</td>
<td>Sex</td>
<td>Gender (M, F)</td>
</tr>
<tr>
<td>11</td>
<td>Class</td>
<td>Class (III, IV, V)</td>
</tr>
<tr>
<td>12</td>
<td>Category</td>
<td>Category (SC, ST, Gen, OBC)</td>
</tr>
<tr>
<td>13</td>
<td>School_type</td>
<td>School type (Govt., Private)</td>
</tr>
<tr>
<td>14</td>
<td>Location_School</td>
<td>Rural &amp; Urban</td>
</tr>
<tr>
<td>15</td>
<td>No_teachers</td>
<td>Number of teachers in school</td>
</tr>
<tr>
<td>16</td>
<td>family_size</td>
<td>Number of members in a student family</td>
</tr>
<tr>
<td>17</td>
<td>Living_zone</td>
<td>Living zone</td>
</tr>
<tr>
<td>18</td>
<td>Father_edu</td>
<td>Father’s Education</td>
</tr>
<tr>
<td>19</td>
<td>Father_occup</td>
<td>Occupation of student father</td>
</tr>
<tr>
<td>20</td>
<td>Mother_edu</td>
<td>Mother’s Education</td>
</tr>
<tr>
<td>21</td>
<td>Mother_occu</td>
<td>Occupation of Student’s Mother</td>
</tr>
<tr>
<td>22</td>
<td>Family_income</td>
<td>Family income</td>
</tr>
<tr>
<td>23</td>
<td>Private_tuition</td>
<td>Are student take private tuition?</td>
</tr>
<tr>
<td>24</td>
<td>Attendance_School</td>
<td>Attendance of student in a class</td>
</tr>
<tr>
<td>25</td>
<td>Previous_result</td>
<td>Previous year result of student in percentage</td>
</tr>
<tr>
<td>26</td>
<td>Grade_Previous_Result</td>
<td>Previous year result of student in Grad</td>
</tr>
</tbody>
</table>

### Tools and Techniques

Classification trees are widely used in different fields such as medicine, computer science, botany and psychology [11]. These trees readily lend themselves to being displayed graphically, helping to make them easier to interpret than they would be if only a strict numerical interpretation were possible.

Primary school student [12], which is one of the classification tree algorithms, is the name given to one version of the Automatic Interaction Detector that has been developed for categorical variables. In fact, student is a technique that recursively partitions (or splits) a population into separate and distinct segments. These segments, called nodes, are split in such a way that the variation of the response variable (categorical) is minimized within the segments and maximized among the segments. After the initial splitting of the population into two or more nodes (defined by values for an independent or predictor variable), the splitting process is repeated on each of the.

### Conclusion

We used students’ data from database course of Betul District. We collected all available data including their usage of Model learning facility of education. We applied determining techniques to discover knowledge. Even if discovered Classification rules based and we sorted the rules using lift metric then we visualized than generated different rules. Then we discovered classification rules using decision tree. Experiments could be done using more data mining techniques such as neural nets, Classification rules, genetic algorithms, nearest Neighbor, Naive Bayes, support vector machines, the used preprocess and data mining algorithms could be embedded into learning system so that anyone using the system can benefit from Classification rules the data mining.

### References


[3]. Hijazi ST, and Naqvi RSM. “Factors Affecting Student’s Performance: A Case of Private Colleges”, Bangladesh e-Journal of Sociology, 2006;3(1).


