Chapter 4
Ensemble Model to Identify the Students at-risk

4.1 Introduction

To discover unseen information and independent decision making Data Mining offers various tools for data analysis in many application domains. It equally offers a relevant response to the data collected and renders authentic information. One of the most important fields in which the data mining tools are applied is the educational setting. Incidentally and ironically, the increasing amount of educational data generally remains unused. It further creates a challenging task for using the data to provide knowledge by using data analysis tools for improving the educational settings. The challenges faced by students in daily routine in Indian Education System such as updating of the syllabus, provide more practical knowledge, liberty to students in selecting their field for future. These challenges helped to solve a few common challenges faced by students

The academic performance of students can be improved by using Data Mining techniques such as prediction, clustering, classification which are often used these days. These techniques help in various other fields such as banking, retail and marketing etc. to take important decisions. As these techniques are quantitative in nature, they render questionable variables for future research. The techniques practised in this study helped in answering the pertinent questions related to the Indian education system vis-à-vis the teacher-student relationship.

However, building a Classification Model (CM) using Data Mining (DM) techniques results better than individual techniques applied to the data. CM concludes the target value from the input values provided for training. CM predicts unseen values such as new class labels and categorising new data. To predict risk in various fields such as loan, prediction of the academic performance of the students, predicting the weather based on training data, CM helps in rendering accurate results.

On the same note, there is an urgent requirement of early prediction and analysis of the students ‘academic performance in their early years of graduation as this
would help strike a balance between the risk factors prevalent and the future prospects. The capability of early detection of students’ academic performance plays an important role in predicting their performance in today’s educational settings which is more modern and competitive. The major concern of any educational institute is thus to evaluate and enhance the quality of education and produce a better future for the youth. An Ensemble model was created using the best performing DM techniques to solve the purpose of this study. Predicting the academic performance of students using Ensemble Model helped this study make better decisions based on the results provided by Ensemble.

To achieve the task of fulfilling a bright future for the students, there are a number of challenges to observe as shown in Figure 4.1:

![Figure 4.1: Challenges in Achieving Task](image)

i. To study and analyse the information related to students’ various economic, social and educational factors that are helpful in observing their academic performance, as these factors often influence a students’ academic performance. An exhaustive literature review was done to extract the information about influential factors that affect the academic performance of the students. These factors will also help in the early detection and prevention of the students at-risk.
ii. Based on influential factors a questionnaire was designed to collect the data. The questionnaire comprised of demographic, academic, social and behavioural factors. The respondents of this questionnaire were students of Bachelor of Computer Applications (II\textsuperscript{nd} Semester students of Goswami Ganesh Dutta Sanatan Dharma College, Sector-32, Chandigarh).

iii. To find out the most appropriate Data Mining Techniques various classification techniques were compared on the basis of their higher accuracy and results based on Naïve Bayes, Decision Tree, Support Vector Machine, Gradient Boost, Random Forest, Clustering, Neural Network etc. The results were compiled after extensive study and analysis.

iv. To find the most influential factors that influence the academic performance of students, Decision Tree, Naïve Bayes, Support Vector Machine was found most appropriate. These techniques helped in finding out the positive and negative impact on the academic career of students. Decision Tree focused on the categorical and numerical values, Naïve Bayes focused on probability and the Support Vector Machine on binary values thus providing accurate values.

v. Considering the accuracy of data mining techniques, an idea came to light that an Ensemble Model of these techniques could be created. An Ensemble Model is a classification model under the machine learning technique that focuses on producing one optimal predictive model. Ensemble Model will help to extract the influential factors from the data set with better accuracy.

vi. Based on these influential factors a student Corrective Action and Preventive Action (CAPA) form was designed to validate the study. The CAPA will be suggesting the preventive measures that will help students’ improve their academic performance and give them an idea about the risk factors as it relates to the future.

The data was collected through a questionnaire during the classes. The questionnaire was distributed manually in the classrooms so that honesty could be
ensured. The collected data was converted into a digitized form using excel. The digitized data was imported into Rapid Miner Studio 9.0.1 for further processing. Rapid Miner is a platform of data science used for data preparation in machine learning used in deep learning, pattern mining and text mining. Rapid Miner is used to create predictive models in various fields for decision making [69] [74]. The objective of this study is to predict the student academic performance based on various influential factors. As compared to other statistical tools and tools used in data science like WEKA [41, 84], Rapid Miner has the ability to handle large data sets [3, 27, 74].

In the following study the data set questionnaire.csv contains attributes like percentage in previous semester, reappear in the previous semester, class attending behaviour, assignment completion attitude, family income, influence of friend circle and habit of consuming alcohol and tobacco, type of family, age of student, qualification of father and a class variable performance as shown in Figure 4.2.

Statistics is the field under mathematics deals with the organized form of collected dataset that corresponds to a statistics matrix. Statistics considers all the aspects of
data that is used in surveys and experiments. The statistics of the data set represents the numerical value in pictorial form using charts. These values can be described as properties of the data set. It is applicable in most of the fields as have been used since centuries such as mathematics, computing, applied sciences and number theories etc. This data set shows the statistics of the attributes in data set classification of various attributes such as Performance, Gender, Category and Stream of Secondary education into classes like at-risk of failing, Number of Males and Females, categories of SC, ST, General and Others, Stream of Secondary Education, respectively. Statistics of the data is shown in Figure 4.3.

![Figure 4.3: Statistics of Data Set in Rapid Miner](image)

### 4.2 Feature Selection

Feature Selection is a major selection method used extensively in a research study. Feature Selection is also often called Variable Selection or Attribute Selection. Feature Selection helps in simplification of models and places variables/data selectively to avoid any repetition/replication. It thus avoids ambiguity and renders a selected honest result. It further helps in removing any redundant or irrelevant variable.
To select the best features from the attributes in a data set, Information Gain was performed to select the best feature using Rapid Miner Studio 9.0.1. Attribute with higher Information Gain was selected that gives the best split. Attributes with less value of Information Gain was removed from the data set. After data cleaning and consolidation, 499 responses were found to be correct and complete post initial scrutiny, and therefore included in the study. Weight of various attributes is shown in Figure 4.4.

The bar chart below represents the weight of various attributes, wherein previous semester percentage has the highest weight gain i.e. as compared to all the attributes included in the study. This attribute helped to predict the performance of the students in a better way.

![Weights of Data Set in Rapid Miner](image)

**Figure 4.4: Weights of Data Set in Rapid Miner**

Data Mining Techniques are considered suitable for creating predictive models [12, 15, 33, 60, 104]. Various data mining techniques were studied and compared by applying to the data set. The results were compared and Decision Tree, Naïve Bayes, Support Vector Machine [27, 44, 47, 55, 58, 72, 79, 84, 87, 92] were found to be the most appropriate techniques for the study. These Data Mining Techniques were preferred over others based on their results and the accuracy of prediction. The learning capabilities of these Classification Techniques were found commendable as compared to other techniques on the basis of their results and performance.
4.3 DATA MINING TECHNIQUES

4.3.1 Decision Tree

Decision Tree operates on numerical as well as categorical values. When Decision Tree applied to the data set, it helped to extract important factors for at-risk students. Most influential factors found was previous semester’s percentage, mother’s income, a major mishap in the family (accident and death), and mother’s qualification, trying to cover topics if missed classes, health-related issues in the family or self, and location of residence from the institute.

The factors that have the highest correlation values and direct towards positive x-axis imply that they have a positive impact on students’ academic performance. On the other hand, if they have less correlation value and move towards negative x-axis then they have a negative impact on students’ academic performance. In contrast, factors like mother’s qualification and health issues in the family or self could be controlled and could leave a negative impact on students’ academic career. Decision Tree provides better accuracy, 98.2% when applied to the data set as compared to other techniques.

![Important Factors for At-risk](image)
Figure 4.5: Important Factors for at-Risk Students using Decision Tree

Based on the influencing factors extracted after applying Decision Tree on the data set, a model was created that reflected the importance of the previous semester’s percentage. As the weight given to this attribute was highest of all, with the weight value of 1.0, it gave the best split to create a decision tree.

The Figure below depicts that if a student scores less than 60% of marks, then that student is at the risk of failing or being a drop-out. Students who scored above 60% marks was not at-risk of failing.

Figure 4.6: Decision Tree Model

4.3.2 Naïve Bayes

Naïve Bayes is based on Bayesian theorem that works on posterior probability. The model gave unique influential factors as compared to DT for students’ academic performance. It facilitated the researcher to take out important factors that put a positive effect on students’ performance such as previous semester’s percentage, mother’s income, and a major mishap in the family, the percentage in mathematics and ease of scoring good grades in the semester. On the other hand factors like the habit of
consuming alcohol and tobacco could influence negatively on students’ academic performance. Naïve Bayes came out to be 98.0% accurate.

The factors that have the highest correlation values and directing towards a positive x-axis imply that they have positive impact on students’ academic performance. On the other hand, if they have less correlation value and move towards negative x-axis then they have a negative impact on students’ academic performance. The habit of consuming alcohol could be handled and treated with care, as this was a sensitive issue and could affect the students’ psychologically.

Figure 4.7: Important Factors for Student at-Risk using Naïve Bayes

A model considering the important factors and after mining them with Naïve Bayes techniques was created. This model endows the researcher with a variety of factors that lead to important decision making helps students’ as well as mentors. Students who participate in the class actively
can score well; on the other hand students’ who miss the classes and do not participate actively in the class have fewer chances of success.

![Naive Bayes Model](image)

**Figure 4.8: Naïve Bayes Model**

### 4.3.3 Support Vector Machine

Support Vector machine takes numerical values to process the results, or it automatically changes the Yes/No values to 0 and 1. Thus, based on SVMs processing new important factors came out to light that put a negative or positive impact on the academic performance of the student. The factors that affect the performance negatively surely bring down the results of students. If these factors can be controlled at the early stage, the drop-out rate can be minimized and the retention rate can be maximized. For example, reappear in the previous semester; solving assignments with help, studying in exams, a major accident during last semester etc. negatively affects the academic performance of the students. Having pressure to appear in previous semester exam distracts student’s attention from present semester exams and the student is not able to perform well in the present semester exams.
Copying assignment from others or taking help to solve them leads to lower grades in assignments, and due to lack of knowledge student scores less in the final exams. In contrast, completing assignment on a daily basis left a positive impact on their academic performance. Studying in the exam period leaves the students in a confused state and they feel lost, as time limitations force them to cover the topics but they are not able to complete it at length. The result of this rush is lower grades in the exams.

Result accuracy of SVM came out to 98.5%.

**Important Factors for Atrisk**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reappear in the previb...</td>
<td>-0.75</td>
</tr>
<tr>
<td>Solving the assignments....</td>
<td>-0.50</td>
</tr>
<tr>
<td>Gender = Male</td>
<td>-0.25</td>
</tr>
<tr>
<td>Study in exams</td>
<td>0.00</td>
</tr>
<tr>
<td>Major mishap(Accident...)</td>
<td>0.25</td>
</tr>
<tr>
<td>Age = 19 years</td>
<td>0.50</td>
</tr>
<tr>
<td>Complete assignments daily</td>
<td>0.75</td>
</tr>
</tbody>
</table>

- **Supports ‘Atrisk’**
- **Contradicts ‘Atrisk’**

**Figure 4.9: Important Factors for Student at-Risk using Support Vector Machine**

### 4.4 Accuracy

To depict the performance of a classification model based on some test data where true values are identified, a table is created referred to as confusion matrix. It is also referred to as an error matrix. Using confusion matrix performance of an algorithm can be depicted. Confusion between classes at the time of prediction can be identified using confusion matrix. It is used while summing up the prediction results of a classification problem. The number of acceptable and unacceptable prediction are summarized by calculating the values and broken down by each class. The confusion matrix demonstrates the confusion created at
the time of building a classification model when it makes the prediction. It also illustrates the errors being made by a classifier. The confusion matrix shown in Table 4.1 illustrates how the confusion matrix works.

Table 4.1: Confusion Matrix for Classification Model

<table>
<thead>
<tr>
<th>Actual Values</th>
<th>Positive (1)</th>
<th>Negative (0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive (1)</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>Negative (0)</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

Description of terms used for confusion matrix:

- True Positive (TP): Positive Observation and Predicted Positive.
- False Negative (FN): Positive Observation but Predicted Negative.
- True Negative (TN): Negative Observation and Predicted Negative.
- False Positive (FP): Negative Observation but Predicted Positive.

Classification Rate/ Accuracy

Classification Rate or Accuracy can be calculated as:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

4.1

Recall:

Recall can be referred to like the ratio of the total number of correctly classified positive values divided by the total number of positive values. High Recall designates the class is correctly recognized (a small number of FN).

Recall can be calculated as:

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

4.2

These techniques were implemented using Rapid Miner Studio 9.0.1. Various models were created and compared on the basis of accuracy achieved for early
performance prediction of students. Result of comparison of the most effective Data Mining Techniques is depicted in Table 4.2.

Table 4.2: Comparison of DT, NB and SVM

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>98.0%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>98.2%</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>98.5%</td>
</tr>
</tbody>
</table>

Figure 4.10: Accuracy of NB, DT and SVM

SVM came out with the highest accuracy as compared to DT and NB. SVM projected the highest accuracy because the data set was best suited as per the input required by SVM. Considering the different scenarios, the accuracy of these data mining techniques may differ in other cases so there was a need to create an
ensemble of these techniques to achieve the same level of accuracy with different data sets in the educational setting. Table 4.2 shows the classification error of the data mining techniques.

<table>
<thead>
<tr>
<th>Model</th>
<th>Classification Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree</td>
<td>1.8%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>2%</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>1.5%</td>
</tr>
</tbody>
</table>

Table 4.3: Classification Error

4.5 Ensemble Model

Ensemble Model allows taking a trial of various features from different techniques and creating a final predictor on the basis of collective results of various techniques with better efficiency. The various popular Ensemble methods in data mining are Decision Tree, Random Forests, and Gradient Boost Trees etc. Creating an ensemble gives the privilege to create a better prediction model as compared to a single model. The idea behind creating Ensemble was to combine the best features of models. This study proposed a heterogeneous ensemble of Decision Tree, Naïve Bayes and Support Vector Machine. Based on the Classification Error of the three techniques and results produced after applying them on the training data set the three techniques were combined to create an Ensemble Model. The accuracy achieved after creating ensemble was much appreciable. The accuracy produced with the ensemble was 98.5% which is much higher than the existing models. The reason behind creating an Ensemble was to maintain the accuracy of the classification. Techniques which produced the best possible results for the data set used in this study. By creating ensemble the researcher made sure that the accuracy and predicted results should be consistent.
So, by merging the best features of these classification algorithms to produce the best results, the researcher created an Ensemble which could perform better when different data set are fed with the same attributes. Results of the ensemble model are shown in Figure 4.11 in the form of a bar chart.

![Prediction of Academic Performance](image)

**Figure 4.11: Prediction of Student at risk**

The training data set consists of 499 instances with 49 attributes and 1 class attribute. The number of students found on high-risk was 173 in number. To predict students at-risk various classification and regression algorithms were applied to the data set. Support Vector Machine, Naïve Bayes and Decision Tree showed better accuracy than various other Classification and Regression algorithms. These algorithms were selected to create an ensemble model for predicting the students at-risk. Naïve Bayes gave 98.0% accuracy, Decision Tree 98.2% and Support Vector Machine 98.5%. The ensemble of the three techniques proved best with an accuracy rate of 98.5%.

As discussed earlier, the Data Mining Techniques operate on different types of data and result in a variety of factors with their different level of accuracy. To design the Ensemble Model, comparison of various classification techniques were
done on the basis of high accuracy. Techniques with higher accuracy are used to create Ensemble for better prediction level. The ensemble model is shown in Figure 4.12

Figure 4.12: Ensemble of DT, NB, SVM

The influential attributes extracted by applying the Data Mining Techniques are discussed above. A description of these variables along with their negative and positive impact is shown in Table 4.3.

Table 4.4: Description of Important Factors

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Impact on Performance</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous Semester Percentage</td>
<td>Positive</td>
<td>Below 50%(6), 50-60%(167), 60-70%(183), 70-80%(109), 80-90%(34), above 90%(0)</td>
</tr>
<tr>
<td>Ease of Scoring Good Scores in Semester</td>
<td>Positive</td>
<td>Yes (479) No(20)</td>
</tr>
<tr>
<td>Location of Residence</td>
<td>Positive</td>
<td>Urban(334), Semi-Urban (85), Rural(80)</td>
</tr>
<tr>
<td>Reappear in Previous Semester</td>
<td>Negative</td>
<td>No(254), 1(101), 2(108), 3 and above (36)</td>
</tr>
<tr>
<td>Major Mishap</td>
<td>Negative</td>
<td>Yes(136), No(363)</td>
</tr>
<tr>
<td>--------------</td>
<td>----------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Habit of Consuming Alcohol and Tobacco</td>
<td>Negative</td>
<td>Yes(36), No(463)</td>
</tr>
</tbody>
</table>

The description of the influential attributes has a positive and negative impact on students’ academic performance as shown in Table 4.3. This table consists of student data that was collected through a questionnaire. 499 instances were recorded from the respondents. It describes the influence of factors on students’ academic performance. Previous semester percentage depicts that if a student scored more than 60% they had chances to improve but if percentage comes down to 50% or below 50% they were at risk of dropping-out or showed low motivation.

Related data shows a number of students lying in a particular category of percentage. In behavioral factor, ease of scoring good grades in the semester lays a positive impact; if a student is eager to score well then they are not at risk of failing. If a student resides in an urban location they have better opportunities, residing in semi-urban, and the rural area did not help students to perform better in their academic carrier. Being an urban resident could increase career options for students. Having a supplementary exam in the last semester put a negative impact on the academic performance of students. It implies that the student is not able to perform well in the present semester, as the pressure of previous semester leads to failure in performing well in the present semester. Any major mishappening in the family could negatively affect the academic performance of the student, the students are not able to concentrate on their studies and have maximum chances of low grades in the present semester, they also may fail to perform in the exams. The habit of consuming alcohol and tobacco could lead to failure in the academic carrier; students addicted to bad habits do not perform well in studies.