Chapter 3

Research Methodology

3.1 Introduction

The research methodology is a structured way of analysing the theoretical analysis of the techniques applied to a field of study. The purpose of the research is to find out answers to a specific problem in a scientific manner. Research interrogates/probes the facts that are hidden for a specific data. Evidence gathered through experiments or empirical studies are considered to be the most powerful support possible for a given problem statement. Empirical research is measured on the basis of observations and knowledge derived through experiments rather than theory or beliefs.

The research methodology adopted for this study is empirical and is based on the data collected through a questionnaire. The researcher interrogated the Literature Review with a view to corroborate their study, created various questionnaires sent online too, rendered the solutions into an excel sheet as it would establish accessible knowledge to the expert and future research scholars, data mining techniques that offered new perspectives to the present study, ascertaining Ensemble Model based on data mining technique where CAPA played a major role in gathering results of the research.

The methods and techniques used in research methodology are concise and terse, and thus offer a direct solution to the problems dealt with herein. As it remains an empirical study, the task of the researcher remained focused on the variables and the questions put thus. The task further focused on creating a study through various tables and charts, such as flow charts and diagrams, as it facilitates the projection of the variables in a clear and concise manner, making it easy for the researcher to provide accessible information for future research. The research methodology thus is not based on assumptions but pragmatic approaches.
A flow chart, Figure 3.1., has been created to illustrate a detailed methodology adopted for the research work:

**Objective 1**
- Literature Review
- Influencing Factors

**Objective 2**
- Creation of Questionnaire
- Validation of Questionnaire
- Respondents to Questionnaire- 588

**Objective 2**
- Data Converted to Excel
- After Data Cleaning and Consolidation - 499
- Feature Selection

**Objective 3**
- Data Mining Techniques (Decision Tree, Naive Bayes, Support Vector Machine)

**Objective 4**
- Ensemble Model
- Factors Identified for Inclusion in CAPA
- CAPA

**Figure3.1: Detailed Methodology for Research**

### 3.2 Data Collection

Collection of data is said to be the process of collecting useful information based upon some measurements and area of interest in a recognised and organised way that facilitate to answer the research questions, in testing hypotheses and evaluating outcomes of the proposed study. Data collection methods are categorised in two types: Primary Data and Secondary data as depicted in Figure 3.1. Secondary data is the data that is already collected, used in studies, published in paper media etc. Secondary data increases the level of accuracy and as well as research validity and reliability. Primary data is the original form of data which is never used for any purpose. Primary data collection can further be classified into
two categories quantitative and qualitative. Foundation of quantitative data collection is mathematical and statistical calculations in various forms. Quantitative data collection incorporates questionnaire and survey-based collection, closed questions and clearly defined objectives on observations, these methods are cheap to apply and can be applied within a short duration of time. Qualitative data collection incorporates face to face interviews and observations and does not include mathematical and statistical calculations. Qualitative data collection is related to real-world data such as: sound, emotions, face to face interaction and open-ended questions that helps to dig deeper information as required.

![Data Collection Diagram](image)

**Figure 3.2: Methods of Data Collection**

The ideal size of the sample is the most concerned issue while preparing for data collection. Lots of misconceptions arose because of the large sample size. In addition to the population size and purpose of the study, the researcher considered criteria variables such as calculation degree of variability between attributes, level of confidence or risk and level of precision.
Degree of Variability is referred to as distribution of attributes in population. The sample size is calculated on the basis of the size of population: the larger the population the larger sample size to be calculated, the smaller the population size is the smaller sample is required.

Level of Confidence or risk is calculated on the basis of Central Limit Theorem that is performed when the population is repeatedly sampled. The average value obtained by these samples distributed normally about true value-that is normally distributed 95% of the sample values for example mean.

Level of Precision is also referred to as sampling error and it is calculated in the percentage range of ±5.

The sample size can differ from one attribute to another on the basis of variance in the attribute. The formula used to calculate sample size in this study is calculated as discussed above.

\[
\text{True Sample Size} = \frac{\text{Sample Size} \times \text{Population}}{\text{Sample Size} + \text{Population} - 1}
\]

The data collection was done based on the above formula depicted in 3.1. The total students enrolled in BCA course are 1000, making the true population size 1000 then the data collection was done from BCA II\textsuperscript{nd} semester students responded to the survey was 499, applying the formula on the population size and sample size it was justified with the result that sample size is correct for the study.

\[
\text{True Sample Size} = \frac{499 \times 1000}{499 + 1000 - 1} = 334
\]

The true sample size resulted in 334 respondents which justifies that the sample size selected for the study is sufficient based on the population size.

Eventually, the formula of sample size yields the required number of respondents for the study. The number of studies often includes 10% of the sample size to balance those responses that the researcher was incapable to include in the study because of a few valid reasons, and to compensate the non-responsiveness etc.
3.3 Questionnaire

A questionnaire is a mode of collecting data in which a respondent is expected to answer a series of questions. It acts as the main instrument in gathering data for the research work. It is essentially a set of standardized questions which follow an order that the researcher has framed for the respondent. The questionnaire that the researcher framed was specific, easy in language as it focused the students, asking one subject per question et al. The questionnaire also did not project any repetition.

Figure:3.3 Sample Questionnaire

Source: https://blog.hubspot.com/service/questionnaire

As shown in Figure 3.1 this study is based on a questionnaire that was prepared on the basis of factors extracted through literature review. The factors with maximum appearances in the literature were selected to prepare the questionnaire. A number of studies have been conducted to predict the academic performance of the students, most of the studies considered are social and academic factors, few of them worked upon through behavioral attributes only [12, 17, 42, 43, 49, 50, 91, 99, 102]. Most of the available literature represents that the studies have been conducted on humanities and secondary school children. A limited number of articles and literature is available on professional courses.

The factors offer a practical approach to the study undertaken as well as highlight the problematic areas. The questionnaires based on the factors undertaken have been studied in-depth.
Table 3.1: Appearance of Attributes in Literature

<table>
<thead>
<tr>
<th>Factors</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family Financial Background</td>
<td>[29] [41] [42] [50] [55] [75] [91] [93] [105]</td>
</tr>
<tr>
<td>Qualification of Parents</td>
<td>[17] [29] [41] [42] [48] [49] [50] [55] [75] [91] [93] [102] [105]</td>
</tr>
<tr>
<td>Family Size</td>
<td>[29] [41] [42] [48] [49] [50] [55] [75] [91] [93] [105]</td>
</tr>
<tr>
<td>Family Type</td>
<td>[17] [29] [41] [42] [48] [49] [50] [55] [75] [91] [93] [105]</td>
</tr>
<tr>
<td>Percentage In Previous Exams</td>
<td>[6] [8] [18] [24] [37] [42] [49] [50] [91] [102]</td>
</tr>
<tr>
<td>Attendance in Previous Class And Present Class</td>
<td>[6] [8] [18] [24] [37] [42] [49] [50] [91] [102]</td>
</tr>
<tr>
<td>Marks in Seminars Assignments</td>
<td>[6] [8] [17] [18] [24] [48] [105]</td>
</tr>
<tr>
<td>Teacher Related Attributes</td>
<td>[18] [21] [37] [55] [76] [78] [107]</td>
</tr>
<tr>
<td>Hint Based Systems For Students/ASSISTments/ITS</td>
<td>[13] [36] [37] [38] [48] [78] [86]</td>
</tr>
<tr>
<td>Online Study Related Behaviour</td>
<td>[38] [43] [57] [65] [88] [101]</td>
</tr>
<tr>
<td>Help Seeking Behaviour Of Students</td>
<td>[6] [12] [13] [48] [56] [75] [86] [93]</td>
</tr>
<tr>
<td>Positive / Negative Impact Of Social Circle</td>
<td>[43] [65] [70] [83]</td>
</tr>
<tr>
<td>Self Study After College Hours</td>
<td>[48] [55] [75] [93]</td>
</tr>
</tbody>
</table>

The proposed study focused on Bachelors of Computer Applications (BCA) course to predict the performance of the students. This course was chosen for study because it is one of the most sought out courses after bachelor in engineering/bachelor of technology after 10+2; pupil seeking to enter the IT hub at a fast pace. The study is based on the academic, demographic, social and behavioral features of the students.
3.3.1 Stages in Preparing for Questionnaire

Selecting a data tool for the study is a major decision based upon a few parameters various tools and data collection methods. The researcher selected questionnaire as a data collection method. Designing a questionnaire is a multistage procedure that demands special attention of many details at once. Collecting information from a different type of respondents require paying attention to each attribute. Since every respondent has different temperament, behavior and vision towards responding to the questions.

![Figure 3.4: Process of Questionnaire Design](image)

The questionnaire was designed on the basis of these influential factors covering all the influential factors mentioned in Table 3.1. This study was conducted to find out the student-at-risk, after applying the data mining techniques on data set collected through questionnaire.

The questionnaire was categorized into four categories based on:

i. Category observed as constant i.e. demographic factors like the qualification of father or/and, mother’s income etc.

ii. Category helping to figure out the academic performance of the student such a previous semester percentage and present semester attendance.

iii. Category elucidating social issues influencing academic performance that included factors such as peer pressure, extracurricular activities and location of the residence.

iv. The last category deals with student’s behaviour towards studies like time spent on self-study after college hours, assignment
completion behaviour, and habit of consumption of alcohol or tobacco and missing classes.

3.4 Validating Questionnaire

Questionnaire validation is referred to as establishing the content validity and construction validity to if the questionnaire is ready to bring out the required information. The validation is performed by experts from the same field with the experience of research.

The contents of the questionnaire were validated by a five member committee comprising of experts from management, statistics and data mining field. A Google form link was sent to the committee members and their suggestions were received through e-mail. Initially, the questionnaire had 77 questions but was reduced to 49 questions as suggested by the validation committee.

The suggestions from the committee included:

(i) Addition of more attributes related to academic performance and decreasing attributes related to social factors.
(ii) To reduce the number of attributes as it will help to perform the study in a qualitative manner.
(iii) To reduce the teacher related attributes as they will not be helpful in extracting the knowledge about students’ academic behaviour.
(iv) To add more student interest-related questions so that personal information can be collected from the students.

The questionnaire was thus modified based on the suggestions from the experts. The questions reduced to 49 representing 200 variables, approximately. The suggestions, however, did not affect the variables or the study in any manner. The questions were put forth after the researcher had formed the logical questionnaire.

The data was collected from the students of Goswami Ganesh Dutta Sanatan Dharma College affiliated to Panjab University, Chandigarh. Panjab University was established in 1882 as University of Punjab at Lahore (now in Pakistan) renamed after partition in 1947, making it one of the oldest universities in India.
588 responses were recorded from students already enrolled in Bachelor of Computer Applications (BCA) for the academic session 2016-2019. BCA is a three years professional degree course divided into six semesters. Students appear in examination twice a year after every six months. Initially, a Google form was created for collecting the responses of the student online but due to infrastructural challenges and low availability of computer labs very fewer responses were received which were not sufficient to conduct the study.

To overcome this issue, offline methods were used to record the response. An offline version of the questionnaire was distributed during the classes and responses were recorded in the presence of teacher thus making it a validated study. The student’s response was equally appreciative and the subjects offered an honest opinion. No help in seeking the answers was provided by anyone else.

Data set of the students’ consisted of a total 49 attributes comprising of academic, behavioural, demographic and social attributes. Table 3.2 shows the various categories of attributes along with its counting in number in a particular category.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total in Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic Variables</td>
<td>19</td>
</tr>
<tr>
<td>Academic</td>
<td>8</td>
</tr>
<tr>
<td>Behavioural</td>
<td>16</td>
</tr>
<tr>
<td>Social</td>
<td>6</td>
</tr>
</tbody>
</table>

Demography relates to the growth of an individual in a socio-cultural environment. It further highlights various characteristics that shape the moral-social and psychological identity of an individual such as income level, education level and size of the family et al. There are total of 19 demographic variables in the data set like demographic variables that consist of age, gender, financial background of the family, educational qualification of parents etc.

Academic attributes highlight the performance of a student that projects their academic performance. It is the extent to which a student, mentor or institute
attains their short or long term educational goals. For the same, a cumulative GPA (previous semester percentage) and successful completion of various benchmarks are considered too. The academic attributes considered during the study were 8 in number consisting of marks scored in the last semester, marks scored in assignments done in the current semester, current semester attendance, gap year, supplementary exam in last semester etc.

Behavioural attributes are the way of describing a range of individual characteristics that can be measured and can be showed to differentiate the effective and ineffective performance of a student; the effective performance being the fruitful conduct and accomplishment of the student whereas the ineffective performance highlights the negative behavioural patterns. Behavioural attributes depict the attitude towards studies such as assignment submission behaviour. There were 16 behavioural attributes like copying the assignment from fellow students, not submitting assignments on time, time spent on self-study after college hours, the habit of consuming alcohol and smoking etc. that describes the behaviour of the students.

Social attributes are formed through social interactions and massively influence an individual’s character. In terms of education, a student often relates to their friends and the peer pressure which often renders them inefficient. This often influences student’s academic as well as behavioural attributes. They are 6 in total such as motivation behind selecting the present course, time spent with friends, indulgence in bad habits etc. During the study, demographic and behavioural attributes were found to be most influential in students’ academic carrier; further shaping their educational choices for the future.

3.5 Characteristics of the Data

It is relevant to highlight the characteristics of the Data collected as it provides a clear and concise view of the variables that affected the results. Following are key characteristics of the data thus gathered:

i. Data is in a unique form (primary data).

ii. Data does not contain duplicate records. Thus no influence on classifiers and performance is affected by duplicate records.
iii. The sufficient number of records exists to perform the experiments, and there is thus no need to divide the data into small proportions to apply the experiment.

3.5.1 Data Pre-processing

The data was collected in an offline method using questionnaire. After collecting the data in manual form, it was converted into a digitized form. Manual entries were done in as Excel Sheet as shown in Figure 3.2.

![Sample Data Set](image)

**Figure 3.5.: Sample Data Set**

The following tasks were done for data pre-processing:-

![Data Pre-Processing Diagram](image)

**Figure 3.6: Data Pre-Processing**
3.5.2. Data Cleaning and Consolidation

Data Cleaning and Consolidation is one of the major tasks as the study tends to gather variables which do not render any meaningful answers. After digitizing data, data cleaning and consolidation was performed. Three criteria were considered to clean the data set:

i. **Missing Values:** In some attributes where missing data was not a problem the response was accepted otherwise it was rejected, for e.g. Father’s missing name was accepted but previous year’s missing percentage was rejected. Hence, the required correction/rejection in the dataset. This problem was faced only while filling the demographic data as it provided extensive data. However, no other modifications were done as regards the missing value.

ii. **Incomplete Data:** There was less probability of incomplete data since most of the answers were on the Lickert Scale/Objective type. Extensive research and questionnaires based on the research left less chance for any incomplete data. Further, the tools applied led to a definite response.

iii. **Format Errors:** There was no chance of formatting errors since all the answers were on the Lickert Scale/Objective type. Some corrections were performed by asking students while others were corrected manually, i.e. the status is employed but the salary is nil. No other data was modified during the correction of the format errors.

The survey conducted focused on the number of instances found correct, complete and considered. The instances considered in the survey was 499, corresponding to 49 attributes and one class variable. The number of students found at-risk after analysing the statistics is 173; data set gives 499 x 50 dimensions. The properties of the data set were based on objective and the Lickert Scale. Additionally, no other changes were made to the survey. The demographic, academic, behavioural and social attributes are reflected through the Number of Attributes. Class attribute
predicts the category of student performance. Table 3.3 shows the description of attributes.

**Table: 3.3.: Description of Data Set**

<table>
<thead>
<tr>
<th>Properties</th>
<th>Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of Instances</td>
<td>588</td>
</tr>
<tr>
<td>Number of Instances Found correct and Included in the Study</td>
<td>499</td>
</tr>
<tr>
<td>Number of Attributes</td>
<td>49</td>
</tr>
<tr>
<td>Class Attribute</td>
<td>1</td>
</tr>
<tr>
<td>Number of Students at-risk</td>
<td>173</td>
</tr>
<tr>
<td>Dimension</td>
<td>499x50</td>
</tr>
</tbody>
</table>

After correcting and completing the entries in the data set, most suitable data mining techniques required analyzing the data set. The number of data mining techniques were studied and compared based on their result calculation ability. Few are the techniques discussed below suitable to the data set for the study.

**3.6 Data Mining Techniques**

**3.6.1 Decision Tree**

Decision Tree is the most popular technique amongst machine learning algorithms. It is used to create classification as well as regression models. Decision Tree is built in Top-Down fashion. It finds the best feature that splits the class label into the purest form of child nodes. In Decision Tree each node symbolizes a characteristic (attribute), each connection (branch) stands for a decision (rule) and each leaf signifies a result (categorical or continues value). The goal of Decision Tree is to build noiseless or homogeneous data that classifies into pure nodes. To achieve its goal, Decision Trees utilize Entropy or Information Gain to build a tree.
3.6.1.1 Entropy

Entropy indicates the noise or diversification in data. To quantify the homogeneity in data, entropy is required to reduce the noise in data. A smaller amount of entropy signifies more purity of an attribute. The smallest possible value intended for entropy is 0. The purity after performing entropy calculation is called Information. Information symbolizes the expected quantity of knowledge required to identify whether a new instance should be classified further to reach the required class label.

\[ \text{Entropy} = \sum_{i=1}^{c} -p_i \log_2(p_i) \] (3.1)

3.6.1.2 Information Gain

Information Gain is also called Mutual Information. It is used to quantify the amount of information given by an attribute about the class. A feature that splits best has a maximum weight. Feature with the highest gain will be selected for the split. Information gain ensures that the noise in the data set should be reduced. It is the entropy of the root node minus the entropy of the leaf nodes. It helps in reducing entropy.

3.6.1.3 Classification Error and Accuracy

Classification error is the amount of error by which the respondent does not provide a true response to a survey item. For nominal categorical data, this can occur in one of the two ways: a false negative response or a false positive response.

\[ \text{Classification Error} = \frac{\text{error total}}{\text{total}} = \frac{FP+FN}{TP+TN+FP+FN} \] (3.2)

\[ \text{Accuracy} = (1-\text{error}) = \frac{\text{correct total}}{\text{total}} = \frac{TP+TN}{TP+TN+FP+FN} \] (3.3)

3.6.2 Naïve Bayes

Naïve Bayes is effective and most commonly used in the machine learning algorithm. Naïve Bayes is a probabilistic classifier that creates classifications using the highest posteriori Decision Rule in a Bayesian setting. Naïve Bayes operates on conditional probabilities. It is based on
Bayes theorem. It calculates Posterior Probability based on Prior Probability and Conditional Probability. It operates on both numerical as well as categorical attributes.

3.6.2.1 Posterior probability

The posterior probability is the probability of an incident that occurs following the evidence or background information that is available. It is calculated after revising the prior probability; the probability calculated before an event will happen before any new evidence is taken into account. Posterior probability can be called as a change in prior probability:

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]

The result calculated is called Posterior Probability.

3.6.3 Support Vector Machine

Support Vector Machine abbreviated as SVM is the most popular among machine learning algorithms; it produces accurate and significant results with less computing power. It is based on the concept of defining boundaries on a decision plane. A decision plane divides a data set into two classes to find a hyper plane on N-dimensional space that increases the margin between two classes and decreases the misclassification. It is used to build Classification, Regression, and Anomaly Detection models, but it is preferred in classification models. It operates on numerical attributes but if Categorical Attribute is encountered then it automatically changes categorical data into binary values 0 (false) and 1 (true). So, based on SVMs processing, new important factors came to light that put a negative or positive impact on the academic performance of the student. For example, reappear in the previous semester; solving assignments with
help, studying in exams, a major accident occurred during last semester etc. negatively affected the academic performance of the students. In contrast, completing assignment on a daily basis left a positive impact on performance.

3.6.4 Gradient Boosting Trees

To solve regression and classification problems Gradient Boosting Technique is popular amongst various machine learning techniques. It constructs an ensemble of weak prediction models for the prediction purposes, typically decision trees. Like other boosting methods it creates the predictive model in phases. Conversion of weak learners into strong learners is called Boosting. Boosting is done by increasing the weight of those attributes that are difficult to classify and decreasing the weight of those that are easier to classify.

Gradient boosting works on three elements [108]:

i. A loss function to be optimized.
ii. A weak learner to make predictions.
iii. An additive model to add weak learners to minimize the loss function.

3.6.5 Ensemble Method

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting) or improve predictions (stacking). It is the aggregation of average outputs of various models to solve complex problems with greater accuracy and higher generalization capacity. According to available literature, it has been proved that ensemble methods are capable of increasing the accuracy by 30% when compared to any best single model [40]. Ensemble models can be constructed using any of the three approaches mainly bagging, boosting and stacking [34]. Ensemble models are capable of increasing performance because they have the ability to minimize classification errors and bias.
3.6.5.1 Bagging

Bagging is used when the purpose is to reduce the variance of a decision tree classifier. The goal is to build numerous subsets of data from the training example selected randomly with replacement. Each group of subset data is used to train their decision trees. As a result, we get an ensemble of different models. Combining the predictions from various trees can be used to build a robust model rather than a single tree [34, 97].

3.6.5.2 Boosting

Boosting is used to generate a group of predictors. In this technique, learners are learned in sequence with early learners fitting simple models to the data and then investigating data for errors. Consecutive Trees (random sample) are fit and at every step, the goal is to improve the accuracy from the prior tree. When an input is wrongly categorised by a hypothesis, its increase in weight helps the next hypothesis to classify it more accurately. This process alters fragile learners into an enhanced model [34, 97].