CHAPTER 4
FRAMEWORK FOR RESEARCH

4.1 Introduction

In order to develop a model for prediction of student potential, the first step is to identify a standard process and use it. The Cross Industry-standard Process for Data Mining (CRISP-DM) (Shearer 2000) has been adapted for use in our study, to develop six classifier models for achieving the objectives of the study. It provides a structured approach to planning any data mining project. It is a robust and well-proven methodology. We discuss it briefly to put it in context.

In this chapter, we have used five phases for development of the model: (i) Problem understanding (ii) data understanding (iii) preparation of data (iv) different classifier models and (v) evaluation. The developed classifier models were applied on the collected dataset and compared for performance, to determine the best classifier.

In addition, this study is concerned with finding the optimal feature subset to be used with the best classifier model. The chapter explains how students' data can be used to build the models. It also discusses the metrics that are used in the classifier models for the evaluation of student's potential and performance.

4.2 CRISP-DM for Education

CRISP-DM is a process that has been used in education sector. It was proposed by (Kurgan and Musilek 2006). Figure 4.1 shows the complete process of the CRISP-DM for the education sector.

Business/Problem Understanding

The problem has to be understood properly and defined clearly. The objectives of the study must be clear and understandable. With this, the aim of the research, the problem statement, and the research design is established.

Data Understanding

This phase starts with the collection of data related to the objectives of the study. The completeness of the data is verified, the missing values are taken care of and redundant data is removed.

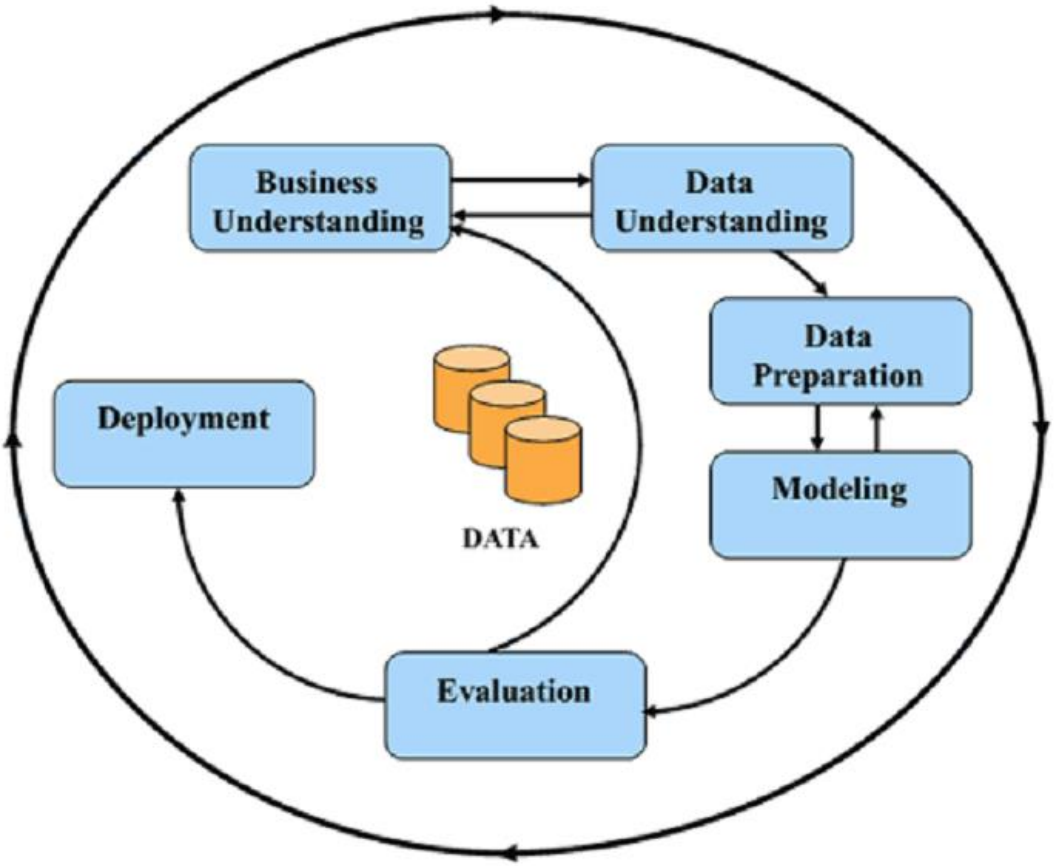

Figure 4.1 CRISP-DM Process

Source: Manasson A., 2019

Data Preparation

This step includes data cleaning, transformation of the data and selecting the optimal feature subset. The aim of this phase is to achieve a dataset that helps in better prediction. Data preparation is most time-consuming phase of this process. It accounts for about three quarters of a data analyst's work. Data reduction is also part of this phase.

Modeling

This step is used for selecting the classifier methods to be applied on the collected dataset to evaluate the performance of the classifiers. The dataset is separated into two subsets, one for training and another for testing. New data, if any, can also be tested.

Evaluation

This step takes the results from the previous step and interprets the results. The interpretation finds interesting patterns and uniqueness in the data. It may also repeat the previous step for refining the results.

Deployment

In this step, the knowledge discovered from the performance model is used for documentation and deploying it to the interested population.

The above phases have been slightly modified/extended in this study, to build a research framework, as explained in the next section.

4.3 Methodology

We have created our own research framework to analyse the student potential and performance in the technical programs, using Machine Lerning. The steps of the framework are illustrated in Figure 4.2. This research has followed the steps for predicting the student potential using individual Classification Algorithms and Ensemble Methods.

Firstly, we have collected the original dataset from different universities and institutions. After that, data preprocessing has been done on collected data for extracting relevant information. This study aims to introduce the student performance model with important features like demographic, academic, and financial.

After data collection and preprocessing, five common classification algorithms were used: Logistics Regression, Naïve Bayes, K-Nearest Neighbor, Decision Tree, and Support Vector Machine. Then Ensemble Methods were applied. These are Bagged Decision Trees, Random Forest, Voting Classifier, Adaboost, and Stochastic Gradient Boosting, based on bagging and boosting. The Ensemble Methods are used to improve the performance of the models, as explained in the previous chapter.

All the phases of this framework are discussed in detail below.

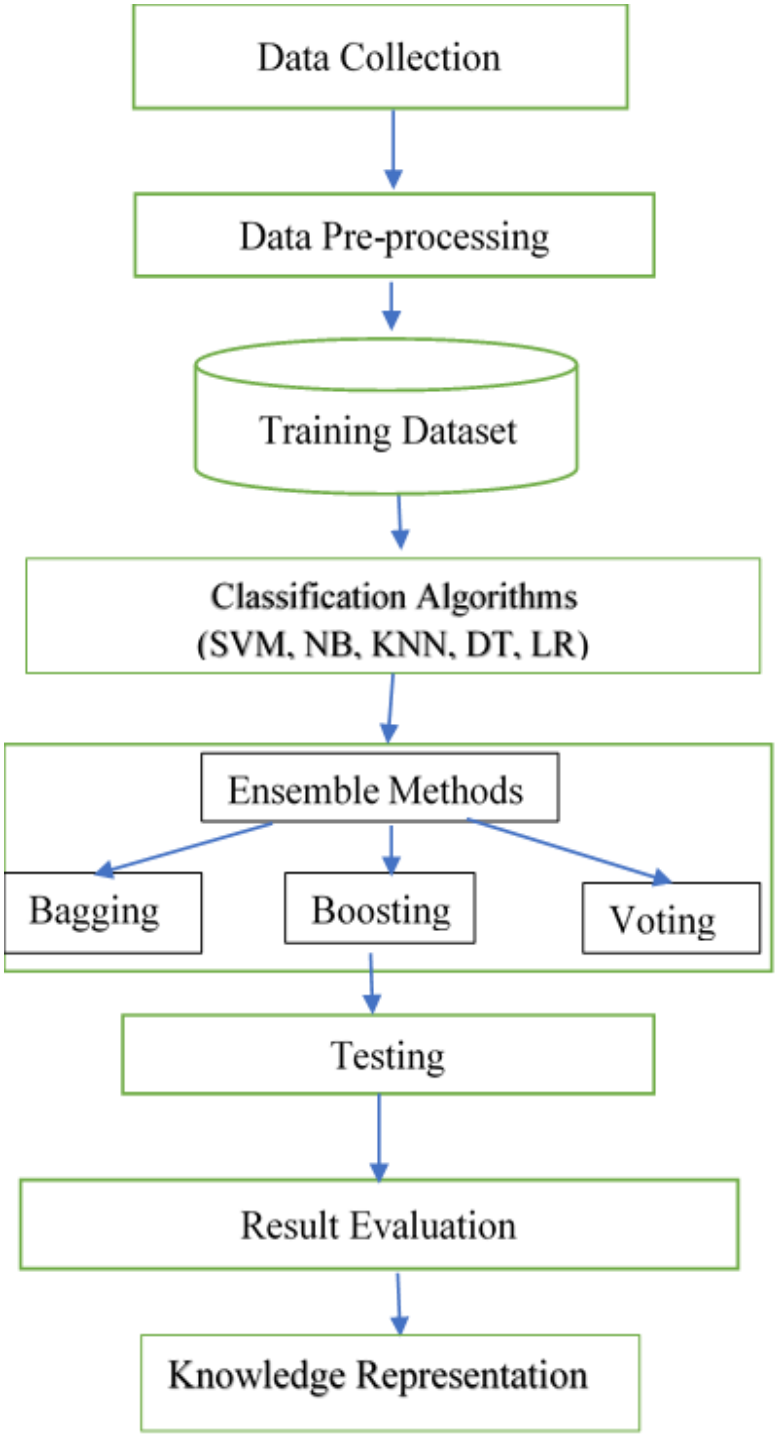

Figure 4.2 Methodology of Research

4.3.1 Data Collection

The aim of our study is to predict whether a student is eligible or not for technical programs, based on their potential and also classify the performance of the students in technical programs. This is a classification problem.

This problem also confirmed by the admission counsellors, teachers and parents in universities. A survey was conducted with the admission counsellors and teachers,

which revealed that it is a difficult task to predict student potential and performance by manual means.

Before analysis, it is important to obtain detailed knowledge about data types of each attribute, to make it suitable for building the prediction model. The source of the data was students passing 12th standard, who took admission in technical and non-technical programs.

It is to be noted that the original dataset that was used to predict student's eligibility in technical programs, had 21 attributes, as shown in Table 4.1. This dataset was re-prepared and an additional attribute was added to it, to indicate the Education Board that the student passed his 12th standard in. This was used for the second study, which predicted the potential of the student, based on the Board. The dataset was further extended for analysis of student performance, by adding five categories: Failure, Poor, satisfaction, good and excellent, for the result of the student.

The data was collected from BSDU and other institutions and universities such as Amity University, Jaipur National University, Vivekananda Global University, Kautilya Institute of Technology, Subhodh College, and Poornima College. The original data includes student details such as id, name, age, dob, address, gender, cluster, father_occupation, mother-occupation, father-qualification, mother_qualification, parent_income along with 10th and 12th Subject marks. The dataset has 21 attributes in total within 3 types (Demographics, financial, academic, and socio-economic attributes).

Attributes	**Data Type**	**Description**
Age	Numeric	Student Age
Gender	Nominal	Student Gender
Fedu	Nominal	Father Education
Foccu	Nominal	Father Occupation
Medu	Nominal	Mother Education
Moccu	Nominal	Mother Occupation
Fincome	Numeric	Family Income
NOS	Numeric	No. of Siblings
Address	Nominal	Belongs to
MoE	Nominal	Medium of Education

10^{th}	Numeric	10^{th} Marks
Stream	Nominal	PCM, PCB, Arts, Commerce
12^{th}	Numeric	12^{th} Marks
StIntforJob	Nominal	Student Interest in job
Health	Nominal	Health
ST	Numeric	Study Time in a week
STS	Numeric	Study time on Sunday
IntInflence	Nominal	Internet Influence while choosing a career
AreaofInterest	Nominal	Area of student
Why this course	Nominal	Why choose this course
Program	Nominal	Technical or Non-Technical

Table 4.1 Original Attributes in Student Dataset

The target population is students passing 12^{th} standard from Rajasthan and also the students who are already pursuing technical and non-technical programs. A total dataset of about 2000 students was used for better prediction of the result through Machine Learning techniques and Ensemble Methods.

The extended dataset, with an attribute for Board passed, has 1000 records each for the two Boards, the Rajasthan Board and the CBSE Board.

Figure 4.3 shows the steps that were used in the data collection process. We discuss each of them next.

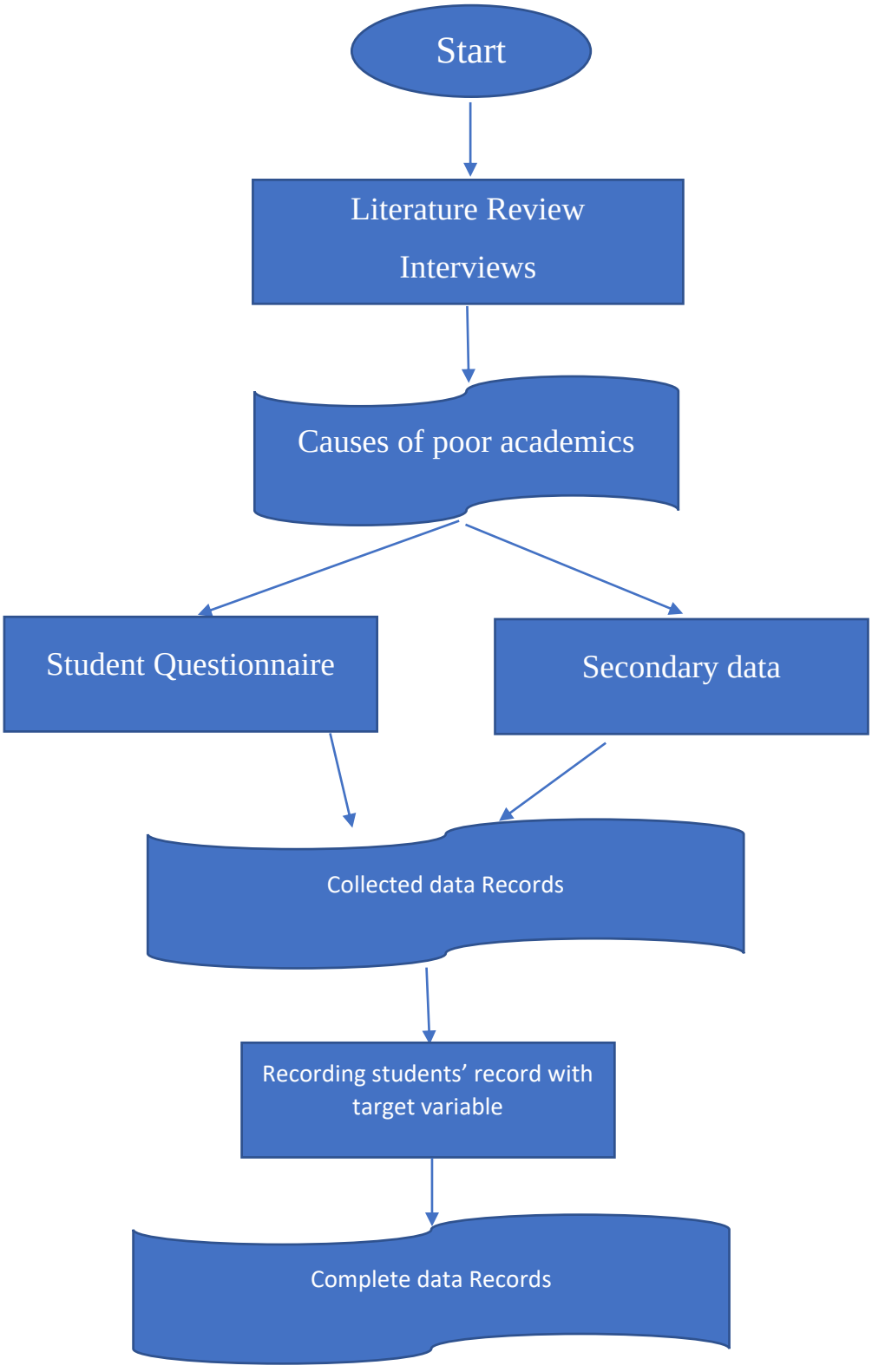

Figure 4.3 The data collection process

Literature Review and Interviews

The comprehensive Literature review that was done, helped us to find out the attributes that have and should be used for such a study. We also interviewed many stakeholders

in the education sector, viz., admission counsellors, teachers, recruiters, Deans and other office bearers, even parents, in order to get their opinion of what attributes are important

Questionnaire for Students

A Questionnaire was then designed with all the 20 attributes and distributed among the students who had passed out from the 12th standard and taken admission in technical and non-technical programs during the year 2018-2020. The data included student's demographic, socio economic, academic and financial information.

Secondary Data Collection

Some data was collected directly from the university's admission department such as 10^{th} and 12^{th} marks. Also marks obtained in different programs were collected from the examination department.

The 20 attributes are the independent variables. These attributes were categorised into 2 types – categorical attributes and numerical attributes. The categorical variables types are gender, age, etc. The numerical variables are in continuous types. Figure 4.4 shows pictorially, how the independent variable student potential and performance influenced the dependent variables. The whole dataset was divided into three groups based on nature. Demographic variable included gender, age, Number of siblings, health, family income and medium of education; Socio-economic variable consists of father education, father occupation, mother education, mother occupation etc. and academic includes all academic-related attributes.

Two universities in the Jaipur district of Rajasthan were initially selected for the pilot study. One was for technical programs and another for non-technical programs. A total of 298 students who took admission in different programs were used as a sample for this study. All information related to the student's demographic, socio-economic and academic variables was obtained from the students directly.

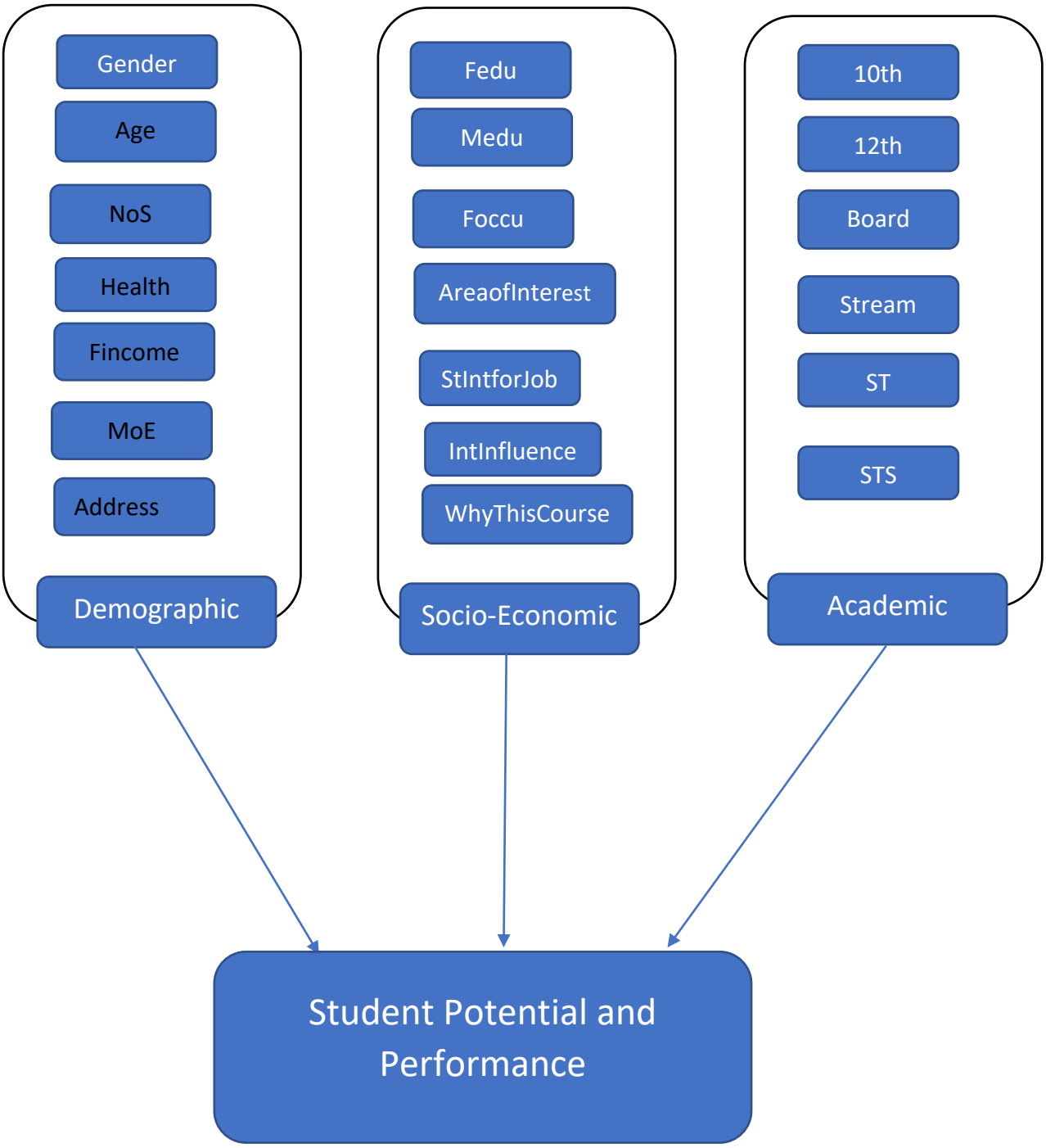

Figure 4.4 Dependency Relation with Demographic, Socio-economic and Academic Variable

The predictive accuracy of the student potential was found to be 45.28%. The outcome of the pilot study showed that there was a strong correlation between the attributes like parents' education, income, student interest and marks obtained at higher secondary level and academic performance.

Base on the outcome of the pilot study, we were ready for a detailed study. The datasets were collected from different universities and institutions. The selection of these

universities in Jaipur was based on the maximum admission taken in technical programs. Preprocessing of data in terms of cleaning or transformation to make modelling easier.

The objective of feature selection is to identify an optimal subset of variables that is sufficient for the construction of a prediction model which gives higher accuracy as well as takes minimum time.

But, the first step in any application based on Machine Learning is to preprocess the data. This means deleting or filling in the missing value and removing the redundant or unrelated attributes from the collected dataset. It increases the reliability of the predictive model and increases the accuracy.

Figure 4.5 shows 30 records from the complete dataset of 2000 records, that was used for analysis of student potential in technical programs. The first row shows the attribute names. It includes 21 attributes with the target variable.

Age	Gender	Fedu	Foccu	Medu	Moccu	Fincome	NoS	address	MoE	10th	Stream	Twelth	S3In3forJo	Health	ST	STS	IntInflence	AreaofInte	Whythisco	Branch
19	0	3	1	1	1	1	3	1	0	1	2	2	1	1	1	2	1	1	1	1
19	0	3	3	2	1	3	1	1	1	4	1	2	1	1	1	1	1	1	1	1
19	0	3	3	3	1	1	1	1	1	5	1	3	1	1	2	1	1	1	1	1
19	0	2	4	0	1	1	3	2	0	3	1	4	1	1	1	1	1	1	1	1
19	0	3	4	1	1	1	1	2	0	5	1	4	1	1	1	1	2	1	1	1
19	0	3	2	3	3	3	1	1	1	5	1	4	1	1	1	1	2	1	1	1
20	0	3	3	3	2	3	1	1	1	5	1	3	1	1	1	2	1	1	1	1
20	0	3	3	3	1	2	1	1	1	4	1	2	1	1	2	1	1	1	1	1
19	0	3	1	2	1	3	1	1	0	3	1	2	1	1	3	3	1	2	1	1
18	1	2	3	3	2	2	1	1	0	4	1	3	1	1	1	1	1	3	1	1
18	1	2	3	1	2	3	3	2	0	4	1	4	1	1	2	1	1	3	1	1
20	0	3	2	3	2	2	1	1	1	4	1	2	1	1	1	1	1	1	1	1
20	0	3	2	2	1	3	1	2	0	3	1	2	1	1	3	3	1	1	1	1
18	0	1	1	0	1	1	1	2	0	1	3	2	1	1	3	3	1	2	1	1
18	1	1	3	1	1	1	2	1	1	4	3	5	1	1	1	1	2	3	1	1
20	0	2	1	0	1	1	1	2	1	3	1	3	1	1	2	1	1	2	1	1
18	1	3	1	3	3	2	1	1	1	6	2	4	1	1	3	1	1	2	1	1
20	0	1	1	1	1	1	1	2	0	5	1	5	1	1	2	2	1	2	1	1
18	0	3	2	3	1	3	1	1	1	1	2	1	1	1	1	1	1	2	1	1
18	0	3	4	3	1	1	1	2	1	5	4	3	1	1	2	2	1	2	1	1
19	0	1	3	1	1	1	1	2	0	2	2	3	1	1	1	1	1	2	1	1
19	1	1	2	1	1	2	1	1	0	3	2	3	1	1	1	1	1	2	1	1
19	0	2	1	0	1	1	1	2	0	3	2	3	1	1	1	1	1	2	1	1
18	1	2	2	1	1	3	3	1	0	3	2	3	1	1	1	1	1	2	1	1
21	0	3	3	1	1	3	1	1	0	2	2	0	1	1	3	1	1	1	1	0
18	0	1	3	1	1	1	1	1	0	3	3	3	1	1	2	1	1	2	1	1
19	0	3	3	3	2	1	1	1	1	1	4	2	2	1	1	1	2	4	1	1
19	0	2	2	0	1	2	2	2	0	2	3	4	1	1	2	2	1	2	1	1
20	0	3	1	1	1	1	1	2	1	1	2	2	5	1	2	2	1	1	1	1
18	0	2	1	0	1	1	1	2	0	5	1	4	1	1	2	2	1	2	1	1

Figure 4.5 Data Collected for Student Analysis Based on Attributes

Figure 4.6 shows the dataset which is used for the analysis of student potential in technical programs, based on Education Boards. The previously used dataset, as shown in Figure 4.5, was extended to add the Board attribute and divided into two datasets based on the different boards, i.e., RBSE and CBSE. Each dataset has 1000 students' These datasets are used to predict the student potential.

Age	Gender	Fedu	Foccu	Fincome	MoE	12th	Branch
19	M	H	P	3	E	3	1
19	M	H	SE	3	E	3	1
19	M	H	SE	1	E	3	0
19	M	S	O	2	H	4	1
19	M	H	O	1	E	4	1
19	M	H	G	3	E	3	1
20	M	H	SE	3	E	3	1
20	M	H	SE	2	E	2	1
19	M	H	P	3	H	3	1
18	F	S	SE	3	H	3	1
18	F	S	SE	3	E	4	1
20	M	H	G	2	E	2	0
20	M	H	G	3	E	3	1
18	M	P	G	1	E	2	0
18	F	P	SE	3	E	5	1
20	F	H	P	2	E	3	1
18	F	H	P	2	E	4	1
20	M	P	G	1	H	5	1
18	M	H	G	3	E	1	0
18	M	H	O	1	E	3	0
19	M	H	SE	3	H	3	1
19	F	P	G	2	E	3	1
19	M	S	P	3	H	3	1
18	F	H	G	3	H	3	1
21	M	H	SE	3	E	2	0
18	M	P	SE	1	E	3	0
19	F	H	SE	1	E	2	0
19	M	S	G	2	H	4	1

Figure 4.6 Dataset Collected for Analysis Based on Boards

Gender	Fedu	Medu	Fincome	MoE	Stream	Tenth	Twelth	ST	STS	Board	Whythisco	Program	Result
M	H	P	1	E	Com	1	2	1	2	RBSE	PI	0	3
M	H	S	3	E	PCM	4	2	1	2	CBSE	PI	1	3
M	H	H	3	E	PCM	5	3	2	3	CBSE	PI	1	4
M	S	N	1	H	PCM	3	4	1	3	RBSE	PI	1	4
M	H	P	1	E	PCM	5	4	1	4	CBSE	PI	1	5
M	H	H	3	E	PCM	5	4	1	4	CBSE	PI	1	5
M	H	H	3	E	PCM	5	3	1	3	RBSE	PI	1	4
M	H	H	2	E	PCM	4	2	2	2	RBSE	PI	1	3
M	H	S	3	E	PCM	3	2	3	2	RBSE	PI	0	3
F	S	H	2	E	PCM	4	3	1	3	CBSE	PI	1	4
F	S	P	3	H	PCM	4	4	2	4	CBSE	PI	1	5
M	H	H	2	E	PCM	4	2	1	1	CBSE	PRI	1	3
M	H	S	3	H	PCM	1	2	3	1	RBSE	PRI	1	2
M	P	N	1	E	ARTS	1	2	3	1	RBSE	OR	0	2
F	P	P	1	E	ARTS	4	5	1	4	CBSE	PI	1	5
M	S	N	1	E	PCM	3	3	2	2	CBSE	PI	1	4
F	H	H	2	E	Com	6	4	3	3	CBSE	PI	1	4
M	P	P	1	E	PCM	5	5	2	4	RBSE	PI	1	5
M	H	H	3	H	Com	1	1	1	1	RBSE	PRI	1	1
M	H	H	1	E	PCB	5	3	2	3	CBSE	PI	1	4
M	P	P	1	H	Com	2	3	1	2	CBSE	PI	1	3
F	P	P	2	E	Com	3	3	1	3	RBSE	FI	0	4
M	S	N	1	E	Com	2	3	1	2	RBSE	FI	0	3
F	S	P	3	H	PCM	1	1	1	1	CBSE	PI	1	1
M	H	P	3	H	Com	2	0	3	1	RBSE	PI	1	1
M	P	P	1	E	ARTS	3	3	2	2	RBSE	FI	0	3
M	H	H	1	H	PCB	1	2	1	1	RBSE	FI	0	2
M	S	N	2	E	ARTS	2	4	2	3	CBSE	PI	0	4

Figure 4.7 Dataset Collected for Analysis of Student Performance

Figure 4.7 shows the third dataset which was used for the analysis of student performance in technical programs. The target variable was categorized into five categories: Failure, Poor, satisfactory, good and excellent, using the numbers 1 to 5. It includes a total of 1850 student records.

4.3.2 Data Preparation

The technique by which raw data is converted into a clean dataset is known as data preprocessing. Datasets are required to be clean to improve their quality and representation, so that the algorithms give accurate results. Selection of data features, data cleaning, data transformation and reduction are all included in data preprocessing.

Our dataset contained 120 missing values in different features in the 2000 records. After removing all the missing values, 1880 records were available. Data transformation was then applied to the dataset. Categorical data type attributes like Gender, address, IntInflence, etc., were transformed to binary data '0' and '1'. Other categorical data type attributes like Foccu, Moccu, Fincome, etc., were transformed to the numerical data type.

We have used the library *sklearn* to pre-process data. The following procedures were applied for preprocessing the dataset:

a) Label Encoding: Label Encoding refers to converting the labels into the numeric form to convert into the machine-readable form. Sklearn library provides a very effective tool by which categorical attributes are converted into numeric values. LabelEncoder is used to encode categorical attributes with values between 0 and n_classes-1. For example, there are two levels of male or female in ‘gender’ attribute.

b) One Hot Encoding: This method is used to divide the column that contains numerical *categorical data* into multiple columns based on the number of categories present in that column. Each column has “0” or “1” corresponding to which it is placed.

c) Data standardization: It is a useful technique in which attributes transform with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a 0 mean and 1 standard deviation. scikit-learn was used with the StandardScaler class for standardizing data.

Data cleaning

Data cleaning is the process of deleting or replacing the missing values and those rows that did not have the target variable. Sometimes students leave blank values in the

questionnaire, which create an error due to missing values. Other processes like sorting, labelling and filtering were used to improve data readability and visualization. These parts are explained next.

Attribute Selection

The attributes were selected for better prediction of student potential and performance. The dataset has to contain only relevant data, so that the size of the dataset can be reduced for easy management and so that it takes minimum time for execution. The following tasks were performed:

- Attributes that had a null value were removed.
- Irrelevant attributes like *student name, DOB, address*, which are not helpful in prediction, were deleted from the data set.
- Redundant attributes which provide identical information, were also eliminated.

Replacing the Missing Values

In the questionnaire, some values in the dependent variables were missing. Some had numerical values and some had categorical attributes. So, numerical values were filled with the mean value of the column, as was done in (Acuna and Rodriguez, 2004) and categorical values were replaced with the most frequently occurring value in the column (Garcia et al., 2015).

Delete Records

The rows that did not have the target variable were identified as insufficient and incomplete for training or test data. They are not used in supervised learning because it is works on labelled data only. So, they were removed from the collected dataset (Han and Xia, 2014). Redundant values were also deleted.

4.3.3 Exploratory Data Analysis (EDA)

To study and observe the behaviour of data, attributes and relationships between independent and dependent variables were graphically visualized, so that the pattern of data is properly studied; as also to explore the dependency and weightage of attributes, to extract reliable features to develop a reliable model. The primary plots that were visualized by doing the following:

- Individual attributes were plotted against their frequency, to observe the data as shown in figures 4.8(a)-4.8(g),

- Attributes were plotted against the target variable, to study the dependency and weightage, as shown in figures 4.8(h)-4.8(n)

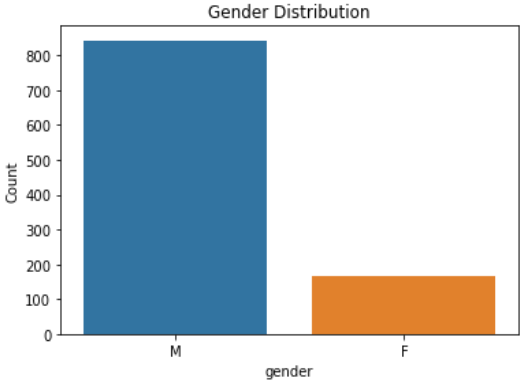

Figure 4.8(a) Gender Distribution

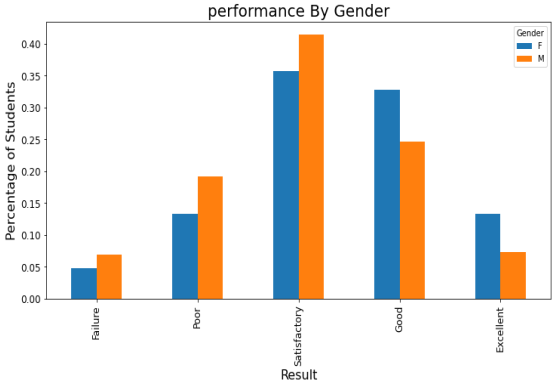

Figure 4.8(h) Performance by Gender

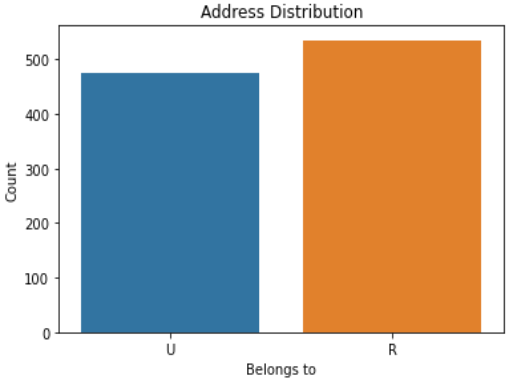

Figure 4.8 (b) Address Distribution

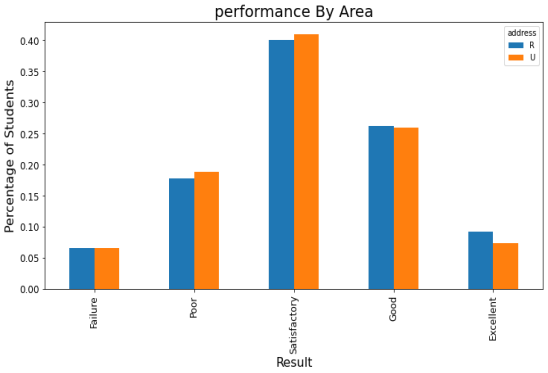

Figure 4.8 (i) Performance by Area

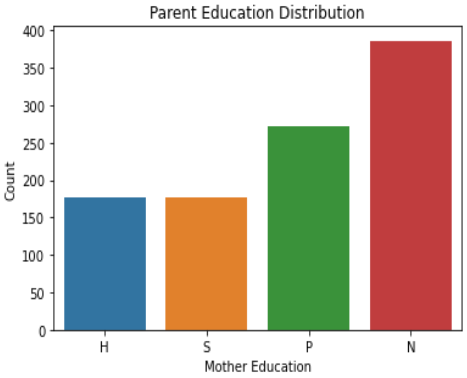

Figure 4.8 (c) Mother Education Distribution

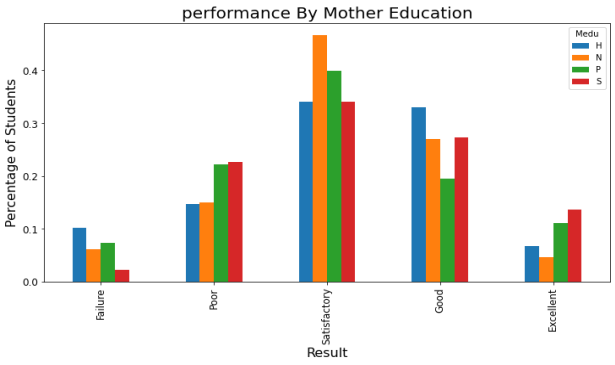

Figure 4.8 (j) Performance by Mother Education

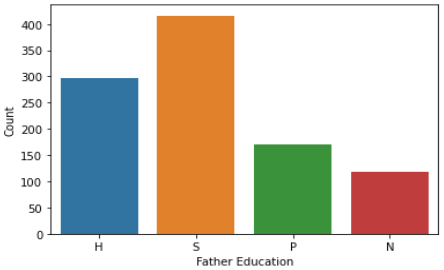

Figure 4.8 (d) Father Education Distribution

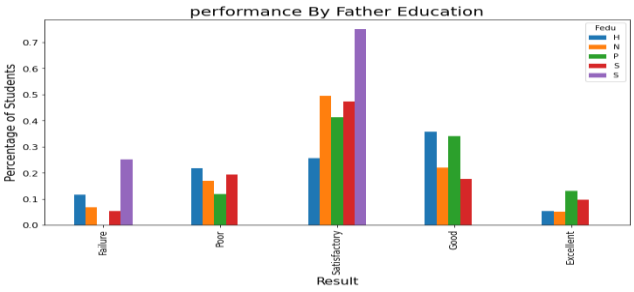

Figure 4.8 (k) Performance by Father Education

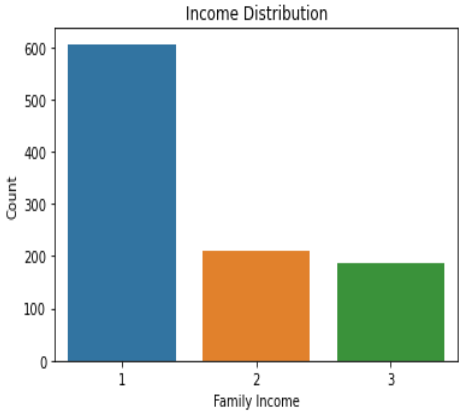

Figure 4.8 (e) Income Distribution

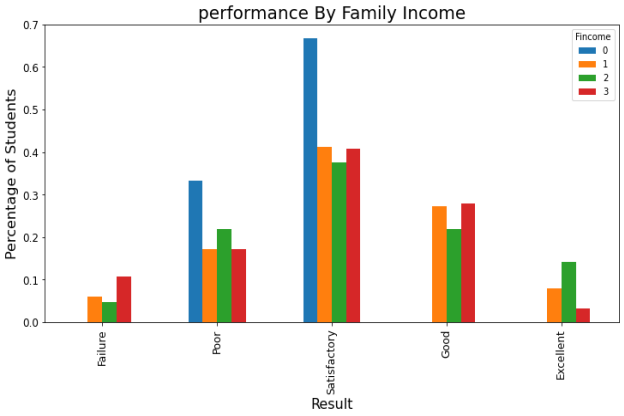

Figure 4.8(l) Performance by Family Income

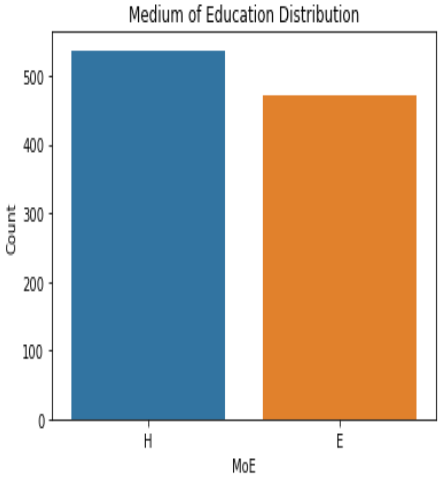

Figure 4.8 (f) Medium of Education Distribution

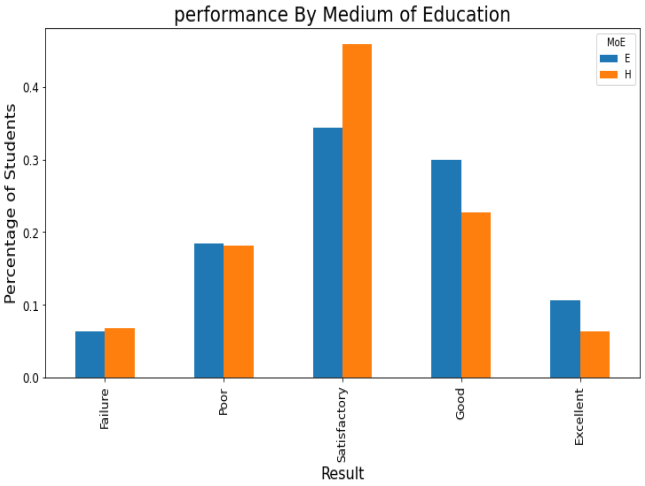

Figure 4.8 (m) Performance by Medium of Education

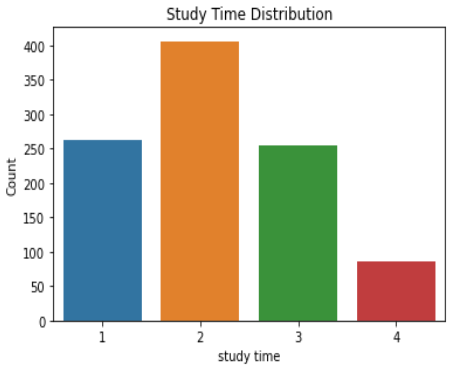

Figure 4.8 (g) Study Time Distribution

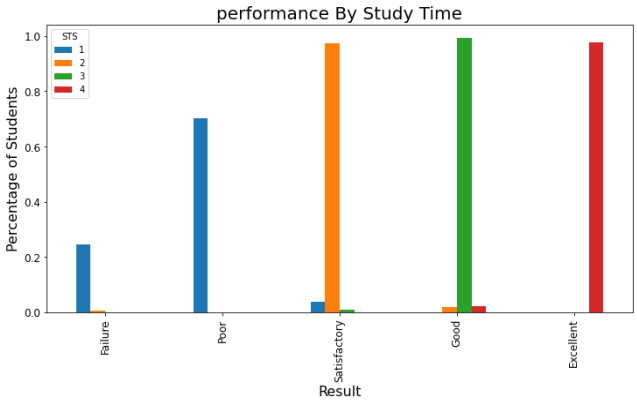

Figure 4.8 (n) Performance by Study Time

Fig 4.8 (a) shows the distribution of students by gender. Fig 4.8 (h) shows the performance, based on gender. It shows that female students performed better than the male students. Fig 4.8 (b) and Fig 4.8 (i) show that urban students give better performance than rural area students.

Parent's education is also an important attribute that affects the performance of students. Those students whose parents are educated, performed well in exams as shown in fig 4.8 (c-d) and (j-k).

Family Income has also an impact on student's performance. Higher-income group students do well in the study rather than lower-income group student,ts as given in fig 4.8 (e) and (l).

Students who studied in a different medium of education such as English and Hindi. It has also affected the performance of students. Fig 4.8 (f) and (m) are showing that English medium students got good marks in the exam than Hindi medium students. Study time attribute is also an important feature that affects the student's performance. Those students who studied 4-5 hours daily got better marks in exams rather than those who studied 1-2 hours. This analysis showed that all attributes collected in the dataset affect the performance of the students.

4.3.4 Feature Selection

Feature selection is a process in Machine Learning, where a subset of the features is selected from the dataset for better prediction. This is an important step of preprocessing for avoiding the curse of dimensionality and helps in getting a better prediction result in a real-world classification problem.

The method used to select a minimum sized feature subset, is given by Dash & Liu, (2007), with the following criteria:

- The accuracy of classification does not decrease and
- The selected features are relevant to the target variables
- Training time is less

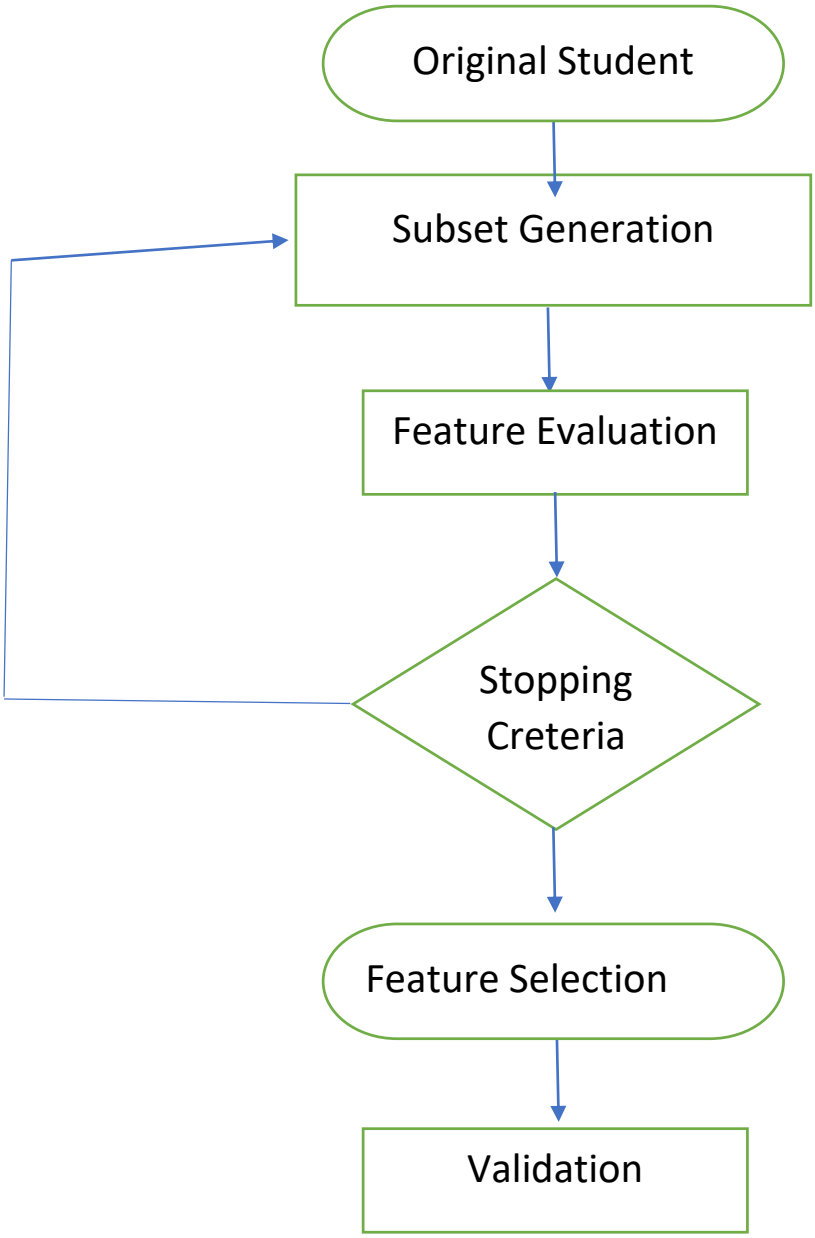

Figure 4.9 Feature Selection Process

The four basic steps as mentioned in Figure 3.9 are:

- A generation procedure used to generate the subset
- An evaluation function used to evaluate the subset
- A stopping criterion to decide when to stop
- A validation procedure to check that the subset is valid or not.

Subset Generation

This procedure is a heuristic search process with each state in search space specifying a subset for evaluation. It has a search starting point and a search strategy.

This starting point influences the search direction. There are two strategies; forward selection and backward selection. In forward selection, we start with no features and add features successively. In backward selection, we start with all attributes and then remove attributes, or start with both ends in the bidirectional selection and add or remove simultaneously.

There exist 2^N subsets and feature selection proceeds for a dataset through the evaluation function, to try to find the best one. But this is too costly for medium size feature subsets. Other methods like sequential, random and complete search may reduce computational complexity.

Subset Evaluation

The goodness of a subset is measured by an evaluation function and this value is compared with the previous one. If its value is found to be better, then it replaces the previous subset. Filter methods select features independently, while wrapper methods are used in greedy search algorithms to estimate all possible combinations of the attributes, to select the combination that produces the best result.

In both the methods, each new subset is required to be evaluated by an evaluation criterion. This criterion is divided into two groups: independent criteria and dependent criteria.

a) Independent criteria

Feature selection algorithms estimate the goodness of an attribute subset, by manipulating the inherent features of the training data, without an ML algorithm. Frequently used classes and distance measures (Liu and Motoda, 1998), dependency measures, information measures and consistency measures are used.

b) Dependent Variable: Any dependent criteria is used in the wrapper model on a classification method to perform feature selection. Then it uses algorithm performance on the subset being evaluated to find which features will be selected. This approach gives better performance, but it is computationally costly and not appropriate for all algorithms.

Stopping Criterion

The feature selection process may run exhaustively without a suitable stopping criterion. So, it must decide when to stop the process of searching for feature subsets. A feature selector may stop adding or removing features, based on the evaluation strategy when no improvement is found in the current feature subset. Some of the criteria are:

1) Search is completed

2) Reached the given boundary

3) A sufficiently good subset is selected

Results Validation

A direct way for validation of results is to measure the results using previous information about the data. If the relevant features are known, then the known set of features are compared with the selected feature subset. The information of the irrelevant or redundant features can also help for not choosing these features.

Feature Selection Models

The success of any algorithm in Machine Learning depends on the attributes of the dataset used for learning. If the data do not fit the statistical regularity, the model is will fail. It is possible to make new data from the old in such a way as to exhibit the statistical regularity, but task complexity may become tractable for the large dataset containing a greater number of features. Unlike the new input data construction, feature selection is the well-defined and fully automatic and a computationally tractable process. The benefit of feature selection includes:

- Reduction in the amount of data
- Predictive accuracy improvements
- Understandable learned knowledge
- Reduced execution time

Feature selection algorithms are divided into two broad categories: filter model and wrapper model. The filter method relies on the general features of the training data to select some attributes independently of any algorithm. The wrapper model tries to use a subset of features and train different models using them.

On the basis of inferences, we decide to add or remove features from the subset. It is based on greedy search algorithms to evaluate all possible combinations of features and select the best result. Filter Method can execute faster than wrappers. Filter does not require re-execution for different algorithms. So, this research uses only filter algorithms.

Before implementing feature selection algorithms on student data, it is important to understand the commonalities and differences among them. Feature selection classification algorithm are used for two main reasons:

- It reveals relationships between different algorithms.

- It enables the focus of the selection of a proper feature selection algorithm for a given task. For example, in the case of the classification method, the accuracy of prediction is a suitable evaluation criterion.

We describe the Filter method in detail below. They are called filter methods because they filter out irrelevant features before induction occurs.

Filter Methods

Filtered algorithms work with the generalized concept in the following form:

Let N be a given dataset with n number of attributes A_1, A_2, ……A_{n-1}, A_n. The algorithm begins the search from a given subset T_0, which may be an empty set, and searches with a specific search strategy. The generated subset T in each step is evaluated by in depended measure I and compared with the previous one. If the present subset is found to be better, it replaces the previous one. This process continues until a stopping criterion is reached as shown in Figure 3.10.

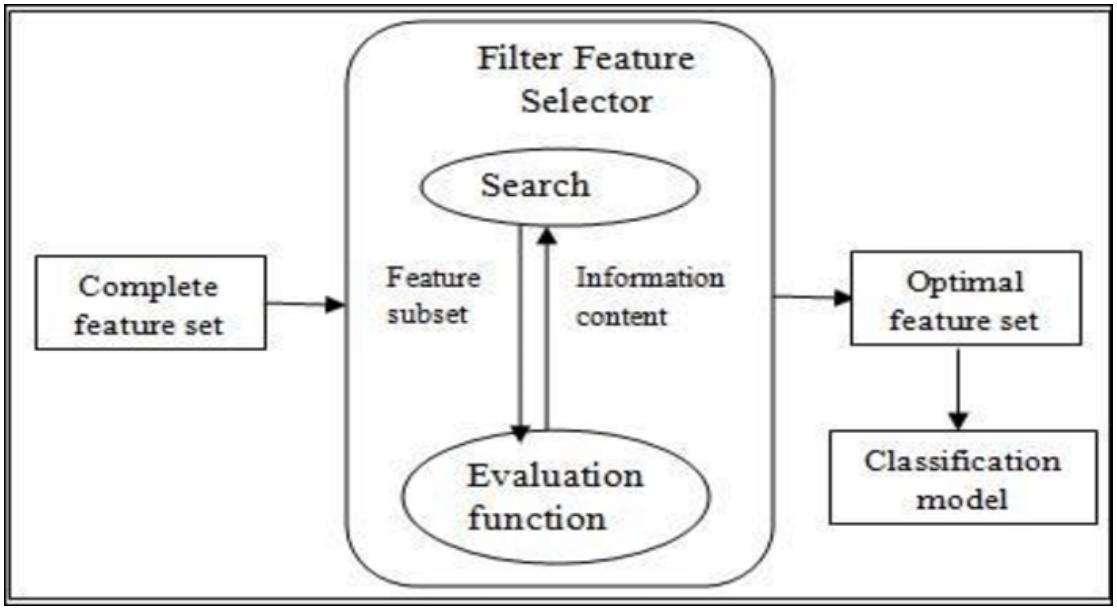

Figure 4.10 Filter method

Source: Rathinasamy, 2017

Input: N ($_1$, A_2, ……A_{n+1}, A_n)

To

Δ

Output: T_{best}

Step 1: Start

Step 2: Initialize T_{best}= T_0

Step 3: Ω_{best} = eval = (T_0, N, I)

Step 4: do Start

Step 5: T = generate(N)

Step 6: Ω = eval (T, N, I)

Step 7: if (Ω is better than Ω_{best})

Step 8: $\Omega_{best} = \Omega$

Step 9: T_{best} = T

Step 10: Stop until termination condition is reached)

Step 11: return T_{best}

Step 12: Stop

Selective Feature Selection Algorithms

Here, we consider the following filter methods:

1. Mutual Information Gain (MIG)

MIG is a type of filter method, where the purpose of the elimination is to reduce the size of the input feature set and at the same time to maintain class discriminant information for classification problems. This method is used to measure whether the information between two random variables is symmetric and non-negative, and can be zero, if and only if, the variables are independent. Reduction of the input features set can be desirable, or essential for several reasons, such as reducing the complexity of building and operating a classifier.

In addition to the computational-cost saving, a feature-space reduction can also reduce the actual cost of feature collection and preprocessing, and even lead to an improvement in classifier accuracy.

The MI is defined as the measure of dependence between random variables. It is always symmetric and non-negative.

U = (u_1, u_2, ……., u_k) and V = (v_1, v_2, ……..., v_d) is defined as

$$I(U,V) = \sum_{u}\sum_{v} p(u,v)\ \log \frac{p(u,v)}{p(u)p(v)}$$

Here (u_1, u_2, ……., u_k) and (v_1, v_2, ……..., v_d) indicate the values of the discrete variable u and v, $p(u,v)$ are a joint density function and $p(u)$ & $p(v)$ are the marginal functions.

2. Univariate (ANOVA) Test

The analysis of variance (ANOVA) can be considered as an extension of the t-test. 2 groups are compared by an independent t-test. The given test, in which a distribution population with same standard deviation are equal, is called a hypothesis. An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. The best-known F-test plays an important role in the ANOVA.

This method is used to analyse group differences in samples, using statistical models and their associated estimation procedures (variance among and between groups). Ronald Fisher (statistician and evolutionary biologist) developed the ANOVA. This method is built on the law of total variance, where the variance observed in a particular

variable is divided into components attributable to different sources of variation. In other words, it applied a statistical test of whether two or more population means are equal, and hence generalizes the t-test beyond two means. The factors of the total deviation are compared by F-test. For example, in one-way, or single-factor ANOVA, a comparison of F test statistic is used to test statistical significance.

$$F = \frac{variability\ between\ groups}{variability\ with\ in\ groups}$$

3. Univariate ROC_AUC Test (ROC)

Feature Selection is based on Univariate ROC_AUC for Classification and MSE for Regression. The Receiver Operator Characteristic (ROC) curve is used for the evaluation of classification performance. The ROC curve has been used as a popular metric for the evaluation of ML models due to its superiority in dealing with cost-sensitive and imbalanced data. This approach for feature selection is simple and effective in the evaluation of individual attributes. The ROC and AUC (area under the ROC curve) have been generally used to find the accuracy of classification methods in supervised learning. Through two-dimensional graph analysis, it is difficult to compare two curves of ROC directly. The AUC, denoted as a quantitative measure, provides a good summary for investigating ROC curves.

4. Fisher Score and Chi2 Test (CHI)

Fisher score is used as a supervised feature selection method. However, it selects each feature independently, based on the Fisher criterion according to their scores, which generates the subset of features. A chi-squared test is also written as $\chi 2$ test. It is a statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution. This method is used to find if there is a significant difference between the expected frequencies and the frequencies observed in one or more categories. Chi-square test measures dependence between stochastic variables.

4.3.5 Jupyter Notebook (ML Tool)

Jupyter Notebook is a python-based tool for ML. It offers different functionalities grouped into the categories of Data visualization, evaluation, classification, clustering and unsupervised learning. This tool supports different tasks such as read csv or other format file, discretize, transformation, preprocessing, train classifier models and others.

Python is a popular language used for the research and development. It is a huge language with several modules, packages and libraries that provide different ways of achieving a task.

Python and its libraries like NumPy, SciPy, Scikit-Learn, Matplotlib are used in data science and data analysis. They are also broadly used for creating scalable machine learning algorithms. Python implements machine learning techniques such as Classification, Regression, Recommendation and Clustering.

There are a number of libraries and packages generally used in performing various machine learning tasks as listed below −

- NumPy − used for its N-dimensional array objects.
- pandas − data analysis library that includes data frames.
- matplotlib − 2D plotting library for creating graphs and plots.
- scikit-learn − the algorithms used for data analysis and data mining tasks.
- seaborn − data visualization library based on matplotlib.

This study used the library *sklearn* to pre-process data, as mentioned earlier. Jupyter Notebook was chosen as a tool for this study due to its ease-of-use and data visualization characteristics. Through this interactive tool, following tasks can be performed:

- Data pre-processing by using different functions and libraries
- Applying different classifier models
- Compare the accuracy of different ML techniques
- Show and compare results of applied ML techniques

4.3.6 Applying various classifiers

In this subsection, diffeent ML classifier techniques are discussed, to find the best classifier model. The reason for selecting classifiers is that they have been frequently used, as seen in the Literature review in the previous Chapter. Additionally, different studies use different classification methods. it is not defined which is the best classification method for a particular situation. We cannot say which classifier performs better than others in all situations (Asif et al. 2014). Therefore, it is essential to select the different classifier models and determine which one performs best.

In this study, an investigation was done to find out the best classifier that gives the best prediction performance using the student dataset. 70% of the dataset was used for training the classifier models and 30% was used as testing data. The following classifier models were built, for determining the best classifier model: Logistics Regression, KNN, Decision Tree, Naïve Bayes, Support Vector Machine, Random Forest and Gradient Boosting, AdaBoost and Bagged DT.

4.3.7 Evaluation of best classifier

Evaluation is an essential part of the building of models and feature selection. This study used 10-fold cross-validation to evaluate the classifier models. In cross-validation, the dataset was divided into 10 different subsets by random splitting. Nine parts were used for training and one portion for testing. Confusion matrix was made from 10 experiment results. This matrix was used to determine the performance of classifier models (Sen et al., 2012).

Classifier Performance

We construct prediction models for student potential prediction using various classification algorithms and evaluate their accuracy and other performance measures. This also includes the effect of different feature subsets performed by various selection algorithms. Classifier performance is used to measure the goodness of the classifiers.

Classifier performance measures

This section describes the metrics that were used to evaluate classifier models. These metrics are calculated from the confusion matrix given in Table 4.2. The classifier model performance is evaluated based on the cross-validation instances, these are correctly classified and incorrectly classified by the model in the confusion matrix (Asif et al., 2014). The confusion matrix is used for the analysis of the model's performance.

Five common measures are used in this research: Accuracy, Precision, Sensitivity, Specificity, and F-measure. Model accuracy is the total number of correct classifications out of the total number of classifications done.

Predicted Label			
Positive			Negative
Actual Label	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 4.2 Confusion Matrix

Some definitions are given below for calculating the metrics used to evaluate the different classifier models.

Accuracy

The accuracy of classification algorithms is one way to measure how often the algorithm classifies a data point correctly. This measure is calculated using the confusion matrix given in Table 4.2. It provides the overall performance of the classifier as given below.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (4.1)$$

Precision

It is a measure of the correctness achieved in true prediction i.e., of the observations labelled as true, how many are labeled true. It calculates the positive predictive value as given below (Thai-Nghe et al., 2009).

$$Precision = \frac{TP}{TP + FP} \quad (4.2)$$

Recall (Sensitivity)

It is a measure of the actual observations which are predicted correctly i.e., how many observations of the true class are labeled correctly. It is also known as 'Sensitivity'. It is the measure of the true positives which are correctly classified as defined below (Sokolova and Lapalme, 2009)

$$Sensitivity = \frac{TP}{TP + FN} \quad (4.3)$$

Specificity

It is a measure of how many observations of the false class are labeled correctly. It calculates the proportion of negatives that are correctly classified as shown below:

$$Specificity = \frac{TN}{FP + TN} \quad (4.4)$$

F-Measure

It combines both precision and recall to obtain an average value that is balanced. It measures the harmonic mean of the classifier model's precision and recall and is shown below in Equation 4.5.

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4.5)$$

4.4 Chapter Summary

This chapter discusses the methodology of the research for achieving the objectives of the study. It describes the six-step framework followed for the analysis of student potential and performance in technical programs. The phases explained are problem understanding, data understanding – data collection, data preparation- data cleaning, transformation and feature selection, building models and evaluation.

This chapter also included an overview of feature selection and techniques used in the literature. The benefit of feature selection is that it reduces the dimensionality of the student data. This is helpful for classifier models to produce better results with high predictive accuracy and operate fast. The effect of feature selection was studied in terms of filter methods on the student data and their results. Therefore, this study presented the methodology for the ML phase.

The next chapter presents the experimental work in which different ML techniques have been applied to the student dataset and evaluate the performance of classifier models.