CHAPTER 1
INTRODUCTION

1.1 Introduction

In recent years, Internet usage has increased greatly, resulting in the availability of large amounts of data, also called 'Big Data'. This data comes primarily from constant use of social media, and also from Internet of Things systems and other day-to-day digital transactions. But most of this data is useless and contains no useful information. Machine Learning techniques are used to extract useful information from big data to be able to analyse it purposefully.

Machine Learning (ML) is a subfield of Artificial Intelligence. This is a study from examples and experience, without an explicit computer program. It uses data in a generic algorithm, instead of a writing program, and it develops logic based on the given data. It focuses on the development of computer programs that can access data and use learning to achieve the relevant task (Mitchell T. M., 1997).

Arthur Samuel defined ML as "Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.'' Recently, Tom Mitchell of Carnegie Mellon University, went further into the engineering concept and said that "A computer program is said to learn from experience E concerning a task T and some performance measure P, if its performance on T, as measured by P, improves with experience E" (Burns Ed).

It should be kept in mind that problems that cannot be solved by numerical means, only ML can solve. Furthermore, the aim of ML is not to guess correctly, because it belongs to domains where there is no such thing. The aim is to make target estimates to be useful (N Mccrea).

The basic idea of ML is to make algorithms that can take input data and use statistical analysis methods to predict output data. ML concepts are based on the principles of statistics. Training samples that are used with machine learning techniques should be sufficient and random.

ML is divided into two categories: supervised and unsupervised, the way humans learn in both supervised or unsupervised ways. A child learns many things in life through his or her experience, without being told. Similarly, a research scholar studies and discovers new

thoughts. These are examples of unsupervised learning. The learning that humans learn from their parents, teachers, elders or friends, is called supervised learning.

Regression and Classification are two kinds of Machine learning algorithms. Classification algorithms can be both individual or their combinations. Ensemble methods are the use of supervised learning algorithms that combine two or more classifiers into a Meta estimator, by taking the voting or averaging of their prediction, for the final estimation. The key objective of using Ensemble methods is to build a model of ensembles that integrates a combination of varied individual classifier techniques with good accuracy. Ensemble methods help to improve predictive results by combining several models. These approaches allow the production of better forecasting performance than a single model.

This study will discuss Classification techniques and Ensemble Methods in detail in Chapter 2.

1.2 Introduction to academics and student performance

Students are the most important component of the education sector, as the progress and status of any University or Institution is based on its students' academic performance. Usasmah et al. (2013) stated that, through learning assessment and extra-curricular activities, student performance can be measured.

Machine Learning algorithms are widely used for predicting educational data. The academic performance of students has been of interest by the researchers and institutes. Mainly GPA is used for defining the student performance for their success. A detailed investigation of student performance is given in Chapter 2. In addition, predicting student potential in technical programs has been an important issue of investigation, since it may help the students identify their strengths in different areas of study.

Digitization has had an immense impact on the working of the education system and the career planning of students. Generally, career counsellors / experts play an important role in evaluating and assisting students in appropriate career selection and planning. These conventional methods and practices are not very impactful and after a point, they prove inefficient and ineffective. Today, many institutions and educational bodies have started using advanced automated solutions, based on Artificial Intelligence.

Automation of the counselling system saves efforts as well as time and has the potential to reach a large and diverse group of people.

This study will explore various technologies that can be used to provide career guidance and counselling to students. The study uses real-time students' data and considers different attributes to find out, using Machine Learning techniques, as to which factors play a major role in choosing a career by students,

1.3 Education Research

Education is very important for human beings and ensures development. Educational research provides benefits to the end-users. The major aim of this research is to investigate behavioural and socio-economic attributes of students, teachers and other participants in any educational system. Educational research uses different techniques which can be categorized into two methods. Both methods are used in different fields, though these methods are different from each other. We discuss these methods in the following section.

1.3.1 Quantitative Research

This method is used to collect a large volume of data, dealing with numbers, which are measurable and finding meaningful information and relationship patterns between them or for their classification. This method aims to develop and employ models or theories. This is used for the investigation of mathematical expressions of quantitative relationships.

Different ML techniques are used to discover meaningful new correlations patterns and trends by using large amounts of data stored. In this, the most investigating aspect is that the complex statistics is represented through visualization on charts that convey a large amount of information easily. The attributes for students' or teachers' information include demographic data, family data, health, academic and socio-economic attributes.

1.3.2 Qualitative Research

This is a free- form technique used to gather information about the problem. It is based on the opinions of people, such as feelings, motivation and perceptions. Hence, the information is gathered in the textual form rather than number or category. Case studies are considered qualitative research.

This method investigates the 'whys' and 'how' of decision making, not the 'what' and 'when'. This method needs small but focused data rather than large samples. But it is

expensive and time consuming for its execution. Many fields, especially social science, use this method.

1.4 Education System in India

Education is a method of communicating or transmitting knowledge through guidance, teaching, or study. It gives skill to everyone to become self-confident, self-sufficient, independent, and able to face challenges in life. It increases the ability to manage daily life problems, improve childcare, and be ready for the future. It allows people to prepare for the upcoming challenges, seek better jobs, and succeed in their lives.

There are many levels of educational study in India. This is shown in Figure 1.1 below.

1.4.1 Levels of school education in Rajasthan

The educational format in Rajasthan is of the Ten+ Two+ Three Pattern. School education is divided into two parts, the first ten years studying general education, followed by two years of higher secondary education. General education consists of primary and secondary education. After middle school, a student studies for two years to complete ten years, for taking a secondary school certificate (SSC). After that the student opts for subjects according to his/her choice and completes two more years of education, for the award of higher secondary education (HSC). In the +3 stage, the student goes for higher education in his/her chosen area in a professional (technical) or non-professional course.

The present study aims to correctly predict the student's potential and performance for choosing a program, which may be either technical or non-technical, based on his/her background, which is captured in a record of attributes. We discuss these in the next section.

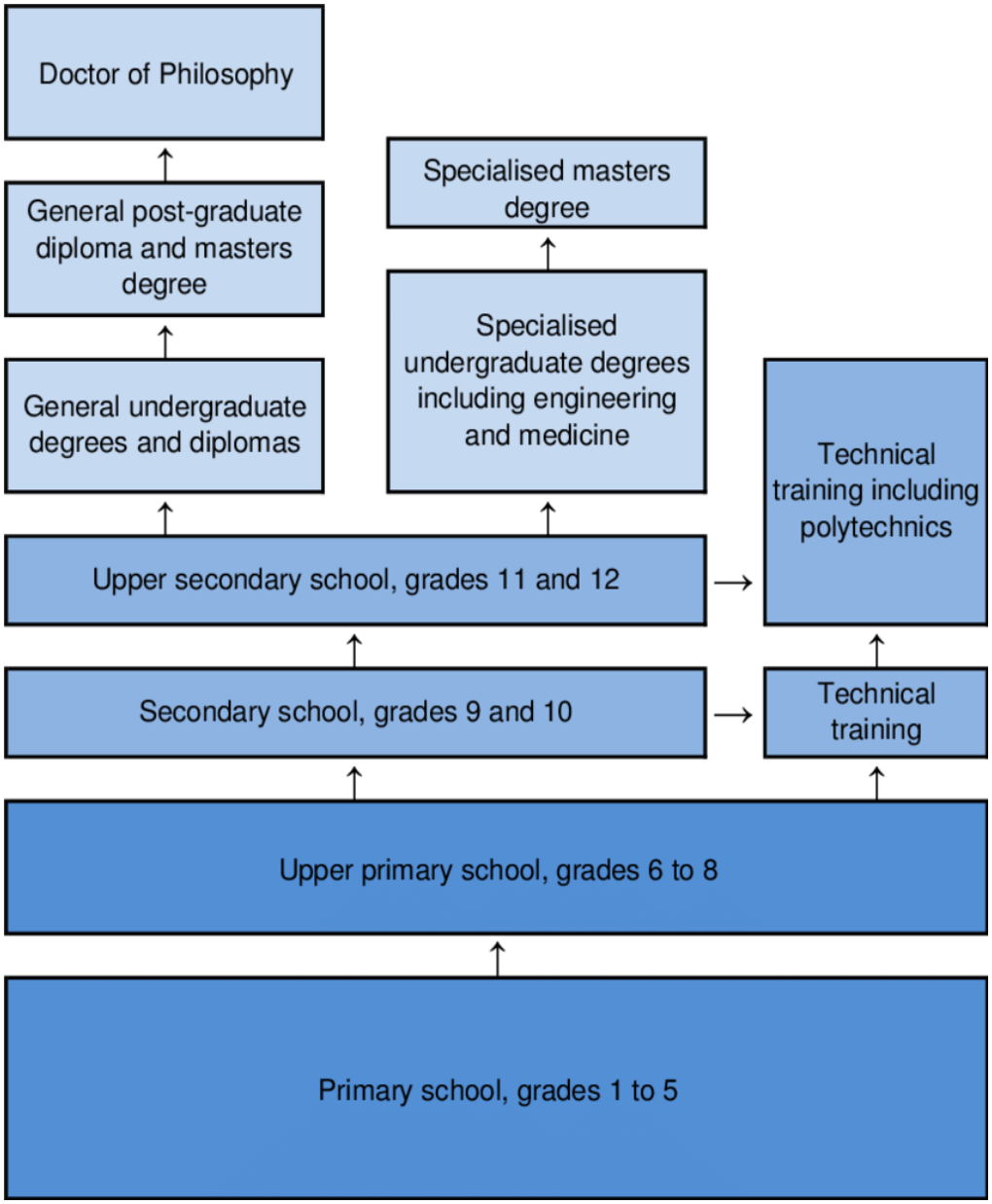

Figure 1.1 Education System in India

Source: Hill and Chaulax, 2011

1.4.2 Factors that influence Student Performance

The main aim of education is for a student to succeed in his/her life and better one's academic achievements. In recent years, many kinds of research have been conducted to find the factors which influence (positively or negatively) student potential and performance. It has been reported that the analysis of student academic potential and performance is very challenging.

This is because a student's academic potential and performance is the product of demographic, socio-economic and academic factors. Hence in this thesis, we report the analysis of student potential and performance based on these three set of attributes. We consider each of these attributes in turn below.

1.4.2.1 Demographic Attributes

Demographic attributes are related to a student's personal information, such as gender, age, sibling structure, medium of Education and location.

Gender is an important factor for analysis. Many researchers have used this factor to demonstrate the pattern of education in boys and girls (Tho, 2004). One research showed that gender is the most important factor in analysis of academic performance. This affects the student degree of motivation and also academic performance.

Sibling Structure shows the number of children in the family and a child's position in the family. It is a less significant factor but it also affects the performance of the child in his/her education.

Medium of Education is another important factor for investigation of the performance of the student. In India, students who have studied in the English Medium are seen to generally take admission in technical programs, as compared to those who have studied in their regional or mother tongue, who mostly prefer to join non-technical programs.

1.4.2.2 Socio-Economic Attributes

Socio-economic factors include information about the family's educational level, parental occupation, interest and social status in the community. High socioeconomic families have more chances to succeed in preparing the child for admission in technical programs because they have enough resources to support child development. They provide a high quality environment, care and books, for learning at home.

1.4.2.3 Academic Attributes

These factors influence directly the student's academic performance at the higher secondary level, which includes type of secondary education, Board of higher education, stream, study time and marks obtained at the secondary and higher secondary level. These factors are responsible for the final academic performance of students.

Thus, the prediction of students potential and academic performance is based on attributes like personal, socio-economic, academic and other environmental factors. Every year thousands of students pass out of the higher secondary education system and compete for getting into the desired course at college level. This necessitates the use of prediction models to analyse the potential of students and to assist them for choosing the correct program, according to their ability and background.

1.5 Motivation for this Study

Around 15-18 Lakh students complete their higher secondary education from Rajasthan Board of Secondary Education (RBSE) and/or Central Board of Secondary Education (CBSE) in Rajasthan every year. Data about these students is collected by Colleges and Coaching Centres, to counsel them and prepare them for admission to higher education institutions.

This study is concerned with the factors that influence the student potential for taking admission in technical programs. Conventional techniques for analysis of student potential and performance in the technical domain are currently used for this purpose. However, these conventional methods of data analysis are very lengthy and most of the steps involved are manual. On the other hand, Machine Learning (ML) techniques are faster and most of the steps involved in the process can be automated by programming the algorithms for machine learning, hence it is less time-consuming. So, machine learning techniques will be used for this research.

A survey was conducted at the start of the study for understanding the concept and problem. The focus of the survey was to take views of the stakeholders in Rajasthan. The survey results showed that the problem has existed for many years. These findings motivated us to develop an ML model that could classify the students according to their ability to take up technical or non-technical programs. The survey results are presented next.

1.5.1 Survey Results

This survey was conducted with admission counsellors, professors and parents of children who took admission in both technical and non-technical programs. The goal was to understand the problem and the participant’s perspective.

A total of 18 participants took part in this survey, 5 admission counsellors and 13 professors. The admission counsellors guided us about the factors that help to analyse the student potential and performance, because they are aware of the problems faced by students. The professors are more informed about the students' academic performance. All the participants had experience of more than 5 years in their field. They identified the problems that have existed for the past many years in this area.

They suggested that all the education stakeholders have to work together to resolve this problem. Some attributes like medium of education, Family Income, Student Interest and Education Board are important and greatly affect and influence the student potential and performance in higher education.

1.6 Problem Statement

In the last few years, researchers have started to use ML techniques to help students and administrators to improve the education system. ML techniques have been applied to predict the student potential in technical programs and also analyze the performance of students studying in different programs, for their evaluation and grading. The aim of the study was to identify the best classifier model for analysing the collected data. It also selects the optimal feature subset from the total attributes.

This study was conducted in Rajasthan. The scope of the study was students passing 12^{th} standard from Rajasthan schools and also those studying in technical programs in Rajasthan. The data was collected in Jaipur Region, Rajasthan.

Previous studies were concerned with student's performance prediction but no study was done on analysing the potential of a student in technical programs and no such study has been done to analyse the student potential, based on students studying in the different Education Boards permitted in Rajasthan.

The problem at hand can thus be stated as follows:

"To study and analyse students' data using ML techniques, to predict the students' suitability for admission to different technical programs. Also, to evaluate the performance of students enrolled in these programs, based on their Education Boards, using appropriate ML techniques".

1.7 Objectives of the Study

This research aims to analyse and predict student potential using Machine Learning techniques. Prediction of potential can help the students to decide for choosing the best programs to further their career. There are some factors like demographic, academic, and socio-economic factors that affect the student's potential for choosing technical or non-technical programs. The purpose of this research is to find the main factors that influence student potential. The main objectives of the research are enlisted below.

1) To study and implement different Machine Learning techniques using Python and select the most appropriate techniques for our use.

2) To analyze data of students eligible for technical programs and predict their potential for admission to different programs.

3) To analyze the performance of students studying in different programs for their evaluation and grading.

4) To use this dataset to predict the student potential, based on the different Secondary Education Boards of the students.

1.8 Research Design

The study uses the (CRISP-DM) model to accomplish the system design and help to complete the objectives of the study. Figure 1.2 shows the complete research design.

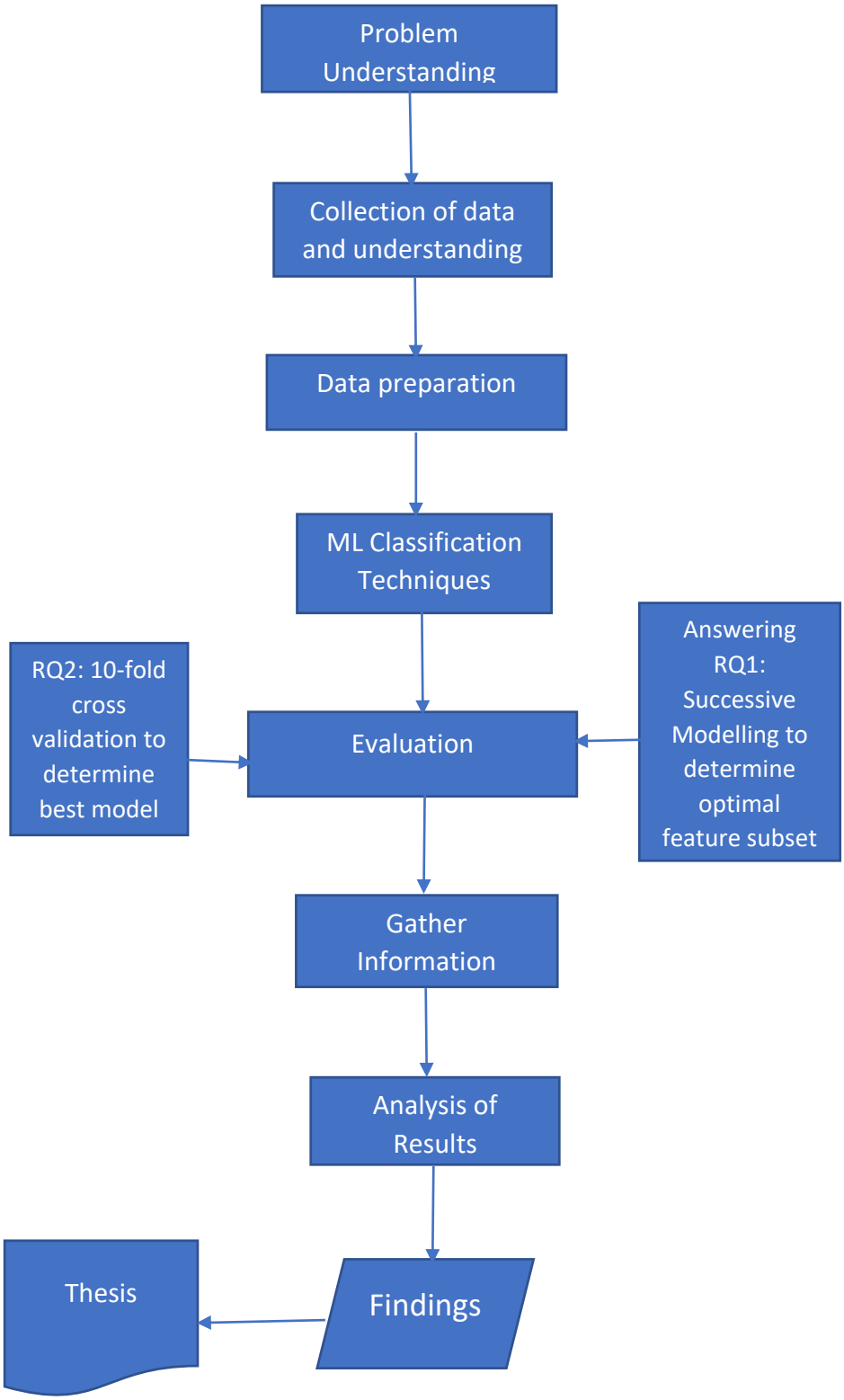

Figure 1.2 Research Design

The steps which are followed by this process: problem understanding, data collection and understanding, data preparation, ML techniques, evaluation and using discovered knowledge (Kungan and Musilek, 2006)

The study aimed to the analysis of student potential and performance in technical programs. The required data attributes were identified through survey and literature.

After that the data was collected through Forms stored in an Excel worksheet to make meaningful datasets.

In the data processing step, data was cleaned and transformed into a format which was used for classifier modelling. Then, four filter methods were used for feature selection for better predictive accuracy. Six common classifiers were selected for analysis of student potential and performance.

The best model was found through 10-fold cross-validation criteria evaluation. Five metrics were used to validate the results. These are Accuracy, Precision, Sensitivity, Specificity and F-measure.

1.9 Organization of the Thesis

The Thesis is organized into seven chapters, including this introductory chapter.

Chapter 2 reviews key concepts of Machine Learning, factors that affect students in Higher Education, different studies related to academic performance that applied Machine Learning prediction techniques, and to draw decisions between different attributes and their effect on students' academic performance.

Chapter 3: This chapter describes the methodology of the research work used to achieve the objectives of this study and discuss the proposed solutions to handle the problems stated above. It includes procedures for data collection and data pre-processing, determination of the optimal feature subset and finding of the best classifier. It also includes an overview of feature selection in terms of filter methods and also applied filter methods to extract optimal feature subset.

Chapter 4 consists of the experimental and analysis work, including preparing data and applying Machine Learning techniques and Ensemble Methods on the data. It discusses the different classifiers on the cleaned dataset for the prediction of student potential and

performance in terms of predictive accuracy, precision, sensitivity, specificity and f-measure.

Chapter 5 shows the results and findings of different classification and Ensemble Methods on the Higher Education data, in order to find the best classifier for predicting the potential of students for study in technical programs.

Chapter 6: This chapter discusses the results of the application of different base and ensemble ML methods on the student dataset, to predict the performance of students studying in a technical program. It also shows how the target variable is divided into three categories to show the optimum result for each.

Chapter 7 is the concluding chapter of the thesis. It discusses the contributions made by the research work and suggests future work related to this area.