CHAPTER 1

INTRODUCTION

1.1 Background

Vast amount of data are being captured and stored regularly in the institutions of learning, but sometimes, construction of predictive model from this massively stored data does not give the desired outputs in term of accuracy. This may be as a result of the model having the problem of over-fitting. A mechanism should, therefore, be put in place to strengthen the technique used in modelling, to pave way for the building of an accurate predictive model from students’ data. This task is vital as information is inevitable for efficient planning and there is need to process data in order to have information.

Predictive modelling of students’ data is a key task in Educational Data Mining (EDM). The EDM is a relatively new field of scientific study. Although, for quite a long time, researchers have been capturing and analysing various data that emanate from educational domain. However, only recently has EDM been established as a field in its own right (Scheuer & McLaren, 2012). The EDM is concerned with the development of models to reveal the uniqueness inherent in the data from the educational domain.

The process of EDM transforms the raw data within the education system into valuable information that could make a great potential impact on research and practice (Romero & Ventura, 2010). The focus of this research is on enhancement of the Feed-forward Neural Networks (FNN) technique and the development of a rule-based
algorithm that is capable of unveiling useful knowledge from students’ data. The word knowledge that is commonly used in data mining has been clarified in the literature. It is not really the kind of knowledge that we carry in our heads, but just a choice of word to refer to the structures that learning methods produce (Witten et al., 2011).

What informed the decision to explore educational data was as a result of the poor academic performance of several newly enrolled undergraduate students in recent times as observed in Nigerian Universities. The review of the literature also shows that, the rate at which students are withdrawn from the university is higher at the early stage of their study. This set of newly enrolled students do not have any continuous assessment or past semester results that can be modelled to predict their future performance. The present research therefore, explored their past academic and demographic achievements.

Identification of predictors of relevant influence has been through the domain experts and findings from the literature. These data were modelled in conformity to the data mining process. Some rationales for seeking information about students at the early stage of their study, is to enable the teachers adopt a teaching methodology that can be of immense benefit to all students irrespective of their learning challenges. It would also help to identify student whose performance is outstanding but in serious need of support. Also, those found to be academically weak could be rightly guided by the counselling unit of the institution by providing them with further assistance that can be tailored directly to their needs.

A good understanding of the students' pattern of enrollment, retention and the rate of dropouts are the key research focus of EDM. The capacity of institutional administrators to clarifies the reason for students drop out, what is responsible for some students that over stay in their study, and so on, is less critical than the ability to accurately predict an instance of such events. The EDM is an emerging interdisciplinary research area, that mainly has to so do with the development of techniques that is suitable for the exploration of dataset that originates from an educational context (Romero & Ventura, 2010).
Exploration of educational data requires building a useful descriptive or predictive model that can be used to describe or predict the hidden information that resides in the data. However, the building of the predictive model from the students’ data is the concern of this research. In addition to the use of an enhanced FNN for this task, further focus was to develop a rule-based algorithm that can achieve the resulting model of high accuracy and capable of generalizing well.

Although, with machine learning techniques, achieving a total accuracy is not usually feasible and according to (Negnevitsky, 2011), the issue of accuracy using machine learning such as neural networks is sometime affected by a number of factors. At times, setting of the parameters may be responsible for inconsistent outputs.

One of the main objectives of educational institutions has been identified as to provide high quality education to students (Sundar, 2013). In view of this, a huge budget is usually set aside in most countries to achieve this goal; as the rate of growth of a country can be measured by the quality of its education system (Mishra et al., 2014). In order to achieve the highest level of quality in the institutions of learning, there is a need to discover knowledge from the students’ data.

Several ways have been identified by which the quality of education can be improved, the study proposed in (Baradwaj & Pal, 2011) identified alienation of the traditional classroom teaching model, detection of ambiguities in the result sheets of the students and prediction of students’ performance at the appropriate time. The embedded knowledge in the educational data set are extractable, this can be achieved by systematic exploration through the data mining techniques.

Data mining is one of the important research fields in computer science and it has been broadly defined as the process of discovering novel and interesting patterns in large amount of data (Jukic et al., 2014). It can also be seen as an area of scientific inquiry that revolves around the development of a series of techniques for making discoveries from educational settings and using those techniques to better understand the students (Baker, 2010).