CHAPTER 7

CONCLUSION AND SUGGESTIONS FOR FURTHER WORK

7.1 Introduction

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. In this research, **i**t is shown that Machine Learning (ML) techniques can be implemented successfully in the education sector. The study has focused on ML techniques and their applications, based on the CRISP-DM process, to predict the student potential for, and performance in, technical programs.

This chapter summarizes the work done in the thesis, while discussing the findings of the research and its contributions, in section 7.2. In section 7.3 it recommends how this research work can be further extended in future.

7.2 Summary and Contributions of Work Done

Predicting student potential and performance has been an important area of the research. Knowing their future academic performance may help students assess their strengths and allow them to choose their careers in their most productive areas. This thesis contributes towards the prediction of student potential and performance, based on actual data collected, using different ML techniques.

Chapter 2 explained different ML techniques which can be used for predicting the student potential and performance. This research used various ML classification algorithms for prediction and analysis. The history of ML, starting with its definition, was also discussed.

Chapter 3 studied and investigated the previous work done by researchers in this area, using different ML techniques, in the higher education sector. It discussed the models and algorithms used to find the relationship between collected student data and the performance of the students.

Chapter 4 explained the research methodology used in this research. First, the dataset collected from various Universities and Institutions of higher learning, using 20

different attributes and their types, was characterized into 3 main parts: student dataset for potential analysis, student datasets based on two different Education Boards and student dataset for student performance. These datasets contained the student's demographic, socio-economic and academic attributes.

Then, data pre-processing was done. The datasets were cleaned, filtered and transformed because this helps to understand the data and give better predictive accuracy when used by different ML techniques.

In chapter 5, we reported on the results of applying the ML techniques on the student dataset for prediction of student potential for technical programs. The dataset was divided such that 70% of the data was used for training and 30% for testing. The six classifiers (KNN, NB, DT, SVM, RF and GB) were trained on the data of students who completed their 12^{th} and took admission in technical and non-technical programs.

A comparison was made of the performance of the six classifier models based on their prediction of student potential, using five metrics, viz., accuracy, precision, sensitivity, specificity and f-measure. The result showed that Gradient Boosting (GB) was the best classifier according to the highest correctly classified 1526 records. GB achieved the highest accuracy of 91.3%, the best precision of 88.3%, the best specificity of 88.6%, the best sensitivity of 94.1% and the best f-measure of 91.1%. It also misclassified the lowest number of records.

Next, feature selection filter methods were applied to find the most important features necessary to predict student potential. Features were ranked using four algorithms: Mutual Information Gain (MIG), Univariate (ANOVA) Test, Univariate ROC_AUC Test and Fisher Score & Chi2 Test. The impact of Filter methods resulted in a marginal increase in the accuracy of the ANOVA and ROC classifiers, when compared to using them with the Full Feature Subset (FFS). The other classifiers underperformed on the filter subsets. For the analysis of student potential, eight features were identified as the most predictive: StIntforJob, Stream, 10^{th}, WhythisCourse, Fincome, Fedu, Twelth and ST. These were ranked the best features, in two out of four algorithms.

These optimal feature subsets help education stakeholders to guide the students for taking admission according to their potential. The findings show that StIntforJob was the most important feature for the analysis of student potential for taking admission in technical programs. They also show that predictive accuracy achieved through the

optimal feature subset using classifier models, was different from classifiers that used full feature selection.

ML techniques were also applied to the analysis of student potential, based on two different education Boards, i.e., RBSE and CBSE. The whole dataset was divided into two separate datasets, one each for the different Boards. These datasets were again split into training and test datasets in the ratio 70:30, with 10-fold cross-validation for estimating the generalization performance. The results indicated that Bagged DT is marginally better than Gradient Boost and gives better accuracy. This study agrees with previous studies that no one classifier performs better in different scenarios and datasets (Asif et al., 2014).

The above study will benefit students in choosing the most appropriate program for study according to their potential. It will also benefit higher education institutions in counselling their students for their future careers. It will be useful for the application of appropriate machine learning techniques in the education field for the evaluation and grading of students, based on their academic, behavioural and financial data.

In chapter 6, Different ML techniques were applied to the student datasets to predict students' performance while studying in technical programs. Six classifier models were used to build the student performance prediction models under three different cases, by varying class variable values. In the first case, the target variable was divided into two labels: fail and pass. In the second case, the target variable was classified into three labels: fail, good and very good and in the third case, five class labels were used: fail, poor, satisfactory, very good and excellent.

Different classifiers were chosen in this research work and comparative analysis of their performance was done using Python. Students' Dataset was pre-processed and later fed to Decision Tree, Random Forest, Support Vector Machine, Logistics Regression, Ada Boosting, and Stochastic Gradient Boosting ML models. Training and testing were done in ten different folds resulting in an accurate Model. The obtained results were measured in terms of Accuracy.

For the two labels class case, SGB classifier method achieved predictive accuracy of 99% while DT achieved lowest accuracy of 98%. The accuracy ranged from 99 to 98%. In the three class labels case, the highest accuracy of 98.86% was achieved by Random Forest. The accuracy marginally come down to the range from 97% to 94%. For the

five class labels case, Random Forest gives higher accuracy of 96%. The accuracy of this case ranges from 96 to 81%.

Thus, it was found that the value of predictive accuracy is highly dependent on the different classes. Accuracy of the two-labels case was greater than the multi-labels case. The accuracy of classifiers decreased by increasing the class labels. No single classifier was found to be the best model for student's performance prediction. Accuracy depends on the problem and collected dataset and the attributes which are used for analysis.

The results of the above study will help the students and their teachers to correctly predict their performance in a particular program before-hand. This will be helpful to students to choose which courses to take in their program of study and to excel in them. It will also help recruiter companies to have a better idea of the student candidates they are planning to hire. This will help educational administrators and policymakers working in this sector in the development of new policies on higher education.

7.3 Contributions of the research work

1. Modifying the educational data pre-processing part to improve the quality of the model outcome.
2. In this thesis, the datasets were cleaned, filtered and transformed because this helps to understand the data and give better predictive accuracy when used by different ML techniques.
3. This research has also contributed towards predicting the student potential and performance, considering more accurate academic data using ML techniques.
4. Feature selection filter methods were applied to find the most important features necessary to predict student potential.
5. The research will benefit students in choosing the most appropriate program of study, according to their potential.
6. It will also benefit higher education institutions in counselling their students for their future careers. The management can improve their strategies and the quality of education using such knowledge. This will help educational administrators and policymakers working in this sector in the development of new policies on higher education that are related to students' future careers.
7. It will be useful for the application of appropriate machine learning techniques in the education field for the evaluation and grading of students in all programs.

7.4 Suggestions for further work

In this section we highlight some important aspects of the future work that can be done in this important area.

1. Current research has focused on some Machine Learning techniques to predict student potential and performance in technical programs. The use of Machine Learning techniques allows us to delete noisy data and find strong correlation patterns between selected attributes. Many other ML techniques and their hybrids can be tested on the datasets, so that more accurate results may be got.
2. The dataset used in this study was collected from different universities and institutions of Jaipur, Rajasthan only. This limitation can be removed and data collected from many other districts, not only in Rajasthan but in neighbouring states to get a general idea about students in the wider North-West region. Thus, many other educational Boards may be added in the dataset.
3. We have considered only those attributes of students which are normally used by educational Institutions, in the dataset. It is thought that many other attributes, like age of student, his/her mental level, in terms of IQ (Intelligent Quotient) or EQ (Emotional Quotient) may also be used to give a more accurate prediction of their potential.
4. Again, the size of the dataset is an important factor in such statistical analysis. Needless to say, more student data should be included in the analysis in order to improve the accuracy of the ML techniques used.
5. We have collected student data using the conventional means of physically moving to different Universities and educational Institutions and giving their students a Google Form to fill. This is a slow process. It is important to include new ways of collecting student information, taking advantage of social media and other on-line communication tools, like LMS, VLE and Wikis. These have low missing values and will thus speed up the data collection process.
6. Finally, this research has been restricted to using only Machine Learning techniques, like Decision Tree, KNN, Naïve Bayes, SVM, since the data sets used were limited. Large datasets, collected as given in the above suggestions, will allow many Deep Learning techniques to be used to predict student potential and performance at much higher accuracies. This will make it possible to build larger, knowledge-based Recommender Systems for use by higher educational Institutions.