APPENDIX-I

List of papers communicated/ accepted/ published/ presented

1) Shashi Sharma, Sunil Kumar Pandey, Kumkum Garg," Machine Learning for Predictions in Academics", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-5, pp.4624-4627, 2020.

2) Shashi Sharma, Soma Kumawat, and Kumkum Garg, "predicting student Potential using Machine Learning Techniques", 4TH International Conference on Innovative Computing and Communication (ICICC 2021), Delhi, will be published in Springer AISC Series (Accepted).

3) Shashi Sharma, Soma Kumawat and Kumkum Garg, "Machine Learning Techniques: A Review of Predicting Student's Performance", 12th International Conference Digitalization as Vehicle for Innovation, Organizational Growth, and Effectiveness, Gwalior (Presented).

4) Shashi Sharma, Soma Kumawat and Kumkum Garg, "Predict Students Potential using Ensemble Methods", International Conference on Recent Trends in Machine Learning, IOT, Smart Cities & Applications 2021(ICMISC 2021), Hyderabad, will be published in Springer (Communicated).

ADDENDUM

QUERIES AND RESPONSES PERTAINING TO EXAMINER 1

Q1. In the objectives it is not clear what are going to be the evaluation criteria or results- is it that the accuracy of your algorithms to predict the % of students passing in exams be the criteria to validate your system? So, if 80% students are passing in real situation and the system says 79 then that's good measure of accuracy for the system? Because all could see were many % tables with lots of AI algorithms everywhere. It is important to setup the objectives to show your scope of research. In the objectives it is mentioned that the performance factors would be studied for engineering education. But in the end, we didn't see any student potential evaluation only algorithms accuracy of output. Please elaborate on this aspect.

R1. This study is divided in three main objectives: student dataset for potential analysis, student datasets based on two different Education Boards and dataset for student performance. One of the objectives of the study has been achieved through applying different classifier models on student data to analysis of student potential based on target variable i.e., technical and non-technical. The purpose of second objective is to examine the student performance prediction models that can be used to classify students based on their results. The six classifier models will also be used to build the student performance prediction models, under three different cases, by varying class variable values. In the first case, the target variable is divided into two labels: *fail* and *pass*. In the second case, the target variable is classified into three labels: *fail*, *good* and *very good*; and in the third case, five class labels: *fail, poor, satisfactory, very good* and *excellent* are used. Results of six algorithms will be evaluated in order to predict the students' performance based on the collected data. Another objective was fulfilled to apply classifier models based on Education Boards i.e., RBSE and CBSE.

These datasets contained student's demographic, socio-economic and academic attributes. every objective has individual results based on accuracy. In ML accuracy is calculated based on confusion matrix. Algorithm accuracy shows how accurate it predicts. Next, feature selection filter methods were applied to find the most important features necessary to predict student potential.

Q2. Then we would want to know how the system understood the performance factors? Were any weights assigned to factors such as past performance or family background or every factor was just YES/No or 0/1 types? Were there any fuzzy range of values to those that were input as a table? Would be good if an example is provided to show how each algo interpreted the sample data and what output come out from each? It is confusing to understand what so many percentages with each algorithm means.

R2. 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. Some attributes have given range. Labelencoder and one hot encoding was used for data preprocessing.

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. Once the data is prepared for modelling, then the learning tasks start off. As a part of it, do the following activities: The input data is first divided into parts – the training data and the test data (called holdout). Consider different models or learning algorithms for selection. Train the model based on the training data for supervised learning problem and apply to unknown data. 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.

Q3. Coming to the impacts and conclusion, the final points is that you are suggesting a set of performance factors that can accurately tell about pass percentage of the engineering students? Are there any factors that are more

important than the rest? Can this set of factors be changed? What if data is not available for some factors? Are these factors the only ones that can correctly predict engineering students' performance?

R3. 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 student's demographic, socio-economic and academic attributes. Yes, some factors are more important than other. In this study StIntforJob, Stream, 10^{th}, WhythisCourse, Fincome, Fedu, Twelth and ST were identified as the most predictive. Yes, it can be changed. We have collected the data through data collection form. If any attribute has missing, delete the student information or take average value of that attribute. No, Other factors can also take for prediction of student's performance like attendance, class assignment.

Queries and Responses pertaining to Examiner 2

Q1. In Section 3.5. research gaps 1 & 2 are not being able to clearly reflect the contributions of the thesis.

R1. We have studied many papers and none of them show any study done in this area of work in Rajasthan or on students from Rajasthan. That's why we have mentioned these gaps. In the thesis, therefore, we have taken the dataset of students from Rajasthan only.

Q2. Works considered for Literature Review could be categorized and an analysis was expected on the same (in tabular form).

R2. We have shown the literature review as per the prevailing University norms. We have categorized the literature based on the attributes and the prediction techniques. The tabular form has been used in the presentation, not in the thesis. We have followed the APA format which was given by the Research Progress Committee (RPC) at the beginning of the PhD program.

Q3. Several figures are found copied from various sources where proper website reference is to be provided.

R3. We had tried to give proper citation for all the figures, but some would inadvertently have been left out. Proper references have now been provided for all such figures.

Q4. Visibility of Figures 4.1 and 6.1 is found poor.

R4. Both the Figures have now been changed and their quality improved.

Q5. There are several methods/ techniques in data preprocessing; out of which one or more have been used in the research work. This need to be mentioned in the preprocessing phase along with proper justification.

R5. There are multiple ways to prepare any data such as cleaning, filtering, categorizing, etc., but in this thesis, there are following ways to prepare the data:

1. Data Cleaning: The dataset contains 120 missing values in different features from 2000 records. All the missing values are removed. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.

Encoding the categorical data: Categorical data is data which has some categories such as, Gender, Address.

Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So, it is necessary to encode these categorical variables into numbers.

LabelEncoder was used to encode categorical attributes with values between 0 and n_classes-1 because most of the algorithms use numerical data. For example, there are two levels of male or female in ‘gender’ attribute.

One hot encoding was used to divide the column that contains numerical *categorical data* into multiple columns based on the number of categories present in that column.

Q6. Conclusion chapter should not contain chapter wise summary, since these are already stated earlier. But it must contain the contributions of the research work (in clear words- point-wise).

R6. Contributions of the research work have now been added separately, point-wise, in the Conclusion Chapter, as Section 7.3. Suggestions for further work is now Section 7.4.

Contributions of 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.

Note: At the time of viva, all the questions were discussed in some or the other way, and I have incorporated all above Changes/Suggestions/Recommendations in the thesis accordingly. I have also discussed with co-guide and incorporated all the changes as per the Co-Guide instructions and all responses checked by Co-Guide. I asked the Co-Guide about the comments she was referring to "Modification" and she clarified that she meant to make changes/incorporate in existing Thesis and that way she would have to be submitted the Updated/Modified Thesis.

APPENDIX-II

BRIEF BIO-DATA OF RESEARCHER

I am doing Ph.D. in Machine Learning & AI from Bhartiya Skill Development University, Jaipur under the guidance of Dr. Soma Kumawat and Prof. Kumkum Garg. I have more than 10 years of experience in academics. I have 12 publications in national and international journals. I guided four M.Tech dissertations. My teaching interests include Machine Learning, Statistics with R, programming in C, C++, Design and analysis of algorithm etc. I have done my M.Tech in computer Science with hons. degree.