Early detection of academically poor performer in first year of Engineering using student's non-Cognitive traits data and employing Machine Learning based Classifier

Bhisaji C Surve[1], Dr. Bhawna Sharma[2]
[1] Research scholar, Amity Business School,Amity University, Panvel, India
[2] Off. Director & HoI, Nodal Officer- International Programs, Amity Business School, Amity University, Maharashtra
Email: [1]bhisaji.surve@s.amity.edu, [2]bsharma@mum.amity.edu

Abstract

Every Individual who go through professional programmes like Engineering, Medicine, Management will have to face not only cognitive challenges but non-cognitive too. In India apart from higher secondary school (HSC) exam. score; there are various entrance examination that respective students have to clear with adequate ranks before being admitted into any prestigious institution like IITs (Indian Institute of Technology), NIT (National Institute of Technology) or high ranking Private Institutions. This filters assures respective spectrum of students being landed in respective programmes of respective institution which indirectly means all the students in respective programmes are at par with respect to their cognitive skills but when we analyse failure rate in first year engineering students and dropout rate; it is quite high and it is major point of concern in many universities. Even though, these students in given batch are more or less in narrow band of marks variation in terms of entrance examination scores or HSC score; the question remain as why the failure rate in first year is high?

It leads to investigation as; it is not only cognitive abilities but non-cognitive skills of individuals which contribute in the student's success in first year. This paper is systematic study and development of machine learning based classifier for early detection of academically poor performer in a batch which can be identified and counselled to improve the failure rate in first year of engineering.

***Keywords**:* Structural Equation Modelling, Non-cognitive traits, Measurement model, Construct validity, factor analysis.

1. Introduction

In a study by means of data from the University of California, Geiser and Santelices (2007) establish that high school grades were a stronger predictor of both college GPA and likelihood of college graduation than students' SAT scores, class rank, and family background. Bowen and colleagues speculate that, beyond measuring content mastery, grades "reveal qualities of motivation and perseverance—as well as the presence of good study habits and time management skills" and "often echo the ability to admit condemnation and benefit from it and the capacity to take a reasonably good piece of one's work and reject it as not good enough" .Ultimately it is these qualities, more so than content knowledge, that indicates which students are likely to excel in their studies and persevere in their Higher studies as well as professional career .

Five General Categories of Non-Cognitive Factors which related to Academic Performance:

1. ACADEMIC BEHAVIORS
2. ACADEMIC PERSEVERANCE
3.. ACADEMIC MINDSETS
4. LEARNING STRATEGIES
5. SOCIAL SKILLS.

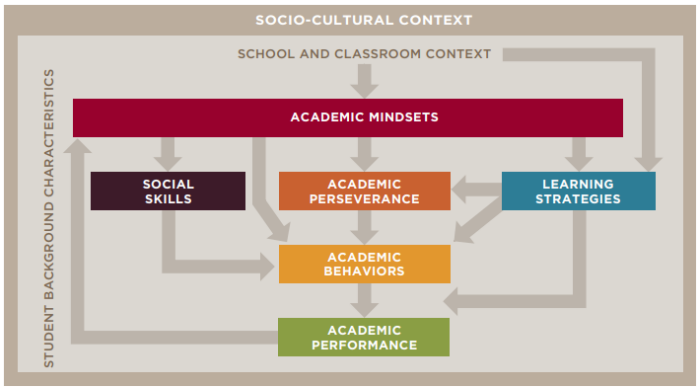


Figure 1. A Hypothesized model of Five Non-cognitive factors and their interaction in Academic domain (source: Literature Review June 2012, Teaching Adolescents to Become Learners, The university of Chicago).

Non-cognitive skill:

They are the unique patterns of thought, behaviours, emotions which socially determined and developed over a period in life. The major non-cognitive skill lists out as self-perception of self-control, metacognitive strategies, social competencies, adaptability, motivation, perseverance, resilience and coping, as well as creativity (Gutman and Schoon 2013).

Research Methodology:

1. **Exploratory study:**

Researchers focused on more flexible, malleable and impactful skills which are vital from student's perspectives. They have identified and restricted six vital non-cognitive skills which are essential for successful professional, family and social life; by doing literature survey and interacting with subject experts.

The paper explore a study based on data captured from students of Engineering Institute of NMIMS University, Mumbai, India which has ABET (*Accreditation Board for Engineering and Technology, USA***)** accredited Engineering program; these Alumni are passed out between 2015 to 2019.

The research is specific about the data set from one Engineering Institution under only one specific programme to ensure homogenous data. As it ensures educational processes parameters like faculties, Examination /Evaluation format, in general cognitive level of students, syllabus etc. being consistent and do not contribute causality toward variance in academic performance and the impact of non-cognitive aspect can be uniformly established. Researcher avoided data to be captured after 2020 as the academic teaching-learning and evaluation process is mostly online and non-standard type; due to Covid-19 Pandemic leading to lock down in country. Final data set used from survey is of 275 responses after cleaning incomplete/erroneous responses.

As per literature review and interactions with subject experts; reserachers first identified following six major non-cognitive skills which are vital from student's point view.

1.1 Definitions of constructs:

- ***Self-efficacy*** towards task refers to an individual's belief (conviction) that they can successfully achieve at a designated level on a task or attain a specific *professional* goal (Bandura, 1997; Eccles & Wigfield, 2002; Linnenbrink & Pintrich, 2002a).
- **Self-motivation** towards achievements is defined by a student's desire (as reflected in approach, persistence, and level of interest) regarding professional subjects when the individual's competence is judged against a standard of performance or excellence (McClelland, et al., 1953).

- ***Test Anxiety*** means under test conditions, individuals have combination of physiological over-arousal, tension along with fear of failure, worry. (Zeidner M. (1998)).
- **Self-control** means in order to achieve longer-term goal; it is the ability to subdue one's impulses, emotions, and behavior (Matt DeLisi (2014))
- **Grit** is the ability to persist in something you feel passionate about and persevere when you face obstacles. Person's passion and perseverance for long-term and meaningful goals (Duckworth, A.L.; Peterson, C.; Matthews, M.D.; Kelly, D.R. (June 2007))
- **Conscientiousness** is one of the Big five personality traits. Individuals who show an awareness of the impact that their own behavior has on those around them. (Costa, P. T. & McCrae, R. R. (1992).)

1.2 Measurement of individual constructs.

In this research context, there are six non-cognitive constructs which are conceptual variables and they constitute our independent variables too and Academic performance which is directly measurable through CGPA which is dependent variable. Hence measurement of such variable is thorough defining a Latent variable which is also called as Construct and indirectly measured through observed variables which are set of questionnaires. Appendix I gives detail questions for each non cognitive traits measurement. All scales used in these measurements are from prior research and relative reference is stated in Appendix I.

In order to perform multivariate analysis; reserachers employed SMART PLS based SEM (structural Equation modelling) tool to understand first causal relationship between six non-cognitive skills and academic performance. Then Predictive modelling is perform using PLSpredic and finally the same data is used for training and testing in machine learning based classifier and implemented through python coding.

The measurement model development work based on pilot data is explained in details through research paper publish in "International Journal on Innovation and Learning, Vol. 30, No. 4, 2021 by Inderscience Enterprises Ltd.". This revised instrument is then deployed through self-developed web site (www.domysurvey.com) to capture final primary data from respondent.

Non-cognitive skills	Cronbach's Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
Test Anxiety	0.73	0.79	0.82	0.61
Consciousness	0.65	0.72	0.80	0.58
Grit	0.70	0.76	0.80	0.51
Academic Motivation	0.81	0.91	0.87	0.62
Self-control	0.84	0.86	0.89	0.67
Self-efficacy	0.77	0.80	0.85	0.59

Table 1. Construct Reliability

The measurement model validity assessment.

In reflective measurement model, the first step for assessment involves examining the indicator loadings. In order to ensure as 50% of the variance in indicator's is being explained by the construct; it is recommended that loadings of indicator items should be greater than 0.708. It is the acceptable criteria for reflective items which is the reflections of latent variable or construct.

1.3 Bootstrapping:

In bootstrapping, subsamples are created with observations randomly drawn (with replacement) from the original set of data. To ensure stability of results, the number of subsamples should be large. Following is the result with subsampling of 5000. When the measurement model assessment is satisfactory, the next step in evaluating PLS-SEM results is assessing the structural model.

Bootstrapping results:

	Original Sample (O)	Sample Mean (M)	Standard Deviation (STDEV)	T Statistics (|O/STDEV|)	P Values
Acd. Motivation -> Acd. Performance	0.151	0.154	0.081	1.872	**0.031**
Conscientiousness -> Acd. Performance	0.067	0.08	0.066	1.02	**0.154**
Grit -> Acd. Performance	**0.188**	0.186	0.055	3.444	**0**
Self Effficacy -> Acd. Performance	**0.28**	0.284	0.051	5.469	**0**
Self control -> Acd. Performance	0.12	0.131	0.058	2.089	**0.018**
Test Anxiety -> Acd. Performance	0.055	0.06	0.071	0.778	**0.218**

Table 2. Bootstrapping results

It is observed as Self efficacy is major construct towards causality followed by Grit and Self-Control.

Standard assessment criteria, which is considered, include the coefficient of determination (R^2), the blindfolding-based cross validated redundancy measure $Q^{2,}$ the statistical significance and relevance of the path coefficients.

As refereeing to figure 1 model diagram; it is observed as Academic Performance construct R^2 as 0.189(19%) to evaluate the portion of variances of the endogenous variables, which is explained by the structural model. For the area of social and behavioral sciences, R^2=2% is classified with a small effect, R^2=13% as a median effect and R^2=26% as a large effect (COHEN 1988).

The coefficient of determination (R^2), which assesses the in-sample model fit of the dependent constructs' composite scores, by using the model estimates to predict the case values of the total sample. The R^2 value, however, only assesses a model's explanatory power.

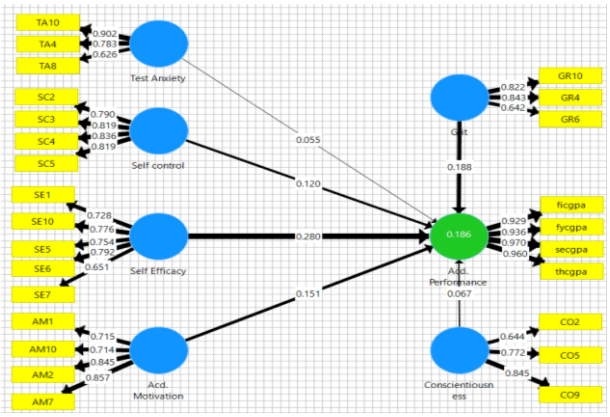


Figure 2: Structural model with loading, path coefficient and R^2

Observations and conclusion before Machine learning model development.

As per the data captured and analysed using SEM; it demonstrates as model in figure 2 indicates even though there six non-cognitive skills studied there are two major skills which covers maximum variance and hence causal effect namely "Self –Efficacy" and "Grit".

Bootstrapping results as per Table 2; these two major skills are highlighted based on P-value and T-statistic figures. Hence in order to develop a classifier which identify the poor performer based on their reporting in non-cognitive skills we have to focus on two skills data verse academic Cumulative Grade Point Average (*CGPA*).

2 Classifier implementation:

2.1 Data preparation and visualization

As discussed above we work with only three variables 1. Grit score 2. Self- efficacy score 3. CGPA for first year (max. is 4). In order to feed data to the model CSV file is compiles for 271 records after remove few outliers. Data visualization is done mainly for first year CGPA distribution.

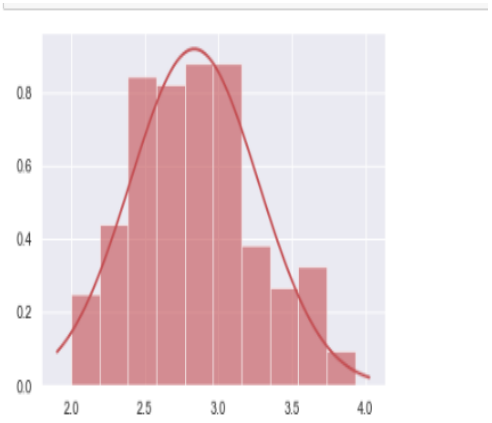


Figure 3: First year CGPA distribution curve

Observed from the graph as it is quite normal distribution with mean around 2.80, hence taking this mid-point for classifier we have dependent binary parameter as 1 for "Good Performer" and 0 as "Bad Performer". This is our Y parameter and we have two independent variables respectively "Grit score" and "Self-efficacy (SE) score" i.e. X1 and X2.

```
In [467]: newdf.head(5)
```

Out[467]:

	ficgpa	SE	GRIT	CLASS
0	3.68	7	18	1
1	2.74	10	11	0
2	2.80	8	12	1
3	3.10	12	18	1
4	3.00	5	20	1

```
In [498]: newdf['CLASS'].value_counts()
Out[498]: 1    148
          0    123
          Name: CLASS, dtype: int64
```

Figure 4: Class wise data counts

It can be noticed from this figure as it is nearly balanced class wise data.

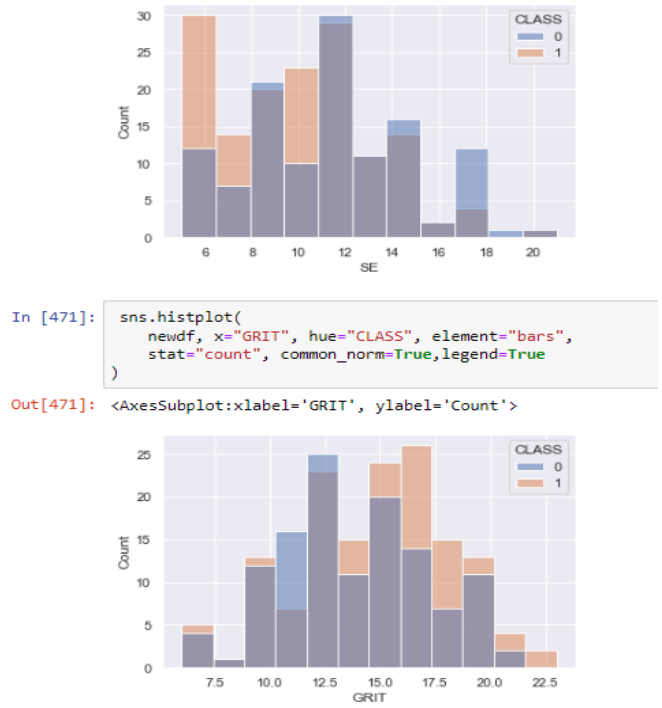


Figure 5: class wise Bar graph in relation to respective independent variable

In the above bar chart for respective predictors i.e. SE and Grit; we can have observation as there is distinct different of red bar which is good performers lies on lower side of SE score and higher side of Grit score. Hence it will help our classifier to get train for predictions.

- Splitting the data for training and testing in ration of 80:20.
- Logistic regression model implementation and validation using ROC (receiver operating characteristic curve) and Confusion matrix.
- SVM (Support Vector Machine) model implementation and validation using ROC and Confusion matrix.

Figure 6: SVM model code and confusion matrix results on test data.

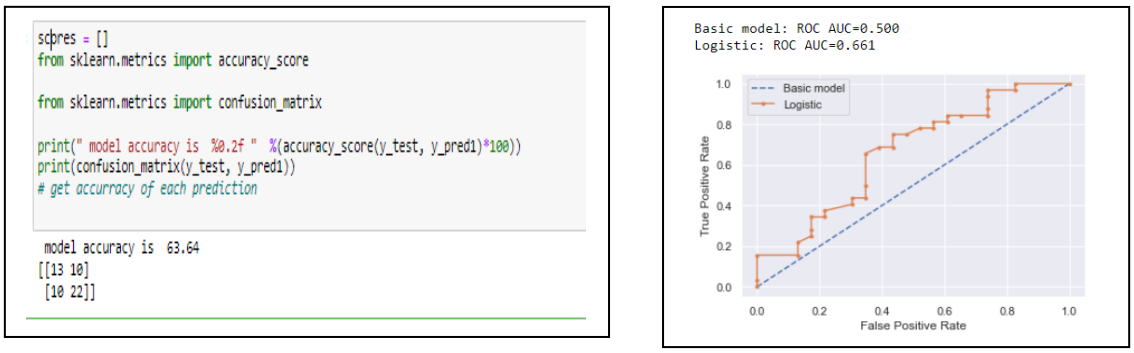
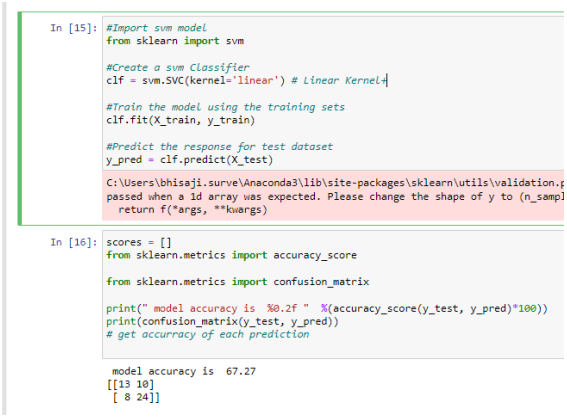


Figure 6: a. ROC (receiver operating characteristic curve) and b. Confusion matrix for Logistic regression model.



```
In [15]: #Import svm model
         from sklearn import svm

         #Create a svm Classifier
         clf = svm.SVC(kernel='linear') # Linear Kernel

         #Train the model using the training sets
         clf.fit(X_train, y_train)

         #Predict the response for test dataset
         y_pred = clf.predict(X_test)

         C:\Users\bhisaji.surve\Anaconda3\lib\site-packages\sklearn\utils\validation.
         passed when a 1d array was expected. Please change the shape of y to (n_sampl
           return f(*args, **kwargs)

In [16]: scores = []
         from sklearn.metrics import accuracy_score

         from sklearn.metrics import confusion_matrix

         print(" model accuracy is  %0.2f "  %(accuracy_score(y_test, y_pred)*100))
         print(confusion_matrix(y_test, y_pred))
         # get accurracy of each prediction

          model accuracy is  67.27
         [[13 10]
          [ 8 24]]
```

Figure 7: SVM model code and confusion matrix results on test data.

Logistic Regression is a predictive analytic technique that is based on the probability. This classification algorithm is used to predict the likelihood of a categorical dependent binary variable.

Support Vector machine is type of supervised learning classifier model which is basically isolate different classes in a hyperplane in multidimensional space. The objective of SVM algorithm is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and this processes take through iterative manner using optimizer techniques.

In order to validate our model, we use Confusion matrix and AUROC (**Area Under the Receiver Operating Characteristics**). A confusion matrix is useful tool as to visualizes and summarizes the performance of a classification algorithm.

		Actual data values	
		Positive (1)	Negative (0)
Model Predicted values	(1)	TRUE +VE (TP)	FLASE +VE (FP)
	(0)	FLASE –VE (FN)	TRUE –VE (TN)

Figure 8: Confusion matrix

In figure 8; the AUC (**Area Under the Curve**) for ROC (**Receiver Operating Characteristics**) curve obtained one of the most important evaluation metrics for checking any classification model's performance.

The True positive rate (***TPR***) gives the proportion of correct predictions in predictions of positive class.

The False Negative rate(**FPR**) also called the miss rate is the probability that a true positive will be missed out by the model.

As per our two prediction model deployed and tested; SVM is having accuracy 67% and Logistic regression model having accuracy about 64% which are reasonably good in context of our application to identify poor performers and the accuracy can be improved with more training data being used to train the model. Even Area

under the ROC curve is above base model indicates as it is useful predictor model and same can be improved my using more kernel techniques deployed in SVM.

3. Conclusion and future scope:

The paper is research journey from concept to valid, useful product which can be used by institution to identify student being poor performers in academic not due to their cognitive abilities but lack of non-cognitive skills.

Paper initially refers to all literature support to the concept of non-cognitive skills in context of academic performance. There are various non-cognitive skills can be listed but in order to restrict scope of research focused study is done by identifying most useful skill from student's point of view. Data is restricted to only one programme and one institute to ensure all other process parameters are constant except student so that the variance caused in academic performance must be more defined by their differences in non-cognitive skills.

Researcher first undertook exploratory study based on SEM; to ensure causal impact of various non-cognitive skills and to identify most prominent skills for predictive modelling. PLSpredict based predictive analysis reconfirms the predictive strength in the data and finally reserachers discussed about implementation using python code.

At every stage of modelling various validation tools are deployed to ensure the results and hence models are meeting the statistical matrices. In the context of this research more institutional data with comparative study and modelling can help us to generalise the model. Even though this concept is tested for engineering but it can applicable for other professional programmes in medical science and management studied.

Nowadays with cloud based AIML (Artificial Intelligence and Machin learning) services provided by AWS (Amazon), Google, Microsoft Azure; it is not great commercial burden for institution to take advantage of these technology based solutions.

Appendix I:

Non cognitive skill	**Questioner used with Likert scale (1-5):**	**Reference Instrument used**
Test Anxiety	1. When I take a test that is difficult, I feel defeated before I even start. 2. I feel under a lot of pressure to get good grades on tests. 3. When I take a test, my nervousness causes me to make careless errors.	Jerrell C. Cassady, W. Holmes Finch, Using factor mixture modeling to identify dimensions of cognitive test anxiety, Learning and Individual Differences, Volume 41,2015, Pages 14-20,ISSN 1041-6080.
Consciousness	Myself… 1. think of myself a lot. 2. constantly thinking about my reasons for doing things. 3. usually aware of my appearance.	Scheier, M. F., & Carver, C. S. . (2013) . Self-Consciousness Scale--(SCS-R) . Measurement Instrument Database for the Social Science.
Grit	1. I have been obsessed with a certain idea or project for a short time but later lost interest. 2. I have difficulty maintaining my focus on projects that take more than a few months to complete. 3. I often set a goal but later choose to purse a different one.	Duckworth, A.L, & Quinn, P.D. (2009). Development and validation of the Short Grit Scale (Grit-S). Journal of Personality Assessment, 91, 166-174.
Academic motivation	Why do you go to Engineering College? 1. Because I experience pleasure and satisfaction while learning new things. 2. Because I thing that a college education will help me better prepare for the career I have chosen. 3. For the pleasure that is experience in broadening my knowledge about subjects which appeal to me. 4. Because my studies allow me to continue to learn about many things that interest me	Alivernini, F., & Lucidi, F. The Academic Motivation Scale: An Italian validation

Self -control	1. I do not seem capable of making clear Plans for most problems that come up in my life. 2. The goals I achieve do not mean much to me. 3. I have learned that it is useless to make plans. 4. The standards I set for myself are unclear and make it hard for me to judge how I am doing on a task.	The Self-Control and Self-Management Scale (SCMS): Development of an Adaptive Self-Regulatory Coping Skills Instrument by Peter G. Mezo
Self -efficacy	I can, 1. Perform experiments independently. 2. Work with tools and use them to build things 3. Work with tools and use them to fix things. 4. Design new things. 5. Master the content in the engineering related courses.	Measuring Undergraduate Students' Engineering Self-Efficacy: A Validation Study Article in Journal of Engineering Education · April 2016.

References:

[1] Camille A. Farrington, Melissa Roderick, Elaine Allensworth, Jenny Nagaoka, Tasha Seneca Keyes, David W. Johnson, and Nicole O. Beechum, LITERATURE REVIEW JUNE 2012, Teaching Adolescents To Become Learners, The university of Chicago

[2] Jihyun Lee and Lazer Stankov, Non-cognitive influences on academic achievement evidence from

[3] PISA and TIMSS. 2016. Sense Publishers, Rotterdam

[4] Natasha Mamaril, University of Illinois, Urbana-Champaign, Ellen L. Usher , University of Kentucky, Caihong Li , University of Kentucky, David Ross Economy Micron Technology, Inc. Measuring Undergraduate Students' Engineering Self-Efficacy: A Validation Study. Article in Journal of Engineering Education · April 2016,DOI: 10.1002/jee.

[5] Peter G Mezo, University of Toledo., The Self-Control and Self-Management Scale (SCMS): Development of an Adaptive Self-Regulatory Coping Skills Instrument ; Article in Journal of Psychopathology and Behavioral Assessment · June 2008,

[6] Angela L. Duckworth University of Pennsylvania, Christopher Peterson University of Michigan; Michael D. Matthews and Dennis R. Kelly United States Military Academy, West Point, Grit: Perseverance and Passion for Long-Term Goals.

[7] Jerrell C. Cassady ∗, W. Holmes Finch,Ball State University, United States, Using factor mixture modeling to identify dimensions of cognitive test anxiety.

[8] Fabio Alivernini, Italian National Institute or the educational evaluation on instruction and training, Fabio Lucidi, University of Roma, The Academic Motivation Scale(AMS): Factorial structure, Invariance and validity in the Italian context. TPM Vol. 15, No. 4, 211-220 – Winter 2008,

[9] Scheier M.F., Carver C.S., The Self-Consciousness Scale: Revised version for use with general populations. Journal of Applied Psychology,15

[10] Bhisaji C. Surve, Researcher, Dr.B.R. Londhe, Professor, Amity School of Management, Amity University, Artificial Intelligence based assessment and development of student's non-cognitive skill in Professional education through online Learning Management system.

[11] Bhisaji C. Surve, Researcher, Dr.B.R.Londhe, Professor, Amity School of Management, Amity University, Non-cognitive constructs measurement model development based on the theory of planned

behaviour in the context of the academic performance of engineering students in International Journal on Innovation and Learning, Vol. 30, No. 4, 2021 by Inderscience Enterprises Ltd

[12] Farrington, C.A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T.S., Johnson, D.W., & Beechum, N.O. (2012). Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review. Chicago: University of Chicago Consortium on Chicago School Research.

[13] Galit Shmueli, Marko Sarstedt, Joseph F. Hair, Jun-Hwa Cheah, Hiram Ting, Santha Vaithilingam, Christian M Ringle , Predictive model assessment in PLS-SEM: guidelines for using PLSpredict., European Journal of Marketing Vol. 53 No. 11, 2019 pp. 2322-2347 © Emerald Publishing Limited 0309-0566 DOI 10.1108/EJM-02-2019-0189.

Book:

[14] Joseph F. Hair,Jr, Kennesaw state University, William C. Black, Lausiana State University, Barry J. Babin University of Southen Mississippi, Rolph E. Anderson, Drexel University, Ronald L. Tatham , Burke Inc., Book: " Multivariate Data Analysis ",Pearson publication,2007.

Web :

[15] www.smartpls.com