IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol-6, Special Issue-4, May 2018 E-ISSN: 2347-2693

Performance Prediction Model for National Level Examinations

P. Shanmugavadivu[1*], P. Haritha[2], Ashish Kumar[3]

[1*]Dept. of Computer Science and Applications, Gandhigram Rural Institute Deemed to be University, Dindigul, India
[2]Dept. of Computer Science and Applications, Gandhigram Rural Institute Deemed to be University, Dindigul, India
[3]Dept. of Computer Science and Applications, Gandhigram Rural Institute Deemed to be University, Dindigul, India

[*]*Corresponding Author: psvadivu67@gmail.com , Tel.: +91-94437-36780*

Available online at: www.ijcseonline.org

***Abstract*—**In the recent years, usage of data mining techniques to statistically analyze the performance of candidates in academics or national level examinations is in increase. The development of predictive analytics tools and their applications are also in the rise. This paper reports on the mechanism of the proposed prediction model that predicts the performance of a candidate appearing for national level examinations. The proposed Performance Prediction Model (PPM) is designed as a framework comprising of data classification and ranking of dataset, computation of correlation coefficient that measures the dependency among the variables and prediction using linear regression. The performance of PPM is validated on UGC-NET (2016) dataset. Based on the observed correlation between Paper-II and Paper-III marks, PPM predicts the score of a candidate in Paper-III with reference to the scored marks in Paper-II. The accuracy of the predicted data is recorded as 88 per cent. The illustrative visualizations presented in this article depict the performance analysisof the candidates in Paper-I, Paper-II and Paper-III.

Keywords*—*Performance Prediction Model (PPM), Classification, Ranking, Correlation Coefficient, Linear Regression Model.

I. INTRODUCTION

Big data plays a major role in Data Analysis, which is a process of cleansing, inspecting, transforming, and modeling datasets to discover hidden patterns, trends and information to formulate decisions, policies etc., Predictive Analytics is a branch of Big Data Analytics that predicts the unknown data from the known set of data. The PPM uses Decision Tree to classify the given dataset, based on some rules. The data are to be reorganized in order to fix ranking on the designated variable(s). Then Correlation and Regression is are appropriately computed, in order to reveal the degree of association among related to subjects that are available in the dataset. Correlation Coefficient is used to find the association among P1, P2 and P3. Prediction was performed by implementing linear regression. Bar plot is used to depict the performance of candidates in P1, P2 and P3 are on selective five subjects.

UGC conducts National Eligibility Test to qualify lectureship (LS) and Junior Research Fellowship (JRF) biannually for 99 subjects. This examination includes three papers namely Paper-I (P1), Paper-II (P2), and Paper-III (P3). Thus, this article proposes a case study on UGC NET Examination results of July 2016 (here after in this paper, candidates marks in P1, P2, P3 in UGC NET July 2016 will be referred as dataset.). Section I contains the introduction of UGC NET, Section II covers the related work of reported in the literature, Section III contains techniques used in the proposed method, and Section IV contains the results and discussion of proposed method. Conclusion of the proposed method is described in Section V.

II. RELATED WORK

A. Classification:

It classifies data according to their respective classes i.e. groups the data that belongs to a common class. It is also called supervised classification [1]. It incorporates several techniques such as Support Vector Machines, Neural Networks, Nearest Neighbour, Decision tree etc. [2, 3]. Support Vector machines aims to design a hyper plane which classifies all training vectors into two classes. Neural Network imitates human's brain and it constructs the

network architecture. Training datasets are processed into the network where their results lead to the classification. Nearest Neighbour classifies the objects based on the closest training examples. Decision Tree, forms a tree like structure in which each child node is the output of a condition that is applied on its parent node. Decision tree is well suited for handling large datasets. It breaks down the entire dataset into smaller subsets. Thus, the proposed method uses decision tree classification. Classification based on decision tree is used in medical imaging, optical video tracking, speech recognition, pattern recognition, biometric identification etc. It is also used for the analysis of complex ecological data [4].

B. Decision Tree:

A decision tree is a decision support system which uses a tree-like structure and makes decisions. A Decision Tree [5], establishes conditions and discovers related outcomes. It is especially used in data mining to perform classification [6]. It is a supervised learning which can handle large amount of data and group them based on conditions. Decision tree is useful for the areas such as text mining, pattern recognition. It addresses classification problem of imperfect data [7] and also used in education to select new students as mentioned in [8]. The proposed Performance Prediction Model (PPM), uses decision tree classification to split the entire dataset based on which is further categorized based on the community.

C. Correlation Coefficient:

Pearson's Correlation Coefficient(*r*) measure the strength of association between two variables [9]. Microbloging Twitter has become faster communicating media than email, instant messaging. Every day massive amount of tweets has been generated. Analytics has been done concerning the tweets. Sentiment analysis are explored in [10]. The answer that was posted as tweets will explore several emotions which will be as negative answers or positive answers. This can be measured by correlation coefficient to conclude the analytics whether it is positive correlation or negative correlation [11]. Applications of Sentiment Analysis is described in [12]. Now, correlation coefficient is used to specify error and to show the correlation between the observed data and distributions. It is also used as a direct index of predictive efficiency [13]. In the proposed method, Correlation coefficient is used to determine the association among three papers.

D. Predictive Analysis:

Predictive Analysis is used to predict future events of unknown. Predictive analysis make use of techniques such as data mining, statistics, modelling, machine learning, and artificial intelligence in order to analyse current data and make predictions about future. It is useful for banking sectors. If a bank plans to provide loan, in order to avoid losses and to yield credits, bank decides to find the customers who are all defaulter and who are all new customers. Such classification leads to predictions [14]. Data mining techniques plays vital role in educational background. [15] Aims to predict and analyse the performance of students in academics and [16] describes the development and implementation of Student Success Program. In the proposed method, the dependency between P2 and P3 leads to prediction of marks of P3 using P2. The idea is to classify the attributes such as P1, P2, P3, Grand Total (GT) and Status (i.e. LS and JRF) based one subject. It is highly complex to perform analysis on 99 subjects, so 4 subjects were taken as a sample and analysis is performed. Here, Decision Tree Classifier is used to classify the dataset. Univariate Decision Tree classifies the dataset based on the subjects. The proposed method, uses linear regression to predict marks of P3. Basically, there are several types of regression such as linear regression, logistic regression, polynomial regression, stepwise regression. Logistic regression finds the probability of success and failure of an event. Polynomial regression applies power to the independent variable. Stepwise regression is used when multiple independent variables are used. Linear regression uses a response variable (independent variable) to predict the predictor variable (dependent variable). Thus, linear regression is used in the proposed method.

III. METHODOLOGY

The proposed Performance Prediction Model (PPM) is designed as framework that comprises of data classification, ranking of dataset, computing correlation coefficient and prediction using linear regression. The dataset is classified using decision tree based on the subjects and categories. Ranking is performed related to the Grand Total that corresponds to their subjects. After the computation of correlation coefficient, the variable with high dependency is identified. The proposed method predicts the target marks of the associated papers.

A. Classification:

Decision Tree Classifier is used in the proposed method to perform initial extraction from the given dataset. The

attributes of the dataset are namely: S.No, Roll No, Subject, Category (GEN, OBC, SC, and ST), PwD (Persons with Disabilities), Applied for (LS or JRF), P1 (Paper-I), P2 (Paper-II), P3 (Paper-III), Gtotal and Status. Initially, classification is applied based on the essential attribute Subject. The predominant condition of decision tree is made by subject codes given in the attribute Subject. Likewise, the other condition is given based on categories such as GEN, OBC, SC, ST. Categories are classified related to their respective subjects.

B. Ranking:

The given dataset consists of qualified candidates along with their marks of P1, P2 and P3 where their Grand Total (GT) is also available. The dataset which was taken for data analysis consists of entire candidates of all 99 subjects. After classifying the dataset, it is ranked which arranges the data in an orderly fashion. The dataset makes several possibilities of ranking, such as ranking can be done concerning GT or it can be done for each of the three papers. In this article, the rank is awarded from larger to smaller. If the total marks are discrete, ranking will also be discrete. If any tie (i.e. some of the candidates having similar total), then the ranking is granted as identical. Conventionally, while tie occurs the ranks will be in floating points. In this work, such floating points are converted into integer values using type conversion. If there comes any need of applying rank correlation, the ties can be left as floating point values.

C. Correlation Coefficient:

Pearson Correlation Coefficient (PCC) is applied in this dataset to determine how the variables are associated. Here, the variables are P1, P2, and P3. The computation of PCC [17] is given as Eqn(1).

$$r(x,y) = \frac{\left(\frac{1}{n}(\sum xy)\right) - (\bar{X})(\bar{Y})}{\sqrt{\left[\left(\frac{1}{n}(\sum x^2)\right) - \bar{X}^2\right]\left[\left(\frac{1}{n}(\sum x^2)\right) - \bar{X}^2\right]}} \quad (1)$$

Correlation Coefficient (PCC) is applied for four different categories depending on the subjects. The dependent and independent variables are paired as

i) P1 and P2
ii) P1 and P2
iii) P1 and P3

The values of PCC lie between -1 and 1. Similarly this process can be applied to rest of the subjects.

D. Prediction:

Computation of CC shows that there is not much significant correlation between P1 and P2 as well as P1 and P3. Since there is dependency between P2 and P3, thus prediction is possible. Besides, the CC of P2 and P3, prediction is performed using Linear Regression (LR). The LR equation [17] is as follows.

$$Y - \bar{Y} = r\left(\frac{\sigma y}{\sigma x}\right)(X - \bar{X}) \quad (2)$$

The above two equations are used to predict the marks of P2 and P3. In Eqn (2) Y is the response variable (Predicted variable) and X is the predictor variable. Having the value of x we can predict the value of y. The prediction model is verified for its accuracy. Eqn (3) is referred from [17]. If the actual values of P2 (i.e. entire column of P2) is similar then, Eqn (4) is used, and otherwise Eqn (5) is used. The equations given below are used to calculate the error rate and accuracy.

$$Mean = \frac{\sum(marks\ obtained\ in\ P3)}{Numberofrows} \quad (3)$$

$$Errorrate = \frac{(Given\ Markof\ P3) - (Predicted\ Mark\ ofP3)}{100} \quad (4)$$

$$Errorrate = \frac{(Mean) - (PredictedMarkofP3)}{100} \quad (5)$$

$$Accuracy(inpercent) = 100\% - Errorrate\% \quad (6)$$

Thus accuracy is ensured for the proposed Performance Prediction Model using Eqn (5).

The Algorithm for PPM is given below

Input: Dataset
Output: Prediction of marks

Step 1: Import dataset
Step 2: Classify the dataset
Step 3: Rank the dataset
Step 4: Find Correlation Coefficient using equation (1)
Step 5: Select high dependency papers
Step 6: Marks prediction using equation (2)

The flow chart of proposed PPM is given below in Fig.1.

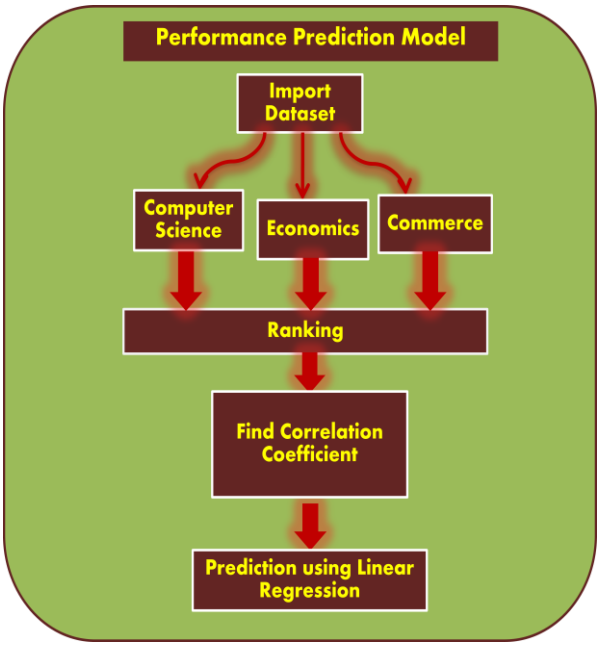

Fig.1 Flowchart of the proposed method

IV. RESULTS AND DISCUSSION

The PPM performs classification using decision tree based on the extraction of subjects and categories. Ranking is performed on the dataset which rearranges the entire dataset in an ordered fashion. Then, the computation of correlation is carried out by PCC which shows in the below Table-I.It shows the high dependency between P2 and P3 from the obtained values of correlation coefficient the high dependent variable are identified and are subjected the respective predictions. Thus the correlation coefficient for each paper is computed with reference to the other two papers. The computed values are shown in Table-1. The positive number denotes the agreement and negative number denotes disagreement. The values of the predicted variables are shown in Table II.

Table-1 Computation of Correlation Coefficient

Subjects	P1 & P2	P1 & P3	P2 & P3
Commerce	0.1586993	0.1026408	0.3267567
Economics	0.1527119	0.1110969	0.4058316
Education	0.0429630	0.0313635	0.2831696
Political Science	-0.065738	-0.0293112	0.1744172
Computer Science	0.1070279	0.1924082	0.3841889

Table-2 Performance Analysis of PPM

X	Y	Actual Value of P3	Error Rate	Accuracy
76	93	96	0.0909	99.90%
68	94	98	3.25	96.75%
56	97	98	1.37	98.63%
64	95	98	3.0	97%
72	106	94	1.75	98.25%

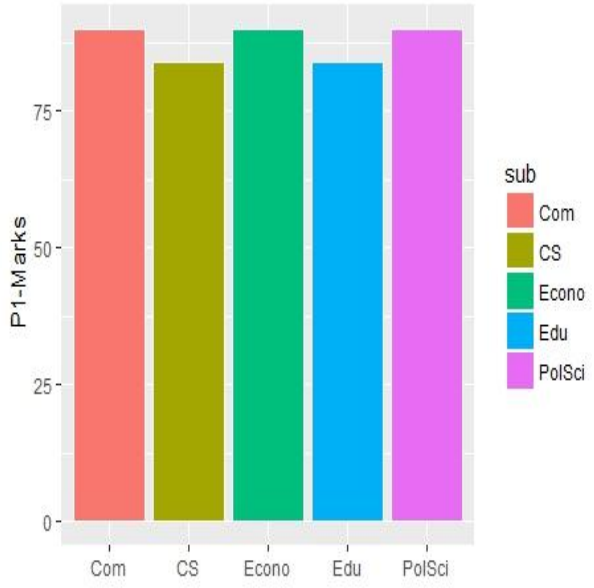

Fig.2 Performance of P1

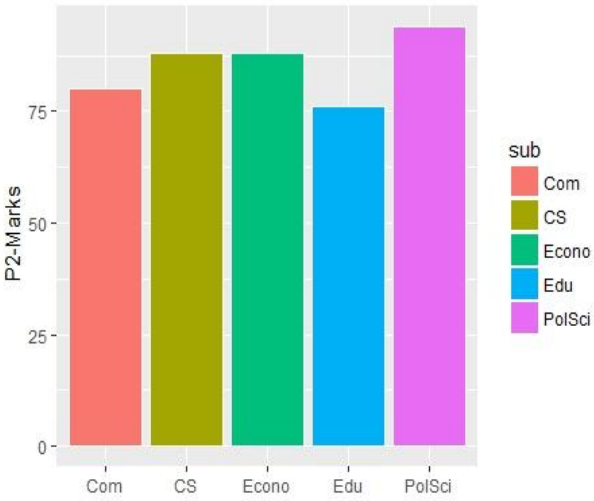

Fig.3 Performance of P2

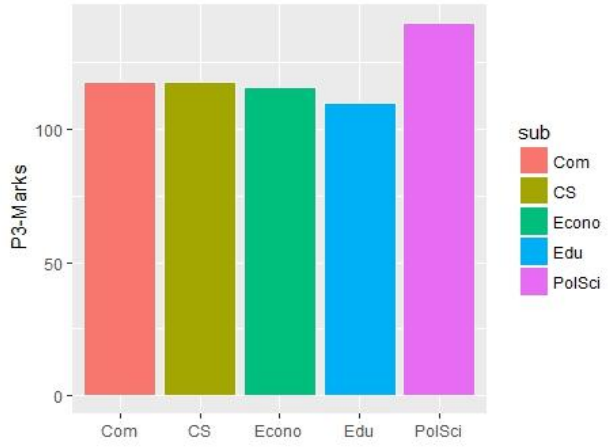

Fig.4 Performance of P3

V. CONCLUSION

This paper presents a framework to predict the marks of Paper-3 based on marks obtained by candidates in Paper-2 using linear regression model and it was validated on UGC NET dataset. The prediction efficiency of proposed model is 88% with UGC dataset. Proposed method can also be applied to predict students external marks based on their performance on internal marks.

REFERENCES

[1] N. Bhargava, G. Sharma, R. Bhargava, M. Mathuria, *"Decision Tree Analysis on J48 Algorithm for Data Mining"*. Journal of Advanced Research in Computer Science and Software Engineering, Vol.3, Issue.6, pp. 1114, 1115 June 2013.

[2] M. Gupta, N. Agarwal, "Classification Techniques Analysis", National Conference on Computational Instrumentation, pp. 128, 129 March, 2010.

[3] S.B. Kotsiantis, *"Supervised Machine Learning: A Review of Classification Techniques"*, Department of Computer Science and Technology, University of Peloponnese, Greece, pp.249, July 16, 2007.

[4] G. De'ath, K. E. Fabricius, *"Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis"*, Ecological Society of America, Vol. 81, No. 11, pp. 3178, 2000.

[5] T.S.Korting, *"C4.5 algorithm and Multivariate Decision Trees."* Image Processing Division, National Institute for Space Research – INPE SaoJose dos Campos – SP, Brazil.

[6] Y.Y. Song and Y. LU, *"Decision tree methods: applications for classification and prediction"*, Shanghai Archives of Psychiatry, Vol.27, No.2.

[7] I.Jenhani, N. B. Amor, Z. Elouedi, "Decision Tree as probabilistic classifiers",In the proceedings of the International Journal of Approximate Reasoning, Vol.48, Issue.3, pp.784, August 2008.

[8] Harwati, A. Sudiya, *"Application of Decision tree approach to Student Selection Model- A Case Study"*. In the proceedings of the IOP Conference Series: Materials Science and Engineering, Vol.105, Conference.1, pp.1, 2016.

[9] V. Bewick, L. Cheek and J. Ball. "Statistics review 7: Correlation and Regression", Published in Critical care, pp.451, 2003.

[10] W. Medhat, A. Hassan, H. Korashy, "Sentiment analysis algorithm and applications: A survey", In the proceedings of Ain Shams Engineering Journal, Vol.5, Issue.4, pp.1093-1113, December 2014.

[11] F. M. B. N. Hanif, P. Saptawati, *"Correlation analysis of user influence and sentiment on Twitter data"* In the proceedings of the International Conference *Data* and Software Engineering (ICODSE), Bandung, Indonesia, 2014.

[12] V. A. Kharde, S.S. Sonawane, *"Sentiment Analysis of Twitter Data: A Survey Techniques"*, International Journal of Computer Applications, Vol.139, No.11, pp.5, April 2016.

[13] H.E. Brogden, *"On the interpretation of the correlation coefficient as a measure of predictive efficiency"*. In the proceedings of the Journal of Educational Psychology, 1946.

[14] G. Sudhamathy, "*Credit Risk Analysis and Prediction Modelling of Bank Loans Using R"* International Journal of Engineering and Technology, Vol. 8, No.5, pp.1954, 2016.

[15] H. Hamsa, S. Indiradevi, J. J. Kizhakkethottam, *"Student academic performance prediction model using decision tree and fuzzy genetic algorithm"*, In the proceedings of the Procedia Technology, Vol.25, pp.326, 2016.

[16] E. Seidal, S. Kutieleh, "Using predictive analytics to target and improve first year student attrition", In the proceedings of Australian Journal of Education, Vol. 61, Issue.2, pp. 200, August 2017.

[17] A. S. Issac, *"Statistics"*, SCITech Publications,Chapter 6, 2011.

Authors Profile

Dr. P. Shanmugavadivu pursed Bachelor of Science from K.N. Govt. Arts College, Thanjavur, 1987.She pursed Master of Computer Science and Applications from Regional Engineering College, Trichy, 1990 and Ph.D from The Gandhigram Rural Institute, India, in year 2008. She pursedMaster of Business Administration from IGNOU, 2015. She is currently working as Professor in Department of Computer Science and Applications, Gandhigram Rural Institute, Dindigul, Tamil Nadu, India. She has published more than 22 research papers in reputed international journals and 31 research papers in edited volumes. Her main research work focuses on Medical Image Analysis, Image Restoration, Image Enhancement, Image Segmentation, Content – Based Image Retrieval. She has 24 years of teaching experience and 15 years of Research Experience.

P.Haritha pursed Bachelor of Sciencefrom Parvathy's Arts and Science college, Dindigul, India in year 2014, Master of Computer Science and Applications from Gandhigram Rural

Institute, India in year 2017. She is currently pursuing M.Phil. in Department of Computer Science and Applications, Gandhigram Rural Institute since 2017. Her main research focuses on Data Analysis.

Mr. Ashish Kumar pursed Bachelor of Science from Chaudhary Charan Singh University, Meerut, India in 2001 and Master of Computer Applications from The Gandhigram Rural Institute – Deemed to be University, India in year 2011. He is currently pursuing Ph.D and working as Project Fellow in Department of Computer Science and Applications, The Gandhigram Rural Institute – Deemed to be University, India since 2012. He has published several research papers in reputed conferences including IEEE, Springer Series and it's also available online. His main research work focuses on Digital Image Proessing, Artificial Intellengence, Machine Learning and Medical Imaging. He has 3 years of teaching experience and 5 years of Research Experience.

