DOI:10.26524/krj.2020.28

*Kong. Res. J.* 7(2): 94-98, 2020
**Publisher**: Kongunadu Arts and Science College, Coimbatore.


https://www.krjournal.com

**RESEARCH ARTICLE**

# PREDICTIVE ANALYSIS OF ACADEMIC PERFORMANCE OF COLLEGE STUDENTS USING ENSEMBLE STACKING

**Kaviyarasi, R.* and Balasubramanian, T.**

Department of Computer Science, Periyar University, Salem-636011, Tamil Nadu, India.

## ABSTRACT

One of the hottest and most popular methods in applied Machine Learning is Ensemble methods. Ensemble combines predictions from different models to generate a final prediction with better performance than any other single model. The research focused on the implementation of Ensemble method for predicting student academic performance based on their personal characteristics, family background, infrastructural environment in the college and external environment, etc...Our study uses RandomForestClassifier, Logistic Regression, and ExtraTreesClassifier as the Base Learners and AdaBoost Classifier as the Meta Learner. This result helps in predicting the accuracy of students' academic performance and also in identifying the poor performers, so that early measures prior to final semester examination can be deployed.

**Keywords:** Educational Data Mining; Ensemble; Stacking; Logistic Regression; Extra Trees Classifier; AdaBoost; Random Forest.

## 1. INTRODUCTION

Education is very important for the overall development of the nation and even it decides one's status. Three focal areas of the education are: - Learners, Learning processes and Learning situations. The academic class is generally not homogeneous but heterogeneous (1). In our earlier education systems, the responsibilities of educators were limited only with teaching the lessons in the classroom alone. But today, the teachers' contribution should be in overall improvement of the students. Hence it is the responsibilities of the academic institutions to provide proper guidance to the students' for choosing the right carrier according to their abilities and aptitudes, so that they can achieve success and obtain personal satisfaction in their life (2). Academic Institutions have severe competition among one another, trying to attract the student who will successfully pass through the educational process and making efforts to handle with student retention. Also the educational institutions are very often forced to take quick decisions, therefore timely and high quality information of student is needed (3). The Educational Data Mining deals with developing new models to explore the data originating from the educational environments. Also through those methods students' are provided with better understating and learning process. The Educational Data Mining researchers set many goals for their research. Few are listed below: i. Predicting students' future academic performance, ii. Analyzing the factors that characterize the performance of students' learning iii. Studying the effects of different kinds of educational support that can be provided by learning software, iv. Advancing the scientific knowledge about learning and learners. The users and stakeholders of EDM are: i. Learners, ii. Educators, iii. Researchers, iv. Administrators.

Student academic performance is one of the important factors in building their future (4). Many factors determine the level of academic performance of the student and some of them are listed here:-

- Student abilities and their personal characteristics
- Faculties abilities and their personal characteristics
- Level of interaction between students and faculties
- Infrastructural facilities available in the college
- External environmental influences on the students'

With the Data Mining techniques, we can discover the hidden factors that affect the students' academic performance as well as we can predict the student performance as early as possible. In previous work, the hidden factors that



*Correspondence: Kaviyarasi, R., Department of Computer Science, Periyar University, Salem-636011, Tamil Nadu, India. E.mail: arasikavi@gmail.com

affect the students' academic performance were identified using the Extra Tree Classifier model.

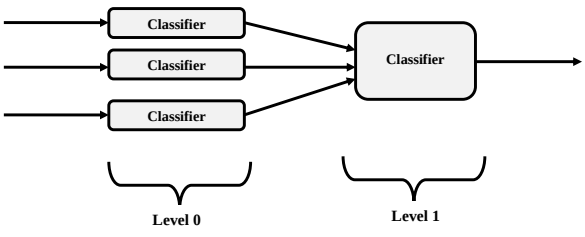
Thus, this paper is organized as follows. Section 2 gives a discussion about Stacking. Section 3 presents some of the related work in the area of Educational Data Mining. Section 4 describes the methodology proposed. Section 5 presents the results obtained. Finally, Section 6 outlines the conclusions.

## 2. STACKING

Stacking, an ensemble learning technique is a way of combining multiple models that minimizes the generalization error rate of one or more generalizers. It is used on both supervised learning and unsupervised learning tasks. Base learners and Meta learners are two types of learners in Stacking. Both are the normal machine learning algorithms like Logistic Regression, Naïve Bayes, Knn, Random Forests, SVM, etc. Stacking can also be used on a single generalizer to improve that (5) Here based on a complete training set, the base level models are trained, and then the meta-model is trained on the outputs of the base level model as features. Often stacking ensembles are heterogeneous because the base level model consists of different learning algorithms. The Stacking has the following steps:

1. *Split the training set into two disjoint sets.*
2. *Train several base learners on the first part.*
3. *Test the base learners on the second part and make predictions*
4. *The predictions from the base learners are used as features to build a new model i.e a higher level learner or meta level learner.*
5. *This model is used to make final predictions on the test prediction set.*

In stacking, the combining mechanism is that the output of the classifiers (Level 0 classifiers) will be used as training data for another classifier (Level 1 classifier) to approximate the same target function.




## 3. RELATED WORK

Eduardo Fernandesa (6) created Classification models based on the Gradient Boosting Machine (GBM) to predict academic outcomes of student performance at the end of the school year for each dataset and the results showed that, though the attributes 'grades' and 'absences' were the most relevant for predicting the end of the year academic outcomes of student performance, the analysis of demographic attributes reveals that 'neighborhood', 'school' and 'age' are also potential indicators of a student's academic success or failure.

Raheela Asif (7) used Data Mining methods to study the performance of undergraduate students. In this work, two aspects of students' performance have been focused, first predicting students' academic achievement and next, studying typical progressions and combining them with prediction results. Low and high achieving students are two important groups of students identified in this work. The results of this study proves that by focusing on a small number of courses that are indicators of particularly good or poor performance, it is possible to provide timely warning and support to low achieving students, and advice and opportunities to high performing students.

Mudasir Asharf (8) made an attempt to investigate pedagogical dataset through more effective ensemble classifier viz. stackingC. Primarily in this study, researchers made a comparison among meta and base classifiers with the intent to examine which classifiers are finest for making predictions in educational backdrop and it was observed that meta classifier viz. stackingC performed with outstanding performance of 95.65% and, among three base classifiers random forest achieved prediction accuracy of 95.76%. In case of base learners, j48 among other classifiers attained admirable accuracy of 92.98%. The Meta classifier viz. stackingC, after subjected to methods of under-sampling and oversampling, attained unprecedented prediction accuracy of 95.96% and 96.11% respectively.

Thakaa Z. Mohammad (9) discusses the clustering of elementary school slow learner students behavior for the discovery of optimal learning patterns that enhance their learning capabilities. In this paper, the development stages of an integrated E-Learning and mining system are briefed. The results show that after applying the clustering algorithms Expectation maximization and K-Mean on the slow learners' data, a reduced set of five optimal patterns list (RSWG, RWSG, RWGS, GRSW, and SGWR) is reached. Actually, the students followed these five patterns reached grades higher than 75%.

## 4. PROPOSED METHODOLOGY

This research is based on the methodology traditional cross- Industry Standard Process for

data Mining (CRISP- DM). CRISP-DM provides a structured approach for planning a research in Data Mining. The six phases in the CRISP-DM is shown in Fig.1 and explained below (10):

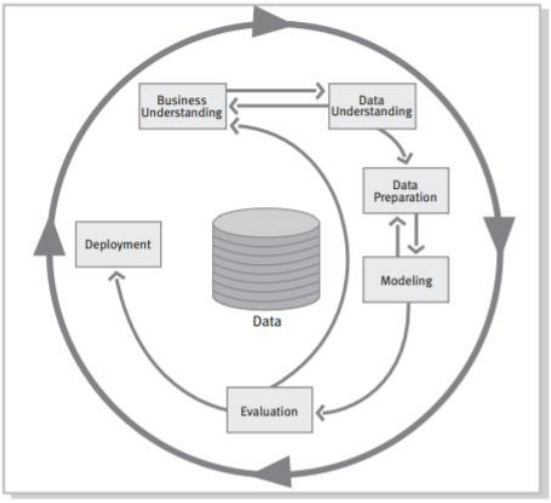



**Fig.1. Phases in CRISP- DM**

*1. Business Understanding.* The first phase focuses on understanding the objectives of the research and requirements from the business perspective, then convert this knowledge into the data mining problem definition to achieve the problem objectives. In this phase, we used student database which is composed of several attributes related to each student, such as personal details, academic background details, internal assessment details, etc., From this dataset, we defined the research goal – to minimize the failure rate of students in the final semester examination as well as finding the factors behind their performance.

*2. Data Understanding.* Data understanding process includes collecting initial data, describing data, exploring data and verifying the data quality. In this phase, we made an analysis on the factors that are affecting the students' academic performance. In this phase, we have conducted an analysis covering 45 variables namely: Student's accommodation- *Home, Hostel*; Taking Care by-*Parents, Guardian*; Living Location- *Town/City, Village*; Parental Status-*Both, Only one*; Parent Cohabitation Status-*Living together, Apart*; Fathers Education-*Educated, Un-educated*; Fathers Job-*Private, Government*, *Own Business*; Mothers Education- *Educated, Un-educated*; Mothers Job-*Private, Government, Others*; Family size- *Less than or equal to 4, More than 4*;10th grade- *More than 70%, Less than 70%*;12th grade- *More than 70%, Less than 70%*; Medium of study - *English, Tamil*; School- *Private, Government*; Secondary syllabus-*State Board, CBSE*; Group at Secondary- *First Group, Others*; Any Part Time-*Yes, No*; Study Interest-*Yes, No*; Reason to choose this college-*College reputation/Course Preference, Others*; Travelling way- *College Vehicle, Private Bus/ By walk/Others*; Travel time- *Less than 1 hr, More than 1 hr*; Have mobile-*No, Yes*; Student Using Mobile-*Others, Own*; Computer/laptop at home-*Yes, No*; Net access-*Yes, No*; Social network id-*No, Yes*; Study hours- *More than 1 hr, Less than 1 hr*; Past arrears- *Yes, No*; Extra college support- *Yes, No*; Extracurricular activities- *Yes, No*; Extra paid classes- *Yes, No*; Going outings- *Yes, No*; Alcohol consumption- *Yes, No*; Health status-*Good, Bad*; Any learning disabilities- *Yes, No*; Place to study-*Yes, No*; Guidance- *Yes, No*; Care at home- *Yes, No*; Interest in course- *Yes, No*; Attention in class- *Yes, No*; Quality of study materials-*Good, Bad*; Attendance percentage- *More than 75%, Less than 75%*; Semester percentage now- *More than 75%, Less than 75%*; Internal test 1-*Pass, Fail* ;Internal test 2-*Pass, Fail*.

*3. Data preparation.* All activities needed for constructing the final dataset from the raw data can be covered in this phase. Selecting Data, Cleaning Data, Constructing Data, Integrating Data and Formatting Data can be carried out here. The variables of the dataset were built in order to be used in next phase models. The goal of this work is to predict the students' academic performance by considering the various factors related to students. Thus the key idea of this phase is to build student dataset, in which it contains personal, social, family background details, academic details, etc.., From our previous work, the high potential factors are identified and listed in the Table 1 (1).

**Table 1. Selected Features based on the Importance Values**

| Attributes | Importance Value |
|---|---|
| Internal test 2 | 0.2492 |
| Internal test 1 | 0.1842 |
| Guidance | 0.0446 |
| Have mobile | 0.0400 |
| Family size | 0.0293 |
| Extracurricular activities | 0.0273 |
| 12th grade | 0.0233 |
| Alcohol consumption | 0.0224 |
| Attention in class | 0.0209 |
| Any ld | 0.0189 |
| Place to study | 0.0189 |
| Travel time | 0.0180 |

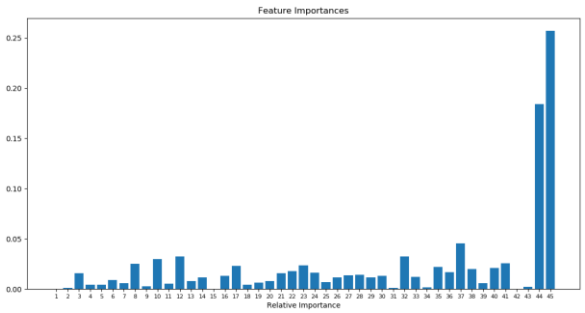


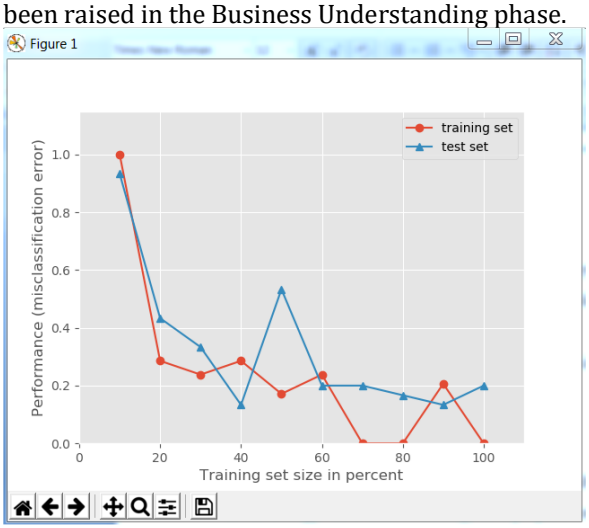

**Fig.2. Feature Importance**

*4. Modeling.* Selecting Modeling Techniques, Generating Test Design, Building Model and Assessing Modeling are the various phases in the Modeling. In this paper, we have chosen the stacking generalization algorithm that produces a predictive model. Here we uses RandomForestClassifier, Logistic Regression, and ExtraTreesClassifier as the Base Learners and AdaBoost Classifier as the Meta Learner.

3-fold cross validation:

Accuracy: 0.84 (+/- 0.02) [RandomForestClassifier]
Accuracy: 0.78 (+/- 0.02) [LogisticRegression]
Accuracy: 0.81 (+/- 0.04) [ExtraTreesClassifier]
Accuracy: 0.85 (+/- 0.02) [StackingClassifier]

*5. Evaluation.* The key ideas of this phase are evaluating results, reviewing process and determining the next steps. Also it evaluates if the generated model solves the problems that have been raised in the Business Understanding phase.




*6. Deployment.* Depending on the requirements, the deployment phase can perform deployment plan, monitoring and maintenance plan, produce final plan and report or implementing the data mining process throughout the enterprise.

Here the CRISP-DM phases are organized, structured, defined and well documented except the deployment phase, which will be the next work.

## 5. RESULTS

In this section, we will discuss the results obtained in few phases of the proposed methodology: data understanding, data preparation and evaluation.

In the *Data Understanding and Data Preparation phase*, we have taken forty five attributes that are related to student personal details, family details, educational details, etc…Among these, the high potential factors that affect students' academic performance have been identified. Next, we used RandomForestClassifier, Logistic Regression, and ExtraTreesClassifier as the Base Learners and AdaBoost Classifier as the Meta Learner. Using stacking, we done predicting analysis students' academic performance and identified the poor performers, so that early measures prior to final semester examination can be deployed.

## 6. CONCLUSION

This paper presented a methodology for analyzing the predictive academic performance of college students. The proposed methodology is based on CRISP-DM and employed a dataset which contains the students' profile details associated with their internal examination details. Initially, we have identified the high potential factors that affect students' academic performance. Moreover, the attributes identified here as relevant factors in whether a student can pass or fail at the semester examination can improve the efficiency of educators support for the student body. This added support can especially benefit students with learning difficulties, thus increasing their chances of passing at the semester examination and reducing the failure rate in general.

## ACKNOWLEDGEMENT

The authors would like to acknowledge the support provided by the affiliated colleges under Periyar University- Salem for providing a valuable dataset of the institutions to carry out the research study.

## REFERENCES

1. Kaviyarasi, R. B. T. (2018). Exploring The High Potential Factors That Affects Students' Academic Performance. *International Journal of Computer Sciences and Engineering*, 128-134.
2. Kaviyarasi, R.B.T. (2018). Predicting the academic performance of college students through machine learning techniques. *Research Journal of Computer and Information Technology Sciences*, 1-10.

3. Kabakchieva, D. (2013). Predicting Student Performance by Using Data Mining Methods for Classification. *CYBERNETICS AND INFORMATION TECHNOLOGIES*, 61-72.
4. Parneet Kaura, M.S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science*, 500 – 508.
5. Wolpert, D. (1992). Stacked Generalization. *Neural Networks*, 241-259.
6. Eduardo Fernandesa, M.H. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, 235-343.
7. Raheela Asif, A.M. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 177-194.
8. Mudasir Asharf, M.Z. (2018). Using Ensemble StackingC Method and Base Classifiers to Ameliorate Prediction Accuracy of Pedagogical Data. *Procedia Computer Science*, 1021-1040.
9. Thakaa, Z. and Mohammad, A.M. (2014). Clustering of Slow Learners Behavior for Discovery of Optimal Patterns of Learning. *International Journal of Advanced Computer Science and Applications*, 102-109.
10. Chapman, P.C. (2000). Crisp–dm 1.0 step-by-step *data mining guide*. SPSS.

## About The License