# STUDENT PASS PERCENTAGE DEDECTION USING ENSEMBLE LEARNINNG

**P Kiran Rao[1], K Giri Kumar[2], T Bala Krishna[3]**

*[1]Assisstant Professor, GPCET (affiliated to JNTUA, Anantapur) Kurnool, India*
*[2]B.Tech Student, CSE Department, GPCET(affiliated to JNTUA , Anantapur),Kurnool, India*
*[3]B.Tech Student, CSE Department, GPCET(affiliated to JNTUA , Anantapur),Kurnool, India*

---

**ABSTRACT**

Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one. Other applications of ensemble learning include assigning a confidence to the decision made by the model, selecting optimal (or near optimal) features, data fusion, incremental learning, nonstationary learning and error-correcting. This article focuses on classification related applications of ensemble learning, however, all principle ideas described below can be easily generalized to function approximation or prediction type problems as well.

## 1. INTRODUCTION

Strengthening the scientific workforce has been and continues to be of importance for every country in the world. Preparing an educated workforce to enter Science, Technology, Engineering and Mathematics (STEM) careers is important for scientific innovations and technological advancements, as well as economic development and competitiveness. In addition to expanding the nation's workforce capacity in STEM, broadening participation and success in STEM is also imperative for women given their historical under representation and the occupational opportunities associated with these fields.

Prediction modeling lies at the core of many EDM applications whose success depends critically on the quality of the classifier . There has been substantial research in developing sophisticated prediction models and algorithms with the goal of improving classification accuracy, and currently there is a rich body of such classifiers. However, although the topic of explanation and prediction of enrollment is widely researched, prediction of student enrollment in higher education institutions is still the most topical debate in higher learning institutions.

The rest of the paper is organized as follows: Section II describes the related works including ensemble methods in machine learning and related empirical studies on educational data mining using ensemble methods. Section III describes the methodology used in this study and the experiment conducted. Section IV presents results and discussion. Finally, section V presents the conclusions of the study.

## 2. ENSEMBLE CLASSIFICATION

Ensemble modeling has been the most influential development in Data Mining and Machine Learning in the past decade. The approach includes combining multiple analytical models and then synthesizing the results into one usually more accurate than the best of its components.

The following sub sections details different base classifiers and the ensemble classifiers.

### 2.1 Base Classifiers:

Rahman and Tasnim describe base classifiers as individual classifiers used to construct the ensemble classifiers. The following are the common base classifiers: (1) Decision Tree Induction – Classification via a divide and conquer approach that creates structured nodes and leafs from the dataset. (2) Logistics Regression – Classification via extension of the idea of linear regression to situations where outcome variables are categorical. (3) Nearest Neighbor – Classification of objects via a majority vote of its neighbors, with the object being assigned to the class most common. (4) Neural Networks – Classification by use of artificial neural networks. (5) Naïve Bayes Methods – Probabilistic methods of classification based on Bayes Theorem, and (6) Support Vector Machines – Use of hyper-planes to separate different instances into their respective classes.

### 2.2 Ensemble Classifiers

Many methods for constructing ensembles have been developed. Rahman and Verma argued that ensemble classifier generation methods can be broadly classified into six groups that that are based on (i) manipulation of the training parameters, (ii) manipulation of the error function, (iii) manipulation of the feature space, (iv) manipulation of the output labels, (v) clustering, and (vi) manipulation of the training patterns.

## 3. RELATED EMPIRICAL STUDIES

Stapel, Zheng, and Pinkwart study investigated an approach that decomposes the math content structure underlying an online math learning platform, trains specialized classifiers on the resulting activity scopes and uses those classifiers in an ensemble to predict student performance on learning objectives.

The study used J48, Decision Table and Naïve Bayes as base classifers and bagging ensemble model. The study concluded that J48 algorithm was doing better than the Naïve Bayesian. Also, bagging ensemble technique provided accuracy which was comparable to J48. Hence, this approach could aid the institution to find out means to enhance their students performance.

## 4. METHODOLOGY

### 4.1 Study Design

This study adapted the Cross Industry Standard Process for Data Mining (CRISP-DM) process model suggested by Nisbet, Elder and Miner as a guiding framework. The framework breaks down a data mining project in phases which allow the building and implementation of a data mining model to be used in a real environment, helping to support

business decisions. Figure I give an overview of the key stages in the adapted methodology.

**Stage 1:** Business Understanding

**Stage 2:** Data Understanding

**Stage 3:** Data preparation

**Stage 4:** Modeling

**Stage 5:** Evaluation

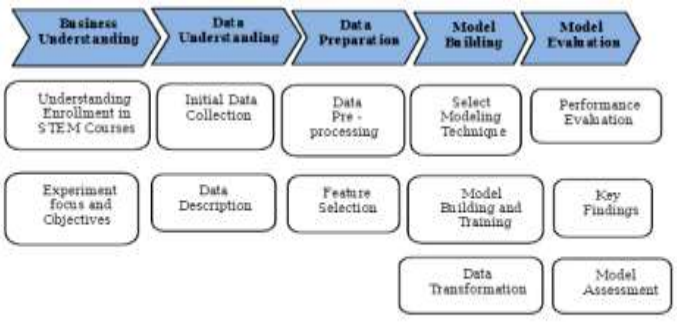


**Figure I: Adapted Methodology for Research Implementation**

### 4.2 Experiment

#### 4.2.1 Data Collection

Data was collected from sampled students through a personally administered structured questionnaire at Murang'a University of Technology, Kenya for the academic year 2016-2017. The target population was grouped into two mutually exclusive groups namely; STEM (Science,Technology,Engineering and Mathematics) and non-STEM Majors.

#### 4.2.2 Data Transformation

The collected data attributes were transformed into numerical values, where we assigned different numerical values to each of the attribute values. This data was then transformed into forms acceptable to WEKA data mining software.

#### 4.2.3 Data Modeling

To find the main reasons that affects the students' choice to enroll in STEM courses the study used three base classification algorithms together with an ensemble model method, so that we can find accurate or exact factors affecting students' enrollment in STEM.

## 5. RESULTS AND DISCUSSION

We collected students' information by distributing structured questionnaire among 220 students and 209 responses were collected. This data was preprocessed and recorded into Microsoft Excel file and then through online conversion tool, the Excel file was converted into .arff file which is supported by the WEKA software tool. We used Weka 3.6 software for our analysis. Table II shows the results obtained from the experiment.

**Table II: Comparison of Algorithms**

| S/No | Algorithm | Correctly Classified instances (%) | Incorrectly Classified instances (%) |
|---|---|---|---|
| 1 | J48 | 84 | 16 |
| 2 | CART | 77 | 23 |
| 3 | Naïve Bayes | 72 | 28 |
| 4 | Bagging | 82 | 18 |

The information on Table II shows comparison details of the algorithms that were used in our analysis. When we compared the models, we found that the J48 Algorithm correctly classified 84% of the instances and 16% of the instances incorrectly classified. The classification error is less compared to the other two baseline classification algorithm, that is, CART (23% Incorrectly Classified Instances) and Naïve Bayes (28% Incorrectly Classified Instances).

## 6. CONCLUSION

There are many factors that may affect students' choice to enroll and pursue a career in STEM in higher education institutions. These factors can be used during the admission process to ensure that students are admitted in the courses that best fit them. To categorize the students' based on the association between choice to enroll in a STEM major and attributes, a good classification is needed. In addition, rather than depending on the outcome of a single technique, ensemble model could do better. In our analysis, we found that J48 algorithm.is doing better than Naïve Bayesian and the CART algorithms.

## 8. REFERENCES

**[1].**https://www.researchgate.net/publication/326241925_Using_Machine_Learning_Algorithm_to_Predict_Student_Pass_Rates_In_Online_Education

**[2].**https://www.hindawi.com/journals/scn/2018/5264526/

**[3].**https://www.emerald.com/insight/content/doi/10.1108/JARHE-09-2017-0113/full/html?fullSc=1

**[4].**https://www.researchgate.net/publication/326241925_Using_Machine_Learning_Algorithm_to_Predict_Student_Pass_Rates_In_Online_Education

**[5].**
https://dl.acm.org/citation.cfm?id=3018896.3065830