STUDENTS PERFORMANCE PREDICTION USING RANDOM FOREST ALGORITHM

Gorinkala Hemasri[1], Kalla.Kiran[2]

[1]M.Tech. student of Ramachandra college of engineering, Eluru
[2]HOD &Associate professor dept of AI&DS, Al&ML and engineering Ramachandra college of engineering Eluru

***Abstract*: Performance analysis of outcome based on learning is a system which will strive for excellence at different levels and diverse dimensions in the field of student's interests. This system developed to analyze and predict the student's performance only. The proposed framework analyzes the students' demographic data, study related and psychological characteristics to extract all possible knowledge from students, teachers and parents. Seeking the highest possible accuracy in academic performance prediction using a set of powerful data mining techniques. The framework succeedsto highlight the student's weak points. The realistic case study that has been conducted on 200 students proves the outstanding performance of the proposed framework in comparison with the existing ones.**

***Keywords* Machine Learning (ML), Random Forest Algorithm.**

I. INTRODUCTION

The economic success of any country highly depends on making higher education more affordable and that considers one of the main concerns for any government. One of the factors thatcontributes to the educational expenses is the studying time spent by students in order to graduate.For example, the loan debt of the American students has been increased due to the failure of many students in getting graduated on time [1]. Higher education is provided for free to the students in Iraq by the government. Yet, failing of graduating on time costs the government extra expenses. To avoid these expenses, the government has to ensure that the student graduate on time. Machinelearning techniques can be used to forecast the performance of the students and identifying the atrisk students as early as possible so appropriate actions can be taken to enhance their performance.One of the most important steps when using these techniques is choosing the attributes or the descriptive features which used as input to the machine learning algorithm. The attributes can be categorized into GPA and grades, demographics, psychological profile, cultural, academic progress, and educational background [2]. This research introduces two new attributes that focus on to the effect of using the internet as a learning resource and the effect of the time spent by students on social networks on the students' performance. Four machine learning techniques, fullyconnected feed forward Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression, have been used to build the machine learning model. ROC index has been used to compare the accuracy of the four models. The dataset used to build the models is collected from the students at the College Of Humanities during 2015 and 2016 academic years using a survey and the student's grade book. The dataset has the information of 161 students. The activities of this research include feature engineering to create the students dataset, data collecting, data preprocessing, creating and evaluating four machine learning models, and finding the best model and analyzing the results.

2. LITERATURE REVIEW

Much research has been done in the area of educational data mining where a predictive model is built to forecast the performance of students to identify the at risk students. This problem can be considered a hard problem because the performance depends on many characteristics related to the students. These characteristics can be categorized into student's GPA and grades, demographics, psychological profile, culture, academic progress, and educational background [2]. The student's GPA is the most important attribute used to predict the performance. The GPA can represent the real value for the future educational and career possibilities and progression. In addition, the academic potentials can be evaluated by the student GPA. The demographics information that consists of the family background, the gender, disability, and age is also considered an important attribute [3]. This research introduces two new attributes that focus on using descriptive features related to the internet and social network usage and their effect on the performance. On the other hand, many machine learning and data mining techniques have been used to predict the students' performance such as: Artificial Neural Network (ANN); K-Nearest Neighbor (KNN); Support Vector Machine (SVM); Linear Regression; Logistic Regression; Decision Tree (DT); Random Forest (RF); Principal Component Analysis (PCA); Naïve Bayes (NB); Neuro-Fuzzy classification (NF); Decision List (DL); Bayesian Network (BN); and Discriminant Analysis (DA). Table 1 shows a summary of the research papers that relate to this study.

Paper	Features	Dataset Size	Machine Learning Algorithms	Best Algorithm
Meier et al, 2015 [4]	Grades	700	New algorithm proposed, KNN Linear regression, logistic regression, SVM	New algorithm proposed
Guleria et al, 2014 [5]	Class Performance, Attendance, Assignment, Lab Work, Sessional Performance	120	DT	DT
Xu et al, 2017 [1]	Grades, Backgrounds	1169	Linear Regression, Logistic Regression, RF, kNN, Proposed Progressive Prediction algorithm	Proposed progressive prediction algorithm
Arsad et al., 2013 [6]	Grades	896	ANN	ANN
Li et al, 2013 [7]	Grades	72	PCA	PCA
Gray et al, 2014 [8]	Aptitude, Personality, Motivation Learning strategies	914	NB, DT, Logistic Regression, SVM, ANN, KNN	SVM, KNN, NB
Buniyamin et al, 2016 [9]	Grades	391	Neuro-Fuzzy classification	Neuro-Fuzzy classification
Alharbi et al, 2016 [10]	student demographics, general performance, students' modules	1789 Testing 898 Training	Logistic regression, ANN, DL, BN, DA, DT , And Ensemble approach	No overall winners
Livieris et al, 2012 [11]	Grades	279	ANN, DT, NB, Rule-Learning SVM	ANN SVM
Hamsa et al, 2016 [12]	Internal grades, sessional grades and admission score	168	Fuzzy Genetic Algorithm and DT	FGA model is less strict than DT
Arsad et al., 2014 [13]	Grades	896	ANN, Linear Regression	ANN, Linear Regression
Sarker et al, 2014 [14]	Personal and demographics information, student satisfaction and integration	149	ANN, Logistic Regression	logistic regression
Huang et al, 2011 [15]	GPA and Grades	239	Linear Regression, ANN, Radial Basis Function NN, SVM.	SVM

Table 1- Summary Of The Related Research Papers

SYSTEM ARCHITECTURE

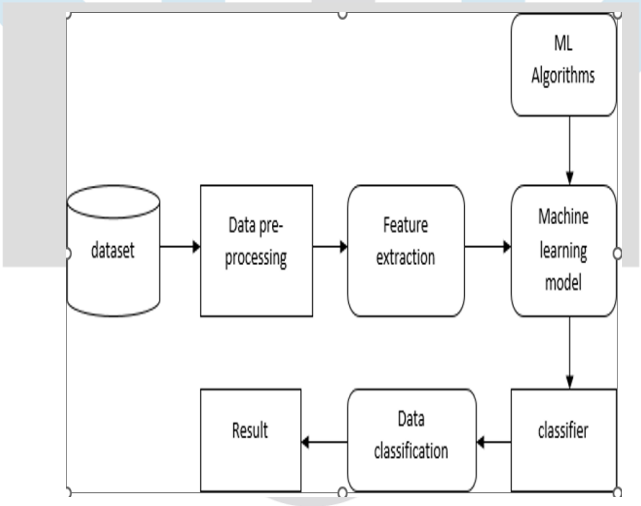


Fig.1. System Architecture

3.METHODOLOGY

1.DATA COLLECTION

Data used in this paper is a set of student details in the school records. This step is concerned with selecting the subset of all available data thatyou will be working with. ML problems start with data preferably,lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called *labelled data*.

2. PRE-PROCESSING

Organize your selected data byformatting, cleaning and samplingfrom it.

Three common data pre-processing steps are:

- Formatting
- **Cleaning**
- Sampling

Formatting: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonym zed or removed from the data entirely.

Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

3.FEATURE EXTRATION

Next thing is to do Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using Classifier algorithm. We use classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered. The rest of our labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random Forest. These algorithms are very popular in text classification tasks.

4. EVALUATION MODEL

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross- Validation to avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance. Performance of each classification model is estimated base on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

1. Random Forest Algorithm

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data, and hence the output doesn't depend on one decis ion tree but on multiple decision trees. In the case of a classification problem, the final output is tak en by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called **Aggregation**.

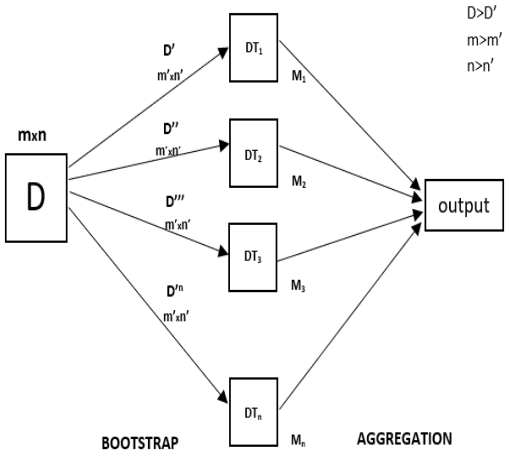
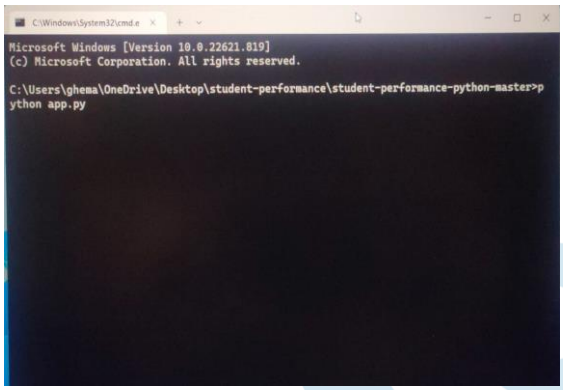


Fig.3. Random Forest Algorithm

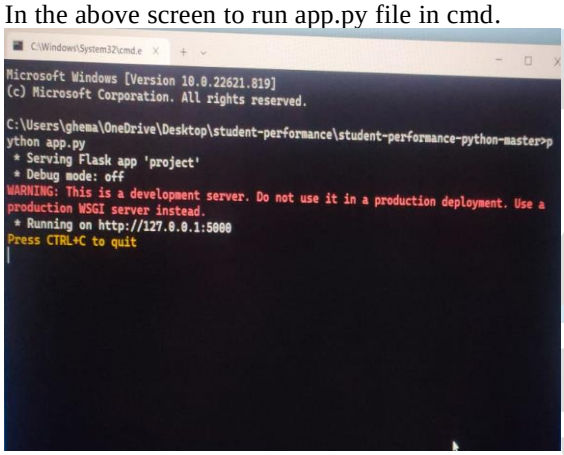
Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.
Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.
We need to approach the Random Forest regression technique like any other machine learning technique

1. Design a specific question or data and get the source to determine the required data.
2. Make sure the data is in an accessible format else convert it to the required format.
3. Specify all noticeable anomalies and missing data points that may be required to achieve the required data.
4. Create a machine learning model
5. Set the baseline model that you want to achieve
6. Train the data machine learning model.
7. Provide an insight into the model with test data
8. Now compare the performance metrics of both the test data and the predicted data from the model.
9. If it doesn't satisfy your expectations, you can try improving your model accordingly or dating your data, or using another data modeling technique.
10. At this stage, you interpret the data you have gained and report accordingly.

4.RESULTS

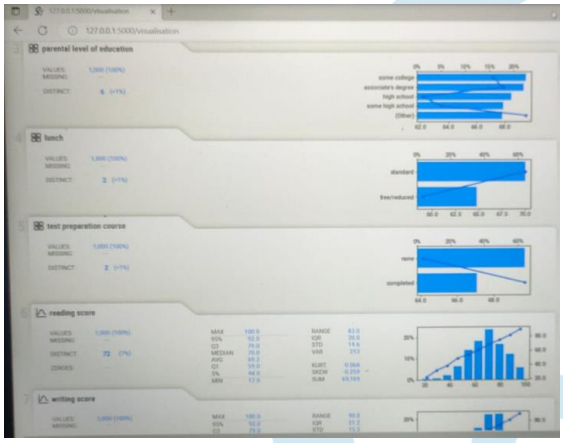


In the above screen to run app.py file in cmd.



In the above screen to get the url after running the app,py

In the above screen to get the prediction results.



In the above screen to get the prediction results.

5.CONCLUSION

Finally, performance analysis for students are a major problem. It is important that they are countered. The work reported in this thesis indicates the machine learning techniques with supervised learning algorithms to understand the performance of algorithm with respect to student records where we analyses the performance of student and categorized it into three classes as high , average, low with the accuracy of 79% .

FUTURE WORK

In the future we provide some technical solution by improve the efficiency of student performance. The user interaction model could be derived for giving the record of student dynamically and it could give staff an alert message about those students who are having low performance. We could build the prediction using Neural Network and can expect improvised results. We can add non- academic attributes along with academic's attributes.

6.REFERENCES

[1] Poza-Lujan, Jose-Luis and Calafate, Carlos T. and Posadas Yague. "As- sessing the Impact of Continuous Evaluation Strategies: Tradeoff Between Student Performance and Instructor Effort", IEEE Transactions on Edu- cation, vol.59, pp.17-23, Feb 2016.

[2] Elbadrawy, Asmaa and Polyzou, Agoritsa and Ren, Zhiyun and Sweeney. "Predicting Student Performance Using Personalized Analytics", IEEE, vol. 49,pp. 61-69, Apr.2016.

[3] Ganeshan, Kathiravelu and Li, Xiaosong. "An intelligent student advising system using collaborative filtering", 2015 IEEE Frontiers in Education Conference (FIE), pp. 1-8, Oct. 2015.

[4] Barney, Sebastian and Khurum, Mahvish and Petersen, Kai and Un terkalmsteiner, Michael and Jabangwe, Ronald. "Improving Students With Rubric-Based Self-Assessment and Oral Feedback." IEEE Transactions on Education, vol. 55, pp.319-325, Aug 2016.

[5] Lopez Guarin, Camilo Ernesto. "A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining", IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 10, pp. 119-125, Aug 2015.

[6]Grivokostopoulou, Foteini. ”Utilizing semantic web technologies and data mining techniques to analyze students learning and predict final perfor- mance” 2014 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), pp. 488-494, Dec 2014.

[7]Huang, Zhifeng and Nagata, Ayanori and Kanai-Pak, Masako and Maeda, Jukai and Kitajima. ”Self-Help Training System for Nursing Students to Learn Patient Transfer Skills” IEEE, vol. 7, pp. 319-332, Oct 2014.

[8]Bai, Samita and Rajput, Quratulain and Hussain, Sharaf and Khoja, Shakeel A. ”Faculty performance evaluation system: An ontological approach”, 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), pp. 117-124, Nov 2014.

[9]Cheng-Yu Hung, Cheng-Yu and Kuo, Fang-O and Sun, Jerry Chih Yuan and Pao-Ta Yu, Pao-Ta. ”An Interactive Game Approach for Improving Students Learning Performance in Multi-Touch Game Based Learning” IEEE Transactions on Learning Technologies, vol. 7, pp.31-37, Jan 2014.

[10]Kaur, Parwinder and Agrawal, Prateek. ”Fuzzy rule based students’ performance analysis expert system”, 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), pp. 100-105, Feb. 2014. [11]Mei-Hua Chen, Mei-Hua. ”An Automatic Reference Aid for Improving EFL Learners' Formulaic Expressions in Productive Language Use” IEEE Transactions on Learning Technologies, vol. 7, pp. 57-68, Jan 2014.

[12]Sa, Chew Li and bt. Abang Ibrahim, Dayang Hanani. ”Student per formance analysis system (SPAS)”, The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M), pp. 1-6, Nov. 2014.

[13]Mustafa, Hassan M. H. ”Dynamical evaluation Of academic perfor- mance in e-learning systems using neural networks modeling (time re- sponse approach)”, 2014 IEEE Global Engineering Education Conference (EDUCON), pp. 574-580, Apr 2014.

[14]Simpson, Jane and Fernandez, Eugenia. ”Student performance in first year, mathematics, and physics courses: Implications for success in the study of electrical and computer engineering”, 2014 IEEE Frontiers in Education Conference (FIE) Proceedings, pp. 1- 4, Oct 2014.

[15]Achumba, I. E. and Azzi, D. and Dunn, V. L. and Chukwudebe, G. A. ”Intelligent Performance Assessment of Students’ Laboratory Work in a Virtual Electronic Laboratory Environment.” IEEE Transactions on Learning Technologies, vol. 6, pp. 103-116, Apr 2013.

[16]Chen, Hsuan-Hung and Chen, Yau-Jane and Chen, Kim-Joan. ”The Design and Effect of a Scaffolded Concept Mapping Strategy on Learning Performance in an Undergraduate Database Course” IEEE Transactions on Education, vol. 56, pp. 300-307, Aug 2013.