Learning from Imbalanced Educational Data Using Ensemble Machine Learning Algorithms

Thingbaijam Lenin
Research Scholar, Martin Luther Christian University, Meghalaya, India.
E-mail: lenin.th@gmail.com

N. Chandrasekaran
Ex IBM and Director, CDAC-India, Visiting Prof., Martin Luther Christian University, Meghalaya, India.

Received October 28, 2020; Accepted November 25, 2020
ISSN: 1735-188X
DOI: 10.14704/WEB/V18SI01/WEB18053

Abstract

Student's academic performance is one of the most important parameters for evaluating the standard of any institute. It has become a paramount importance for any institute to identify the student at risk of underperforming or failing or even drop out from the course. Machine Learning techniques may be used to develop a model for predicting student's performance as early as at the time of admission. The task however is challenging as the educational data required to explore for modelling are usually imbalanced. We explore ensemble machine learning techniques namely bagging algorithm like random forest (rf) and boosting algorithms like adaptive boosting (adaboost), stochastic gradient boosting (gbm), extreme gradient boosting (xgbTree) in an attempt to develop a model for predicting the student's performance of a private university at Meghalaya using three categories of data namely demographic, prior academic record, personality. The collected data are found to be highly imbalanced and also consists of missing values. We employ k-nearest neighbor (knn) data imputation technique to tackle the missing values. The models are developed on the imputed data with 10 fold cross validation technique and are evaluated using precision, specificity, recall, kappa metrics. As the data are imbalanced, we avoid using accuracy as the metrics of evaluating the model and instead use balanced accuracy and F-score. We compare the ensemble technique with single classifier C4.5. The best result is provided by random forest and adaboost with F-score of 66.67%, balanced accuracy of 75%, and accuracy of 96.94%.

Keywords

Ensemble Learning, Random Forest, Adaptive Boosting, Stochastic Gradient Boosting, Extreme Gradient Boosting.

Introduction

Academic Performance of students is considered to be the epicenter for education system and in the last few years there has been a massive shift towards the use of technology in the teaching learning process to improve the student's academic performance. One such technology is the use of educational data mining. Educational data mining with machine learning techniques are found to be useful to bring about hidden pattern in the student's data. A prediction model for early prediction of student's performance using machine learning technique may be a boon to the faculty and the institute. It may be helpful to them to take timely and necessary measures to improve the student's performance.

However, the use of machine learning in the prediction of student's performance highly depends on the data use and the classifier adopted. To develop a good prediction model, there is a need to identify the right classifier and provide appropriate data for training the classifier. Real time student's data are found to be noisy, imbalanced, contain missing values and irrelevant. At the same time a classifier that does significantly well in one given data, may not do so in another data.

In this study, we try to develop a prediction model for identifying the student at risk from three categories of data namely demographic, prior academic record, personality. The data are trained using two ensemble machine learning techniques namely bagging algorithm like random forest (rf) and boosting algorithms like adaptive boosting (adaboost), stochastic gradient boosting (gbm), extreme gradient boosting (xgbTree) and analyzed to identify the best classifiers from these ensemble techniques. The study was conducted using R programming language and we categorized the students as "GD" and "FR", which refers to "Good" and "Fair" respectively. Students predicted as "FR" are considered to be students at risk and are likely to fail or drop out of the University while students predicted "GD" are expected to perform well without any additional assistance from the faculty.

The rest of the paper is organized as follows. In Section II, we have provided a brief summary of various works related to improving prediction performance on educational datasets and in other domains found in literature. Section III describes the methodology used for model development and this is followed by Section IV, in which the results and discussions are presented in detail.

Related Studies

There have been various studies conducted on the educational dataset using various techniques of machine learning and we discuss some of the significant studies that are closely related to our work.

B. Musiliu [1] conducted a study on developing student's performance prediction model using data mining techniques with Students' Essential Features (SEF). The author employed a set of classifiers, viz. Bayes Network, Logistic Regression and REP Tree. He also used ensemble methods of Bagging Boosting and Random Forest. The study showed that there is a strong relationship between student's essential features and their academic achievement. The accuracy of student's predictive model using students' essential features in the case of REP Tree as single classifier and in ensemble methods achieved 83.33% prediction accuracy. In terms of ROC, boosting method of REP Tree achieved best with 0.903.

In the study conducted by Al-Malaise, et al. [2], multi-agent data mining technique was implemented on 155 instances of data collected from e-learning system for prediction of student's performance. The ensemble classifier used were Stagewise Additive Modeling using Multiclass Exponential Loss Function (SAMME), Adaboost.M1 and LogitBoost. Single classifier C4.5 was used to compare the classifier performance with the ensemble classifier. They found that Stagewise Additive Modeling using Multiclass Exponential Loss Function (SAMME) and Adaboost.M1 provided the same accuracy and outperformed single classifier C4.5.

Using demographic, academic and behavioral features, Kumari, et al. [3] conducted a study on predicting student's performance. They used ID3, Nave Bayes, K-Nearest Neighbor (KNN), Support vector machine (SVM). They also used ensemble methods namely Bagging, Boosting and Voting and found that the ensemble method showed 89% accuracy and thus improved the overall accuracy from using single classifier.

With the aim to find the best modeling solution in identifying dropout student predictors Hutagaol and Suharjito [4] analyzed and measure the correlation between demographic indicators and academic performance to predict student dropout. They used three single classifiers, KNN, Naïve Bayes (NB) and Decision Tree (DT) and obtained accuracy of 75.27%, 64.29%, 64.84% respectively. The accuracy of the model was found to improve while using ensemble method with the meta-classifier gradient boosting. They also found that features of student's attendance, homework-grade, mid-test grade, finals-test grade,

total credit, GPA, student's area, parent's income, parent's education level, gender and age are important factor for developing predictive model for student's dropout.

Satyanarayana and Nuckowski [5] found that student data when filtered provided an enormous improvement in the prediction accuracy. The study conducted with the aim to identify factors and rules that influence educational academic outcomes, compared a single filter with ensemble filters. They found that using ensemble filters works better for identifying and eliminating noisy instances. They used decision trees, random forest and naïve bayes and found that using ensemble models not only provided better predictive accuracies on student performance, but also better rules for understanding the factors that influence better student outcomes.

The work of Shet and Gayathri [6] involved developing a predictive model using 150 students data. They used J48, Naïve Bayes, decision table and obtained 85%, 40%, 58 % accuracy. They also used bagging ensemble method and found to provide 82% accuracy. The study also showed the correlation between the attributes and performance and found that attributes like study hour per day, type of study materials, IQ, self-motivation, interest on academics are highly affecting the students' performance.

Classifiers Used in the Current Work

Ensemble methods is a modelling technique that uses multiple weak learners to develop a better learner. It is based on the logic that a group of classifier gives more accurate decisions as compared to a single classifier. It has two main aims namely boosting the overall accuracy as compared to the single base classifier and achieving better generalizability. Ensemble modeling thus combines the set of weak classifiers to create a single model that gives better accuracy. The broad steps involved in the ensemble methods are creating multiple data sets from the original dataset, building multiple classifiers generally weak classifier and aggregating the results of the weak classifiers. Please refer to figure 1 which illustrates the steps [7].

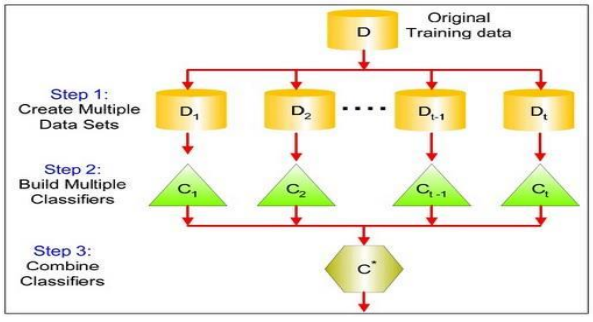

Figure 1 Steps involved in Ensemble Method

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

A. Abbreviations and Acronyms

Bagging [8] is an ensemble model based on bootstrap random sampling with replacement technique. It is an effective technique that decreases error by decreasing the variance in the result due to unstable learners, algorithms (like decision tree) whose output can change dramatically when the training data is slightly changed. The pseudo algorithm for bagging is as follows:

i. Create a set of m independent classifiers by randomly resample the training data
ii. Given a training set of size n, create m bootstrap samples of size n' by drawing n' examples from the original data, with replacement, n' ' usually <n
iii. If n=n', each bootstrap sample will on average contain 63.2% of the unique training examples, the rest are duplicates
iv. Combine the m resulting models using simple majority vote.

1. Random Forest

Random forest [9] is an ensemble of unpruned classification or regression trees. Using random feature selection in the tree induction process, random forest is formed from the bootstrap samples of the training data. Majority vote (classification) or averaging (regression) is used to provide the prediction. In another words, for the given set of inputs, it uses multiple random trees classifications to vote on an overall classification. Random forest generally exhibits an improvement in the performance over the single tree classifiers. The steps involved in creating a random forest with k trees in a dataset are as follows:

Step 1. Take random sampling with replacement "k" subset of data having "i" instances. A tree is constructed with each subset of data
Step 2. Choose "m" predictor feature randomly from all features at each splitting node while constructing a tree

Step 3. Evaluate the predictor using an objective function Select best split and perform binary split on the node
Step 4. Repeat step 1 to 4 to construct a forest with k trees
Step 5. Combine the result of k trees using simple majority vote to obtain the predictive result

Thus, random forest is bagging algorithm where each model is a random tree rather than a single model and each tree is grown according to the bootstrap sample of the training set.

B. Boosting

Also known as sequential ensemble, boosting algorithms learns from an ensemble of weak models with the aim of improving the accuracy of any supervised learning. It does so by creating weak learner from different distributions of data set through sequence of iteration and combining them to form a strong learner, "committee". The accuracy of the model is obtained by taking into account the misclassified instances mainly [10]. The weak learners are determined by applying a base machine learning algorithm on different resampled data set. The pseudo code for boosting algorithm is as follows:

Step 1. Base learner assigns equal weight to each sample observation
Step 2. Determine the misclassification error or false prediction
Step 3. Assign the false prediction to the successive base learner by giving higher weightage
Step 4. Repeat step 2 and 3 until it reaches a higher accuracy

1. Adaptive Boosting (Adaboost)

It is a boosting algorithm developed by Freund and Schapire [11] and used for binary classification. It considers a single feature and makes a single split decision tree known as decision stump. The following pseudocode depicts the algorithm:

Step 1. Assign equal weight to each data point.
Step 2. Perform prediction and determine the misclassification error
Step 3. Provide high weight to the misclassified data points
Step 4. Repeat step 2 and 3 if high accuracy is not obtained else obtained the result

2. Stochastic Gradient Boosting(gbm)

It is one of the most powerful algorithms for developing a predictive model. It has three main elements namely loss function, weak learner and additive model. It trains weak

model sequentially and every subsequent weak model minimize the loss function of the whole system. It follows the basic logic of constructing new base learner which can be maximally correlated with a negative gradient of the loss function associated with the whole ensemble [12],[13]. For any input data (x_i, i=1,2…N), let the loss function be Ψ (y, f) for the base learner h(x, θ) having number of iteration "*I*", the pseudocode are as follows:

Step 1: Calculate the negative gradient $g_t(x)$
Step 2: Fit a new base learner function $h(x, \theta_t)$
Step 3: Obtain the best gradient descent step-size

$$\rho_t = \arg\min_{\rho} \sum_{i=1}^{N} \Psi\left[y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t) \right]$$

Step 4: Calculate
Step 5: Determine $\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x, \theta_t)$
Step 6: Repeat step 1 to 5 for *I* times

3. Extreme Gradient Boosting(xgbTree)

Designed to boost the accuracy and the speed, extreme gradient boosting was developed by Tianqi Chen [14]. It is an ensemble machine learning algorithm that is based on decision tree with gbm. Extreme gradient boosting is different from gbm in that the objective function in extreme gradient boosting has a regularization function and so the objective function is given as:

$$\text{Obj} = \Psi(y, f) + \Omega$$

Where, Ψ(y, f) is the loss function which controls the predictive power, and **Ω** is regularization component which controls simplicity and over fitting. The loss function which needs to be optimized can be log loss for binary classification. The regularization component (**Ω**) is dependent on the number of leaves and the prediction score assigned to the leaves in the tree ensemble model.

Methodology

A. Understanding the Data

Data consisting of 497 instances were collected from our university. It has 19 independent attributes with one dependent attributes and can be categorized into three categories

namely, demographic, prior academic record, personality [15]. The data collected were transformed to make it feasible for the R program to execute. Please refer to table I, II, III for the nature of the data in the dataset.

Table I Demographic Data

Attribute	Notation	Nature
Gender	GND	Nominal
Permanent Location	PL	Nominal
Category	CAT	Nominal
Father's Occupation	FOC	Nominal
Mother's Occupation	MOC	Nominal

Table II Prior Academic Data

Attribute	Notation	Nature
Matriculation Institute Location	SIL	Nominal
Matriculation Performance	MPR	Nominal
Subjects	HSB	Nominal
Higher Secondary Institute Location	HSL	Nominal
Higher Secondary Performance	HPR	Nominal
Performance	PRF	Nominal

Table III Personality Data

Attribute	Notation	Nature
Naturalistic Intelligence	NLT	Numeric
Musical Intelligence	MUS	Numeric
Logical Mathematical Intelligence	LOM	Numeric
Existential Intelligence	EXT	Numeric
Visual Spatial Intelligence	VSP	Numeric
Interpersonal Intelligence	INE	Numeric
Bodily Kinesthetic Intelligence	BDK	Numeric
Verbal Linguistic Intelligence	VLI	Numeric
Intrapersonal Intelligence	INA	Numeric

The dataset is then split into 70% and 30% thereby obtaining 353 training data and 144 testing data.

B. Data Imputation

We find the collected data to have missing values and thereby making it unsuitable to analysis in most of the machine learning algorithms. The attribute of FOC is found to have a 19.52% of missing values followed by MOC with 17.1%. Attribute like HPR,

HSB, HSL, MPR, PL are found to have 4.83%, 4.23%, 3.62%, 2.62% and 0.2% of missing values respectively. In order to assign values to the instances with missing value for these attributes, we use K-Nearest Neighbor imputation technique. In this technique, the imputed value is obtained from an aggregated K-values of the nearest neighbor [16] with the Gower distance [17] between the two neighbors' i^{th} and j^{th} are obtained from the

$$d_{i,j} = \frac{\sum_{k=1}^{p} w_k \partial_{i,j,k}}{\sum_{k=1}^{p} w_k} \text{ and}$$

$$\partial i,j,k = \begin{cases} 0 & \text{if } x_{i,k} = x_{j,k} \\ 1 & \text{if } x_{i,k} \neq x_{j,k} \end{cases}$$

Where w_k is the weight and i,j,k is the contribution of the k^{th} variable

C. Modelling and Evaluation the Model

We process the imputed data using the ensemble algorithm namely bagging algorithm like random forest (rf) and boosting algorithms like adaptive boosting (adaboost), stochastic gradient boosting (gbm), extreme gradient boosting (xgbTree). It has also been explored with some single classifier like C4.5 to compare the result obtained from the ensemble classifier. We use R as a programming language in the RStudio IDE [18] with R Caret package[19]. The model is developed with "FR" as positive class.

The dataset is imbalanced and while accuracy metric provides good prediction of the majority class, it fails to provide desirable result for the minority class. Thus, accuracy metric gives misleading result for the imbalanced data set. We therefore rely on balanced accuracy and F score mainly to compare the models developed using different ensemble algorithms. The other metrics we used are recall, specificity, precision and kappa.

Result and Conclusion

The prediction model to be developed for identifying the students at risk is from the data set which are imbalanced and have both nominal and numeric values. In spite of the diverse nature of the data, the following promising results as shown in table 2 are obtained C4.5 is found to provide a balanced accuracy of 74.63 which is same as that provide by ensemble boosting algorithm of gbm and xgbTree.

Table IV Result of the Performance of the Classifiers

Classifier	**rf**	**gbm**	**xgbTree**	**adaboost**	**C4.5**
Accuracy	96.53	95.83	95.83	96.53	95.83
Sensitivity	50	50	50	50	50
Specificity	100	99.25	99.25	100	99.25
Precision	100	83.33	83.33	100	83.33
Kappa	0.65	0.604	0.604	0.65	0.604
Balanced Accuracy	75	74.63	74.63	75	74.63
F score	66.67	62.50	62.50	66.67	62.50
	Ensemble algorithm				Single

It also provides a precision of 83.33 and the same is also provided by both gbm and xgbTree. Similarly, C4.5 provides the same values for sensitivity kappa, specificity and accuracy as well. This the ensemble boosting algorithm of gbm and xgbTree fail to provide any improvement than that provided by the single classifier C4.5.

Ensemble bagging algorithm, rf and boosting algorithm adaboost, provide 75,66.67,0.65,100,100 values of balanced accuracy, F Score, kappa, precision, specificity respectively which are better than that provided by C4.5. Please figure 2, 3, 4 and 5.

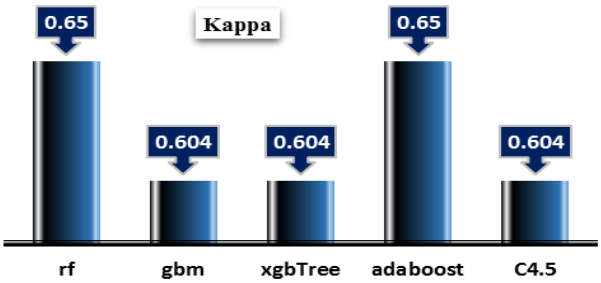

Figure 2 Comparison in terms of Kappa value

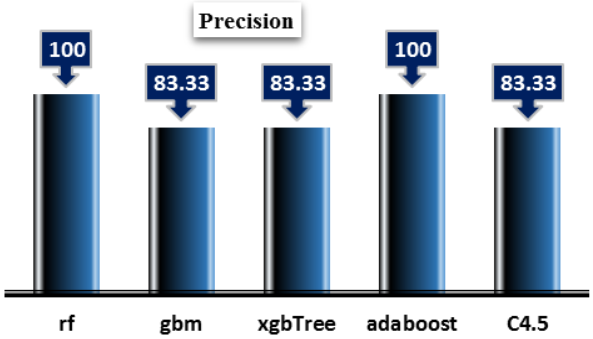

Figure 3 Comparison in terms of Precision

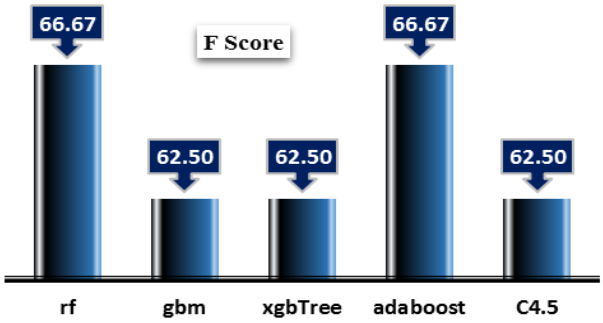

Figure 4 Comparison in terms of F Score

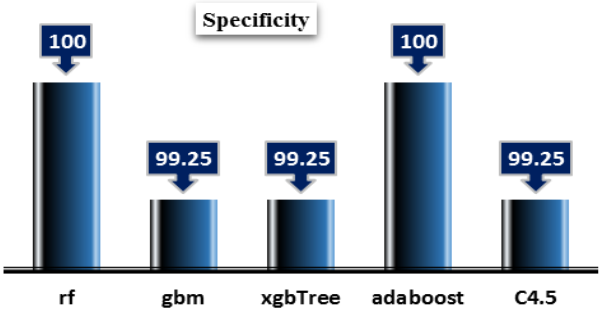

Figure 5 Comparison in terms of Specificity

Though, the models seem to provide a good prediction accuracy of 96.53, they fail to classify the minority class FR. The high accuracy is obtained as the majority class GD is classified correctly. This is again due to the fact that the data is imbalanced. The better evaluation metric is provided by balanced accuracy. In figure 6, we compare balanced accuracy and accuracy provided by the classifiers.

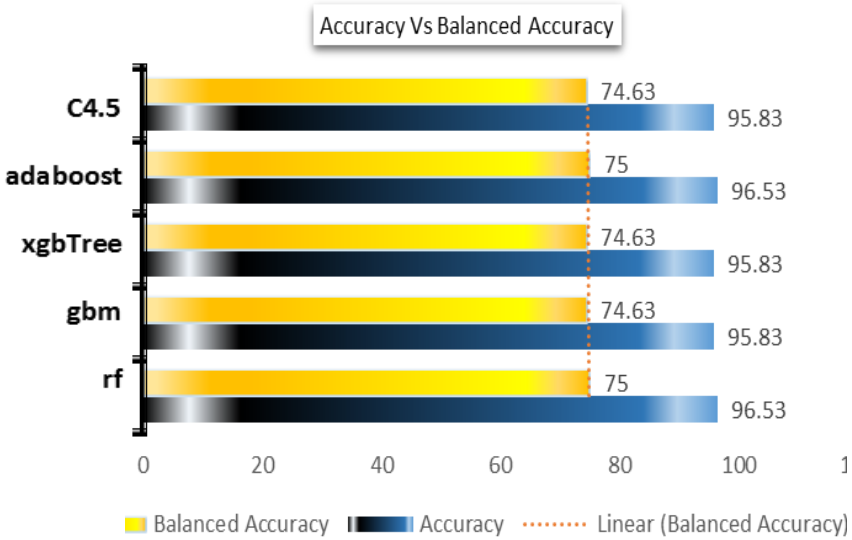

Figure 6 Comparison of Accuracy and Balanced Accuracy

It may be noted that, due to the nature of the problem we want to solve that is predicting student at risk, we are more concern with the correctly classification of the minority class FR. As such, we are interested in the model which provides no or less false negative. We find that the rf and adaboost algorithm (75%) exhibit better performance than gbm and xgbTree. Using ensemble machine learning algorithm though provide marginal improvement in the performance of the model developed for the prediction of student at risk, we need to explore other machine learning methods. We need to explore more suitable machine learning algorithms to improve the performance.

Acknowledgment

The authors are thankful to the administration of Martin Luther Christian University, Meghalaya for providing the necessary data without which this study would have been impossible.

References

Olukoya, & Bamidele M., (2020), Single Classifiers and Ensemble Approach for Predicting Student's Academic Performance. *International Journal of Research and Scientific Innovation, 7*(6), 238-243.

Abdullah, A.L., Malibari, A., & Alkhozae, M. (2014). Students' performance Prediction System Using Multi Agent Data Mining Technique. *International Journal of Data Mining & Knowledge Management Process*, *4*(5), 1–20. http://doi.org/10.5121/ijdkp.2014.4501.

Kumari, P., Jain, P.K., & Pamula, R. (2018). An efficient use of ensemble methods to predict students academic performance. *In 4th International Conference on Recent Advances in Information Technology (RAIT)*, 1-6. http://doi.org/10.1109/RAIT.2018.8389056.

Hutagaol, N., & Suharjito, (2019). Predictive modelling of student dropout using ensemble classifier method in higher education. *Advances in science, technology and engineering systems 4*(4), 206–211. http://doi.org/10.25046/aj040425.

Satyanarayana, A., & Nuckowski, M. (2016). Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance. *Spring' 2016' Mid. Atlantic 'ASEE' Conference, April'8.9,'2016'GWU*.

Shet, S., & Gayathri, J. (2014). Approach for Predicting Student Performance Using Ensemble Model Method. *International Journal of Innovative Research in Computer and Communication Engineering, 2*, 161-169.

Ensemble Learning. https://www.inf.u-szeged.hu/~tothl/ML/10. Ensemble learning.ppt

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32.

Breiman, L. (1997). Arcing the edge. *Statistics (Ber), 4*, 1–14.

Freund, Y., & Schapire, R. (1997). A decision theoretic generalisation of online learning. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics,* 1189-1232.

Mr. R. Senthil Ganesh. (2019). Watermark Decoding Technique using Machine Learning for Intellectual Property Protection . International Journal of New Practices in Management and Engineering, 8(03), 01 - 09. https://doi.org/10.17762/ijnpme.v8i03.77

Prof. Barry Wiling. (2018). Identification of Mouth Cancer laceration Using Machine Learning Approach. International Journal of New Practices in Management and Engineering, 7(03), 01 - 07. https://doi.org/10.17762/ijnpme.v7i03.66