STUDENT ENROLLMENT PREDICTION USING MACHINE LEARNING TECHNIQUES

***Akinode, J. L. and Bada, O. A**
Department of Computer Science, The Federal Polytechnic, Ilaro, Ogun State, Nigeria.
*Corresponding author: john.akinode@federalpolyilaro.edu.ng

ABSTRACT

Higher Institutions of learning are constantly looking for factors that maximize enrollment of students in their citadel of learning. These factors provide academic management, information on the applicants that will likely enroll at their institutions. This paper explores the effect of the various pre-admission factors (WAEC grades, JAMB Scores etc.), that may influence the enrollment of student in a Federal Polytechnic in South west Nigeria. The research employed the field survey approach. A data set of 560 students enrolled in various courses at a Federal Polytechnic in South-West Nigeria from 2017 to 2018 was used to validate the proposed methodology The study adopts Machine learning methods to analyse the correlation of different factors on student's enrolment. Decision tree algorithm (ID3) and support vector machine (SVM) techniques were used for the analysis. The pre-processing, processing and experimenting was conducted using Scikit-learn tool. Results obtained by comparing ID3 Decision Algorithm with other ML Algorithms such as Artificial Neural Network, Logistics Regression shows that ID3 algorithm outperforms other ML Algorithms. The Decision Tree has the highest accuracy of 97% while SVM, KNN and Naïve Bayes has an accuracy of 95%, 85% and 88% respectively. The results demonstrate that applicants' suitability for admission can be predicted based on certain pre-admission criteria (high school grade average (WAEC and NECO) and JAMB Scores). This study will help academic administrators in higher institutions of learning in admissions decision making.

Keywords: Enrolment; Data Mining; Decision Theory; Machine Learning; Prediction.

INTRODUCTION

In recent times, enrollment in public as well as private institutions have soared. Despite steadily rising enrollment rates in Nigeria post-secondary institutions, higher institutions of learning (HEI) are facing with several challenges which hampered smooth operations in schools. These include meagre resources, infrastructural deficit, policy etc.

One of the greatest challenges faced by higher institutions of learning is students' enrolment to various courses. For instance, the management of these institutions would like to know, which students will enroll in particular program. Dorina (2013) reiterates that higher institutions are operating in a sophisticated and highly completive environment. The author identified performance measurement and strategy development as the challenges of most modern Universities. Higher Educational Institution (HEI) is greatly concern on the student's enrolment data to understand the influence on student's decision to attend their institution (Haris et al., 2016). HEI requires to predict on enrolment for analyzing the current trends, understanding the significant impact on the enrolment and revenue outcomes. The information would be used to influence the future strategy and resource decisions (Luan, 2002). The student enrolment prediction will provide HEI important information for future planning and decision making. Esquivel & Esquivel (2020) reiterates that Enrolment directly impacts success factors of Higher Education Institutions. A handful number of theoretical models have been developed to unravel factors that influence student's enrollment in various academic institutions.

Recently, this challenge has been successfully addressed by various institutions through the analysis and presentation of data or data mining. Data mining enables organizations to use their current reporting capabilities to uncover and understand hidden patterns in vast databases. Esquivel & Esquivel (2020) describe data mining as the process of uncovering and analysis of large sum of data that targets to reveal hidden patterns and rules. These patterns are then built into data mining models and used to predict individual behavior with high accuracy. As a result of this insight, institutions are able to allocate resources and staff more effectively. Data mining may, for example, give an institution the information necessary to take action before a student drops out, or to efficiently allocate resources with an accurate estimate of how many students will take a particular course.

Data mining process is known as analytical process of exploration and research in the huge and enormous data to extract useful patterns and find relationships and the extent of the correlation between the elements. Data mining usually deal with data have been obtained for another purpose other than the purpose of data mining, for example, database transactions in a bank, which means that the method of data mining is not at all affect the way the of data collection.

In educational landscape, Undaiva, Patel, Shah and Nikhil (2015), in their work, confirmed that the main goal of Educational Data Mining (hereafter referred to as EDM) is to reveal hidden patterns, association and relations to discover the hidden knowledge through different data mining techniques. The outcome of the process would enable the higher learning institutions to improve their educational processes which include making better decisions, having more advanced planning in directing students, predicting individual behaviours with higher accuracy, and enabling the institution to allocate resources and staff more effectively

Haris et al. (2014) proposed forecasting techniques in the management of student enrolment in Higher Institutions. The authors submit that student's enrolment forecasting helps management of Higher Education Institutions (HEI) in planning and decision making (Haris et al., 2014).This paper focus on the capabilities of data mining and its application in higher education of learning. The study will provide an insight into various Data mining techniques in producing enrolment prediction accuracy using decision tree Algorithm.

REVIEW OF RELATED WORKS

Romero & Ventura (2020) provide a comprehensive review on how Educational Data Mining and Learning Analytics have been applied over educational data. The paper presented the current state of the art by reviewing the main publications, the key milestones, the knowledge discovery cycle, the main educational environments, the specific tools, the free available datasets, the most used methods, the main objectives, and the future trends in this research area.

Esquivel & Esquivel (2020) in their research work, analyzed different characteristics of freshmen applicants and how it affected their admission status in a Philippine university. A predictive model was designed using Logistic Regression to evaluate the probability that an admitted student will pursue to enroll in the Institution or not. The dataset used was acquired from the University Admissions Office. The results of the study showed that given limited information about prospective students, Higher Education Institutions can implement machine learning techniques to supplement management decisions and provide estimates of class sizes, in this way, it will allow the institution to optimize the allocation of resources and will have better control over net tuition revenue.

Mohammed M.E (2015) developed a predictive model to help student select the most suitable faculty using their grades in different subjects in high school. The model was implemented on selective case study of student enrolment in a University in Egypt. The eventual results revealed that the model is adequate for assisting faculty management in identifying major features in individual student, therefore, filter applicants based on the criteria of the predictive model

Haris et al., (2014) in their research work, X-ray the factors that affect the students' enrolment and various forecasting methods that have been applied in different studies. The study considered the forecasting techniques to determine the techniques that could provide the best result and contribute informative knowledge for management to make decisions in students' enrolment in Higher institutions

Dorina (2013) developed data mining techniques and methods for obtaining new knowledge from data collected from universities. The research highlights the potential of data mining applications in the context of University Management.

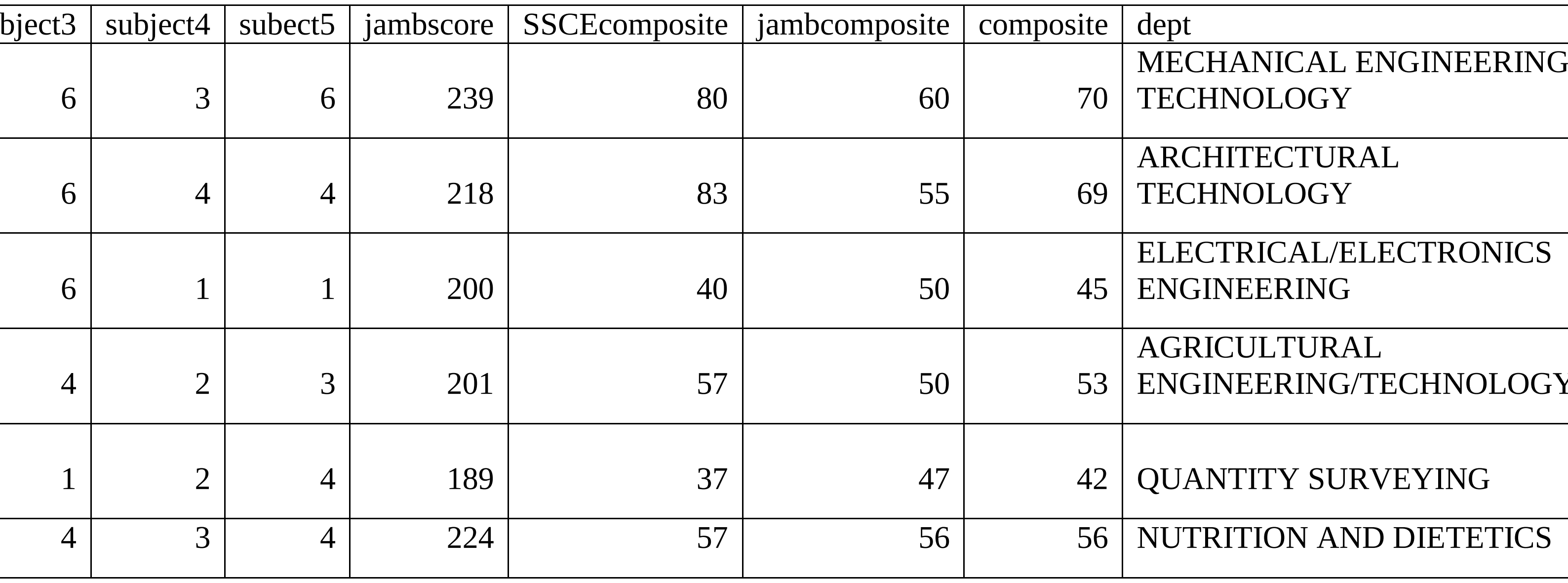
Fong et al., (2009) applied C4.5 algorithm and back-propagation algorithm to predict student admission process. They proposed a study hybrid model of neural network and decision tree classifier that predicts the likelihood of which University a student may enter, by analysing his academic merits, background and the University admission criteria from that of historical records.

Aksenova et al., (2006) in their research work carried out a study on enrolment prediction using support vector machine (SVM) and predictive models based on rules.

METHODOLOGY

The research employed the field survey approach in which the researchers administered structured questionnaire to 500 applicants which was distributed to students in different departments at the Federal polytechnic Ilaro who were admitted between 2019/2020 academic sessions.

Figure 1: Sample of Data after Cleaning and Feature Extraction



bject3	subject4	subect5	jambscore	SSCEcomposite	jambcomposite	composite	dept
6	3	6	239	80	60	70	MECHANICAL ENGINEERING TECHNOLOGY
6	4	4	218	83	55	69	ARCHITECTURAL TECHNOLOGY
6	1	1	200	40	50	45	ELECTRICAL/ELECTRONICS ENGINEERING
4	2	3	201	57	50	53	AGRICULTURAL ENGINEERING/TECHNOLOGY
1	2	4	189	37	47	42	QUANTITY SURVEYING
4	3	4	224	57	56	56	NUTRITION AND DIETETICS

1	5	2	174	33	44	38	HOSPITALITY MANAGEMENT
6	6	6	195	97	49	73	OFFICE TECHNOLOGY AND MANAGEMENT
5	1	1	189	30	47	39	URBAN AND REGIONAL PLANNING

Source: Field Survey,2020.

DATA PREPROCESSING

The research employed the field survey approach in which the researchers administered structured questionnaire to students across all departments at a Federal in South-West, Nigeria who were admitted between 2017/2018 academic sessions. The factors were based on the on the stipulated objectives and research questions which are as below:

i. O-level result: according to the procedure of admission into higher institution in Nigeria a student needs Mathematics, English and other 3 subjects based on the particular course the student is applying for and it was scaled on scale 1-5 i.e A1-6, B2-5, B3-4, C4-3, C5-2, C6-1 and 0 for others

ii. Jamb score: This includes the jamb score of each respondent

iii. SSCE composite: this was computed by adding all the Olevel score and multiplying it by 1/3

iv. Jamb composite: this was computed by dividing the jamb score by 4

v. Composite: this was calculated by adding SSCE composite and jamp composite together and dividing it by 2.

vi. Dept: this is the department each student applied for

vii. Dptcutoff: This is the departmental cut off mark

viii. Verdict: this is our target variable and it was mainly computed by setting an hypothesis such that if composite is ≥ 50 and dptcutoff is $\geq$ jamb score then the student is admitted else the student is denied admission.

A minimum of 500 students were targeted which was mainly 100 level students and must include all the faculties and departments in the school however, we got a response from 570 respondents. The fraction of the targeted study population responding to the questionnaire constituted the sample size. The questionnaire Responses was structured in a format that has the advantage of flexibility for several choice responses. Therefore, a primary data which is refers to first-hand information was obtained from the surveys which was further applied to the research.

However, all missing values are imputed using the mean of the corresponding feature for each continuous variable while for the categorical features, the missing value was imputed using the mode.

DATA MODELING

Before modelling, the data was divided into two set which are the training set and the test set. 80% of the data set is dedicated to training and 20% of data is dedicated to testing the algorithms for Accuracy.

This prediction is done on the basis of training data we have fed to the machine algorithms. For instance, 570 observations are being observed in our dataset. And, this means on the basis of [80, 20] partition, 456 observations (80% training set) are being fed to machine for the learning of the pattern of the admission process, while the remaining 114 (20%) are used as the test set. For the purpose of this research, four (4) different machine learning models were built, these include Decision tree, KNN, Support Vector Machine and Naïve Bayes. multiple models.

The major reason for using more than one model selection procedure is due to the fact that classification accuracy might not be adequate to describe the best model and since one of researcher's major objectives is to reduce chances of over-fitting and select the best model, we decided to use multiple models.

Table 2: Result of the performance of different Machine Learning Algorithms

Model	**Accuracy**	**Precision**	**Recall**	**FScore**
Decision Tree	97%	93%	94%	94%
KNN	85%	79%	82%	80%
Support Vector Machines	95%	95%	89%	92%
Naive Bayes	88%	81%	86%	83%

Table 2 above shows the performance of the student enrolment modelling that was conducted. Looking at Table 1 closely, it was observed that Decision Tree outperform the other three (3) models in modelling the enrolment of students. Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. F1 Score is the weighted average of Precision and Recall.

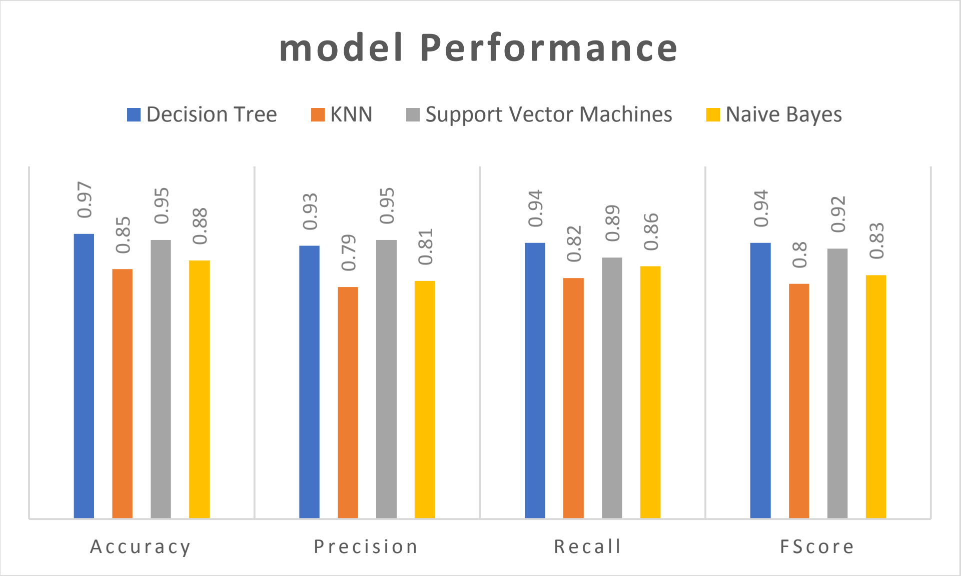


Figure 1: Model Performance

This can be justified further by looking closely into figure 1as it is shown that Decision Tree has the highest accuracy of 97% while SVM, KNN and Naïve Bayes has an accuracy of 95%, 85% and 88% respectively. Similarly, considering the other metrics, decision tree has proven to be a better model in predicting the enrolment of students in this research work.

Furthermore, to evaluate the performance of the model, we applied the confusion matrix to understand the number of the number of correctly and incorrectly predicted classes just as shown in figure 2 below:

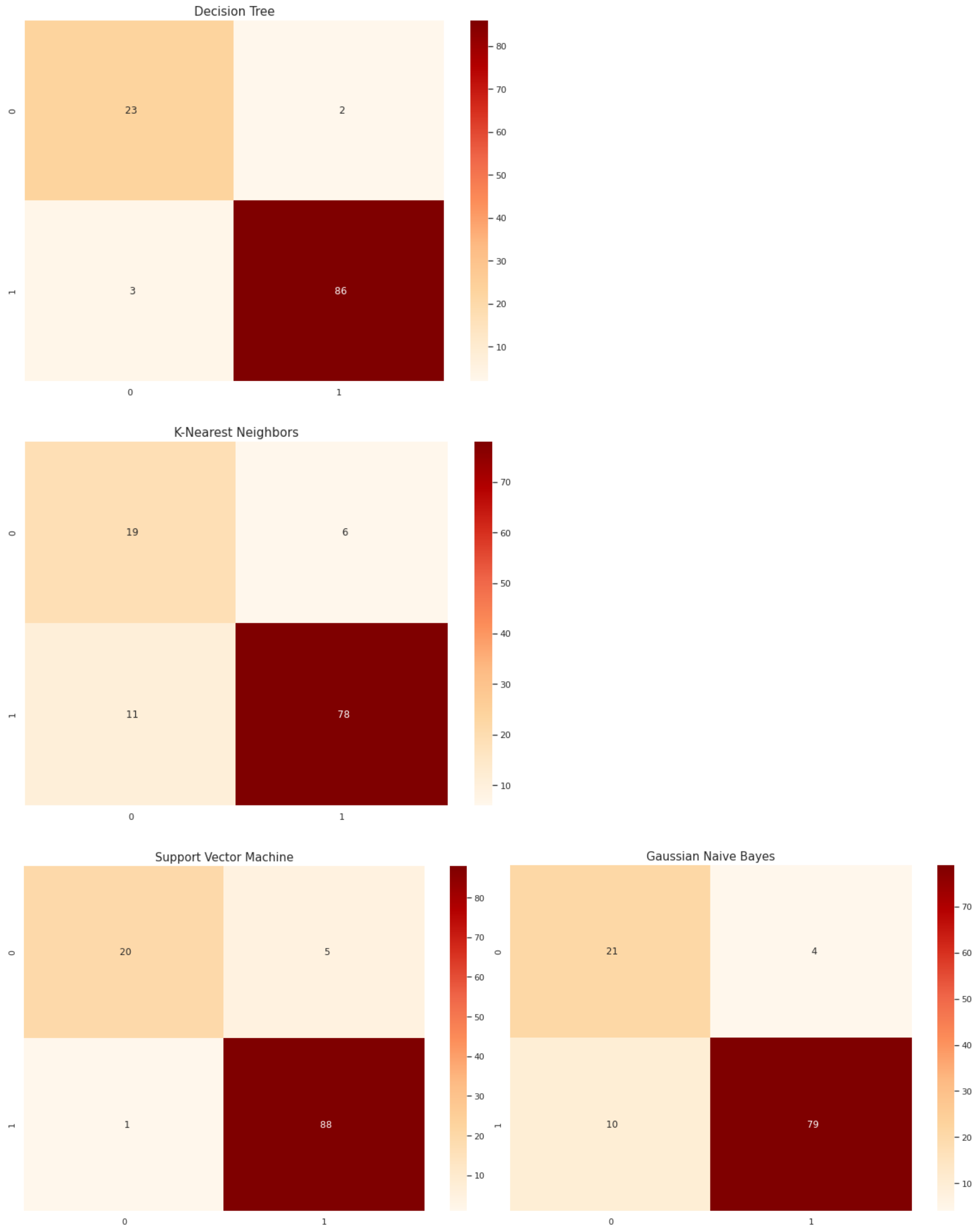


Figure 2: Confusion Matrix of the Model

Looking at the figure 2 above we can see that out remaining 144 (20%) of the test set, we can see that five (5) classes where incorrectly predicted using Decision tree, seventeen (6) classes are incorrectly predicted using K-Nearest Neighbors, six (6) classes were incorrectly predicted using Support Vector Machine while fourteen (14) classes were incorrectly predicted using Gaussian Naïve Bayes.

CONCLUSION

This study considers the enrolment process of students into first-year in a federal Polytechnic in South-West Nigeria based on certain pre-admission factors such as O- level performance (high school final result), the weight of the choice of course and the performance of this student in the UTME (an entrance assessment to higher institutions) and post UTME examinations. Based on our research, it was observed that Decision Tree Algorithm outperform the other three machine learning models used and could be adopted in the future for predicting the enrolment of students. The knowledge discovered by the techniques would enable the higher learning institutions to improve their educational processes which include making better decisions, having more advanced planning in directing students, predicting individual behaviors with higher accuracy, and enabling the institution to allocate resources and staff more effectively. For the future work, we believe that this model can perform better with the addition of more variables to the dataset. Also, the use of boosting algorithms and ensemble algorithms should be considered in the process.

REFERENCES

Aksenova S.S., D. Zhang and M. Lu (2000) .Enrollment prediction through data mining. In International Conference on Information Reuse &Integration, Waikoloa, HI, USA.

Dorina K (2013). Predicting Student Performance by Using Data Mining Methods for Classification. Bulgarian Academy of Sciences Cybernetics and Information Technologies • Volume 13, No 1

Esquivel, J. A., & Esquivel, J. A. (2020). Using a Binary Classification Model to Predict the Likelihood of Enrolment to the Undergraduate Program of a Philippine University. International Journal of Computer Trends and Technology. https://doi.org/10.14445/22312803/ijctt-v68i5p103

Fong, S., Si, Y. W., & Biuk-Aghai, R. P. (2009). Applying a hybrid model of neural network and decision tree classifier for predicting university admission. ICICS 2009 - Conference Proceedings of the 7th International Conference on Information, Communications and Signal Processing. https://doi.org/10.1109/ICICS.2009.5397665

Haris, A., Abdullah, M., Othman, A. T., & Rahman, F. A. (2014). Application of Forecasting Technique for Students Enrollment. Knowledge MAnagement International Conference (KMICe).

Haris, N. A., Abdullah, M., Hasim, N., & Abdul Rahman, F. (2016). A study on students enrollment prediction using data mining. ACM IMCOM 2016: Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication. https://doi.org/10.1145/2857546.2857592

Luan, J. (2002). Data Mining and Its Applications in Higher Education. New Directions for Institutional Research. https://doi.org/10.1002/ir.35

Mohamed M. Ezz(2015) . Advisory System for Student Enrollment in University Based on Variety of Machine Learning Algorithms. International Journal of Computing Academic Research

(IJCAR) ISSN 2305-9184 Volume 4, Number 2 (April2015), pp. 34-45

Romero C, Ventura S.(2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining Knowl Discov. 2020;10:e1355. https://doi.org/10.1002/widm.1355

Siraj, F., & Abdoulha, M. A. (2009). Uncovering hidden information within university's student enrollment data using data mining. Proceedings - 2009 rd Asia International Conference on Modelling and Simulation, AMS 2009. https://doi.org/10.1109/AMS.2009.117

Siraj, F., & Ali, M. (2011). Mining Enrollment Data Using Descriptive and Predictive Approaches. In Knowledge-Oriented Applications in Data Mining. https://doi.org/10.5772/14210.

Undavia J.N., Pranshat. M. Dolia and Nikhil. P. Shah, (2013). "Education Data Mining in Higher Education - A Primary Prediction Model and Its Affecting Parameters" International Journal of Current Research, 5(5), 1209–1213. https://doi.org/10.13140/RG.2.1.4514.1840