IOSR Journal of Research & Method in Education (IOSR-JRME)
e-ISSN: 2320–7388, p- ISSN: 2320-737x Volume 11, Issue 3 Ser. IV (May – June 2021), PP 52-59
www.iosrjournals.org

Effect Of Cohabilitation Status Of Parents On Child's Academic Performance: A Machine Learning Approach

[1]Oduntan, Odunayo Esther and [2]Abimbola, Olawale Victor
[1]*Department of Computer Science, The Federal Polytechnic, Ilaro. Nigeria.*
[2]*Data Science, Nigeria.*

Abstract
Performance is a key factor that is used in all sectors of human endeavours to measure the effectiveness of policies and achievements. In the educational sector, the performance of a student in learning and skill acquisitions have either a positive or negative impact on the well being of the student; and may have resulted from many features. This research focuses on analysing the effect of cohabitation status of parents for modeling student performance using machine learning approach. Some features that are related to cohabitation status of parents were extracted from the UCI machine learning repository consisting of about 32 features. Machine learning algorithms such as Random forest, Extreme Gradient Boosting, Support vector machines were used to analyse the data. Data imbalance was resolved by applying Synthetic Minority Oversampling Technique (SMOTE) for data augmentation. The efficiency of the algorithms on the model were determined using metrics such as Accuracy, F1 score, Precision and Recall which was implemented using Python programming language. Results show that Random forest has the highest accuracy of about 79.6%, precision of 79%, recall of 78.3% and an F_Score of 78.6%. It is obvious that Random forest perform better than Decision tree, Support Vector Classifier and XGboost in predicting the cohabitation status of students. Confusion matrix was applied to determine feature importance of the model. Results shows that the age of the student, reason for selecting the school, mother's education, father's education mothers' job and fathers job has a higher contribution on the model than any other features.

Keyword: *Academic Performance, machine learning, Synthetic Minority Oversampling Techniques,*

Date of Submission: 21-05-2021 Date of acceptance: 06-06-2021

I. Introduction

In an educational system, the development of students in all endeavors is paramount, this is the major goal of any institution where future leaders are being modelled. There are many factors responsible for the behavior's observed in any student. Some of these factors may be the parental status, society, immediate environment of a student, way of life of both the students and their parents (Mushtag and Khan, 2012); all these are some of the variables that can affect the performance of a student in career development.

Performance can be said to be as how well or badly something is done as well as the workability of a process. it is also defined as the act or process of performing a task, an. action, etc. while the verb perform means to work or function well or badly. The good or bad of the concept 'performance' depends on the context on which it is being examined.

There are various instances to which performances can be checked; for example, there exist academic, dance, events. Academic Performance is the extent to which a student, teacher or institution has attained their short or long-term educational goals (Gregory, 2019). Since students are at the core of learning process, a study tailored to their motivations and strategies and factors hindering their learning is imperative as students themselves play pivotal roles in shifting their own learning and acquiring enhanced academic achievement.

A study conducted by (Aqsa Shoukat et al, 2018) on the impact of parent's education on the academic performance of students, reviles that there exists a relationship between parents' education and the academic performance of their children. Also, according to (Micha G.et al 2016), there is a stronger effect on children's academic performance based on the cultural status of the parent. However, (Adeyemo and Kuyoro 2010) who examines the effect of students socio-economic/family background on students' academic performance in tertiary institutions using decision tree algorithm was able to justify the point that students whose parent does not have much time for due to work or any other important things to them are likely to perform woefully in exams. In addition to this their results was in line with that of (Bakare 1975) study justifies that factor affecting students' performance can be traced to the society schools and family of the students.

There have been different studies on the prediction and modelling of student's performance however, the approach of measuring the cohabitation effect of parents has been less studies most especially with the approach of machine learning.

Modelling students' performance is a task that human being's knowledge alone is not sufficiently elaborate to extract a feature that could aid decision on a particular student by the stakeholders. Machine learning programs and algorithms learn from their experiences and aims to achieve satisfactory results, once exposed to sufficient training examples. Machine learning tools programs whose behaviors adapts to their input data offer a solution to such issues; they are, by nature, adaptive to changes in the environment they interact with. The component of experience, or training, in machine learning often refers to data that is randomly generated.

It becomes obvious that there are treasures of meaningful information buried in data archives that are way too large and too complex for humans to make sense of. Learning to detect meaningful patterns in large and complex data sets are a promising domain in which the combination of programs that learn with the almost unlimited memory capacity and ever-increasing processing speed of computers opens up new horizons.

One of the challenges in machine learning is to select the right algorithm for the intended problem. According to the popular No Free Lunch Theorem (Wolpert and William, 1997), there is no golden machine learning algorithm that can outperform all the other machine learning algorithms in solving all possible problems.

In this study, we will be selecting some features that are related to cohabitation of parents to predict the academic performance of the students. The efficiency of the algorithms on the model were determined using metrics such as Accuracy, F1 score, Precision and Recall which will be implementation was carried out using Python programming language.

II. Materials And Methods

2.1 DATASET

The dataset used for this research work was adopted from the UCI machine learning repository. The dataset extracted from the repository consists of about 32 features. However, the data in table 1 below shows the considered features we will be considering which are capable of measuring the effect of cohabitation of parents on students' performance.

Table 1: Description of the Dataset

Attribute	Description (Domain)
sex	Student's Sex (binary: 'F' - female or 'M' - male)
age	Student's Age (numeric: from 15 to 22)
address	Student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	Family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	Parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
Medu	Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2- 5th to 9th grade, 3-secondary education or 4- higher education)
Fedu	Father's (numeric: 0 - none, 1 - primary education (4th grade), 2- 5th to 9th grade, 3-secondary education or 4- higher education)
Mjob	Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	Reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
rdian	Student's guardian (nominal: 'mother', 'father' or 'other')
commute_time	Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
Mjob	Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

The description of the fourteen (14) features extracted from the UCI machine learning repository data on student's performance which provided the cohabitation features we will be using for the modelling.

2.2 MACHINE LEARNING MODELS

2.2.1 Decision Tree Algorithm

Decision Tree (DT) is one of the most popular and straight forward methods of machine learning. Decision tree analysis is a predictive modelling tool that can be applied across many areas. Decision trees can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions. Decision's tress are the most powerful algorithms that falls under the category of supervised algorithms. They can be used for both classification and regression tasks. The two main entities of a tree are decision nodes, where the data is split and leaves, where the outcome is derived. The implementation of the algorithm is illustrated in figure 1 below:

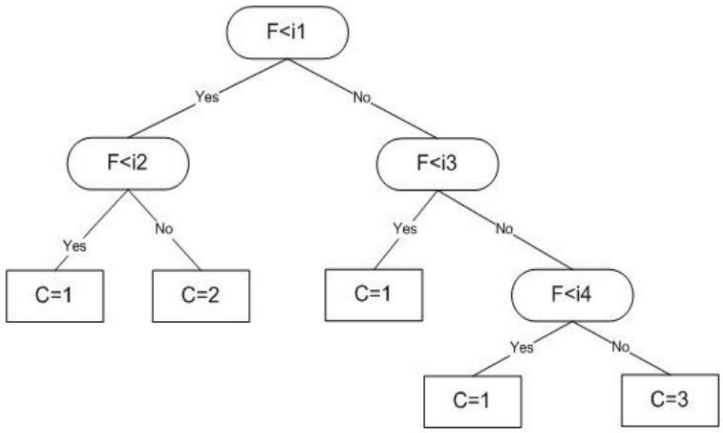

Figure 1: Implementation of Decision Tree Algorithm

2.2.2 Random Forest Algorithm

Random forest is a supervised learning algorithm which is used for both classification as well as regression. Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result. Random Forests are a aggregate of tree predictors in which every tree relies upon at the values of a random vector sampled independently with the same distribution for all trees in the forest. The fundamental precept is that a group of "weak learners" can come collectively to form a "strong learner". Random Forests are a extremely good tool for making predictions considering they do not overfit due to the law of large numbers.

2.2.3 Xgboost Algorithm

XGBoost (Chen, T., & Guestrin, C. (2016)) stands for "Extreme Gradient Boosting". XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.

One of the major features of the Xgboost model is the "Regularized learning" which helps to smooth the final learning in other to avoid overfitting. Xgboost is a faster algorithm when compared to other algorithms because of its parallel and distributed computing which is capable of computing a scalable portable and accurate library for model prediction.

2.2.4 Support Vector Classifier (Svc)

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. SVMs have their unique way of implementation as compared to other machine learning algorithms.

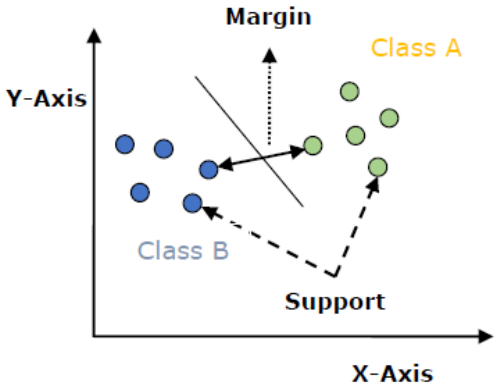

Figure 2: An illustration of Support Vector Machine/Classifier

The support vectors are the datapoints closet to the hyperplane, the hyperplane is the decision plane or space which is divided between a set of objects having different classes and the margin is the gap between two lines on the closet data points of different classes. An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) by generating hyperplanes iteratively so that the classes can be separated in best way.

2.3 DATA AUGUMENTATION METHOD

2.3.1 Synthetic Minority Oversampling Technique (SMOTE)

Over the years, researchers have applied different types of methods for data augmentation most especially in the case of imbalance data modelling and one of the most used methods are the under sampling and oversampling techniques. For the purpose of this research work we will be applying SMOTE whereby the minority classes will be oversampled.

SMOTE is an oversampling method where the synthetic samples are generated for the minority class. This algorithm enables to overcome the overfitting issues posed through random oversampling (Chawla, N. V., et al; 2002). For each of the samples in the minority class, the distance from all sample points is calculated for each sample of x_i to get its K-nearest neighbor where the sampling ratio N is set according to the sample imbalance ratio. Technically, each of the selected neighbor are been selected from the original samples as it is shown in equation 1 below.

$$x_{new} = x + rand(0,1) \times |x - x_n|, new \in 1,2,3 \dots N \quad (1)$$

By repeating the above-mentioned steps N times, to synthesize N new samples. If the minority class has a total T samples, then NT new samples can be synthesized.

III. Results

3.2 Preprocessing

The initial dataset consists of 1044 rows and about 33 features, however, the major features which will be used for the data modeling were selected from the original dataset and was presented in table 1 above with a shape of 1044 rows and 13 columns. This made the dataset we will be considering for the modeling of the effect of cohabitation status in student's final grade.

Based on the objective of the of this research work, we will be working on a classification modeling to predict if the cohabitation status in student's final grade and the final grade is a continuous variable. Thereafter, we wrote a function in python to classify the final grades that are between 0-9 to **fail** and 10-20 to **pass** which was plotted in figure 3 below.

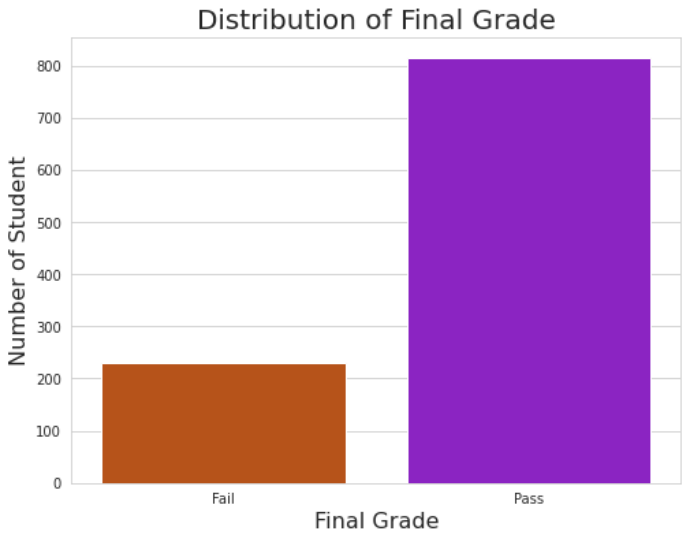

Figure 3: The distribution of the Final Grade.

We could observe that the classes for the final grade are not evenly distributed i.e imbalanced.
Checking for missing values, we found out that there exist no missing values in the dataset.
In trying to understand the data more we plotted a correlation matrix on all the continuous features in other to check the relationship between them.

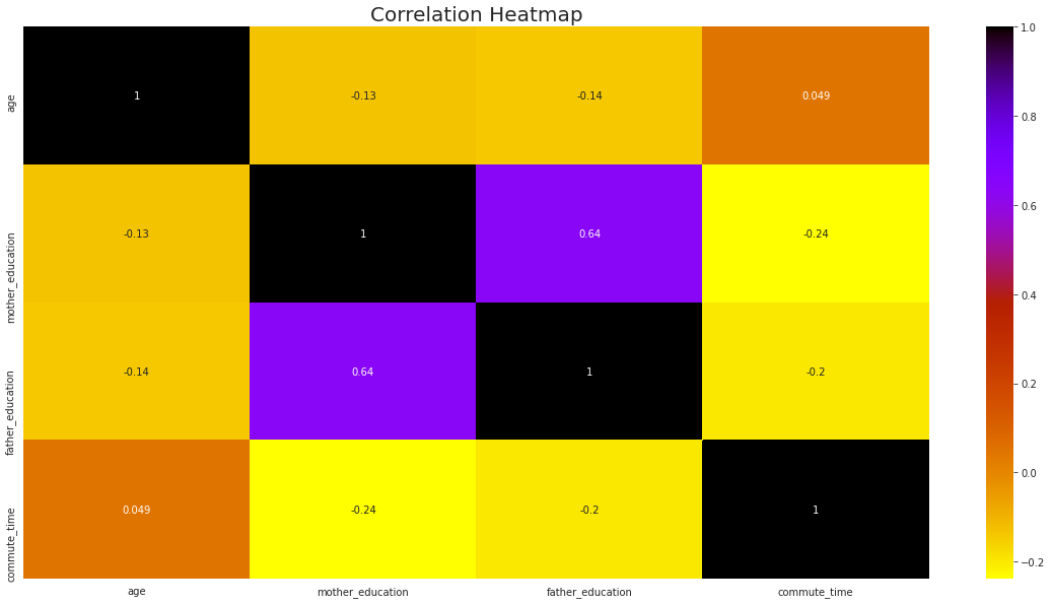

Figure 4: Correlation Map of the features.

The result of the correlation map as shown in figure 4 shows that there exists a negative relationship between mothers and father's education level on the age students and 'commte_time' i.e average number of times in minutes the students will take to get to school. The relationship between 'commute time and age' with that of 'fathers' education and mothers' education' are however positive.

3.2 Modeling Results

Since we have an imbalance dataset, we have no other choice than to apply a resampling method called **Synthetic Minority Oversampling Technique (SMOTE)** to our data in which we will be oversampling the minority class of the model and we will be applying the machine learning modeling on the resampled dataset.
After a proper data cleaning and resampling, we are left with 1628 occurrence and 12 features for modeling. However, the initial number of students that pass was 814 while those that fail are 230 but after resampling, students that pass were 814 while those that fail are also 814.

The resampled data was splitted into training (X_train and y_train) set and the test (X_test, y_test) set of 70% to 30% respectively.
The training set "X_train and y_train" was fed to the model, in other to learn the patterns in the data, thereafter, we used for the X_test for predicting the y_test.
The values of our predictions were then used to compute the metrics that was shown in table 2 and the graphical representation of the performance of the model is shown in figure 5.

Table 2: Classification Results

Classifier	Auc	Accuracy	Recall	Precision	Fscore
Random Forest Classifier	0.795	0.796	0.783	0.790	0.786
Decision Tree	0.749	0.748	0.762	0.728	0.744
Support Vector Classifier	0.675	0.677	0.634	0.674	0.654
XgBoost	0.769	0.769	0.762	0.758	0.760

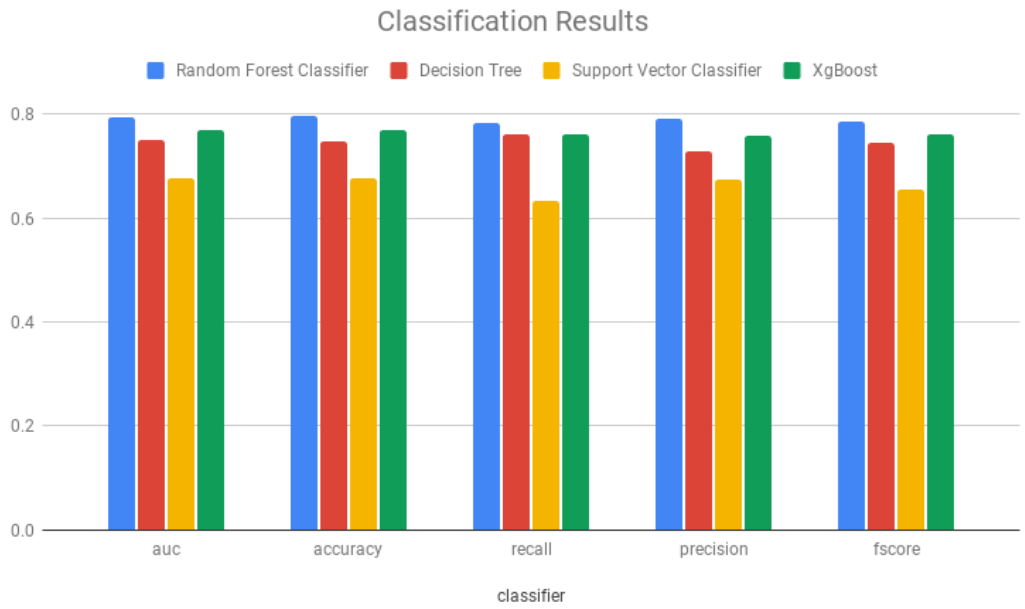

Figure 5: Classification Results

Looking at the result closely, we could see that Random forest classifier outperform other models used for the modelling of the cohabitation status modeling. As seen in table 2, we found out that Random forest has the highest accuracy of about 79.6%, precision of 79%, recall if 78.3% and an F_Score of 78.6%. With this, it is obvious that Random forest perform better than Decision tree, Support Vector Classifier and XGboost in predicting the cohabitation status of students.

For further investigation and in understanding the total number of correctly predicted class for each of the model, we plotted the confusion matrix of each of the model. As seen in figure 6, the sum of values in the leading diagonal indicates the total number of values that are correctly predicted while the sum of values from the antidiagonal shows the total number of classes that are wrongly predicted.

The result from the confusion matrix shows that the total number of correctly predicted class for, Random Forest Classifier is 389, while the correctly predicted class for Decision tree, Random Forest and Xgboost are 366, 331 and 376 respectively.

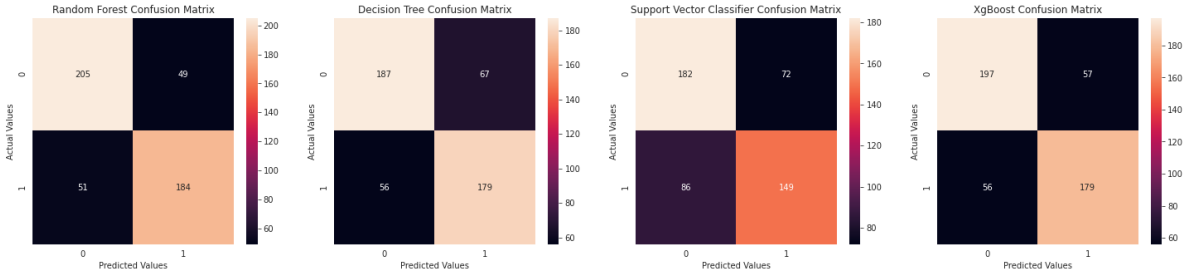

Figure 7: Confusion Matrix of the Model

The feature importance of the model as shown in figure 8. It shows that the age of the student, reason for selecting the school, mother's education, father's education mothers' job and fathers job has a higher contribution on the model than any other features.

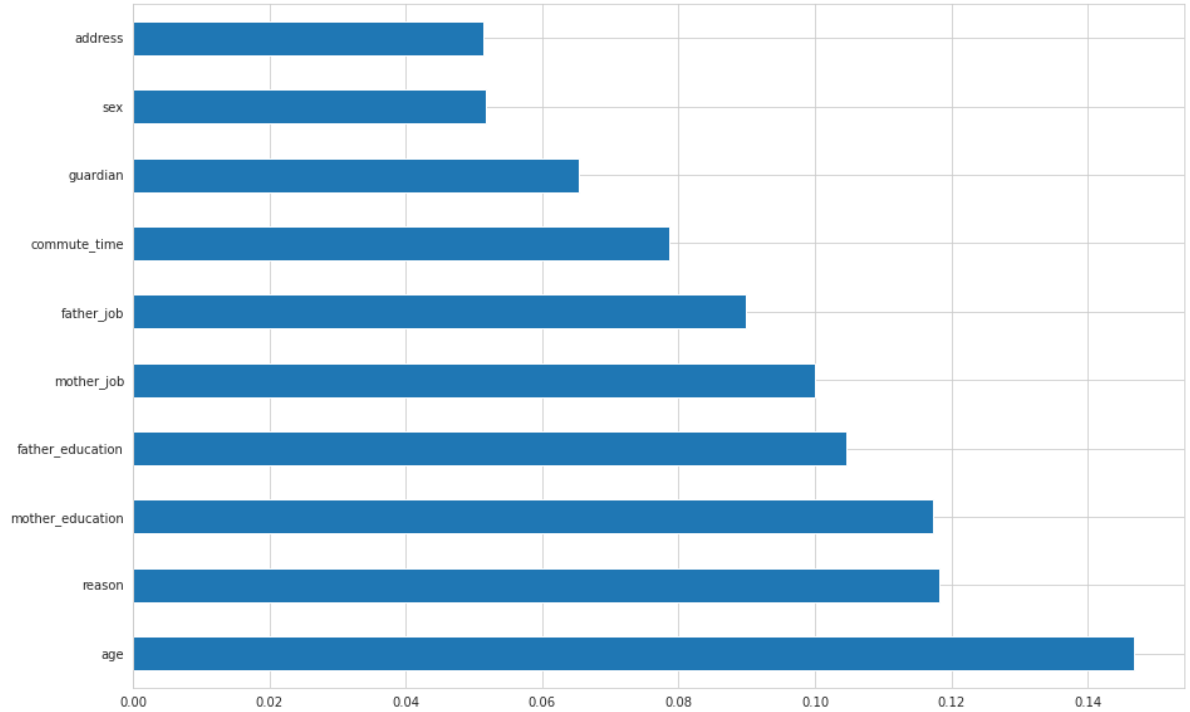

Figure 8: Feature Importance of the Model

IV. Conclusion

This paper presented the effect of cohabitation status of parents for modeling student performance. An open-source dataset containing students and parent information's was employed in this study. The cohabitation features were selected from the whole dataset which was used for the model training. However, the data was highly imbalance in other to resample the data, we applied SMOTE to resample the data.

Random forest, Decision tree, Support Vector Classifier and Xgboost were used for the data training and the performance evaluation was performed on the model using Accuracy, F1 Score, Recall and Precision for the respective machine learning algorithms.

This study has contributed to knowledge by applying machine learning algorithms to model student performance, considering various features as highlighted in the dataset.

References

[1]. Ayon D.(2016) Machine Learning Algorithm: A Review International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 1174-1179

[2]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357. https://doi.org/10.1613/jair.953

[3]. Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

[4]. Gregory Arief D. Liem (2019) Academic performance and assessment, Educational Psychology, 39:6, 705-708, DOI: 10.1080/01443410.2019.1625522

[5]. Hoi, S. C., Wang, J., & Zhao, P. (2014). Libol: A library for online learning algorithms. *Journal of MachineLearning Research*, *15*(1), 495.
[6]. Mushtag, I and Khan, S.N.(2012) Factors Affecting Students' Academic Performance, Global journal of Management and Bussiness Research, Global Journals Inc.(USA) ISSN: 2249-4588
[7]. Mahesh B.(2020) Machine Learning Algorithms-A Review, International Journal of Science and Research Vol. 9(1). 381-386. ISSN: 2319-7064
[8]. Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138.
[9]. Wolpert, David H., and William G. Macready. "No free lunch theorems for optimization." IEEE t ransactions on evolutionary computation 1.1 (1997): 67-82.

1Oduntan, Odunayo Esther, et. al. "Effect Of Cohabilitation Status Of Parents On Child's Academic Performance: A Machine Learning Approach." *IOSR Journal of Research & Method in Education (IOSR-JRME)*, 11(3), (2021): pp. 52-59.