# Comparison of Feature Selection Techniques for Predicting Student's Academic Performance

Olukoya, Bamidele Musiliu

*Ph.D Student, Federal University Oye-Ekiti, Nigeria (FUOYE)*

***Abstract:*** **In recent time, educational data mining (EDM) has received substantial considerations. Many techniques of data mining have been proposed to dig out out-of-sight knowledge in educational data. The Knowledge obtained assists the academic institutions to further enhance their process of learning and methods of passing knowledge to students. Powerful tools are required to analyze and predict the performance of students scientifically. This paper focuses on comparing two feature selection techniques in identifying major factors among the numerous affecting students' academic that could give accurate prediction. Student educational data was retrieved from Kaggle data repository and feature selection on is done by applying Information Gain Attribute Evaluator and Correlation Based Features Selection (CFS) using WEKA as an Open Source Tool. Further a comparison is made among these two feature selections algorithm to select best attributes for prediction among all.**

## I. INTRODUCTION

The progress of a country is attached to the quality of its education system. There have been tremendous changes in educational setting across the globe in its functioning (Mishra, 2014). Like any other sector, education sector is facing challenges. The major challenges faced by higher education is abysmal students' academic performance. To make the matter worse, some students leave school without completing their programs. One of the major objectives of any educational institution is to provide quality education to concerned students. Today, educational institutes and organizations seek to improve their systems by developing robust Information and Communication Technology (ICT) solutions to help the concerned managements in decisions making process. When this is done, it goes a long way to add value to the organizations objectively (Phil & Shoba, 2017).

In last decade, the number of higher education universities/institutions have proliferated manifolds. Large numbers of graduates/postgraduates are produced by them every year. Universities/Institutes may follow best of the pedagogies; but still they face the problem of dropout students, low achievers and unemployed students. Understanding and analyzing the factors for poor performance is a complex and incessant process hidden in past and present information congregated from academic performance and students' behavior. Powerful tools are required to analyze and predict the performance of students scientifically. Although, Universities, Colleges/institutions collect enormous amount of students' data before and after admitted into the school, but most of these data are unutilized. The data stored and growing in a certain period will result in the accumulation of data. This condition will be useless if it is not managed to be extracted information in it. When the student's data collected are properly analyzed with the aid of machine learning concept, it improves the teaching methods, enhances quality of teaching, identifies feeble students, identify factors that influence student's academic performance (Roy & Garg, 2017).

One of the techniques of extracting information on a large amount of data is data mining. This technique is able to find information in the form of patterns, features, or rules known as knowledge. Data mining has several methods of data processing such as, classification, regression, or clustering. Educational Data Mining (EDM) has been investigated in diverse studies such as Bendangnuksung & Prabu (2018) and Amrieh et al. (2016a) with single classifiers and ensemble framework respectively.

In order to provide better results, there are several techniques that can improve the accuracy of the results tested on data processing methods. The information gain techniques tested on the decision tree decision algorithm, Random Forest, ANN, SVM, and Naive Bayes for predictive student academic performance were able to improve the performance of the algorithm (Sari, 2016). The comparative results of the algorithm's performance before and after the feature selection show that the technique is capable of affecting the accuracy level of the machine learning classification algorithm.

Wahyuni applied the F-Score and Rough Set attributes selection techniques in improving breast cancer diagnosis (Wahyuni, 2016). From the several algorithms tested, there is a significant classification performance, but there are also algorithms that do not show any improvement in both feature selection techniques. Naive Bayes is on the best algorithm result in the diagnosis of breast cancer.

There are multiple different feature selection techniques used in Knowledge Discovery and data mining. Every method or technique has its advantages and disadvantages. Thus, this paper uses Correlation Based Features Selection (CFS) and Information Gain technique to compare and verify the results. In the end, the best result could be selected in terms of accuracy and precision.

## II. REVIEW OF THE LITERATURE

The high increase of drop out in schools motivated Gulati (2015) to investigate the drop out feature of students in

schools. At the end of the study where data mining was used, the major factors that influenced the drop out in the school were revealed. Conclusively it was submitted that demographic factors, socio-economic factors, family factors, etc. can be the reasons for dropping out of students from their various studies.

Devasia and Hegde (2016) examined the hidden cause of student's failure to get employment after leaving school. With deployment of data mining techniques, the study was able to monitor academic performance of students.

In a similar study by Dekker (2009), the study was to predict student drop-out using different classification and the cost-sensitive learning approach for different data sets. The study found out that the decision tree classifiers such as C4.5 gave better results compared to Bayes Net and Jip rule classifiers.

The embedded method as the feature selection technique using Bagging, Boosting, and Random Forest. From all experimental results, the highest accuracy is ANN using Boosting feature selection. Accuracy using behavioral features can increase as much as 22.1%, while after the implementation of feature selection, the accuracy increases 25.8%. Accuracy gained reached 80%.

The study conducted by Wati et al. (2017) was carried out to unveil the causes of poor learning attitude of students in the school. The study focused on comparison of the performances of two data mining algorithms to predict student learning based on the student records (data set). After the experiment, the result showed that average percentage of both classifiers was above 60%, whereas Naïve Bayes has higher precision average.

*Data Set*

The dataset used in this study was retrieved from Kaggle online machine learning repository and it is readily available for data mining. This educational dataset was originally collected from students through learning management system (LMS) called Kalboard 360. The dataset contains records of 480 students with 16 attributes.

Table 1.0: Overview of dataset

| S/N | Category | Attributes | Details |
|---|---|---|---|
| 1 | Demographical | Nationality | This shows the student country of origin. Kuwait = 179, Saudi Arabia = 11, Jordan = 172 ,USA = 6, Lebanon = 17, Iran = 6,Venezuela = 1, Egypt = , Tunisia = 12, Morocco = 4, Syria = 7, Palestine = 28, Iraq = 22, Libya = 6 |
| | | Gender | Student sex status [Male = 305; Female = 175] |
| | | Place of Birth | Kuwait = 180, Saudi Arabia = 16, USA =16, Jordan = 176, Lebanon = 19, Iran = 6, Venezuela = 1, Egypt = 9, Tunisia =9, Morocco = 4, Syria =6, Palestine = 10, Iraq = 22, Libya = 6. |
| | | Parent responsible for student | Father or Mummy |
| 2 | Education background | Educational Stages (School levels) | High Level: 30, Middle level: 250, and Low Level: 200. |
| | | Grade Levels | G-01 to G-12 |
| | | Section Identity | A,B,C |
| | | Semester | First or Second |
| | | Topics | Math = 21, English = 45, IT = 95, Arabic = 56, Science = 51, Quran = 22, French = 65, Biology =30, Spanish =25, Chemistry = 24, Geology =24, History = 19. |
| | | Student absence | Rated: < 7 or > 7. Student Absent Days: > 7 days: 191 < 7 days: 289 |
| 3 | Parents Participation on learning process | Parent Answering Survey | Presenting if parent answer question provided by school (Yes or No) option |
| | | Parent School Satisfaction | This seek for the parent satisfaction level from school as (Good or Bad) |
| 4 | Behavioral attribute | Discussion groups | The level of student behavior interaction with the e-learning system |
| | | Visited resources | |
| | | Raised hand on class | |
| | | Viewing announcements | |

*Concept of Feature Selection*

Feature selection is also called variable selection or attribute selection, this can be used interchangeably for the purpose of this work. Feature selection is the process of selecting a subset of the relevant features for use in model construction (Brownlee, 2016). Feature selection method helps in creating accurate prediction models. It can be used to identify and eliminate unnecessary, irrelevant and redundant attributes from data that do not contribute to the predictive model's accuracy or may even reduce model accuracy. Fewer attributes are used because they reduce model complexity, and simpler models are easier to understand and explain. Removing irrelevant features will not affect learning performance. Many people are so confused to differentiate between feature selection and feature extraction, the key difference between feature selection and extraction is that feature selection keeps a subset of the original features while feature extraction creates brand new ones. Both feature extraction and feature selection are capable of improving performance, lowering computational complexity, building better generalization models, and decreasing required storage.

According to (Sari, 2016), there are three general classes of feature selection algorithms: filter methods, wrapping methods, and embedded methods. Filter method, applying statistical measures to assign scores to each feature. Features are ranked by score and selected to be stored or deleted from the data set. Wrapping method use the pre defined learning algorithm to evaluate the performance, which will be returned to the feature search component for the next iteration of feature subset selection. The feature set with the best performance will be chosen as the final set. The Embedded method, it studied which features most were contributing to the accuracy of the model when the model was being made. The most common type of embedded feature selection method is the regularization method.

Supervised feature selection is usually used for classification tasks. The availability of the class labels allows supervised feature selection algorithms to effectively select discriminative features to distinguish samples from different classes. A general framework of supervised feature selection is shown in Fig. 1.0.

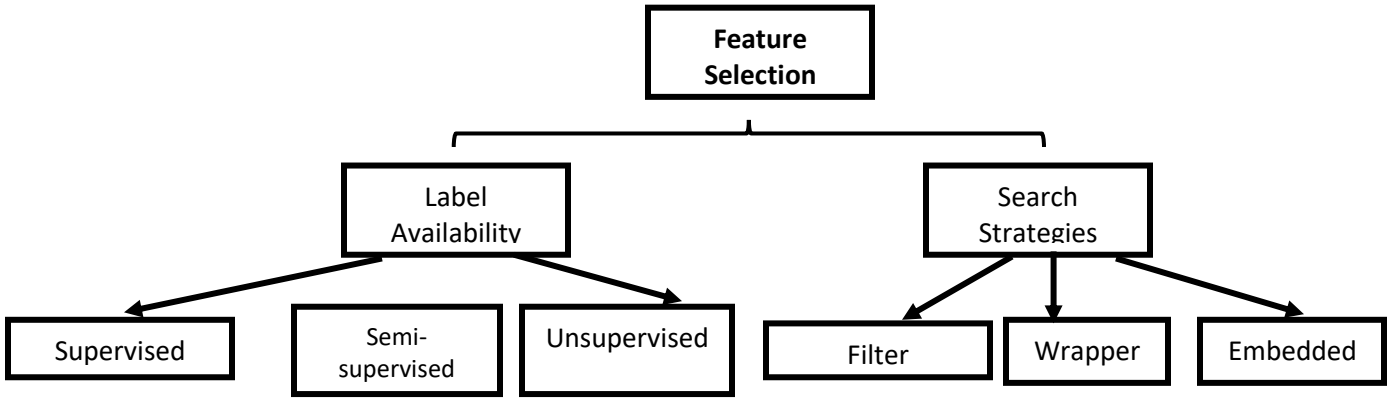


Fig 1.0: General frameworks of supervised and unsupervised feature selection.

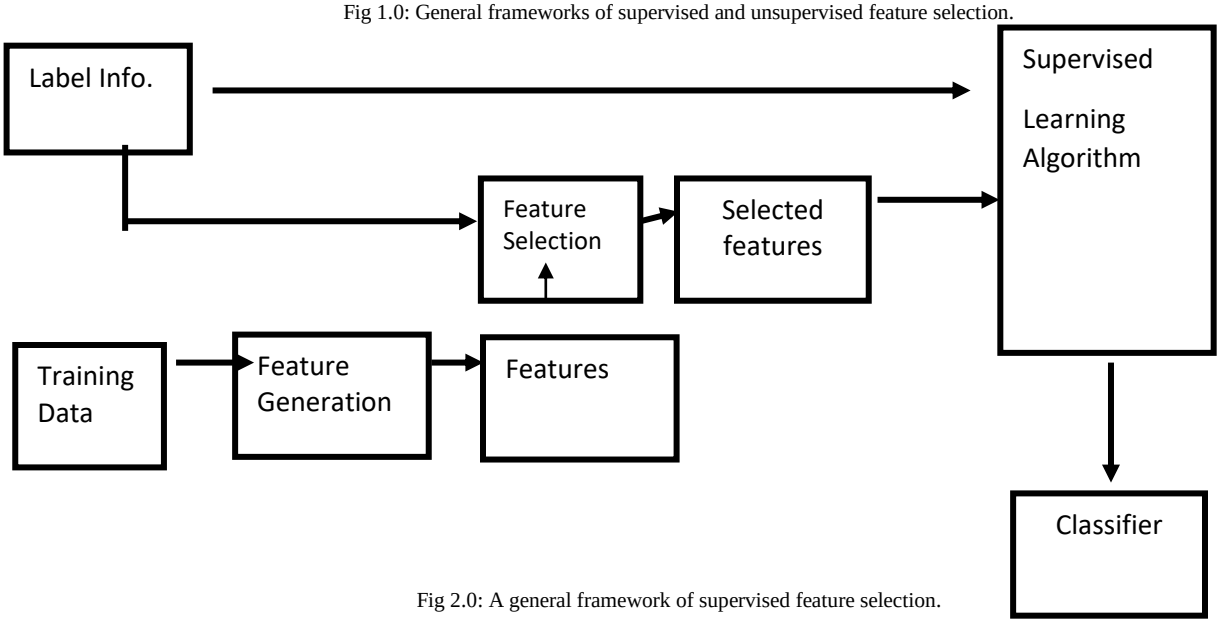


Fig 2.0: A general framework of supervised feature selection.

*Correlation Based Features Selection (CFS)*

Correlation Based Features Selection (CFS) is referred to as Correlation Attribute Evaluator on WEKA. It evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average.

*Information Gain Attribute Evaluator*

Information Gain (IG) is an entropy-based feature evaluation method, widely used in the field of machine learning. As Information Gain is used in feature selection, it is defined as the amount of information provided by the feature items for the text category. Information gain is calculated by how much of a term can be used for classification of information, in order to measure the importance of lexical items for the classification. The formula of the information gain is shown below.

$$G(D_t) = -\sum_{i=1}^{m} P(C_i)\log P(C_i) + P(t)\sum_{i=1}^{m} P(C_i \backslash t)\log P(C_i \backslash t) + P(\bar{t})\sum_{i=1}^{m} P(C_i \backslash \bar{t})\log P(C_i \backslash \bar{t}) \quad . \quad . \quad . \quad (1)$$

From Formula above, C is a set of data collection, in which there is the feature t. The value of $G(D,t)$ is greater; is more useful for the classification for C.This t should be selected. If the greater value of $G(D,t)$ is wanted, it should make the value of $P(t)$ and $P(\bar{t})$ smaller.

## III. MATERIALS AND METHODS

The methodology adopted for this study is presented in this section. This is briefly outlined in Figure 3.0 below.

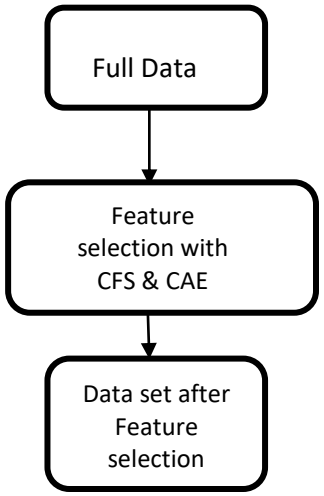


Fig 3.0: Feature Selection Model

*Development of Feature Selection*

The system used in the development of ensemble models in this research is the Waikato Environment for Knowledge Analysis (WEKA). WEKA is a machine learning system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA programming language. WEKA is a contemporary tool for developing machine learning techniques and their deployment to actual domains of interest like crime detection and so on. Many feature selection techniques and mining tasks are implemented and they are readily available for use in WEKA.


Fig 4.0: WEKA Graphical User Interface

## IV. RESULT

*Effective Features*

Correlation Based Feature Selection (CFS) is applied to decide the most important attributes in predicting student's performance. Table 2.0 presents the results of selected features for CFS. Search method of best first was used. CFS started set was with no attributes. Forward Search direction was employed in the search. Stale search after 5 node expansions was recorded. Total number of subsets evaluated stood at 121. Merit of best subset found set at 0.564.

Table 2.0: Result of CFS feature selection on the dataset

| SN | Selected Attributes |
|---|---|
| 1 | Relation |
| 2 | Raisedhands |
| 3 | VisITedResources |
| 4 | Discussion |
| 5 | ParentAnsweringSurvey |
| 6 | StudentAbsenceDays |

As shown in the table above, 6 attributes are selected as most important out of total 16 attributes. Hence, these attributes returned by this feature selection method is referred to as Students' essential features (SEF) in this study.

*Information Gain Attribute Evaluator*

Information Gain Attribute Evaluator is applied to decide the most important attributes in predicting student's performance. Table 3.0 presents the results of selected features for Info Gain Attribute Eval, it uses attribute ranking search method. Forward Search direction was employed in the search. It is obvious that Info Gain Attribute Eval considered to use all the student's attributes. There will definitely be redundancy, irrelevant and unnecessary among the attribute selected. Series of experiments will be needed to be carried out in order to have best result. In this case, the average is 8, i.e. testing from attribute 1 – 8, record the result, test again from 1 – 10, record the result, repeat this process until best and worst results are unveiled.

Table 3.0: Result of CAE feature selection on the dataset

| SN | Ranked | Selected Attributes | Initial Attribute S/N |
|---|---|---|---|
| 1 | 0.45801 | VisITedResources | 11 |
| 2 | 0.39745 | StudentAbsenceDays | 16 |
| 3 | 0.37337 | raisedhands | 10 |
| 4 | 0.2578 | AnnouncementsView | 12 |
| 5 | 0.1504 | ParentAnsweringSurvey | 14 |
| 6 | 0.12773 | NationalITy | 2 |
| 7 | 0.1261 | Relation | 9 |
| 8 | 0.12292 | PlaceofBirth | 3 |
| 9 | 0.11393 | Discussion | 13 |
| 10 | 0.10676 | ParentschoolSatisfaction | 15 |
| 11 | 0.07611 | Topic | 7 |
| 12 | 0.05178 | gender | 1 |
| 13 | 0.04748 | GradeID | 5 |
| 14 | 0.01182 | Semester | 8 |
| 15 | 0.01058 | StageID | 4 |
| 16 | 0.00703 | SectionID | 6 |

## V. CONCLUSION

From the works and result, concluded that feature selection in data mining especially Correlation Based Feature Selection (CFS) and Information Gain Feature Selection on evaluating student academic performance, Correlation Based Feature Selection is better in keeping optimal and best attributes for testing, this in turn have grater impact on the accuracy and prediction result. This paper also shows us that mental level that gained by relation, raise hands, visited resources, student's participation in group discussion, parent answering survey and student absence days have some impact on student's academic performance and graduation time.

## REFERENCES


[1] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's Academic Performance Using Ensemble Methods. *International Journal of Database Theory and Application*, *9*(8), 209–213. https://doi.org/10.14257/ijdta.2016.9.8.13

[2] Brownlee, J. (2016). "Machine Learning Algorithms," Machine Learning Mastery, Available: http://machinelearningmastery.com/. [Diakses 15 May 2017].

[3] Devasia, M. T., P, M. V. T., & Hegde, V. (2016). Prediction of Students Performance Using Educational Data Mining. In *Department of Computer Science Amrita Vishwa Vidyapeetham University,Mysuru Campus Mysuru, Karnataka, India* (p. 190).

[4] Gulati, H. (2015). Predictive Analytics Using Data Mining Technique. *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on*, 713–716.

[5] Phil, M., & Shoba, S. A. (2017). Educational data Mining & Students Performance Prediction Using SVM Techniques. *International Research Journal of Engineering and Technology (IRJET)*, *4*(8), 1248–1254.

[6] Roy, S., & Garg, A. (2017). Predicting Academic Performance of Student Using Classification Techniques. In *4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics* (Pp. 568–572).

[7] Sari, B. N. (2016). "Implementasi Teknik SeleksiFitur Information Gain pada AlgoritmaKlasifikasi Machine Learning untukPrediksi Performa AkademikSiswa," dalam Seminar Nasional TeknologiInformasi dan Multimedia, Yogyakarta.

[8] Tama,B. A. (2015). Learning to Prevent Inactive Student of Indonesia Open University. *Journal of Information Processing Systems (JIPS)*, *11*(2), 165–172.

[9] Wahyuni, E. S. (2016) "PenerapanMetodeSeleksiFituruntukMeningkatkan Hasil Diagnosis KankerPayudara*," Journal SIMETRIS*, vol. 7, pp. 283-294, 2016.

[10] Wati, M., Indrawan, W., Widians, J. A., & Puspitasari, N. (2017). Data Mining For Predicting Students ' Learning Result. In *Dept. of Computer Science and Information Technology Universitas Mulawarman Samarinda, Indonesia* (p. 28).