Journal of Theoretical and Applied Information Technology
15th January 2019. Vol.97. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EDM PREPROCESSING AND HYBRID FEATURE SELECTION FOR IMPROVING CLASSIFICATION ACCURACY

[1] SAJA TAHA AHMED, [2] PROF. DR. RAFAH SHIHAB AL-HAMDANI, [3]DR. MUAYAD SADIK CROOCK

[1]Informatics Institute of Postgraduate Studies, University of Information Technology and Communication, Baghdad, Iraq

[2]Informatics Institute of Postgraduate Studies, University of Information Technology and Communication, Baghdad, Iraq

[3]Computer Engineering Department, University of Technology, Baghdad, Iraq
[1]sajataha@ymail.com, [2]Rafah_hamdani@yahoo.com, [3]muayadkrook@yahoo.com

ABSTRACT

Educational Data Mining (EMD) is in charge of discovering useful information from educational datasets. In recent years, the data is mounting rapidly due to the ease access to the websites of e-learning intakes extraordinary enthusiasm from different colleges and instructive foundation. High dimensionality, irrelevant, redundant and noisy dataset can affect the knowledge discovery during the training phase in a bad way as well as degrading machine learning performance accuracy. All these factors often rise demand for dataset preparation, analysis, and feature selection. The fundamental aim of research is to enhance the precision of classification by information preprocessing and expel the unessential information without discarding any vital data by means of feature selection.This paper proposes EDM dataset preprocessing, and hybrid feature selection method by combining filter and wrappers techniques. In the filter-based feature selection, the statistical analysis is based on the Pearson correlation and information gain. In the wrapper method, the accuracy of the feature subset is tested using a neural network as a baseline algorithm. The obtained results show an enhancement in performance accuracy toward selecting minimum feature subset with high predictive power over using all features.

Keywords: *Educational Data Mining, Hybrid Feature Selection, Neural Network, Data Preprocessing, Accuracy.*

1. INTRODUCTION

Modern education comes on the scene rapidly in the ongoing years. The internet has contributed significantly to the advancement of e-learning systems. This leads to encouraging numerous universities and educational institutions to adopt a web-based educational system to prompt the expansion in the volume of students information. In this context, Educational Data Mining (EMD) has emerged as a new field of research to investigate educational data in order to determine different instructive research issues, such as identifying successful students of a given course and recognizing students who can drop out or fail for more attention within course headway [1].

In general, EDM helps in early student's performance prediction for enhancing teaching, learning and decision making. Such that it has the ability to explore, envision and analyze a huge amount of student's dataset. Therefore, it determines factors that can influence a student's academic performance and extract useful patterns of student's learning habits.

Researches have appeared that no single learning approach is remarkably predominant in altogether cases. In addition, extraordinary learning calculations regularly create comparable outcomes[2]. The nature of the data, utilized to recognize the task to be learned, is only the issue that can have a colossal effect on the achievement of a learning algorithm. If the data fails to acquire statistical regularity, the learning can fail. It is possible that data is transformed to a new data exhibits valuable patterns to facilitate learning, but the difficulty of this task is the intractability of a fully automatic method. If data has useful

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

regularities, then it can be suitable for machine learning with less time-consuming [3].

Preprocessing step is an important stage in DM that includes cleaning, transformation, reduction and discritization[4]. It also takes a substantial amount of processing time to do an attribute importance analysis. Feature engineering is the process of mutating dataset to attributes that best to describe machine learning problem, enhance the predictive power of model to unseen data and reduced execution time. Feature selection is successfully taken into account if the dimensionality of the data is reduced and the accuracy of a learning algorithm improves or remains the equivalent [4]

This research aims to explain the dataset preparation steps essential for a machine learning algorithm and analyze the most relevant subset features to get high-performance accuracy. This is done by introducing the hybrid feature selection method (i.e embedded method) which fuses the following approach:

i. Filter method: it is based on two criteria correlation feature selection(i.e Pearson correlation) and information-gain attribute evaluation,
ii. Wrapper method: it uses a neural network as a basis for comparing the effects of feature subset selection.

The proposed mechanism is examined by two datasets: Iraqi student performance, and UCI student performance. The experimental results show that the smaller dataset has the ability to improve prediction accuracy comparable with all features as inputs to the algorithm (i.e. no feature selection).

2. RELATED WORK

The research problem of EDM preprocessing and feature selection can be presented through various directions. Within the current literature, a brief presentation of variety approaches offered a baseline for such issue is introduced.

In [6], the most relevant features subset was proposed to give the highest accuracy subset investigated using six filtered feature selection techniques, and evaluated feature subset in terms of F-measures and Receiver Operating Characteristics (ROC) values. These values produced by the NaïveBayes algorithm as base-line classifier method. The outcomes were confirmed that there is a reduction in processing time and constructional cost for both training and classification phases with minimum feature subset. For the early academic failure prediction of rural college students in the introductory courses, the authors of [7] proposed a hybrid feature selection approach using filter methods and ensemble classifiers based on a voting technique. In [8], the proposed method preprocessed and analyzed student performance with a new type of features concerned to the interactivity with Kalboard 360 e-learning system, called behavioral features. The classification was done using NN, NB, and DT. The conducted results showed that a strong relationship existed between student behaviors and his scholarly accomplishment. In addition, there was an improvement up to 29% with the various classification algorithms accuracy based on behavioral features, the last two paper, were not clear up on which filter based feature selection method was done and without indicating features ranking to proof feature significant degree with target class. In [9], feature selection was applied based on Correlation-based Feature Selection (CFS), Chi-Squared Attribute Evaluation, Information Gain Attribute Evaluation, and Relief attribute evaluation, compared the accuracy of different feature selection based on classification algorithms. In particular; J48, Naïve Bayes, Bayes Net, IBk, OneR, and JRip were adopted and the obtained results appeared that IBK had the highest accuracy with CFS subset evaluator as compared with other classifiers. In addition, CFS subset evaluator had high accuracy than other feature selection algorithms.

3. DATA PREPROCESSING AND HYBRID FEATURE SELECTION

EDM problem passes through many stages in order to improve the quality of the data set and solve the machine learning problem. In the proposed model the processing steps include:

- Gathering data.
- Exploring data.
- Encoding data
- Normalization
- Feature selection.
- Defining model.
- Training, testing model and predicting the output.

The proposed model consists of two basic stages:

- Data preparation includes steps (i-v),
- The model building includes the steps (vi-vii).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The overall framework for generating an optimal feature subset is illustrated in Figure 1 as a workflow. The features selection deals with the encoding dataset, while the dataset normalization is an important process before training and testing data, fed into the neural network. The following subsections explain each processing step in more details.

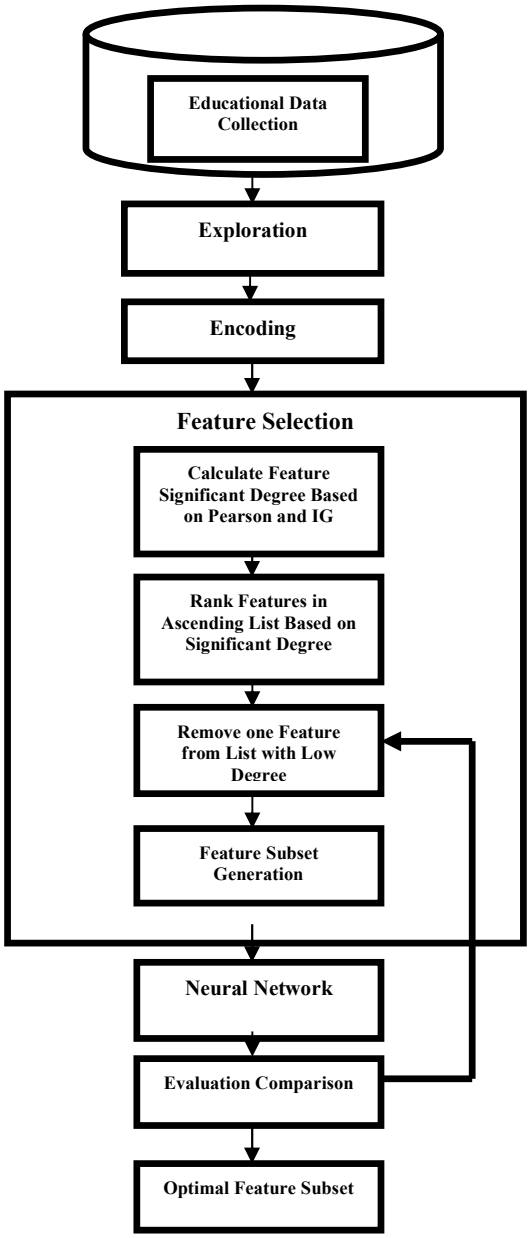

Figure 1: Workflow of the Overall Framework

The proposed EDM preprocessing is compared to the previously published work [7]; this work has the following differences:

- The proposed framework manages datasets incorporate understudies in city and ruler.
- Feature selection mechanisms involve two criteria based on IG and Pearson correlation.
- All features are tested and compared in ranking list.
- Evaluation based on accuracy of neural network.

3.1 Dataset Collection and Description

As mention earlier, this study incorporates two data sets. The first dataset is called Iraqi dataset that is collected through applying (or submitting) questionnaire in three Iraqi secondary schools for both applicable and biology branches of the final stage during the second semester of the 2018 year. Initially, the questionnaire was contained 56 questions in three A4 sheets and it was answered in a class by 250 students (samples). Latter, 130 samples are discarded due to lack of information since pre-processing is applied to obtain the most complete information of students. After removing inconsistencies and incompleteness in the dataset, this study considers 120 samples instances with 55 features for experiment purposes. The features are distributed into five main categories: Demographic, Economic, Educational, Time, and Marks. Table 1shows the dataset's attributes/features and their description. As illustrated in this table, new features are introduced, such as holiday and worrying effects. The relationships between parents with schools and use of books and references by the student are also considered.

Table 1: Iraqi Dataset

Feature Category	#Question	Feature	Description
Demographic	Q0	Gender	Binary (Female ,Male)
	Q1	Social status	Nominal (single, married, apart)
	Q2	Age	Numeric (1:<17, 2:17-19, 3:19-21, 4:>21 year)
	Q3	Governorate	Binary (Baghdad, other)
	Q4	Living	Binary (City, Rural)
	Q5	Mother education	Ordinal (Illiterate, Medium, Secondary, B.A., Higher)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	Q6	Father education	Ordinal (Illiterate, Medium, Secondary, B.A., Higher)
	Q7	Family member Education	Binary (Yes, No)
	Q8	Father Alive	Binary (Yes, No)
	Q9	Mother Alive	Binary (Yes, No)
	Q10	Family Size	Numeric (0:<4,1:4-8, 2:>8 member)
	Q11	Parent Apart	Binary(Yes, No)
	Q12	The Guardian	Nominal (mother, father, other)
	Q13	Family Relationship	Ordinal (Bad, Good, Vgood, Excellent)
Economic	Q14	Father Job	Nominal (No, Employee, Other)
	Q15	Mother Job	Nominal (No, Employee, Other)
	Q16	Education Fee	Binary (You, Family)
	Q17	Secondary Job	Binary (Free job, No)
	Q18	Home Ownership	Binary (Own, Rent)
	Q19	Study Room	Binary (Yes, No)
	Q20	Family Economic Level	Ordinal (Poor, Good, Vgood, Excellent)
	Q21	You chronic disease	Binary (Yes, No)
	Q22	Family Chronic Disease	Binary (Yes, No)
Educational	Q23	Specialization	Binary (Applicable, Biologist)
	Q24	Study willing	Binary (Yes, No)
	Q25	Reason of study	Nominal(You, Average, Family)
	Q26	Attendance	Ordinal (Poor, Good, Vgood)
	Q27	Failure Year	Binary (Yes, No)
	Q28	Higher Education Willing	Binary (Yes, No)
	Q29	References Usage	Binary (Yes, No)
Time	Q30	Internet Usage	Numeric (0:<2, 1:2-4, 2:>4 hour)
	Q31	TV Usage	Numeric (0:<2, 1:2-4, 2:>4 hour)
	Q32	Sleep Hour	Numeric (0:<5, 1:5-7, 2:7-9, 3:>9 hour)
	Q33	Study Hour	Numeric (0:>2, 1:2-4, 2:4-6, 3:>6 hour)
	Q34	Arrival Time	Numeric (0:<hour, 1: other)
	Q35	Transport	Binary (Foot, Car)
	Q36	Holiday Effect	Binary (Yes, No)
	Q37	Worry Effect	Binary (Yes, No)
	Q38	Parent Meeting	Binary (Yes, No)
Marks	Q39-Q45	Materials Degrees for First Semester	Numeric (0-100)
	Q46	Avg1	Numeric (0-100)
	Q47-Q53	Materials Degrees for Second Semester	Numeric (0-100)
	Q54	Avg2	Numeric (0-100)

The second dataset (Student Alcohol Consumption Data Set), obtained from UCI Portugal [10] is used in this study. This data set was collected during the 2005-2006 year from two public schools depending on two sources: school reports for the three-period grades and number of school absences, and questionnaires. The dataset consists of two classes: student-mat.csv (Math course which holds 395 instances) and student-por.csv (Portuguese language course which holds 659 instances). Both of these datasets, consisting of 32 attributes, are shown in Table 2.

Table 2: UCI Machine Learning Dataset

#Question	Features	Description
Q0	school	student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
Q1	sex	student's sex (binary: 'F' - female or 'M' - male)
Q2	age	student's age (numeric: from 15 to 22)
Q3	address	student's home address type (binary: 'U' - urban or 'R' - rural)
Q4	famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Q5	Pstatus	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Q6	Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
Q7	Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
Q8	Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Q9	Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Q10	reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
Q11	guardian	student's guardian (nominal: 'mother', 'father' or 'other')
Q12	traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
Q13	studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
Q14	failures	number of past class failures (numeric: n if 1<=n<3, else 4)
Q15	schoolsup	extra educational support (binary: yes or no)
Q16	famsup	family educational support (binary: yes or no)
Q17	paid	- extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
Q18	activities	extra-curricular activities (binary: yes or no)
Q19	nursery	attended nursery school (binary: yes or no)
Q20	higher	wants to take higher education (binary: yes or no)
Q21	internet	Internet access at home (binary: yes or no)
Q22	romantic	with a romantic relationship (binary: yes or no)
Q23	famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
Q24	freetime	free time after school (numeric: from 1 - very low to 5 - very high)
Q25	goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Q26	Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Q27	Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
Q28	health	current health status (numeric: from 1 - very bad to 5 - very good)
Q29	absences	number of school absences (numeric: from 0 to 93)
Q30	G1	first-period grade (numeric: from 0 to 20)
Q31	G2	second period grade (numeric: from 0 to 20)
Q32	G3	final grade (numeric: from 0 to 20, output target)

3.2 Dataset Exploration

Dataset can be described in statistical mode, in addition, it can be visualized using graphical diagrams. This is a remarkable data mining pre-processing step, and it helps in the understanding dataset before progressing furthermore in a complex data mining processing. In order to catch characteristics and relationships amongst dataset attributes, the relationship between the GENDER and two classes of the first-semester average (Success and Fail) can be taken as an example. The Iraqi dataset contains some statistics about GENDER. Figure 2 views the values of GENDER attribute in graphical representation since there are 36 male and 84 female.

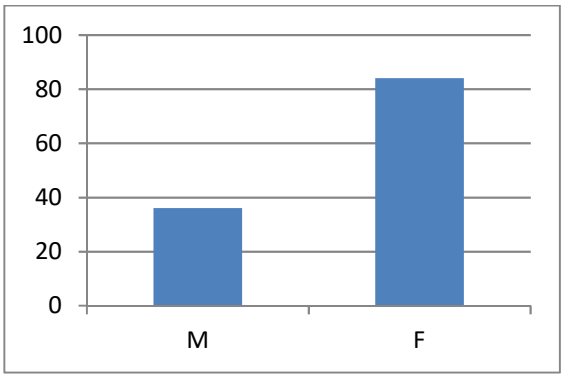

Figure2: Histogram of GENDER Attributes.

The frequency of two target classes, success and fail students (first-semester average), with respect to gender can be shown in Figure 3. The statistical graph of Figure 3 computes the success and failure rate for female 89%, 11 %, and male 69%, 31 %, respectively.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

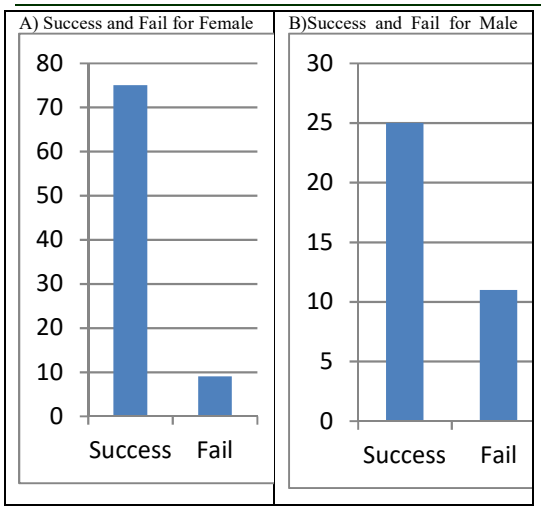

Figure 3: Gender Frequency with Respect to Target Classes.

3.3 Dataset Encoding

Some of the machine learning algorithms need the data to be in the numerical formulation, such as a neural network. While the others require the data to be in the categorical formulation, such as decision tree. There is no informal definition for Feature engineering, but it can be summarized as raw data transformation into attributes that are better described the machine learning problem and resulting in enhanced performance accuracy. The dataset contains features of various data types, for example: Binary, Interval, Numeric and Categorical (Nominal, Ordinal). In addition, there are many feature encoding methods for transforming categorical data to numeric ones, such as label encoding or integer encoding, one hot encoding, binarized and hashing. In this research, the two datasets are encoded using Label Encoder, which is the most straightforward method to transform categorical features into numerical labels. Numerical labels are always between 0 and (#attribute_value-1). The proposed label encoder algorithm follows these steps:

Input: CSV file
Output: Code matrix (0…#student,0…#attributes)

1. **Open a window to select CSV file name, and then open this file as an array of string dataset [student, attributes]**
2. **Build lookup table for all attributes keywords in CSV file as follows:**

```
Get attributes type into attribute_type
For each attribute in the dataset
 If  attribute_type is  ordinal, then
   Set ord_cat to the ordinal category
     For each ord_cat
Set c to zero
       For each student in dataset
 If dataset [student, attribute] not equal to ord_cat,
 then  Set lookup[attribute, c] to ord_cat
       Increment c by one
      Else continue
Else Ifattribute_type is  nominal, then
       Set nom_cat to the nominal category
      Sort nom_cat in an alphabetic ascending order
     Set c to zero
    For each student in the dataset
 If the dataset [student, attribute] not equal to
nom_cat,  then
      Set lookup [attribute, c] to nom_cat
      Increment c by one
      Else continue
Else If  attribute_type is  Binary, then
       Set bin_cat to binary category
       For each student in dataset
       If dataset [student, attribute] equal to
       bin_catp[0], then
       Set lookup [0, attribute] to zero
        If dataset [student, attribute] equal to
       bin_catp[1], then
         Set lookup [1, attribute] to one
Else Ifattribute_type is  Integer, then
      For each student in the dataset
     Set Lookup [student, attribute] to the dataset
     [student, attribute]
```

3. **Build code array from the lookup table, as follows:**

```
For each attribute
For each student
 Set n to zero
While (lookup [n, student] not equal to NULL)
   If dataset [student, attribute] equal to lookup [n,
     student], then
         Set  [student, attribute] to n, break.
  Else  Increment n by one
End while
End for
End for
```

3.4 Dataset Normalization

The experience has proven that the machine learning algorithm is efficiently trained well when datasets are normalized, or scaled, which leads to a better predictor and speed up processing or training. Normalization is the process of scaling attribute values within a specific range (such as 0 to 1), in a manner that all attributes have approximately similar magnitudes. This research normalizes the attribute values using Min-Max normalization at the [-1, 1] range as illustrated in Equation (1) [16]:

$$z = \left\{\frac{x - min(x) * (h - l)}{max(x) - min(x)}\right\} + l \qquad (1),$$

Where x is feature value in a dataset, $min(x)$ and $max(x)$ is the smallest and largest value in that feature, h and l is a high and low boundary, respectively. According to experimental tests and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

implementation, dataset normalization has important issues to be noted. Normalization considers extra overhead; thus datasets must not have few instances with fewer features. In addition, if a desirable care is no longer taken, the dataset may lose the internal structure which leads to lower accuracy.

3.5. Feature Selection

In general, feature selection is the process of identifying a subset of features, considered as highly correlated to the target class, even if these features uncorrelated to each other. A feature subset, which contributes to the predictive task and has the ability to enhance performance accuracy, is finally chosen as independent variables to machine learning algorithm. This is to predict the dependent variable or class label.

Features selection method can be in one of three categories: filter based feature selection, Wrapper, and Embedded. Filter approaches find out the individual strongest feature related to specific class label (i.e. discover inherent relationship) using statistical information. On the other hand, the wrapper approach evaluates the worth of features using the learning algorithm, embedded or hybrid feature selection that combines two earlier methods [11], [12]. Table 3 shows a comparison between filter and wrapper approaches. The proposed hybrid feature selection mechanism occupies advantages of both the filters and the wrappers methods, in such a way that it guarantees the obtaining of optimal feature subset in terms of accuracy, with reasonable computation complexity.

Table 3: Comparison of Filter and Wrapper Methods

Filter based feature selection	**Wrapper-based feature selection**
It does not depend on any performance criteria (i.e accuracy).	It uses a classification accuracy criteria during training based on performance evaluation.
It does not guarantee that the selected feature subset to be the most significant in terms of performance and accuracy.	It ensures that the selected subset could achieve a better predictive performance, in terms of, the final classification accuracy.
It is computationally less overhead.	It is computationally expensive and may be uncompromising for a big dataset with huge attributes.

Filter-based feature selection uses the specific measurements to recognize irrelevant attributes and filter out them by predictive power. The proposed model computes two statistical measurements that are: Pearson correlation and information gain for each feature with the target class. Then, the selected metrics are ranked by their feature scores.

Pearson's correlation coefficient is also known statistical models as the R-value. It computes a value that indicates the strength of the correlation for any two features. Pearson's correlation coefficient is computed by taking the covariance of two features and dividing by the product of their standard deviations. The coefficient is not affected by changing the scale of the two features. Correlation between features ***x*** and y is computed using equation (2) [13], as follows:

$$r_{xy} = \frac{\sum(xi- \ )(yi- \ )}{\sqrt{\sum(xi-x)^2(yi-y)^2}} = \frac{SXSY}{\sqrt{SXSY}} \qquad (2)$$

Correlation matrix amongst features in datasets is computed using the following algorithm, written in pseudo-code:

```
Input: code matrix [student, attributes]
Output: correlation matrix [attributes ,
 attributes]
For each attribute in code
Set sum to zero
For each student in code
    Add  sum into code [attribute, student]
Compute mean by dividing sum into
 student
Set i to zero
Set k to i+1
While i is less than attributes -1
While k is less than attributes
  Set sum to zero
   Set sx to zero
   Set sy to zero
   For each student in code
     Set x to code [student, i]- mean[i]
      Set y to code [student, k]-mean[k]
      Increment sum with x*y
      Increment sx with x^2
      Increment sy with y^2
Compute correlation matrix [i,k] dividing
sum by square root of sx*sy.
```

When the value of ***R*** is near 1, then the relationship between variables is positively correlated. If the ***R*** is 0, then there is no correlation. While if R-value is near to -1, then features are negatively correlated. For example, in the student-mat.csv dataset, the Pearson correlation between G1(first-period grade) and G3(final grade) is 0.801468 and between

Journal of Theoretical and Applied Information Technology
15th January 2019. Vol.97. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

G2(second-period grade) and G3is 0.904868, indicates that G1, G2, G3 are strongly correlated. Figure 4 illustrates the relationship between G3 and G2. In addition, ***R*** between absence and G3 is 0.034247, signs to poorly correlated features.

In the information theory, the entropy measures the unpredictability or impurity properties. Entropy has value ranges between 0 and 1 and can be calculated using equation 3 [14]:

$$H(S) = -\sum_{I=0}^{n} p(i) log p(i) \qquad (3)$$

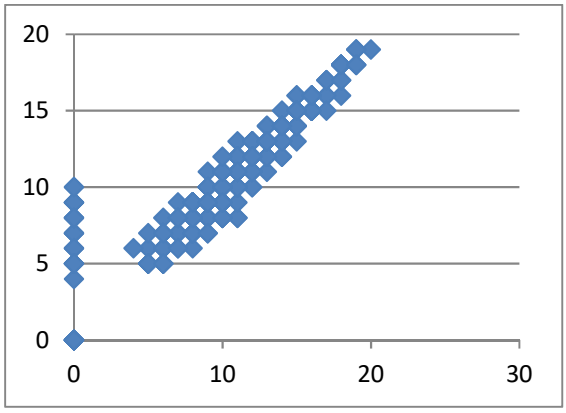

Figure 4: G1 and G2 Correlation.

Information gain is the reduction in the entropy, in the proposed model, information gain used for ranking the features. Usually, a feature with high information gain should be ranked higher than other features because it has stronger power in classifying the data. Information gain is computed using equation 4 [14]:

$$IG(F) = H(S) - \sum_I \frac{Si}{s} H(Si) \qquad (4)$$

An attribute is pertinent if and only if removing it from a feature set can adversely affect the prediction power of the feature set [15]. This implies a relevant feature has useful information that can be led to more predictive classification. In this research attributes sorted in ascending order according to their Pearson and information gain. Step by step, the proposed system can prune one feature with less merit from the feature set to observe which feature subset results in high-performance accuracy. Information gain can be computed following the algorithm, shown in pseudo code as follows:

Step1: Input: Compute target class put it in result [student]
Output: gain [attribute]
Set success to zero
Set failure to zero
For each student
If result [student] equal to true, then increment success by one
Else increment failure by one
Goto step2 to calculate entropy for the result or the goal, set to goal_entropy
Step2: // entropy
Set total to success + failure
Set ratio_success to success divided by tota
Set ratio_failure to failure divided by total
If ratio_success not equal to zero, then
Set ratio_successto -(ratio_success) * Log2(ratio_success)
If ratio_failure not equal to zero, then
Set ratio_failure to -(ratio_failure) * Log2(ratio_failure)
Set entropy to ratio_success + ratio_failure
Step 3: // information gain
For each attributes
Set n to zero
While lookup [student, attribute] not equal to NULL // calculate the number of categories
for each attribute from lockup table
Increment n by one
For each n
For each student in code
If dataset [student, attribute] equal to lookup [n, attribute] and result [student] equal to true, then increment success [n] by one
Else if dataset[student, attribute] equal to lookup[n, attribute] and result[student] equal to false, then increment failure [n] by one
Calculate Entropy [n] by going to step2
Increment sum with - (success [n] + failure [n]) divided by r * Entropy [n]
Set Gain [attribute] to goal _entropy increment with sum

4. EXPERIMENTS AND RESULTS

The experiments and the application system in this study are developed based on visual studio C# 2010. Moreover, the neural network is created using ENCOG library based on BasicLayer and BasicNetwork objects. The network has an input layer of several neurons depending on the number

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of the selected feature. A hidden layer with sigmoid activation function is adopted and an output layer of one neuron represents a target class.

Because neural networks always begin with random values (weight values, learning rate,..), very distinct outcomes show up from two runs of the similar program. Some random weights grant a better starting point than others. Sometimes random weights can be far enough off that the network can fail to learn. For this situation, the weights ought to be randomized again and the procedure restarted [16]. In this study, and for the purpose of results comparison among datasets, all initial weights are randomized between [-1,1] and stored in a CSV file to be used as initial weights at each neural network initialization. These weights are refined to values that can provide the desired output via the training phase.

The datasets are dividing into two groups: 70% training and 30% test. The proposed study focuses on hybrid feature selection techniques to obtain advantages of both and to overcome disability of them. Filter method is one of the most important and frequently used one in data preprocessing for data mining. Both the Pearson correlation and information gain are used to analyze the two datasets. The examined features are ranked in a list according to their merits in ascending order. At each iteration, the proposed system prunes one least significant feature from the list and observes the performance of the neural network in terms of its accuracy. The selected feature subset that results in high accuracy can be considered as an optimal feature subset for further classification.

The evaluation on the basis of Accuracy (ACC) value is executed repetitively on the subsets ranked in the list. It starts from all features ending with two most significant features by eliminating one feature from subset in each iteration. Accuracy measures the degree to which the instances correctly classified by machine learning algorithm and can be computed using a confusion matrix with equation (5) as follows [8]:

$$ACC = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ Population} \quad (5)$$

Confusion matrix of Iraqi dataset,student-pro.csv and student-mat.csv can be illustrated in Tables of 4, 5, and 6 based on the Pearson correlation for iteration 1. It has all the features as input, and the achieved accuracies are 0.72, 0.72, 0.83, respectively.

Table 4: Iraqi Dataset Confusion Matrix

Total Population=36		**Actual Calss**	
		SUCCESS	**FAIL**
Prediction Class	SUCCESS	26	4
	FAIL	6	0

Table 5: Student-pro.csv Confusion Matrix

Total Population=195		**Actual Calss**	
		SUCCESS	**FAIL**
Prediction Class	SUCCESS	135	52
	FAIL	1	7

Table 6: Student-mat.csv Confusion Matrix

Total Population=119		**Actual Calss**	
		SUCCESS	**FAIL**
Prediction Class	SUCCESS	72	10
	FAIL	10	27

From Tables 7, and 8, it can be observed that no agreement is found among features ranked methods since each method sorts features of three datasets differently. Partially exception is evaluated, where Q30 and Q31 have the highest degree of significance in terms of both the Pearson correlation and IG of both UCI datasets(math and pro).

Table 7: Features are Ranked Based on the Pearson Correlation

#	0	1	2	3	4	5	6	7	8	9	10
Iraqi	19	0	18	16	30	2	28	9	4	22	29
Math	14	2	25	22	12	15	11	28	5	26	27
Pro	14	0	26	27	1	12	24	2	28	29	22
#	11	12	13	14	15	16	17	18	19	20	21
Iraqi	32	3	35	15	14	34	31	38	1	21	7
Math	0	16	24	18	29	9	23	19	4	13	21
Pro	25	11	15	17	5	19	4	9	16	18	23
#	22	23	24	25	26	27	28	29	30	31	32
Iraqi	10	36	11	12	5	24	13	6	26	37	20
Math	17	8	1	3	10	7	20	6	30	31	
Pro	10	8	21	3	7	6	13	20	30	31	
#	33	34	35	36	37	38					
Irqi	25	27	8	17	33	23					

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 8: Features are Ranked Based on Information Gain.

#	0	1	2	3	4	5	6	7	8	9	10
Iraqi	7	22	36	24	35	15	11	38	34	29	3
Math	19	18	0	4	5	3	12	16	21	27	1
Pro	5	19	15	16	18	17	4	12	9	1	22
#	11	12	13	14	15	16	17	18	19	20	21
Iraqi	32	28	21	1	9	13	10	31	14	8	5
Math	23	9	17	13	28	22	15	24	10	11	8
Pro	28	21	11	27	24	3	26	8	25	2	23
#	22	23	24	25	26	27	28	29	30	31	32
Iraqi	16	37	26	2	27	4	12	6	19	17	23
Math	26	6	7	20	25	2	14	29	30	31	
Pro	10	13	6	7	29	20	0	14	30	31	
#	33	34	35	36	37	38					
Iraqi	18	20	0	25	33	30					

Neural network accuracy of the student-mat.csv dataset for both Pearson and IG versus a number of attributes is illustrated in Figure 5. There is partially harmonize in terms of selected attributes (feature subset), which give the highest accuracy. In general, it can be shown that the accuracy of NN is enhanced with less number of attributes when a significant degree of Pearson and IG are strongly related to the target class.

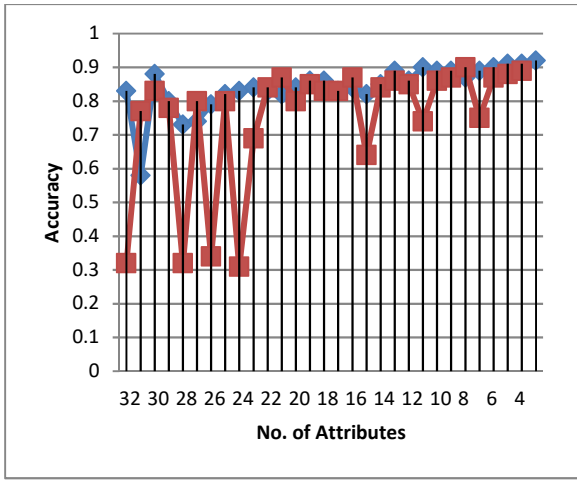

Figure 5: Pearson and IG Based Accuracy vs. Attributes Number

Table 9 explains the highest accuracy with optimal features subset. The results obviously show that the social factors in combination with marks are obtained the highest performance accuracy. Iraqi dataset prediction is based on all category of attributes without any previous marks to forecast the first-semester average. While UCI dataset includes a combination of the first period and second-period average along with all factors to predict final average. In this research, even if the adopted subset encompasses a larger number of features, it gives the same outcomes compared with less number feature subset. The feature subset, with the minimum number of attributes, is chosen to be an optimal subset. Optimal features subset gives the highest accuracy with less number of features. For example, the Iraqi dataset has a seven combinations of feature subsets (Q37 Worry Effect, Q20 Family Economic Level, Q25 Reason of study, Q27 Failure Year , Q8 Father Alive , Q17 Secondary Job , Q33 Study Hours, and Q23 Specialization) which give similar accuracy as optimal subset (Q23 Specialization, and Q33 Study Hours). Finally, it must be noticed that the obtained results in this research are confined by selecting initial random weights and other parameters of the neural network, so terrible decisions of NN parameters can constrain came about precision.

5. CONCLUSIONS

This research presented module that prepared the EDM dataset based on the collection, exploring, encoding, normalization, and feature selection of two sets of data. The attribute significance analysis findings had been indicated as a very important factor in recognizing the features that contribute to a degree of prediction accuracy and can help to support early decision making for poorly students performance. The hybrid feature selection was proposed to evaluate features by measuring Pearson correlation and information gain with respect to the target class. The proposed method ranked the features of the adopted datasets in an ascending order. In addition, it removed features with a less significant degree, and finally evaluated features subset after each pruning based on neural network. Based on the results of the optimal subset, shown in Table 9, it can be inferred that information gain and Pearson correlation on student-mat.csv gave optimal features subset which has 0.92 accuracies with three and two feature subsets. Therefore, the proposed mechanism successfully chose minimum features subset with improving the accuracy of classification over using all features.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 9: Accuracy with Optimum Feature Subset

Datasets	Highest ACC	Pearson	Highest Acc	IG
Iraqi	88	2 (Q23Specialization, Q33 Study Hours)	88	2 (Q30 Internet Usage, Q33 Study Hours)
Math	92	2,3 (G1,G2, Mjob)	92	2 (G1,G2)
Pro	82	3 (G1,G2, INTERNET)	83	3 (G1,G2, STUDYTIME)

REFERENCES

[1] Anal Acharya, Devadatta Sinha, "Application of Feature Selection Methods in Educational Data Mining", International Journal of Computer Applications, Vol.103, No.2, 2014.

[2] Lior Rokach, "Decomposition methodology for classification tasks: a meta decomposer framework", Springer link, Vol. 9, Issue 2–3, pp 257–271, 2006.

[3] C. J. Thornton., "Techniques in Computational Learning", Chapman and Hall, London, 1992.

[4] Jiawei Han, MichelineKamber, Jian Pei, "Data Mining Concepts and Techniques, 3rd edition, Elsevier Inc, 2012.

[5] Mark A. Hall, "Correlation-based Feature Selection for Machine Learning", Ph.D. thesis, The University of Waikato, 1999.

[6] M. Ramaswami and R. Bhaskaran, "A Study on Feature Selection Techniques in Educational Data Mining", Journal of Computing, Vol. 1, No. 1, 2009.

[7] R. Krithiga, Dr.E.Ilavarasan, "A Hybrid Feature Selection Based Framework for Early Prediction of Rural College Students Academic Failure in the Introductory Courses", International Journal of Pure and Applied Mathematics, Vol. 115, No. 6, 2017.

[8] Elaf Abu Amrieh, ThairHamtini, Ibrahim Aljarah, "Preprocessing and Analyzing Educational Data Set Using X-API for Improving Student's Performance", IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2015.

[9] T.Velmurugan, C. Anuradha, "Performance Evaluation of Feature Selection Algorithms in Educational Data Mining", International Journal of Data Mining Techniques and Applications, Vol. 05, No. 2, 2016.

[10] P. Cortez and A. Silva.," Using Data Mining to Predict Secondary School Student Performance", The 5th Future Business Technology Conference (FUBUTEC 2008), pp. 5-12, Porto, 2008.

[11] Hui-Huang Hsu et al, "Hybrid Feature Selection by Combining Filters and Wrappers", Expert Systems with Applications, Vol. 38, No. 7, 2011.

[12] Ivan Kojadinovic, Thomas Wottka, "Comparison between a filter and a wrapper approach to variable subset selection in regression problems", ESIT 2000, Germany, 2000.

[13]A. G. Asuero, A. Sayago, and A. G. Gonz´alez, "The Correlation Coefficient: An Overview", Critical Reviews in Analytical Chemistry, 2006.

[14]Bangsheng Sui," Information Gain Feature Selection based on Feature Interactions", Master thesis, Department of Computer Science, University of Houston, 2013.

[15] Kilho Shin, Tetsuji Kuboyama, Takako Hashimoto and Dave Shepard, " SCWC/SLCC: Highly Scalable Feature Selection Algorithms", MDPI, information 8, 159; doi: 10.3390, 2017.

[16] Jeff Heaton, "Programming Neural Networks with Encog3 in Java", Heaton Research Inc., 2011.