CIT. Journal of Computing and Information Technology, Vol. 28, No. 3, September 2020, 165–181
doi: 10.20532/cit.2020.1004996

A WGFS Based Approach to Extract Factors Influencing the Marketing of Korean Language in GCC

Luai Al-Shalabi and Yousra Tahhan

Arab Open University, Kuwait

This research proposed an approach that is intended to determine the minimal set of important factors that influence the desire of learning Korean language in the Gulf Cooperation Council (GCC). Those factors will then influence marketing of the Korean language in GCC by guiding interested people to increase their commercial abilities, improve their information about Korean drama, and prepare them to study or travel to Korea. A total of 500 responses out of 526 questionnaires were used for the analysis process. Merging the weight by SVM and the weight guided feature selection (WGFS) techniques were proposed to build a strong hybrid model of reduction for the investigated dataset. Five different classifiers were used to test the results. Empirical results have showed that the generated factors (the reduct) are very significant to test the ability/inability of learning the Korean language. SVM was shown as the best with accuracy value of 94%. This research contributed to the literature by highlighting the importance of the Korean language in the GCC and by presenting the important factors that influence learners of the Korean language: encouragements and obstacles. Moreover, current research presented the best classifier which yields to the high performance of classification.

ACM CCS (2012) Classification: Computing methodologies → Machine learning → Classification

Keywords: Korean language, machine learning, feature selection, WGFS, weights, classification

1. Introduction

Learning languages is important to motivate human minds and keep them energetic and healthy. Specifically, Korean language is nowadays attracting particular interest, because of a number of reasons we state below. Authors in [1] and [2] stated that "Korean language is attractive to many people, especially the females". They also stated that "many people like to learn the Korean language for different reasons as their interests". Korea is already one of the strongest economies in Asia and these trends continue. With companies like Samsung, LG, and Hyundai, Korea is the 13th largest economy in the world [3]. Korean dramas have become famous in the GCC, and this is another reason for people, especially the teenagers, to learn the Korean language. The MBC channel is the first channel in the GCC that views a Korean TV series. Korea has become one of the destinations that tourists like to visit, hence knowing Korean can definitely be helpful and give them the opportunity to speak with business owners, neighbors, and tour guides who will help them find their way around this country and make them feel more comfortable. Other reasons could also make this language important if students intend to complete their higher study in Korea as it is a technologically advanced country. In the literature, very limited studies are conducted in the Korean language such as in [4], [5] and, to our knowledge, there are no studies linking the Korean language and the GCC. Hence the incentive to highlight the importance of Korean language in the GCC and in other Arab countries, and to find the obstacles its learners may face. In this sense, information technology and especially machine learning come as very helpful.

Machine learning techniques are used to form data. They apply mathematical algorithms and approaches to structural and unstructured datasets. Many machine-learning methods are used to extract information and knowledge such as classification, clustering and regression. Each method uses its own algorithms to create a proper model for the given dataset. Specifically, the classification method is used to map the records of a dataset into predefined goals; practicing and understanding different classification algorithms helps in building a strong application for a given task. Dimensionality reductions are other machine learning techniques used to discover the significant features of a dataset that influence the classification process. Classification and feature selection techniques are used in this research in order to build a classification model that is able to evaluate the appropriateness for a person to learn the Korean language. They are also used to show the obstacles that may hamper the learner.

The main contributions of this article are summarized below:

- presenting a literature review about the Korean language;
- identifying the main factors (attributes) that influence the wish of GCC people to learn Korean language using the proposed approach;
- generating a classification model that may determine, by identifying common problems and complications, the positive category of people who can learn the Korean language without problems and complications and the negative category of people who will face them;
- adopting deep analysis for the important factors identified (the reduct) and their influence on the ability/inability of learning the Korean language in the GCC. This is done by studying the behavior of participants;
- recommendations will be introduced to overcome the difficulties some people may face during the process of learning Korean language.

2. Literature Review

Motivation to learn languages is regularly studied and measured [5]. *E.g.*, Shin [6] claimed that "the reasons for studying foreign language may be for cultural, educational and practical purposes". The U.S. Department of Education [7] and the U.S. Department of Education and Office of Postsecondary Education [8] stated that "The Korean language has long been nationally considered as one of the major critical languages". In its report, MLA [9] discussed the growth of interest in studying Korean language around the world concluding that it has been steadily on the rise since 1970. Goldberg *et al.* [10] found that the growth of the Korean enrollments is 44.7% , which is the highest amongst the top 15 foreign languages in the world between the years 2009 and 2013.

Much research has been conducted to discover the reasons for studying the Korean language. Results show that the main reason is the interest in Korean culture [11], [12], [13]. According to [4], Korean drama is another reason for studying the Korean language, as it has been touching the teenagers' personality through physical, emotional, and mental state. It has been noted that Korean dramas are undeniably addictive [14], while having the biggest impact in the culture, etiquette, and mannerisms shown [15].

Other researches found that many people love to watch Korean dramas [1], and especially women [2] (the same results can be found in [16]). The most frequent factors motivating people to learn Korean language thus identified include [17]: interesting, important language, future career benefits, and knowledge of Korean heritage.

In the continuation, a review is provided on the machine learning techniques used in our research. It has been noticed in [18] that machine learning algorithms were built and designed to match human intellect by learning from the surroundings. Hence, machine learning techniques have been applied effectively in different fields such as medicine, education, pattern recognition, astronomy, computer vision, finance, and entertainment, learning from previous experiences in order to develop future performance. *E.g.*, in their research [19], Beam and Kohane claimed that machine learning could change

most aspects of modern life. Based on human search history, they showed that Google recognizes what people need to know, while Netflix recognizes what kind of movies people prefer to watch. They also stated [*ibid.*] that Google has started to replace much of its non-machine learning technology with machine learning algorithms. In supervised machine learning, the training dataset which consists of conditional as well as classification attributes is trained and a model is generated, while the result can be used to map a new object to the matching diagnosis [20]. Furthermore, in [21] many machine learning methods are reviewed that have been employed to improve performance of cancer modeling. On the other hand, NRC focuses in its report [22] on classification predictive techniques and presents a predictive machine learning approach that uses known attribute(s) to predict unknown value.

Classification is defined as a supervised technique whose class label is known in advance [23]. Effectively, data classification is a two-step process. In the first step, a model is built by analyzing the data tuples from the training data having a set of attributes. For each tuple in the training data, the value of the class label attribute is known. In the second step, the generated model is applied to the testing dataset in order to test its accuracy. The model can be used to classify new tuples if its accuracy percentage is acceptable. The classification was used in learning [24], medicine [25] and crime [26].

This research is aimed to study the key factors that affect the ability/inability of learning the Korean language. These factors were modeled to evaluate the ability/inability of a new learner to learn the Koran language. The model saves the learner's time and effort by highlighting the obstacles that he/she may face during the learning process. It minimizes the learning time by recommending the resources needed to improve the learner's ability to learn the Korean language. This allows the learner to participate in different segments of life which may need the Korean language.

3. Methodology

The methodology used to study the problem of identifying the key factors affecting the ability of learning Korean language consists of different steps: data collection, data preprocessing, classification model, and evaluation measurements. The following subsections explain the methodology in details.

3.1. Data Collection

This research utilized an online questionnaire to evaluate the desire to learn the Korean language in GCC. Designing a suitable questionnaire is important to collect representative datasets. It is an oriented questionnaire that consists of different categories and questions. The first category consists of demographic questions (related to gender, age, and ethnicity group). The second category contains questions that can measure the degree of desire to learn the Korean language (such as: Why are you learning the Korean language? Do you think that the Korean language is a popular language nowadays? How much do you know about the Korean language? Would you recommend others to learn the Korean language?). The third category consists of questions about the difficulties that may be faced by learners of the Korean language (such as: Do you face difficulties during learning the Korean language? Do you face difficulties in finding adequate resources? Do you face difficulties in how to start learning?). Additionally, some interviews were undertaken with Korean language learners. The learners were asked about the difficulties they faced during the learning process. The purpose of the questionnaire and the interviews was to highlight these difficulties in order to provide recommendations for new learners. The number of participants in this questionnaire was 523 but, as we will see in the next subsection, it dropped down to 500 after removing all the records with missing values. Each participant had to answer 37 questions.

A representative analysis of the dataset is presented below.

From the 500 participants, 216 (43.2%) preferred to learn the Korean language rather than other languages such as Chinese, French, and Turkish. Of these, 111 (51.4%) preferred it as a hobby, 38 (17.6%) preferred it in order to learn about the Korean culture, and only 11 (5.1%) needed it for travel. Many other participants

who do not prefer the Korean language still wish to learn it. We focus on the study of the reasons behind this desire. Overall, out of 500 responses, 340 (68%) expressed the desire to learn the Korean language. The following results are related to these 340 participants:

- participants with good, fair, or poor knowledge about the Republic of Korea are 167 (49.1%), 127 (37.4%), and 46 (13.5%) respectively. Out of the 500 participants, 296 (59.2%) showed a significant desire to learn Korean language with reference to their knowledge about the Republic of Korea. Results show that the Korean language is rising to an important position among other languages known in GCC;
- participants with good, fair, or poor knowledge about the Korean culture are 196 (57.6%), 100 (29.4%), and 40 (11.8%) respectively. Out of the 500 participants, 296 (59.2%) showed a significant desire to learn Korean language with reference to their knowledge about Korean culture. This is a very good indication that the Korean language has become important in the GCC and we must take important steps to simplify learning processes and provide better materials;
- participants with good, fair, or poor knowledge about the Korean language are 137 (40.3%), 126 (37.1%), and 77 (22.6%) respectively. Out of the 500 participants, 263 (52.6%) showed a significant desire to learn Korean language with references to their knowledge about the Korean language. This is also a very good indication that the Korean language has become important in the GCC;
- of these 340 participants, 209 (61.5%) recommend others to study the Korean language.

Out of the 500 responses, the following results were also obtained:

- participants having no difficulty with knowing their real studying level in the Korean language, 65 persons (13%), experience about five times less difficulty in learning Korean than those with such difficulty, 283 persons (56.6%). There should be a procedure to help students realize their studying level before they start learning. Such procedure would help to minimize students' efforts and learning time;
- participants having no difficulty as to where to start learning, 67 persons (13.4%), experience about four times less difficulty in learning Korean than those with such difficulty, 285 persons (57%). The starting point is the key as to where to start learning the Korean language and a high percentage of the participants face problems with that. Relevant institutions should have some way to attract students in the first shot;
- participants having no difficulty with finding adequate learning resources, 64 persons (12.8%), experience about four times less difficulty in learning Korean than those with such difficulty, 284 persons (56.8%). This implies that lack of resources is one of the significant factors hindering spread of the Korean language in the GCC. Steps should be taken to solve this problem;
- participants having no difficulty with knowing the Korean culture, 121 persons (24.2%), experience about two times less difficulty in learning Korean language than those with the difficulty in understanding Korean's culture, 284 persons (56.8%). Short lessons about Korean culture before starting to learn the Korean language is recommended.

3.2. Data Preprocessing

Data preprocessing is translation of data from undesired shape or form into usable and desired shape or form. It is the way to formulate data for machine learning. It includes different procedures such as data cleansing, treating missing data, encoding data, and implementing dimensionality reduction. In order to avoid wrong results, data in the dataset must be sensibly tested, processed, and treated before the training process is employed. The data in our original dataset (ODS) was tested and cleaned by removing all noise such as errors and redundancy. Treating missing data and generating the best reduct of the original dataset are explained below. The cleaned dataset (CDS) which is free of errors,

redundancy, and missing data will be used as input to the dimensionality reduction process.

3.2.1. Missing Data

Datasets usually contain missing data because the data either does not exist or is unknown. To resolve this problem, many methods exist in the literature such as inputting the missing data or deleting the objects that have missing data. In this research the latter method was used because the number of records with missing data was small (only 23 records with the percentage almost equal to 4.4%) and they would not seriously affect the accuracy.

3.2.2. Encoding Data

Most of machine-learning algorithms need numbers to work with. Collected data can be categorical or numerical. Categorical data is the data that generally takes a limited number of possible numerical or textual values. Since our data is categorical, it is advised to preprocess it before feeding it to the machine learning algorithm.

The simplest coding of categorical data could be done by what we call the normal coding (NormC) way using 1, 2, 3, ... for different levels. Another way is to use dummy variables coding (DummC), which is also called (one-hot encoding). For the dummy variables, 1 is used if a particular observation is true and 0 otherwise. For example, if one of the categorical variables is Salary which takes one of the three values: High, Average, or Low, one-hot encoding interprets that as three separate dummy variables where each is given a value of 1 or 0. Table 1 shows the results. Both ways mentioned here will be tested against the Korean dataset and the classifiers used in this research.

Table 1. Dummy variables.

High	Average	Low
1	0	0
0	1	0
0	0	1

The main benefit of encoding categorical variables is model efficiency. As every researcher knows, there is no best machine learning algorithm for all datasets. One algorithm could work best for one dataset but underperform for another, maybe similar one. The nature of the dataset is one of the reasons. For this, we do not agree that the use of normal coding (1, 2, 3, …) for categorical variables is completely wrong or that it does not improve the model efficiency. It is good to test the dataset using different ways of coding models and then decide which one is suitable to use based on the efficiency it provides. In the end, we are looking for a model of high performance.

3.2.3. Feature Selection

The size of data has become larger in the past few years. Therefore, machine learning methods face serious challenges when trying to process such data. Feature selection is one of the most common and important techniques in data preprocessing. It works on removing noisy data, deletes the data which does not contain important information, and removes redundant features [27]. Feature selection can be performed in many ways, out of which we used WGFS (Weight Guided Feature Selection) in this study. WGFS uses input attribute weights to determine the order of features added to the feature set. The highest weight features are the primary to be added to the feature set. The algorithm stops if adding the last n features does not increase the performance or if all features were added [28]. Cross-validation is used to estimate how accurately a model will perform in practice. As for performance evaluation, 10-fold cross-validation of a learning scheme was used. Regarding model selection, we used SVM as the learner with WGFS in order to have a performance metric that can stop the algorithm. Namely, SVM provides good results for many learning tasks, also being a fast algorithm, and additionally works with both linear and quadratic functions.

Weight by SVM (WSVM) and weight by information gain (WIG) were separately applied as a preprocessing step to the WGFS method. Each of them calculates the input weights and then feeds them to the WGFS algorithm. We compare these two methods and then decide which

one to use in our proposed approach, based on their performance. Below we explain each of these weights methods and describe how we feed their results into the WGFS.

Weight by SVM. SVM was developed by Vapnik and others in 1963 [29]. The Weight by SVM operator uses the coefficients of the normal vector of a linear SVM as attribute weights. The attributes with higher weight are considered more relevant.

The input is the dataset samples including the input sample attributes x_1, x_2, ..., x_n, and the output result y whereas the output is the set of weights w. There will be one weight (w_i) for each attribute whose linear combination predicts the value of y.

To calculate the weight, the following formulas will be used:

$$w \cdot x + b = 0, \tag{1}$$

where w is a weight vector, x is the input vector, and b is the bias. Using the classification value, we may rewrite it as:

$$w \cdot x + b = y_i, \tag{2}$$

where y_i is the classification (= 1 or −1).

α is the contribution of the i^{th} training sample to the final solution w. Higher value of α means that the sample has a higher contribution to the weight vector. The weight formula will finally be:

$$w = \sum_{i=1}^{n} (y_i \alpha_i x_i). \tag{3}$$

For the NormC dataset, each attribute has one weight whereas for the DummC dataset several weights will be generated for each attribute which is represented by several other dummy attributes (one weight for each dummy attribute). The average of all these weights was calculated and considered as the weight for the original attribute. The reduct was generated two times based on the shape of the Korean dataset: one time as the input dataset is coded by NormC and the other time when it is coded by DummC. The reduct of the Korean dataset when NormC was used is (Q10, Q14, Q17,and Q21) whereas it is (Q14 and Q21) when DummC was used. Those questions are the most important and significant ones that affect the decision of learners of the Korean language using WGFS with weight by SVM for NormC and DummC datasets, as shown in Table 2 and Table 3 respectively. Notice that Q14 and Q21 represent the intersection between the two reducts. So, we called the set of {Q14, Q21} the WSVM (NormC, DummyC)_core. It represents the high quality of these two questions when we use a different shape of the dataset (NormC, DummyC) but apply the same technique (WSVM).

Weight by Information Gain. Information theory was developed by Claude Shannon [30, 31]. It provides means to derive good methods for deciding how relevant a particular attributive is. In our context its application helps find the amount of information gained about a random attribute(the weight of the attribute) with respect to the class attribute. Also, the weight by information gain ratio (IGR) is used for generating attribute weights which we will use here.The higher the weight of an attribute, the more relevant it is considered.

In order to find the above ratio, we start by calculating the entropy value (E) of all data (which

Table 2. WGFS reduct of NormC dataset using WSVM.

Question	Description
Q10	Why people prefer to learn the Korean language?
Q14	What reasons make people attracted to the Korean language?
Q17	People know/do not know about Korean culture.
Q21	Would you recommend others to learn the Korean language?

Table 3. WGFS reduct of DummC dataset using WSVM.

Question	Description
Q14	What reasons make people attracted to the Korean language?
Q21	Would you recommend others to learn the Korean language?

Table 4. WGFS reduct of NormC dataset using WIG.

Question	Description
Q9	Do you know the basics of the Korean Language?
Q10	Why people, in general, prefer to learn the Korean language?
Q11	What is your specific reason to learn the Korean language?
Q13	What are the difficulties you may face if you want to learn the Korean language?
Q14	What reasons make people attracted to the Korean language?

is a way to measure impurity) based on the sum of the entropy value of the result of the dataset (DS):

$$E(DS) = \sum_{i=1}^{n} -p(a_i) \cdot \log_2 p(a_i) \qquad (4)$$

where DS is the set of n values and $p(a_i)$ is the probability of getting the i^{th} value when randomly selecting one from the set. Then, the IG value of each attribute is calculated based on the reduction of the overall entropy value of the selector by the amount of data value of each alternative multiplied by the entropy value of each result such that:

$$IG(DS, A) = E(DS) - \sum_{i=1}^{n} \left(\frac{|DS_i|}{|DS|} \right) \cdot E(DS_i) \quad (5)$$

where A is subset attribute, $|DS_i|$ is the size of the subset of the dataset that belongs to the attribute on A partition $-i$, and $|DS|$ is the number of sample in the dataset. After that, split information (SI) of each attribute is calculated based on the number of values of each alternate attribute.

$$SI_A(DS) = -\sum_{i=1}^{v} \left(\frac{|DS_j|}{|DS|} \right) \cdot \log_2 \left(\frac{|DS_j|}{|DS|} \right) \quad (6)$$

where v is the number of partition attributes A, $|DS_j|$ is the size of the subset of the dataset belonging to the attribute on A partition $-j$ and $|DS|$ is the number of sample size in the dataset. Finally, the IGR value is calculated for each attribute by dividing the value of IG with SI.

$$IGR(A) = \frac{IG(A)}{SI(A)} \qquad (7)$$

The reduct of the NormC dataset which was generated by WGFS using WIG is (Q9, Q10, Q11, Q13, and Q14). Such a reduct contains the most important and significant questions that affect the decision of a learner of the Korean language and they are shown in Table 4. The core of the WGFS reduct of the NormC dataset using WSVM and WGFS reduct of the NormC dataset using WIG is {Q10, Q14} and is called WSVM_WIG(NormC), whereas the core of the WGFS reduct of the DummC dataset using WSVM and the WGFS reduct of the NormC dataset using WIG is {Q14} and is called WSVM(DummC)_WIG(NormC). The core {Q10, Q14} shows great importance of Q10 and Q14 when the same shape of the dataset (NormC) is used but the different techniques (WSVM and WIG) are applied. The core {Q14} shows great importance of Q14 when the different shapes of the dataset (DummC and NormC) are used but the different techniques (WSVM and WIG) are applied. Among all the previous cores, the most important question is Q14 which is part of all of them.

3.3. Classification Models Used

The goal of the classification is to correctly predict the target category in each case for the data. This is done by training a dataset and then building a classification model for the dataset. There are different important classifiers for machine learning. However, we focused on five of them as explained below.

Decision Tree method. This is a supervised learning method usually used for tasks such as clustering, classification, and regression. Each node represents a test on an attribute value. The leaves represent classes that can predict classification models. The branches display chances of attributes which go to classes. Input to a decision tree is the set of records described by the set of fields creating yes/no output decision [32]. Tree keeps dividing the training dataset into root and leaf nodes partitions until the entire dataset has been considered.

Logistic Regression. This is also a supervised machine learning algorithm that is used to determine whether a variable supports a specific result or not. It usually answers the kind of yes or no questions by analyzing a dataset. It has many types such as ordinal, multinomial, binary, or binomial [33].

Deep Learning. This is a machine learning technique that introduces its importance and prominence in different fields of interest such as robotics, artificial intelligence, NLP, and image classification and recognition. Its importance comes from its capabilities of providing better performance for the given problem and from its easy way of solving problems. In 2016 and 2017, Kaggle (the online community of data scientists and machine learners) was controlled by gradient boosting and deep learning techniques that guide researchers to be successful in applied machine learning [34]. More information can be found in [35].

Gradient Boosted Trees. This method is a nonlinear regression method that is used to improve the accuracy of the sequentially generated trees. As stated in Natekin and [36], "while boosting trees increases their accuracy, it also decreases speed and human interpretability". They also explained that gradient boosted trees are suitable for structured data and they are called "shallow learning" because they use only two layers.

Support Vector Machines (SVMs). This is a well-known classification technique that divides data into at least two columns by the identified hyper-plane. Classification of the points is correct if the points are far from the hyper-plane. This technique works well for small data of less than 1000 objects [37].

3.4. Evaluation Measures

It is essential to determinethe best classifierfor ahe given problem. Accuracy could not be sufficient measurement to assess the classification results. Al-Shalabi [38] explained the importance of other measures for the classification quality including recall and precision.

Accuracy measurement is calculated by the summation of True Negative (TN) and True Positive objects (TP) divided by the total number of objects in the dataset. It is a ratio of correctly classified objects to the total objects.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

The second measurement is called recall. It is defined as the number of TP objects divided by the summation of TP and False Negative objects (FN). It is the capacity of the classifier to return positive objects.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

The last metric is precision. It is defined as the number of TP objects divided by the summation of TP and False Positive objects (FP). It is the capacity of the classifier to classify negative objects as negative and not as positive.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

4. The Proposed Approach

In this section, the proposed approach will be explored by listing its steps which should be executed sequentially. These steps are:

1. Collect the dataset for investigation which is the original dataset (ODS); it is the Korean dataset for our case.
2. Clean the dataset by removing the noise which is represented by the duplicate rows or missing values.
3. Use normal coding (NormC) to code all categorical attributes of the cleaned dataset.
4. Apply the weight by SVM method to the coded dataset for calculating the weight of each attribute.
5. Input the weights to the WGFS method in order to find the dataset reduct (RDS).
6. Find the performance of the reduct by applying each of the mentioned classifiers individually.
7. Compare the performance results.
8. Choose the one with highest performance and name it as the most suitable one for the given dataset.
9. Use the chosen classifier to build the classification model which will be used to classify the ability/inability of learning the Korean language.

The algorithm was constructed from the proposed approach as shown below.

Algorithm 1. Proposed algorithm.

Step 1.	**for** each column in the ODS Ci, i: = 1. . p
Step 2.	**for** each row R_j,j: = 1. . m
Step 3.	**if** (no value exist)
Step 4.	Delete the row
Step 4.	**else**
Step 4.	Perform NormC
Step 5.	ODS:= the cleaned coded dataset
Step 9.	**for** each column in the ODS C_i, i: = 1. . p
Step 10.	Calculate WSVM[i]:= W_{Ci}
Step 11.	Compute RDS RDS:= WGFS(WSVM[i..p])
Step 12.	Generate the classification model CM:= Train(RDS)

The steps of the proposed approach are shown in Figure 1.

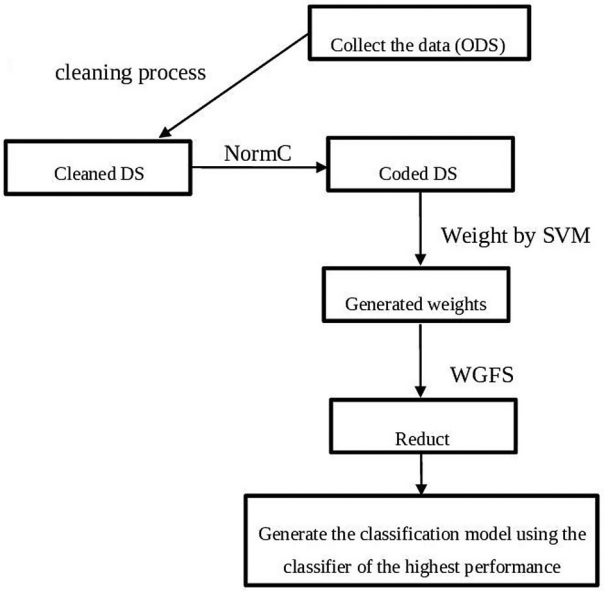

Figure 1. Proposed approach.

5. Results

The data collected were stored in an Excel sheet known as the Korean dataset. Question number 37 was chosen as a decision attribute that influences the ability/inability of learning the Korean language. The significant features (reduct) generated by the proposed approach showed important results. The 36 original conditional attributes were minimized to the minimal subset. The attributes in the minimal set are able to classify any new object faster and with similar or better accuracy than the original complete dataset.

The machine learning classification model for evaluating the ability/inability of people to learn the Korean language in GCC was built from the reduct. Decision tree, logistic regression, SVM, Naive, and random forest are some of the common classification techniques and they were used in this research. Accuracy, recall, and precision are the classifiers' evaluation metrics used. High significance was set to the accuracy. The model that gives the uppermost accuracy value will be selected to predict any new unknown incident as ability/inability to learn the Korean language, then providing the recommendations which show how to overcome the possible learning difficulties. The experiments have been conducted using the RapidMiner software tool, as previously discussed in [26].

First, the accuracy of the original dataset for each classifier was calculated. After that, four experiments were conducted, based on the generated reduct. Accuracy, recall, and precision were then calculated from the four experiments as follows:

1. for reduct generated by WGFS with weight by SVM and NormC dataset,
2. for reduct generated by WGFS with weight by SVM and DummC dataset,
3. for reduct generated by WGFS with weight by IG and NormC dataset,
4. for reduct generated by WGFS with weight by IG and DummC dataset.

Table 5 shows the accuracy of the original dataset whereas Table 6, Table 7, Table 8, and Table 9 show the accuracy, recall, and precision of the four experiments mentioned earlier.

Table 5. accuracy of each classifier based on the original dataset.

Model	Accuracy (%)
Decision Tree	88.1
Logistic Regression	87.4
SVM	88.8
Deep Learning	91.6
Gradient Boosted Trees	88.8

Table 6. WGFS with weight by SVM and NormC dataset: accuracy, recall, and precision of the reduced dataset.

Model	Accuracy (%)	Recall (%)	Precision (%)
Decision Tree	92	94.12	90
Logistic Regression	94	93.93	92.60
SVM	94	95.59	92.11
Deep Learning	87.4	85.94	85.92
Gradient Boosted Trees	89	89.26	87.3

Table 7. WGFS with weight by SVM and DummC dataset: accuracy, recall, and precision of the reduced dataset.

Model	Accuracy (%)	Recall (%)	Precision (%)
Decision Tree	88.2	86.19	86.78
Logistic Regression	88.2	86.19	86.78
SVM	88.2	86.19	86.78
Deep Learning	88.6	87.15	87.12
Gradient Boosted Trees	88	86.05	86.58

Table 8. WGFS with weight by IG and NormC dataset: accuracy, recall, and precision of the reduced dataset.

Model	Accuracy (%)	Recall (%)	Precision (%)
Decision Tree	85.8	84.1	83.66
Logistic Regression	88.2	87.02	86.45
SVM	86.4	84.54	84.89
Deep Learning	86.8	88.14	84.8
Gradient Boosted Trees	86	87.22	83.92

Table 9. WGFS with weight by IG and DummC dataset: accuracy, recall, and precision of the reduced dataset.

Model	Accuracy (%)	Recall (%)	Precision (%)
Decision Tree	87.4	91.7	90.3
Logistic Regression	87.4	91.7	90.3
SVM	87.4	91.7	90.3
Deep Learning	87.4	91.7	90.3
Gradient Boosted Trees	87.4	91.7	90.3

Among the four experiments conducted and through a close look at the previous results, the best performance was generated for WGFS with weight by SVM and NormC dataset (which is our proposed approach) by all classifiers used in this research.

Analysis results showed that logistic regression and SVM classifiers are the best, with the highest accuracy (94%), whereas deep learning is the worst, with the lowest accuracy (87.4%). Decision tree and gradient boosted trees are good classifiers with 92% and 89% accuracy values respectively. Logistic regression and SVM are able to achieve 2% higher accuracy than the decision tree which comes in second place.

Recall is another evaluation metric. It showed that the maximum value is achieved for SVM with (95.59%), which is followed by decision tree (94.12%), logistic regression (93.93%), gradient boosted trees (89.26%), and deep learning (85.94%). The recall values of logistic regression and decision tree are almost the same with only 0.19% extra for decision tree.

Precision is the third evaluation metric used in this study. Logistic regression showed the highest result as it had the maximum precision value (92.6%). SVM resulted with 92.11% precision value, followed by decision tree (90%). All the previous three classifiers are important since they all reached precision values equal to or more than 90%. Gradient boosted trees (87.3%) and deep learning (85.92%) are at the end of the list with good performance. Logistic regression deserves to be the best among all other classifiers as it has the uppermost precision value, 0.49% higher than the nearest classifier's precision value which is SVM.

SVM satisfies the peak of both accuracy and recall, but it comes to the second place after logistic regression in the precision percentage value. Therefore, it is the pioneer classifier for the Korean dataset.

Our results showed strong indications that the feature selection process has an impact on classification accuracy. Comparison between the accuracy of the original dataset and the accuracy of each of the four experiments on the reduced datasets is illustrated in Figure 2, Figure 3, Figure 4, and Figure 5. In Figure 2, results showed that all the classifiers except deep learning gave better accuracy for the reduced dataset than the accuracy of the original dataset. Figure 3 showed that only decision tree and logistic regression accuracies are better than the accuracy of the original dataset whereas Figure 4 highlighted that only logistic regression model resulted in higher accuracy value than that of the original dataset. Figure 5 showed that none of the classifiers has better accuracy than the accuracy of the corresponding classifier that was applied to the original dataset, but logistic regression gave the same accuracy value.

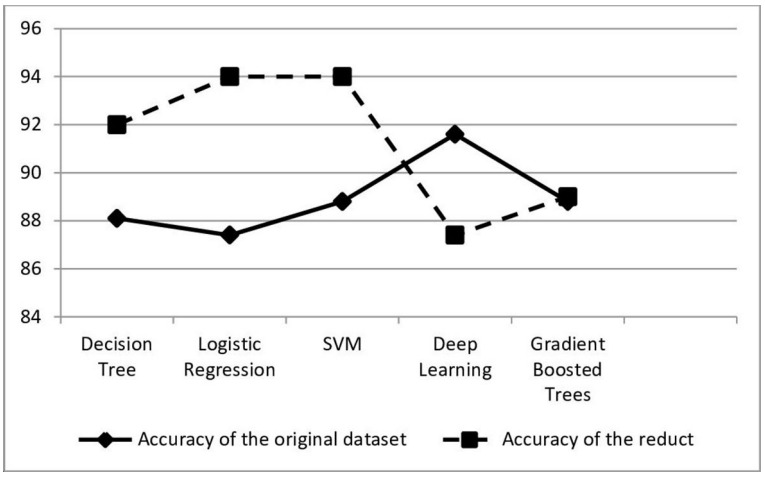

Figure 2. Accuracy comparison between ODS and the reduced dataset generated by WGFS with weight by SVM and NormC dataset for the five classifiers.

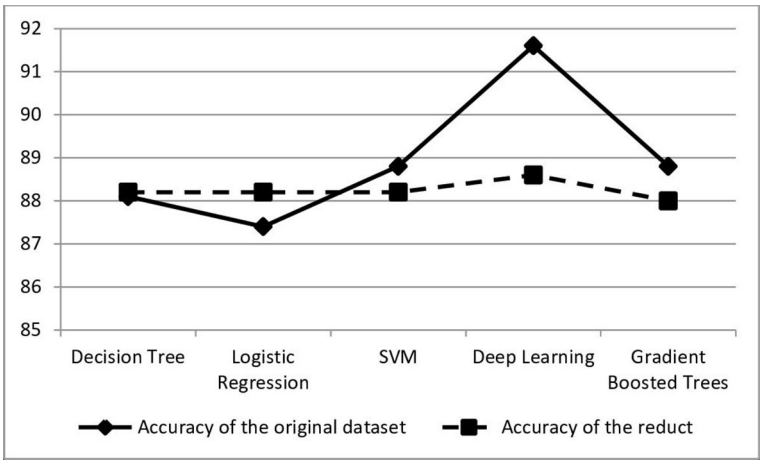

Figure 3. Accuracy comparison between ODS and the reduced dataset generated by WGFS with weight by SVM and DummC dataset for the five classifiers.

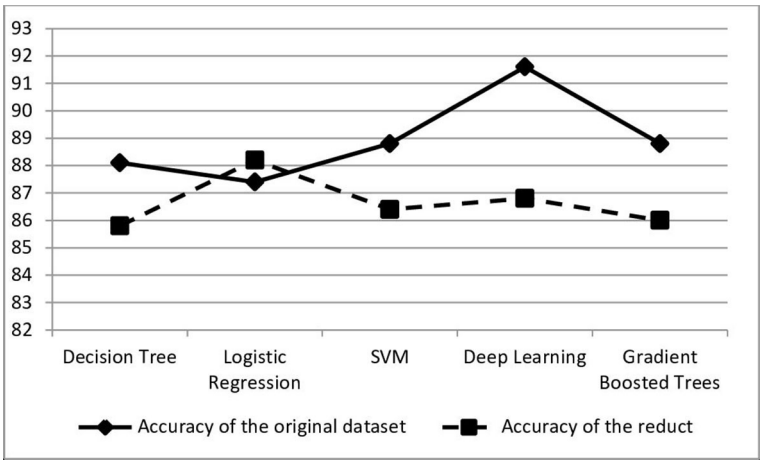

Figure 4. Accuracy comparison between ODS and the reduced dataset generated by WGFS with weight by IG and NormC dataset for the five classifiers.

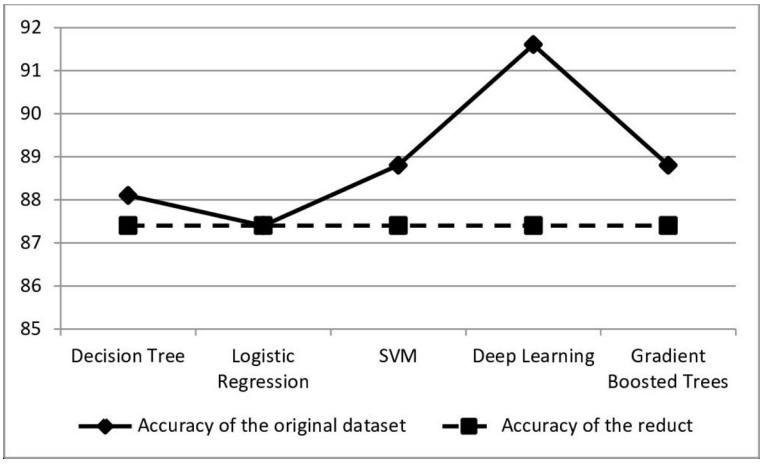

Figure 5. Accuracy comparison between ODS and the reduced dataset generated by WGFS with weight by IG and DummC dataset for the five classifiers.

Figure 6 illustrates accuracy comparison between the five classifiers for the reduced dataset generated by the proposed approach. The figure shows that logistic regression and SVM are able to predict new cases with higher predictive rate than other models but closely similar to the decision tree. Deep learning and gradient boosted trees are excluded as they are slightly far from the best classifiers' accuracy with 6.6% and 5% respectively. Among the five classifiers used, logistic regression and SVM are superior and they are recommended by the researchers to be used with the Korean dataset.

Since three metrics were used for performance measurement and in order to make the decision of choosing the best classifier fair enough to all classifiers, points were given to each classifier for each evaluation measure. The five classifiers were evaluated from 1 (best) to 5 (worst), based on their measurement values. Accuracy measurement, logistic regression and SVM were given one point each, which represents the best accuracy classifier, followed by decision tree, gradient boosted trees and deep learning with 2, 3, 4, and 5 points respectively. The same procedure was carried out for recall and precision. After that, the total points (score) were calculated for each classifier. The classifier with minimum score is the best while the one with maximum score is the worst. This analysis, as summarized in Table 10, shows that the best classifier for the Korean dataset is the SVM, and it is followed by logistic regression, decision tree, gradient boosted trees, and deep learning respectively. SVM and logistic regression classifiers are very close to each other whereas deep learning is the worst. This order is almost the same as the order of accuracy measurement, which means that the accuracy measurement is the most important for the Korean dataset and that we can ignore other measurements so as to save time and effort. Figure 7 shows the importance of the classifiers used for the classification of the Korean dataset. To make the figure readable, the following formula was used to come up with new readable values representing the importance of each classifier compared to other classifiers. The values were rounded to the nearest decimal point:

$$Importance = \frac{1}{score} \cdot 100. \qquad (11)$$

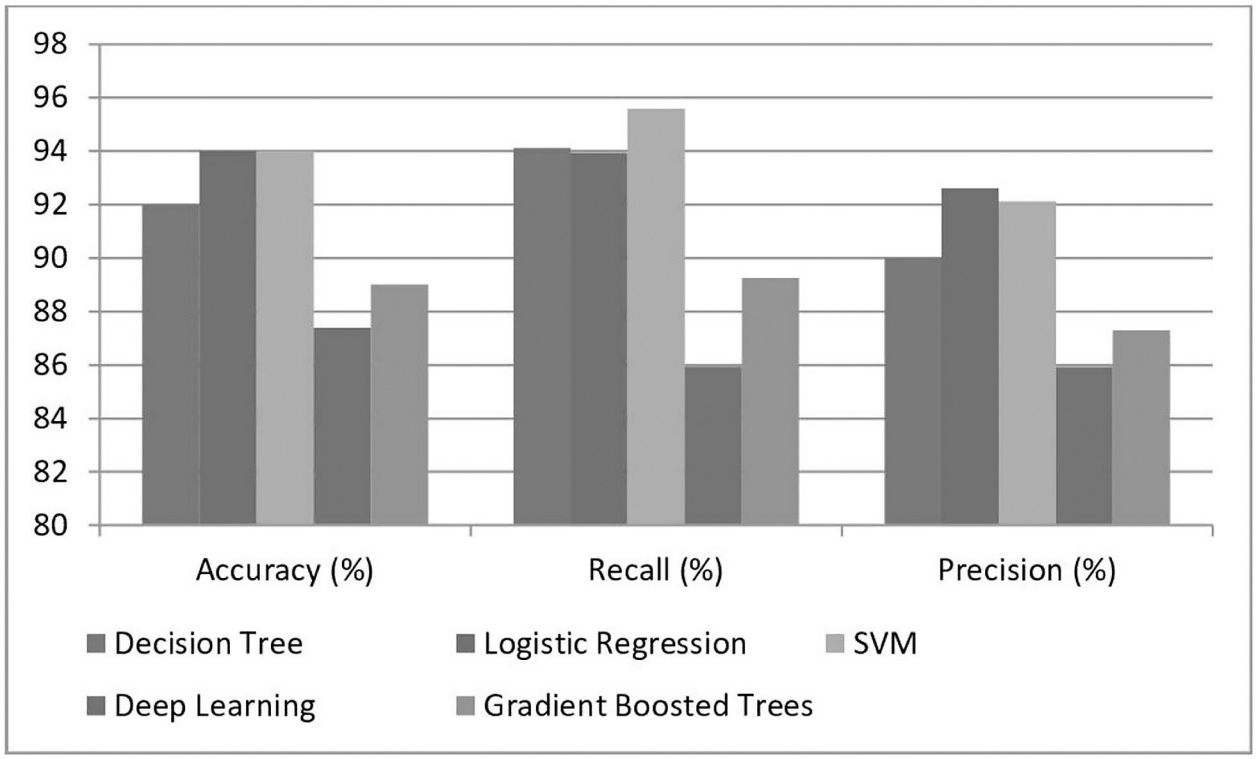

Figure 6. Performance comparison between the five classifiers.

Table 10. Ranking of the points of importance.

Model	Accuracy ordering	Recall ordering	Precision ordering	Score	Importance of the classifier	Order-based importance
Decision Tree	2	2	3	7	14.3	3
Logistic Regression	1	3	1	5	20	2
SVM	1	1	2	4	25	1
Deep Learning	4	5	5	14	7.14	5
Gradient Boosted Trees	3	4	4	11	9.1	4

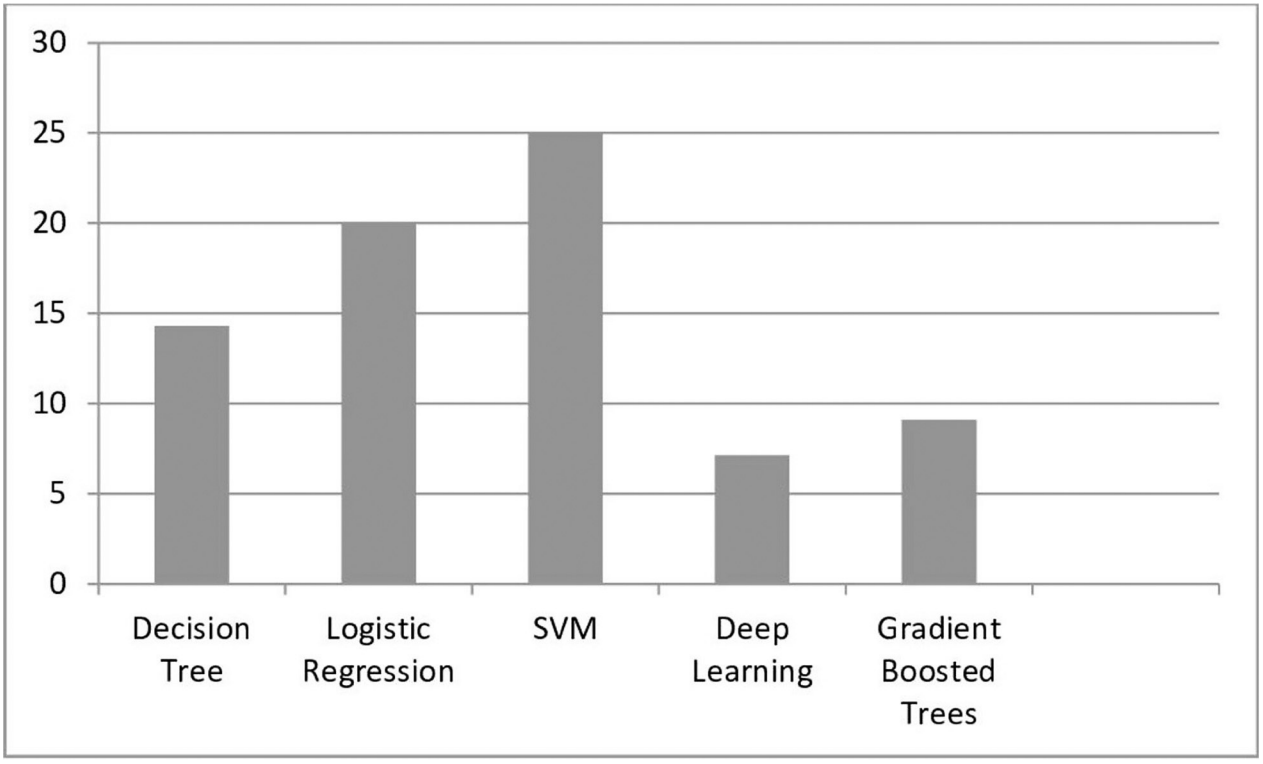

Figure 7. Importance of the classifiers based on the ranking score.

6. Conclusion

In this work, the WGFS method was used for dimensionality reduction with two different methods for weights and two different shapes for the dataset. After the analysis, WSVM using NormC was the best and therefore chosen for the proposed approach. It minimized the dimension of the Korean dataset from 36 dimensions (attributes) to 4. Such 4 attributes have an impact on the classification accuracy. Five different classifiers were used for testing the performance of the proposed model, achieving better accuracy by using the 4 attributes than by using the original 36 conditional ones. SVM is the best classifier with the highest accuracy value (94%), followed by logistic regression (94%), but with less recall value than SVM. Decision tree with accuracy (92%) comes at the third place and is followed by gradient boosted trees

and deep learning respectively. The new classification model of 4 attributes is recommended to measure the ability and inability of people in GCC to learn the Korean language. The model will also save training and prediction time.

Using our system as a starting point for any person who would like to learn the Korean language is recommended since it can give a summary of the person's ability/inability to learn the Korean language. Relevant institutions should put efforts into marketing the Republic of Korea by popularizing its history, culture, and attractive language. This will increase the possibility to spread the Korean language across the GCC. These institutions should also work hard on guiding the learners and making the learning resources available to them. This will considerably simplify the process of learning the Korean language in GCC.

In the future, we may test different methods of calculating weights and then feed these weights to WGFS. Moreover, other algorithms than WGFS could be used to generate the best reduct. Besides these, the comparison between different algorithms is being made. Furthermore, other classifiers can be included to test the performance of newly proposed approaches. Applying the proposed approach to other datasets similar to the one used in this research is also planned.

Acknowledgment

The authors would like to thank all the people who motivated, guided and helped to successfully complete the research described in this paper. We would also like to extend our gratitude to our families for their patience and support.

This research was funded by the Research Sector, The Arab Open University-Kuwait Branch under Decision No. 20002.

References

[1] P. Chomphungam, "A Content Analysis of Theses Related to Korean Studies in Thailand During 1988–2009", Doctoral dissertation, Chulalongkorn University, 2010. http://dx.doi.org/10.14457/CU.the.2010.1159

[2] S. Naidu, "A Quantitative Study on the Impact of the Korean Dramas among Youth in Chennai", Bachelor's thesis, Madras Christian College, Chennai, India, 2015. [Online]. Available: https://www.academia.edu/31998057/A_Quantitative_Study_on_the_Impact_of_the_Korean_Dramas_among_Youth-in-Chennai.

[3] U. Heo *et al.*, "The Political Economy of South Korea: Economic Growth, Democratization, and Financial Crisis", *Maryland Series in Contemporary Asian Studies*, vol. 2008, no. 2, p. 1, 2008.

[4] W. Reyes *et al.*, "Understanding the Embrace of Filipino Teenagers on Korean Dramas, Manila: Far Eastern University Manila", 2017.

[5] A. S. Thompson and J. Lee, "The Motivational Factors Questionnaire in the Korean EFL Context: Predicting Group Membership According to English Proficiency and Multilingual Status", *The Language Learning Journal*, vol. 46, no. 4, pp. 398–414, 2018. http://dx.doi.org/10.1080/09571736.2015.1130082

[6] S. C. Shin, "Students' Motivation, Learning Experiences and Learning-Style Preferences: A Survey on Australian College Students of Korean", *The Language and Culture*, vol. 5, no. 2, pp. 289–319, 2009.

[7] U.S. Department of Education, "Consultation with Federal Agencies on Areas of National Need", 2016. [Online]. Available: https://www2.ed.gov/about/offices/list/ope/iegps/consultation-2016.pdf.

[8] U.S. Department of Education, 2016 & U.S. Department of Education and Office of Postsecondary Education, "Enhancing Foreign Language Proficiency In the United States: Preliminary results of the National Security Language Initiative", Washington, D.C: U.S. Department of Education, 2016. [Online]. Available: https://nsep.gov/sites/default/files/nsli-preliminary-results.pdf.

[9] The Modern Language Association of America, "Language Enrollment Database", 2013. [Online]. Available: http://www.mla.org/flsurvey_search.

[10] D. Goldberg *et al.*, "Enrollments in Languages Other than English in United States Institutions of Higher Education", 2015. [Online]. Available: http://aut.ac.nz.libguides.com/APA6th/reports.

[11] J. Park, "Learner-Center Korean Education Through Song", Master's thesis, Korea University, Seoul, South Korea, 2008. [Online]. Available: http://m.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=e04db79c07631860ffe0bdc3ef48d419#redirect.

[12] L. Sattathamkul, "A Study on Learning Korean Language of Thai Students", Master's thesis,

Hankuk University of Foreign Studies, Seoul, South Korea, 2008. [Online]. Available: http://m.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cd09&control_no.

[13] M. Yim, "Needs Analysis for Learner-Centered Cultural Class in Korean Language Teaching: Based on the Needs Analysis of Japanese and Chinese Learners of Korean", Master's thesis, Yeonsei University, Seoul, South Korea, 2005. [Online]. Available: http://m.riss.kr/search/detail/DetailView.do?p_mat_type=54d9b8bc7cdb09&control_no=3758ca7ba9c8106d.

[14] S. J. Casasola, "The Side Effects of Watching a Good Korean Dramas", 2017. [Online]. Available: http://www.psst.ph/side-effects-watching-good-k-drama/.

[15] H. Salmeen, "What are the Impacts of Korean Dramas", 2017. [Online]. Available: https://www.quora.com/What-are-the-impacts-of-Korean-dramas.

[16] C. L. A. Semilla and P. R. Soriano, "The Impact of Korean Dramas Among Senior High Schools Students of the Marinduque Midwest College", Senior High School thesis, Marinduque Midwest College, Marinduque, Philippines, 2017. [Online]. Available: https://www.academia.edu/36105193/THE_IMPACTS_OF_KOREAN_DRAMAS_AMONG_SENIOR_HIGH_SCHOOL_STUDENTS_OF_THE_MARINDUQUE_MIDWEST_COLLEGE.

[17] J. Damron and J. Forsyth, "Korean Language Studies: Motivation and Attrition," *Journal of the National Council of Less Commonly Taught Languages*, vol. 12, no. 1, pp. 161–188, 2012.

[18] I. El Naqa and M. J. Murphy, "What is Machine Learning?", Machine Learning in Radiation Oncology, pp. 3–11, 2015. http://dx.doi.org/10.1007/978-3-319-18305-3

[19] A. L. Beam and I. S. Kohane, "Big Data and Machine Learning in Health Care", *Jama*, vol. 319, no. 13, pp. 1317–1318, 2018. http://dx.doi.org/10.1001/jama.2017.18391

[20] N. H. Shah *et al.*, "Making Machine Learning Models Clinically Useful", *Jama*, vol. 322, no. 14, pp. 1351–1352, 2019. http://dx.doi.org/10.1001/jama.2019.10306

[21] K. Kourou *et al.*, "Machine Learning Applications in Cancer Prognosis and Prediction", *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015. http://dx.doi.org/10.1016/j.csbj.2014.11.005

[22] NRC, "Frontiers in Massive Data Analysis", The National Academies Press, 2013.

[23] R. K. Roul and J. K. Sahoo, "Classification of Research Articles Hierarchically: A New Technique", Computational Intelligence in Data Mining, pp. 347–361, 2017. http://dx.doi.org/10.1007/978-981-10-3874-7_32

[24] L. Al-Shalabi, "Data Mining Application: Predicting Students' Performance of ITC Program in the Arab Open University in Kuwait-The Blended Learning", *International Journal of Computer Science and Information Security*, vol. 14, no. 12, pp. 827–833, 2016.

[25] L. Al-Shalabi, "Improving Accuracy and Coverage of Data Mining Systems that are Built from Noisy Datasets: A New Model", *Journal of Computer Science*, vol. 5, no. 2, pp. 131–135, 2009. http://dx.doi.org/10.3844/jcssp.2009.131.135

[26] L. Al-Shalabi, "Perceptions of Crime Behavior and Relationships: Rough Set Based Approach", *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 15, no. 3, pp. 413–420, 2017.

[27] V. Kumar and S. Minz, "Feature Selection: A Literature Review", *The Smart Computing Review*, vol. 4, no. 3, pp. 211–229, 2014. http://dx.doi.org/10.6029/smartcr.2014.03.007

[28] M. Ringsquandl *et al.*, "Semantic-Guided Feature Selection for Industrial Automation Systems", in The Semantic Web. ISWC 2015, (M. Arenas *et al.*, Eds.), *Lecture Notes in Computer Science*, vol. 9367, Springer, Cham, 2015. https://doi.org/10.1007/978-3-319-25010-6_13

[29] E. B. Boster *et al.*, "A Training Algorithm for Optimal Margin Classifiers", in *Proc.of the fifth annual workshop on computational learning theory*, 1992, pp. 144–152. http://dx.doi.org/10.1145/130385.130401

[30] I. James, "Claude Elwood Shannon 30 April 1916 – 24 February 2001", *Biographical Memoirs of Fellows of the Royal Society*, vol. 55, pp. 257–265, 2009. http://dx.doi.org/10.1098/rsbm.2009.0015

[31] S. Jimmy and G. Rob, "A Mind At Play: How Claude Shannon Invented the Information Age", Simon and Schuster, pp. 63–80, 2017.

[32] M. J. Aitkenhead, "A Co-Evolving Decision Tree Classification Method", *Expert Systems with Applications*, vol. 34, no. 1, pp. 18–25, 2008. https://doi.org/10.1016/j.eswa.2006.08.008

[33] C. Y. J. Peng *et al.*, "An Introduction to Logistic Regression Analysis and Reporting", *The Journal of Educational Research*, vol. 96, no. 1, pp. 3–14, 2002. http://dx.doi.org/10.1080/00220670209598786

[34] F. Chollet, "Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek", MITP-Verlags GmbH & Co. KG, 2018.

[35] S. Zhang *et al.*, "Deep Learning Based Recommender System: A Survey and New Perspectives", *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–38, 2019. http://dx.doi.org/10.1145/3285029

[36] A. Natekin and A. Knoll, "Gradient Boosting Machines, a Tutorial", *Frontiers in Neurorobotics*, vol. 7, p. 21, 2013. http://dx.doi.org/10.3389/fnbot.2013.00021

[37] E. M. Gertz and J. D. Griffin, "Support Vector Machine Classifiers for Large Data Sets (No. ANL/MCS-TM-289)", Argonne National Lab. (ANL), Argonne, IL (United States), 2006. http://dx.doi.org/10.2172/881587

[38] L. Al-Shalabi, "Comparative Study of Data Mining Classification Techniques for Detection and Prediction of Phishing Websites", *Journal of Computer Science*, vol. 15, no. 3, pp. 384–394, 2019. http://dx.doi.org/10.3844/jcssp.2019.384.394

Received: April 2020
Revised: December 2020
Accepted: January 2021

Contact addresses:

Luai Al-Shalabi
Arab Open University
Kuwait
e-mail: lshalabi@aou.edu.kw

Yousra Tahhan
Arab Open University
Kuwait
e-mail: yousratahhan@hotmail.com

Luai Al-Shalabi is an Associate Professor of data mining at The Arab Open University, Kuwait Branch. He completed his PhD in computer science in 2000 with a focus on data mining. His areas of interest include data mining, data science, knowledge discovery, and machine learning. He has published over 25 publications in reputable local and international conferences and journals, mostly on data mining and its applications.

Yousra Tahhan was born in Kuwait in 1996. She received her BSc in information technology and computing from The Arab Open University, Kuwait Branch, in 2019. Her areas of interest include databases and languages.