*International Journal of Internet, Broadcasting and Communication Vol.12 No.3 102-107 (2020)*
*http://dx.doi.org/10.7236/IJIBC.2020.12.3.102*

**IJIBC 20-3-13**

# A research on the key factors for classification of diabetes based on random forest

Yong sub Shin[*], Namju Lee[**], Chigon Hwang[***]

[*]*Graduate School of Smart Convergence Kwangwoon University, Seoul, Korea*
[**]*Visiting Professor, Department of Physical Education, Institute of Information Technology, Kwangwoon University, Seoul, 01897, Korea*
[***]*Visiting Professor, Department of Computer Engineering, Institute of Information Technology, Kwangwoon University, Seoul, 01897, Korea*
*e-mail : iceboy724@kw.ac.kr, namju1210@gmail.com , duck1052@kw.ac.kr*

***Abstract***

*Recently, the number of people visiting the hospital is increasing due to diabetes. According to the Korean Diabetes Association, statistically, 1 in 7 adults over the age of 30 are suffering from diabetes. As such, diabetes is one of the most common diseases among modern people. In this paper, in addition to blood sugar, which is widely used for diabetes awareness, BMI, which is known to be related to diabetes, triglycerides and cholesterol that cause various complications in diabetics it was studied using random forest techniques and decision trees known to be effective for classification. The importance of each element was confirmed using the results and characteristic importance derived using two techniques. Through this, we studied the diabetes-related relationship between BMI, triglyceride, and cholesterol as well as blood sugar, a factor that diabetic patients should pay much attention to.*

***Keywords:*** *Decision Tree, Random Forest, Supervised Learning, Diabetes*

## 1. Introduction

Due to the recent westernized eating habits and lack of exercise, the number of obese people is increasing [1]. Diabetes is a very common disease that cannot be easily considered and overlooked, according to a survey by the National Statistical Office, which ranks sixth among the causes of death in Korea[2]. In recognizing and managing diabetes, blood sugar levels are mainly checked. Diabetes causes arteriosclerosis and other complications, so studies of various other factors are also necessary [3]. Therefore, separate studies have been conducted on the association of diabetes with BMI [4], triglycerides and cholesterol [5]. In this paper, we propose to classify diabetes by combining blood sugar levels, BMI, triglyceride and cholesterol levels. Several classification models of artificial intelligence are used in the treatment technique of factors for discriminating and preventing diabetes. Among them, data mining, which systematically finds rules or

Manuscript Received: June. 6, 2020 / Revised: June. 10, 2020 / Accepted: June. 16, 2020
Corresponding Author: duck1052@kw.ac.kr
Tel: +82-10-3959-9490, Fax: +82-2-916-4751
Visiting Professor, Dept. of Computer Engineering, Institute of Information Technology, Kwangwoon University, Seoul, 01897, Korea

patterns in large-scale data, is frequently used [6]. The representative classification techniques of data mining are decision trees and random forests. The decision tree is visual and easy to understand. [7]. As a disadvantage, in decision trees, each single decision tree is highly classifiable, but tends to be over fitting to some data. Overfitting refers to a phenomenon in which the error rate of data decreases initially when the number of branches in the tree gradually increases, but the error rate increases when the number of branches exceeds a certain level [8]. In order to prevent this, proper pruning is required at the time when the error rate of data increases. [9, 10]. In this paper, we try to use the Random Forest technique to prevent overfitting and improve accuracy. Chapter 2 briefly describes the techniques and algorithms required for research and diabetes, and Chapter 3 describes the accuracy of each data element and algorithm. Chapter 4 describes the tests and test results, and finally Chapter 5 describes the conclusions.

## 2. Related Work

### 2.1 Classification technique

Decision trees are one of the most widely used machine learning algorithms. It can be applied to both classification and regression problems. Because it mimics human thinking, data understanding and interpretation is concise, and the flow of logic to interpret actual data can be judged [7]. When making a decision, a question is asked and the answer to the question is made up of trees. Basically, learn by asking Yes/No questions to reach the final decision. In general, when constructing a tree by teaching and learning a tree, the decision tree should be constructed in a direction that increases the homogeneity and decreases the impurity and uncertainty [11]. The algorithm evaluation measures include the entropy used by the ID3 algorithm in the decision tree and the information gain used by the random forest. Entropy is a numerical value quantified for impurity, and Information Gain is the value of entropy change before and after division. The ID3 algorithm constructs a tree by dividing branches based on entropy. If the entropy of a group is high, it means that it is difficult to find the characteristics of the group. Therefore, when dividing the branch of the decision tree, it can be said that the best classification is to classify in the direction where the entropy is minimized [7]. When sorting in the direction of low entropy, dividing down the branch creates a tree until the single node, the last node of the decision tree, becomes a single node. At this time, if the number of branches becomes too large, the phenomenon of overfitting a part of data may occur [8]. When overfitting occurs, the training data in the existing decision tree shows high accuracy, but if new data is provided, the accuracy may decrease. Therefore, the Random Forest technique is used to prevent overfitting [9]. The Random Forest method is one of several methods to prevent overfitting of the decision tree, and the ensemble is a technique that combines several machine learning models to create a strong model. Random Forest, used in this paper, is a method of making several trees through training and collecting classification results obtained from multiple trees to get a conclusion. Random Forest has a characteristic that the trees produced by randomness are slightly different. Due to this characteristic, the relationship between each tree is lowered, and as a result, generalization performance is improved. It is a model that works very well without parameter tuning because of its excellent performance and reduces variability [10, 12]. Therefore, in this paper, the decision tree and random forest are applied and compare.

### 2.2 Algorithm

ID3, C4.5 and Random Forest, which are representative algorithms of decision trees, were used for the recognition of diabetes patients using decision trees.

**Table 1. Classification Algorithm**

| Algorithm | Special feature | Evaluation index |
|---|---|---|
| ID3 | Categorical | Entropy |
| C4.5 | Numerical | Information Gain |
| Random Forest | Numerical | Information Gain |

The ID3 algorithm used for the first time is a representative decision tree classification algorithm, which is the basis of several classification algorithms [13]. The ID3 algorithm creates a root node based on a representative element that classifies the entire data, and becomes a leaf node when the elements other than the representative element can no longer be separated. If not, it is an algorithm that selects the element with the lowest entropy value by changing the entropy and gradually completes the tree by extending the branches. It can only be used for categorical attributes, and the attributes used by the parent node are not reused. Numeric attributes cannot be used. At this time, the categorical type refers to data divided into several categories. Blood type, gender, and so on. This includes the nominal form that has no meaning in order and the order form that has meaning in order. There are two types of numbers: discrete types with discrete values and continuous types with continuous values. In the entropy, the formula for obtaining the entropy for area A where m records belong is defined as follows.

$$E(A) = -\sum_{k=1}^{m} p_k \log_2(p_k) \tag{1}$$

Here, $P_k$ is the ratio of records belonging to category k among records belonging to area A. Since the decision tree is an algorithm that can be used for classification, the entropy after classification must also be calculated. The formula for obtaining entropy after classifying A region as subset R is defined as follows.

$$E(A') = \sum_{i=1}^{d} R_i \left(-\sum_{k=1}^{m} p_k \log_2(p_k)\right) \tag{2}$$

Ri is the ratio of records in the area of i after division among the records before division. The next used C4.5 algorithm and Random Forest can be said to be the algorithm that complements the disadvantages of ID3. Numerical data can be used and the problem of overfitting caused by meaningless attributes is eliminated through pruning. Random forest is different from the fact that C4.5 tree is a single tree. Information gain is used as the evaluation index of the two algorithms, C4.5 and Random Forest.

$$information\ Gain = E(A) - E(A') \tag{3}$$

The information gain can be easily obtained by subtracting the calculated value of Equation 2 from the calculated value of Equation 1.

### 2.3 Diabetes

Diabetes is a name given to the fact that glucose increases in the urine as blood sugar in the blood increases. Diabetes occurs when the pancreas has problems with its ability to secrete insulin or cells in the body can't respond to the insulin it makes. Diabetes is divided into type 1 and type 2 [1]. Type 2 accounts for most of Korean diabetes. It is mainly related to the elements of lifestyle and genetics. It progresses relatively slowly compared to type 1 and is associated with age and obesity, lack of exercise, lack of diet and stress [5]. In general, women have a slightly higher incidence than men because of changes in the hormonal environment of pregnancy. Currently, the criteria for separating patients with diabetes glycated hemoglobin, which is classified as diabetes with glycated hemoglobin not greater than 6.5% [1,14]. In this paper, we study additional factors to recognize such diabetes.

## 3. Main text

### 3.1 Data Elements

In this paper, the data elements used are FBS, BMI, gender, cholesterol, and triglycerides. The first factor, fasting blood sugar (FBS), is most closely related to diabetes in previous studies [4]. The second factor, Body Mass Index (BMI), causes fat to accumulate in the cells when obese, and makes inflammatory substances in the cells. As a result, free fatty acids gradually increase and insulin resistance occurs. BMI was used because increased insulin resistance increases the risk of complications such as diabetes [4]. The third factor, gender, differs in the incidence of diabetes due to the biological characteristics of men and women. In addition, more than 50% of diabetics suffer from hyperlipidemia. Therefore, in this paper, the amount of total cholesterol was applied as a factor to test how much it affects diabetes . Lastly, triglycerides were added. When the amount of triglycerides increases, fatty acids increase in the blood, and the action of insulin gradually decreases as fatty acids increase, making blood sugar control difficult. Difficulty controlling blood sugar was associated with diabetes, so it was applied as a factor to test the effect of triglycerides.

**Table 2. Input variables and data elements for diabetic awareness**

| Input variable | Factor | Explanation |
|---|---|---|
| x1 | FBS(Fasting Blood Sugar) | Diabetes Diagnosis Critical Factors, Diabetes blood glucose level is based on fasting blood glucose level of 126 mg/dL or higher, |
| x2 | BMI (Body Mass Index) | Obesity determination method, Height = t / Weight = w / BMI = $\frac{w}{t^2}$<br>Normal : BMI 20~25<br>Overweight : BMI 25~29.9<br>Obesity : BMI 30 or higher |
| x3 | Gender | Gender |
| x4 | Cholesterol | (1) High density cholesterol (HDL) (2) Low density cholesterol (LDL) (3) The total amount of cholesterol in three triglycerides. Baseline : Less than 180mg/dL |
| x5 | Triglyceride | A form of fat synthesized in the body. High levels of triglycerides can cause various problems in the body. In particular, elevated triglyceride levels in the blood can cause cardiovascular problems. Baseline : Less than 180mg/dL |
| xd | Patient or not | Diabetes disease Yes / No |

### 3.2 Entropy

For the production of decision trees, the entropy before classification and the entropy after classification were obtained using the entropy described in 2.2.

**Table 3. Entropy before and after classification**

| Entropy before classification | Entropy after classification | | | | |
|---|---|---|---|---|---|
| 0.77555 | FBS | 0.77141 | 0.33729 | 0.61505 | 0.90303 |
| | Cholesterol | 0.77549 | 0.31789 | - | - |
| | Triglyceride | 0.76653 | 0.33127 | 0.61347 | - |
| | BMI | 0.73546 | - | - | - |
| | Gender | 0.77524 | 0.77542 | 0.62054 | 0.90471 |

The item with the highest entropy is located at the top, and the decision tree is produced and analyzed in Chapter 4 through the result.

## 4. Experiment and Evaluation

First of all, since diabetes is not clearly classified in the types of attribute variables [15], the validity and effectiveness of the factors were verified by comparing the accuracy of various factors mentioned above. In this paper, a total of 5 elements were used.

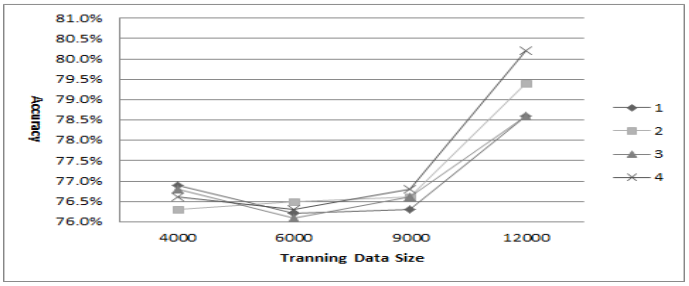


**Figure 1. Accuracy graph according to data size**

T-1 is a form of accuracy when the Random Forest is composed of three elements, FBS, BMI, and Gender, which are the most frequently used elements in diabetes-related papers [3, 16]. As the humidity data increased to 4000, 6000, 9000, and 12000, the accuracy was 76.9%, 76.2%, 76.3%, and 78.6%, respectively. As the number of training data increased, the accuracy increased gradually. Next, the accuracy of T-2 with cholesterol elements added to T-1 was 76.3%, 76.5%, 76.6%, and 79.4%. Likewise, when the triglyceride was added as in the T-3, the above result was obtained. Finally, adding all the elements you want to study in this paper, the accuracy was highest with 76.6%, 76.3%, 076.8% and 80.2% as follows. That is, it can be seen from this study that the elements of Cholesterol and Triglyceride can also be used as a test index for diabetic patients. Next, using the Random Forest, a tree was created with the five items Gender, Triglyceride, Cholesterol, FBS, and BMI mentioned above. The size (number of trees) of the forest, the most important parameter in the Random Forest technique, was set to 100. Smaller forests require less time to construct and test the tree, but have less generalization ability and less accuracy. On the other hand, as the size of the forest increases, the test time increases slightly, but a regular tree of random trees is created by averaging the results generated from multiple tree sets, so it is more continuous and has better generalization. [10]. Therefore, 100 trees were synthesized to make the most efficient tree. In addition, in order to confirm how much the factors of each item influenced diabetic patients' perception, the importance of characteristics of each factor was also checked. Characteristic importance is a measure that determines which of the elements in a tree in any forest, on average, reduce impurities. After the end of the training, the score that reduced the impurity for each element is calculated and the result is displayed. Finally, the result is normalized so that the total sum of importance is 1. The characteristic importance of BMI was 0.59071599, followed by Triglyceride with the importance of 0.16245313. The following were in the order of Fasting Blood Sugar (FBS), Cholesterol, and Gender, respectively, in order of 0.10175725, 0.09772863, and 0.047345. The figures show the relative importance and relevance of each element to be studied in this paper. As a result, it can be seen that triglycerides and cholesterol, which are known to be related to diabetes, are closely related to diabetes.

## 5. Conclusion

In addition to FBS and BMI, which are mainly used for diabetic management and diabetic patients, this study attempted to investigate the factors that are not used, such as Triglyceride and Cholesterol, are also related to diabetes and how much they affect it[3,16]. Tests confirmed that Triglyceride and Cholesterol are closely related to diabetes. It can be seen that not only diabetics, but also the general population of the risk group of

diabetes needs attention regarding FBS and BMI, which are measures of diabetes, but should pay attention to the management of Triglyceride and Cholesterol. Finally, using data mining such as Random Forest, which was tested and studied in this paper, it is possible to train not only diabetes but also other disease data, and the learned data can identify and verify new factors affecting disease in advance. The verified data can identify the association between each element, and once identified, classify the disease and manage important factors in advance. In addition, the disease can be prevented through the management of the data element, and the possibility of disease classification and prevention using data mining is high, which will be useful in the future.

## References


[1] The Institute of Internet, Broadcasting and Communication, Submission of manuscript. *http://www.iibc.kr.*

[2] Krishnasamy, S., & Abell, T. L. (2018). Diabetic gastro paresis: principles and current trends in management. *Diabetes Therapy*, *9*(1), 1-42. DOI: https://doi.org/10.1007/s13300-018-0454-9

[3] 2018_ Cause of death statistics (2019) Statistical Office

[4] Sung-ha Lee, & Hoon Jin. (2013). Analysis and Prediction of Diabetic Patients using Decision Tree. *Korean Society of Electronics Engineers Conference Academic conference*, 829-833.

[5] Minjin Lee, & Sang soo Kim. (2017). Obesity management in diabetics. *Journal of Korean Diabetes*, *18*(4).

[6] Korean Diabetes Association *https://www.diabetes.or.kr/general/class/index.php?idx=2*

[7] Jae kyu Lee, Soonbeom Kwon, Gyu-geon Lim , Management Information Systems, bubyoungsa. pp.534, 2005.

[8] Müller, A. C., & Guido, S, *Introduction to machine learning with Python: a guide for data scientists*, O'Reilly Media, Inc, 2016

[9] Deng, H., Runger, G., & Tuv, E. (2011, June). Bias of importance measures for multi-valued attributes and solutions. In *International conference on artificial neural networks* (pp. 293-300). Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-642-21738-8_38

[10] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140. DOI: https://doi.org/10.1007/bf00058655

[11] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[12] Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *35*(4), 476-487, DOI: https://doi.org/10.1109/tsmcc.2004.843247

[13] Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, *6*(3), 21-45, DOI: https://doi.org/10.1109/mcas.2006.1688199

[14] Jin, C., De-Lin, L., & Fen-Xiang, M. (2009, July). An improved ID3 decision tree algorithm. In *2009 4th International Conference on Computer Science & Education* (pp. 127-130). IEEE.

[15] Kyunghee University Hospital, https://www.khuh.or.kr/04/01.php?hospitalpath=md&table=mdlecture&page=5 &command= view_article&key=348&s_key=&keycode=&keycode2=

[16] Sunjoo Boo. (2012). Glucose, Blood Pressure, and Lipid Control in Korean Adults with Diagnosed Diabetes. Korean J Adult Nurs, 24(4), 406-416. DOI: https://doi.org/10.4028/www.scientific.net/amr.962-965.2842