Implementation of Menstrual Irregularity Prediction Model for Big Data Healthcare System

Ha-Ra Oh, Jung-Hyok Kwon and Eui-Jik Kim*

Department of Convergence Software, Hallym University, South Korea; hara9544@hallym.ac.kr, jhkwon@hallym.ac.kr, ejkim32@hallym.ac.kr

Abstract

Objectives: This paper presents an implementation of a menstrual irregularity prediction model based on a decision tree for a big data healthcare system. Methods: To build the menstrual irregularity prediction model, we use personal health data classifying people into the menstrual irregularity or normal groups as the training dataset. For accurate prediction, we use various attributes that affect menstrual irregularity, such as age, irregular meals, childish diseases, and accidents or serious trauma. For data classification, we create a decision tree by selecting the most influential attribute in each decision phase. The modeling and performance evaluation are performed through R Studio Version 3.2.4. Findings: We evaluate the performance of the menstrual irregularity prediction model through a confusion matrix. The evaluation results show that the menstrual irregularity prediction model exhibits 84.0 %, 82.9 %, and 85.5 % of accuracy, precision, and recall performance, respectively. Improvements/Applications: We expect that our menstrual irregularity prediction model will be a reference guideline for realizing the big data healthcare system.

Keywords: Big Data, Classification, Decision Tree, Healthcare System, Menstrual Irregularity Prediction

1. Introduction

Menstrual irregularity is a common symptom experienced by 80% of women during the childbearing years. It is caused by many factors, such as stress, irregular meals, lack of exercise, drinking, and smoking, and it requires continued treatment and care. If ignored, it may lead to uterine hemorrhage, anemia, cell modification and endometrial cancer, and sterility1-3.

Recently, the use of the big data healthcare system is being considered for the treatment and care of menstrual irregularity, since it can provide a personalized healthcare service. In the system, the collected data is analyzed to obtain meaningful data. For data analysis, classification is commonly used, since it enables decisions to be made more easily and it can be utilized in a wide range of applications4. For instance, it is used as an aid in decision-making, such as finding the best customers, diagnosing diseases, and judging situations.

This paper focuses on the implementation of a menstrual irregularity prediction model using the decision tree classification model. The decision tree classification model represents the relationships between attributes in the “IF–THEN–” format, and it compares attribute variables to perform the classification task. Therefore, it makes it easier to analyze the causes of specific events and patterns. We use R Studio Version 3.2.4 for modeling and performance analysis. The menstrual irregularity prediction model is created based on health data collected from 1,000 people. We evaluate the performance of the created model by comparing the predicted values and actual values. The evaluation results show that the menstrual irregularity prediction model shows 84.0% accuracy, 82.9% precision, and 85.5% recall performance.

The rest of this paper is organized as follows: In Section 2, the menstrual irregularity prediction model is described in detail. The performance evaluation is presented in Section 3. Finally, we conclude this paper in Section 4.
2. Menstrual Irregularity Prediction Model Based on Decision Tree

The decision tree has been mainly used for classification. It analyzes the distribution of the collected data by grouping the data having similar attribute values. The most important part of a decision tree is the attribute selection for classifying the collected data into groups. If all data in each group have the same attribute value, the purity of the classification model is 100%. In other words, to create a decision tree model, the process of selecting the most influential attribute is repeated in every decision phase to increase the purity of the group. Then, the creation of the decision tree model is completed when the group of all items assigned the same class or the size of the tree reaches a predefined maximum size.

The menstrual irregularity prediction model is generated based on 1,000 personal health data records provided in comma separated value (CSV) format. In this paper, we use R Studio Version 3.2.4 for data analysis. Moreover, to create the decision tree model, we use the rpart() function and plot() function provided by R Studio. To create the decision tree, we classify the symptoms into menstrual irregularity or normal based on the attribute value of the personal health dataset. To build an accurate prediction model, we use the following diverse attributes: age, irregular meals, childish diseases, accidents or serious trauma, surgical interventions, high fevers in last year, frequency of alcohol consumption, smoking habit, and number of hours spent sitting per day.

In this paper, we study women aged 18 to 36 in their childbearing years. For age, we represent 18-year-olds with 0 and 36-year-olds with 1. For irregular meals, we represent respondents who eat regular meals with 0 and those who eat meals at different times (more than two hours from the usual time) with 1. For childish diseases (e.g., chickenpox, measles, mumps, and polio), we represent respondents who suffered from childish diseases with 1 and respondents who did not with 0. For accidents or serious trauma, we represent respondents who experienced them with 1 and respondents who did not with 0. For surgical interventions, we represent respondents who experienced them with 1 and respondents who did not with 0. High fevers in last year are represented with three values (i.e., −1, 0, 1 for no fever in last year, fever more than three months ago, and fever less than three months ago, respectively). Frequency of alcohol consumption is divided into six values (i.e., never: 0, hardly ever: 0.2, once a week: 0.4, several times a week: 0.6, every day: 0.8, and several times a day: 1). Smoking habit is divided into three values (i.e., no smoking:−1, occasional smoking: 0, and daily smoking: 1). Lastly, number of hours spent sitting per day is represented with 0 for 0 hours and 1 for 16 hours (represented with decimals for values between the hours of 0 and 16). Tables 1 and 2 show the ratios of the attributes (i.e., irregular meals, childish diseases, accidents or serious trauma, surgical interventions, high fevers in last year, smoking habit).

Figures 1–3 show the personal health data distributions of the attribute variables (i.e., age, number of hours spent sitting per day, frequency of alcohol consumption) of the menstrual irregularity and normal groups. From these figures, we see that old age, frequent alcohol drinking, and sitting for a long time are associated with menstrual irregularity.

Table 1. Ratios of attributes having two values

<table>
<thead>
<tr>
<th>Attribute</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irregular meals</td>
<td>48.7%</td>
<td>51.3%</td>
</tr>
<tr>
<td>Childish diseases</td>
<td>18.4%</td>
<td>81.6%</td>
</tr>
<tr>
<td>Accident or serious trauma</td>
<td>49.9%</td>
<td>50.1%</td>
</tr>
<tr>
<td>Surgical intervention</td>
<td>49.6%</td>
<td>50.4%</td>
</tr>
</tbody>
</table>

Table 2. Ratios of attributes having three values

<table>
<thead>
<tr>
<th>Attribute</th>
<th>−1</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>High fevers in last year</td>
<td>17.7%</td>
<td>60.4%</td>
<td>21.9%</td>
</tr>
<tr>
<td>Smoking habit</td>
<td>56.5%</td>
<td>21.8%</td>
<td>21.7%</td>
</tr>
</tbody>
</table>
We can obtain the decision tree based on the collected dataset using the `plot()` function provided in the `rpart.plot` package. Figure 4 shows the created decision tree supported by R Studio. As mentioned earlier, the created decision tree is represented by “IF-THEN-...” format for indicating the relationship between the attributes. For example, if a woman has both a smoking habit value over 0.5 and surgical intervention value over 0.5, she is classified into the menstrual irregularity group according to the figure. In particular, the figure shows that smoking respondents are more likely to experience menstrual irregularity.

3. Performance Evaluation

In this section, we create the confusion matrix, which lists the predicted menstrual irregularity data and original menstrual irregularity data, to evaluate the performance of the menstrual irregularity prediction model. The confusion matrix is a result of the classification, which involves matching the predicted and actual values by the matrix. Table 3 contains the following four cases: (1) The prediction of menstrual irregularity is correct (i.e., True positive, TP), (2) The prediction of menstrual irregularity is not correct (i.e., False positive, FP), (3) The prediction of normal is correct (i.e., False negative, FN), (4) The prediction of normal is not correct (i.e., True negative, TN).

The first column of Table 3 shows the number of respondents predicted as belonging to the menstrual irregularity group (i.e., 513 people), while the second column shows the number of respondents predicted as belonging to the normal group (i.e., 425 people). In the confusion matrix, the rows represent the actual values. The first row in Table 3 indicates that 425 people are accurately predicted as belonging to the menstrual irregularity group among the actual menstrual irregularity group (i.e., 497 people). In addition, 72 people are mispredicted as belonging to the menstrual irregularity group. The second row indicates that the 415 people are accurately predicted as belonging to the normal group among the actual normal group (i.e., 503 people). In addition, 88 people are mispredicted as belonging to the normal group.

Table 4 shows the accuracy, precision, and recall value that can measure the reliability of the comprehensive classification of the menstrual irregularity prediction

Table 3. Confusion matrix

<table>
<thead>
<tr>
<th>Measured menstrual irregularity or normal</th>
<th>Predicted menstrual irregularity or normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Menstrual irregularity</td>
<td>425</td>
</tr>
<tr>
<td>Normal</td>
<td>72</td>
</tr>
<tr>
<td>Normal</td>
<td>88</td>
</tr>
<tr>
<td>Menstrual irregularity</td>
<td>415</td>
</tr>
</tbody>
</table>
model. Accuracy means that the actual data and the data that the model predicted have the same ratio. In other words, the accuracy is the ratio that is correctly predicted. We calculate accuracy as the sum of the ratios that the prediction of menstrual irregularity is correct and the prediction of normal is correct. Precision means that the ratio of same portion as the actual attribute among the specific attribute data. We calculate the precision as the ratio of the actual menstrual irregularity cases among predicted menstrual irregularity cases. Recall refers to the ratio that the specific attribute is correctly predicted. We calculate the recall as the percentage of people that are predicted by the model to suffer from irregular menstruation among people who suffer from actual menstrual irregularity. Accuracy, precision, and recall are defined in Equations 1–3.

\[
Accuracy = \frac{TP+TN}{TP+TN+FP+FN}
\]

\[
Precision = \frac{TP}{TP+FP}
\]

\[
Recall = \frac{TP}{TP+FN}
\]

4. Conclusion

This paper focused on the implementation of a menstrual irregularity prediction model using the decision tree classification model. We considered various attributes for accurate prediction. To create the decision tree, we selected the most influential attribute in every decision phase. The attribute values of the decision tree were determined by the personal health data distributions. To evaluate the performance, we used the confusion matrix. The results showed that the implemented menstrual irregularity prediction model exhibited 84.0% accuracy, 82.9% precision, and 85.5% recall.

5. Acknowledgments

This research was supported by the Leading Human Resource Training Program of Regional Neo Industry through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2016H1D5A1910427). This work was also supported by the Korea Foundation for the Advancement of Science & Creativity (KOFAC) grant funded by Ministry of Education (MOE).

6. References