Machine learning classification of infectious disease distribution status

Irzal Arief Wisky¹, Musli Yanto², Yogi Wiyandra³, Hadi Syahputra³, Febri Hadi²

¹Department of Information Systems, Faculty of Computer Science, Universitas Putra Indonesia "YPTK", Padang, Indonesia
²Department of Informatics Engineering, Faculty of Computer Science, Universitas Putra Indonesia "YPTK", Padang, Indonesia
³Department of Computer Systems, Faculty of Computer Science, Universitas Putra Indonesia "YPTK", Padang, Indonesia

ABSTRACT

Infectious diseases are common diseases and are caused by microorganisms such as viruses, bacteria, and parasites. Indicators of the spread of this disease can be seen based on the population level and the number of confirmed cases. This study aims to develop a machine learning (ML) analysis model using the K-means cluster, artificial neural network (ANN), and decision tree (DT) methods. The dataset used in this study was obtained based on the number of confirmed patients and the distribution of the population. The analysis process is divided into two stages, namely preprocessing and the classification process. The pre-processing stage aims to produce a classification pattern that can describe the level of distribution status. The classification pattern will be continued at the classification analysis stage using ANN and DT. Classification analysis gave significant results with an accuracy rate of 99.77%. The results of the classification analysis can also describe the level of knowledge distribution based on the decision tree. Overall, the contribution of this research is to develop a classification analysis model that presents the latest information and knowledge. The results of the research presented also have an impact on the control process in environmental management and public health.

Keywords:
Classification
Infectious diseases
Information dan knowledge
Machine learning
Public health

1. INTRODUCTION

Analysis of the status of the spread of infectious diseases is used as a tool for the public health management process [1]. In general, these infectious diseases consist of influenza, hypertension, diarrhea, tuberculosis and others [2]. Infectious diseases cause pain, paralysis, and even death with a fairly high percentage rate of 69.91% [3]. The transmission rate is spread nationally so that it is one of the main health problems today [4]. To help solve these problems, the classification analysis process can play an active role in developing a model to provide the best alternative solution.

Classification analysis has been developed in various problems to provide the desired results [5]. These various models use several methods in conducting classification analysis [6]. The analysis model can be seen in the concept of machine learning (ML). The model has been able to contribute quite effectively to the classification process [7]. ML works optimally to present output with a fairly good level of accuracy [8]. The development of ML in several studies shows a significant graph in solving problems in the world of health [9]. These problems can be seen from the process of identification, classification, and prediction [10], [11]. In this...
case, ML will also be used to carry out a classification analysis process on the problem of the status of the spread of infectious diseases.

The method that will be used in the classification analysis process involves K-means clusters, artificial neural networks (ANN), and decision trees (DT). This method can work more effectively in presenting the desired results. K-means cluster is a method that can categorize data based on mathematical calculations [12]. K-means is a very popular method used in the big data concept [13], [14]. This method works on pre-processing to produce a classification pattern [15]. The pattern obtained can be proven to be effective in carrying out the classification analysis process [16]. The results given from pre-processing using K-means clusters will be forwarded to the analysis using the ANN concept.

ANN is a concept that is widely used in ML [17]. This method is a supervised learning concept that can produce fairly good accuracy results [18]. ANN performs an analysis based on the weighted value obtained to produce the output [19]. This concept continues to develop as many problems have been well resolved [20]. In the process of analyzing the classification of the spread of infectious diseases, ANN is expected to provide optimal performance. To get these results, the stages of the training and testing process in learning will be maximized to produce outputs [21]. ANN performance can be seen based on the level of accuracy and error in the output presentation [22]. The outputs obtained in the process will later be re-analyzed in order to present the information and knowledge needed. The ANN output results obtained will be re-analyzed using the DT concept.

In general, DT works by conducting analysis based on previously obtained patterns to present knowledge-based [23]. The DT analysis process will refine the analysis of the spread of infectious diseases in the form of a decision tree. The results represented in the DT represent a previously hidden information and knowledge [24]. Several analytical models have been developed with DT such as the classification process for a disease that aims to support a decision [25].

Overall, this study presents the novelty of the classification analysis model. The model was developed through pre-processing and classification processes on the ML concept. The up-to-date model also provides a structured and systematic analysis process to provide precise and accurate output. With this, this research can to present new knowledge and information that describes the spread of infectious diseases. Furthermore, that this research will also be useful for related parties in environmental and community health management.

2. RESEARCH METHOD

The classification analysis process using the machine learning concept has 2 stages, namely the pre-processing stage and the classification analysis stage. The methods and algorithms used consist of the K-means cluster, ANN, and DT. The description of the research, stages can be seen in Figure 1.
Figure 1 explain the analysis process starting with data analysis based on population size and infectious disease. The classification stage starts from pre-processing using the K-means cluster algorithm aimed at obtaining patterns in classification analysis. With the analysis pattern obtained, the classification process will be carried out using an ANN. ANN learning using a feedforward algorithm aims to get the maximum classification results. The classification analysis stage will be continued by using the DT method to obtain information and knowledge. The results of the analysis based on numbers present the output in the form of a DT in the status of the spread of the disease seen from the population and the number of infection cases.

2.1. Data collection

The discussion of this study uses population data and infectious disease figures for 3 periods, 2018, 2019, and 2020. The source of the data used comes from the Pesisir Selatan District Health Office. The data will be analyzed previously to be used as variables in conducting classification analysis. The variables used will be seen based on 2 indicators based on population and the number of cases of infectious diseases. The variables used in the analysis process can be seen in Table 1.

<table>
<thead>
<tr>
<th>Population number</th>
<th>Variable</th>
<th>Infected number</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>X1</td>
<td>Ispa</td>
<td>X8</td>
</tr>
<tr>
<td>Female (1-12 Year)</td>
<td>X2</td>
<td>Influenza</td>
<td>X9</td>
</tr>
<tr>
<td>Female (1-30 Year)</td>
<td>X3</td>
<td>Gastritis</td>
<td>X10</td>
</tr>
<tr>
<td>Female (31-45 Year)</td>
<td>X4</td>
<td>Hipertensi</td>
<td>X11</td>
</tr>
<tr>
<td>Female &gt; 45 Year</td>
<td>X5</td>
<td>Diarrhea</td>
<td>X12</td>
</tr>
<tr>
<td>Common cold</td>
<td>X6</td>
<td>Rheumatism</td>
<td>X13</td>
</tr>
<tr>
<td>Asthma</td>
<td>X7</td>
<td>Fever</td>
<td>X14</td>
</tr>
<tr>
<td>Dengue fever</td>
<td>X8</td>
<td>Commond cold</td>
<td>X15</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>X9</td>
<td>Asthma</td>
<td>X16</td>
</tr>
<tr>
<td>Dispepsia</td>
<td>X10</td>
<td>Dengue fever</td>
<td>X17</td>
</tr>
<tr>
<td>Skin Allergies</td>
<td>X11</td>
<td>Tuberculosis</td>
<td>X18</td>
</tr>
<tr>
<td></td>
<td>X12</td>
<td>Dispepsia</td>
<td>X19</td>
</tr>
<tr>
<td></td>
<td>X13</td>
<td>Skin Allergies</td>
<td>X20</td>
</tr>
</tbody>
</table>

K-means cluster is an initialization algorithm for grouping data [26]. The implementation of K-means clusters can provide results in the form of analysis patterns for recommendations for classification, determination, and prediction processes [27]. K-means works by looking for and finding similar patterns in the data with the output of information and knowledge [28]. This algorithm is an exploratory analysis concept that can be applied in supervised machine learning [29]. The concept of the K-means cluster algorithm can be seen in (1) [30].

$$\sum_{j=1}^{k} \sum_{x \in \mathcal{E}} ||x_i - \mu||^2_2$$

(1)

2.2. Artificial neural network (ANN)

ANN is a method that is widely used in machine learning [31]. The ANN method in machine learning gives promising results to produce a comprehensive review [32]. The implementation of this concept can carry out learning in the classification analysis process with better output [33]. ANN is a non-linear concept with mathematical calculations on a modeled problem to produce the output [34]. ANN performance results provide a fairly high level of sensitivity based on network output [35].

2.3. Decision tree (DT)

DT is a classification analysis concept developed to provide decisions based on data filters [36]. The development of this method is used in the classification process by validating the tests carried out [37]. DT is used in solving problems on complex data to produce information and knowledge that is presented in the form of a DT [38]. The performance process still uses mathematical calculations in the development of decision-making systems [39]. The equations in the DT method can be seen in (2) [40].

$$\text{Entropy}(S) = - \sum_{i=1}^{c} \text{PS}(ci) \log \text{PS}(ci)$$

(2)
3. RESULTS AND DISCUSSION

3.1. Pre-processing analysis

The pre-processing analysis stage aims to maximize the classification process that will be carried out [41]. This process can provide a better and structured analysis presentation to get better output results [42]. In this pre-processing analysis, the algorithm used is the K-means cluster. This algorithm can group data based on the level of closeness of the relationship in the data [43]. The results of the pre-processing analysis using K-means clusters can be seen in Table 2. Table 2 shows that the results of the cluster provide a classification pattern based on the population of data on the status of the spread of infectious disease numbers. The cluster results show the level of distribution with high status (C1) as many as 8 items, moderate as many as 2 items, and low as 5 items. From Table 1, it can be seen that there are 3 categories of infectious disease distribution status, namely high, medium, and low status. With the results of the pre-processing, a classification process will be carried out for the spread of infectious disease numbers.

<table>
<thead>
<tr>
<th>Population (X1-X7)</th>
<th>Infected Number (X8-X20)</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>150 788 720 332 252 209 564 21 14</td>
<td>15 16 46 56 20 96 26 9</td>
<td>3 0 71 High</td>
</tr>
<tr>
<td>215 109 105 473 351 306 805 26 75</td>
<td>48 87 57 12 16 34 67 8</td>
<td>4 0 56 High</td>
</tr>
<tr>
<td>137 684 694 303 219 201 516 32 15</td>
<td>86 71 10 29 89 35 74 6</td>
<td>9 0 28 High</td>
</tr>
<tr>
<td>151 758 757 333 242 219 567 32 65</td>
<td>16 92 55 15 0 10 67 0</td>
<td>0 0 0 High</td>
</tr>
<tr>
<td>264 134 130 582 430 378 991 52 22</td>
<td>237 29 0 31 54 24 46 4</td>
<td>4 0 18 High</td>
</tr>
<tr>
<td>160 807 794 352 258 230 599 71 47</td>
<td>14 16 0 68 17 0</td>
<td>0 0 0 High</td>
</tr>
<tr>
<td>451 225 225 994 722 655 169 30 13</td>
<td>11 11 52 12 0</td>
<td>0 0 0 Low</td>
</tr>
<tr>
<td>303 147 155 666 471 451 113 73</td>
<td>0 26 28 73 25 45 49 88</td>
<td>3 8 3 29 High</td>
</tr>
<tr>
<td>525 257 268 115 823 779 196 80 34</td>
<td>69 54 86 63 38 0</td>
<td>13 5 0 25 Midd</td>
</tr>
<tr>
<td>505 252 253 111 806 734 188 86</td>
<td>0 41 23 10 36 22 0</td>
<td>98 4 6 0 17 Midd</td>
</tr>
<tr>
<td>314 154 160 692 495 464 117 57 31</td>
<td>56 41 0 30 0</td>
<td>91 0 0 0 11 Low</td>
</tr>
<tr>
<td>465 230 234 102 736 680 173 36 23</td>
<td>28 19 38 11 18 0</td>
<td>24 0 0 0 0 Low</td>
</tr>
<tr>
<td>367 176 190 809 565 553 137 40 69</td>
<td>26 14 87 31 21 75 81 6</td>
<td>5 1 72 Low</td>
</tr>
<tr>
<td>727 347 380 160 111 110 272 27 0</td>
<td>12 72 56 56 34 87 87 6</td>
<td>1 0 12 High</td>
</tr>
<tr>
<td>485 240 244 106 770 709 181 29 56</td>
<td>18 16 79 16 24 47 68 0</td>
<td>0 0 13 Low</td>
</tr>
</tbody>
</table>

3.2. Classification analysis

The classification process in the discussion aims to see the status of the spread of infectious diseases based on infection numbers and population. In this case, the analysis process begins by using the ANN method with a feedforward algorithm. The ANN method is a concept that can carry out learning with better outputs [44]. ANN can also be implemented in the case of the classification of a disease by using the concept of feedforward learning. The results given have a fairly high level of accuracy [45]. Basically, this method learns the pattern of network architecture formed by the training and testing process [46]. The study aims to obtain the best network architecture pattern that will be used in the classification analysis process [47]. The results of the best classification of the ANN network architecture pattern can be seen in Figure 2.

Figure 2 is the result of the best classification ANN network architecture through the learning process by training and testing the previous classification pattern. The ANN architectural pattern has 3 layers, namely the input layer, the hidden layer, and the output layer [48]. The architecture is shown in Figure 2. consists of a layer of 20 units of the input layer, 5 layers of hidden layers of five units namely (50, 35, 25, 15, and 10), and 1 layer of the output layer of one unit. This architectural pattern will be used to carry out the classification process on the status of the spread of infectious diseases. The results of the classification process using ANN can be seen in the learning output graph in Figure 3.

Figure 3 describes the results of the classification analysis using ANN which has a fairly good output. These results can be seen from the performance value of 0.0731% so that the ANN learning process approaches the maximum results in the classification process. ANN output can also be proven the level of relationship based on the linearity value of the input used [49]. In this case, the level of relationship between input and output units is 96.98%. These results are sufficient to illustrate that ANN is able to perform classification analysis on the status of the spread of infectious diseases.

The analysis process will still be continued with the aim of exploring knowledge based on the classification pattern that has been analyzed with ANN. The DT method can present output in the form of knowledge-based [50]. In concept, DT performs analysis to find information and knowledge hidden in a pile of data [51]. The classification analysis process using the DT concept will focus on two directions, namely based on the population level and the number of distribution figures. The purpose of this two-way classification analysis is to find information and knowledge from a different perspective. The results of the
analysis provided by DT can be used as a reference to follow up the handling process for related parties. The results of the DT classification analysis based on the population level can be seen in Figure 4.

Figure 4 explains that DT is capable of generating information and knowledge in the form of a DT image. The classification results presented can be seen that the population with the age category > 45 years has the highest risk for transmission. Then for the population aged 31-45 and under 30 years, it also has a relatively moderate level of probability. To ensure the results obtained in Figure 4, the analysis process will also be seen based on the rate of spread of infectious diseases. The results of the analysis can be described in Figures 5 and 6.

Figure 5 is the result of a DT that describes information and knowledge about the status of the spread of infectious diseases. These results are based on indicators that have been analyzed previously. Figure 6 is a form of classification rule that presents knowledge in the spread of infectious diseases. Overall the results of the classification analysis developed in the new model are able to provide significant results.
Figure 4. Results of DT classification analysis based on population level

Figure 5. Decision tree of classification analysis results
Figure 6. Classification rules for the spread of infectious diseases

The analytical model presented is also quite effective in presenting the up-to-date process of ML classification analysis. The update of the model can be seen based on the output of the analysis stages that have been carried out. The overall analysis results have been validated by measuring the level of accuracy and error as well as testing the performance and sensitivity of the analytical model. With these results, the proposed analytical model is able to provide an update on the previous model in describing the classification of the status of the spread of infectious diseases.

4. CONCLUSION

The development of classification analysis using ML gives quite good results. Overall, this study presents an updated analysis model for the classification of the status of the spread of infectious diseases. The analysis process provides output in two directions, namely classification based on data on the number of infected cases and population distribution. These results are obtained through pre-processing in order to obtain a precise and accurate analysis pattern. Classification analysis provides an accuracy rate of 99.77% and an error rate of 0.33%. Furthermore, the output of the classification results is also able to describe the DT with an accuracy of 91.67%. The DT will be used as an information and new knowledge for related parties. The knowledge gained can also be useful in carrying out environmental and community health management processes.

REFERENCES


BIOGRAPHIES OF AUTHORS

Irzal Arief Wisky was educational background comes from the information systems study program. Further education for the post-graduate program in Information Technology, Faculty of Computer Science, was completed in 2013. In my job, I am a lecturer at Putra Indonesia University, YPTK Padang, majoring in information systems. The history of research includes decision support systems and expert systems. He can be contacted at email: irzal.arief12@gmail.com

Musli Yanto He is a lecturer at Universitas Putra Indonesia YPTK Padang. Educational background since 2014 has completed postgraduate studies in the field of informatics engineering. His skills include Data Science Analysis, Algorithms and Programming, Big Data and Artificial Intelligence. He can be contacted at email: musli_yanto@upiyptk.ac.id

Machine learning classification of infectious disease distribution status (Irzal Arief Wisky)
Yogi wiyandra educational background comes from a computer system study program. Advanced education for the postgraduate program in Information Technology, Faculty of Computer Science, was completed in 2014. In my work, I am a lecturer at the Putra Indonesia University, YPTK Padang, majoring in computer systems. Research history includes expert systems, decision support systems, and several branches of artificial intelligence. He can be contacted at email: yogiwiyandra@upiyptk.ac.id

Hadi Syahputra, a graduate of Putra Indonesia University YPTK. In 2013 he obtained a Master's Degree in Computer with a concentration in Information Technology. His career as a lecturer at Putra Indonesia University YPTK began in 2013 until now. 2015 until now he has been trusted as a sports coach at Putra Indonesia University YPTK. Working in the field of networking and Micro Controllers with several published articles. Monitoring DNS Query with Pi-Hole Firewall Using Raspberry B+ Integrated with Mikrotik Router RB 931-2nd is an article that has been published in an international journal in 2021. He can be contacted at email: hadisaputra@upiyptk.ac.id

Febri Hadi educational background comes from a Informatic of Engineering study program. Advanced education for the postgraduate program in Information System, Faculty of Computer Science, was completed in 2013. In my work, I am a lecturer at the Putra Indonesia University, YPTK Padang, majoring in information systems. Research history includes expert systems, decision support systems, and data mining. He can be contacted at email: febri_hadi@upiyptk.ac.id