INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 11, Issue 9, (2021) E-ISSN: 2222-6990

Predicting Common Diseases among Students using Decision Tree (J48) Classification Algorithm

Maslina Abdul Aziz[1], Amirah Jasri[2], Mohd Razif Shamsudin [3], Ruhaila Maskat [4], Nurulhuda Noordin[5] and Mohd Izuan Hafez Ninggal[6]

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Malaysia, Department of Computer Science, Faculty of Computer Science and Information Technology, Putra University Malaysia

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v11-i9/11030 DOI:10.6007/IJARBSS/v11-i9/11030

Published Date: 12 September 2021

Abstract

Predictive analysis is very useful in the process of decision making. It discovers useful information by predicting the future outcome. Nevertheless, it is essential to understand an appropriate technique before the predictive analytical model should be developed. This research will compare two predictive methods which are decision tree technique (using J48 algorithm) and rule induction technique (using JRip algorithm). The aim of this research is to build the predictive model for health datasets of students in one of the universities in Selangor. By analyzing medical profiles such as gender, diseases, the symptoms of the diseases, the organs of the diseases and the body systems of diseases, the model can predict the likelihood of disease that may occur in the future. It can offer significant and important insights, such as the patterns and relationships between medical attributes related to the diagnosis datasets. In this study, we are also able to identify the most common illness that may have infected the students in the past five years. This data analysis could be beneficial specifically to the health center to plan, coordinate tasks and make better decisions.

Keywords: Predictive Analysis, Algorithm, Decision Tree, Rule Induction, Diseases.

Introduction

Big data analytics has become a trend as many organizations use it as a prediction tool to gain profits in the future. Some even reported that if data analytics is used correctly, it may also be able to increase the profit margin significantly. Big data analytics plays a huge role to improve comprehensive understanding of hidden values and incurs more new opportunities for many industries such as health care, financial, business, manufacturing and so on. There are a number of research papers that surveyed different types of big data sources and techniques (Ariffin et al., 2020; Fang et al., 2015; Mohsen et al., 2017). There are also solutions

to efficiently deal with big data storage. Data analytic is used to replace the traditional statistical methods of data analysis. Dong et al., (2013); Noor, et. al (2018) stated that data analytics using data mining are being used by previous researchers and experts to replace the traditional manual data analysis which has already become insufficient in the health area.

Big Data can be streamlined as the data that is large sized in quantity and volume. Normally this situation occurs when the data exceeds the processing capacity that would usually use traditional existing technologies to be processed. Big data can also be referred to as a new term which is in large size and complex. This large and complex data cannot be managed by using current technologies or data mining software tools (Morales et al., 2017) (Zhang et al., 2018). Some scholars also concluded that Big Data should be defined as simply a large amount of data that can't be managed normally. There is a need for new technologies and architectures that should enable us to extract the value from these data. Big data cannot be managed by using the existing traditional techniques. This extraction value is handled by capturing and analyzing the process of big data. Generally, big data can be divided into three different categories. These categories are structured data, unstructured data and semi-structured data. According to Abawajy (2014), he stated that the structured data referred to the data that is entangled with a predefined formatting. These data relationships are generally simple and known, so it leads to easy management and processing. The relational database in MainFrame, Oracle, SQL server, DB2 etc. is an example of structured data.

Next is the unstructured data category. It is defined as the data which is absent from any form of structure (Liu et al., 2014). Approximately, there were 500 quadrillion files – unstructured data. The quantity of these data is reported to increase double in every two years. It is difficult and expensive to work with unstructured data. In addition, for the processing purposes, the conversion process from unstructured data to the structured data is also not practical. The examples of unstructured data are the comments on social media, tweets, likes, tags, chatter, videos and pictures. It is, however, very crucial that we should understand the meaning and context of unstructured data in order to make better decisions.

In the healthcare industry, to identify the symptoms, health conditions and the records of the patients in the medical field, there were many digital healthcare solutions introduced. They are very important and necessary. Digital healthcare solutions promised to convert the whole healthcare process to become more efficient, less expensive and higher quality (SAS Insights, n.d.). Knowledge discovery by using data analytics in data mining can help to convert the massive healthcare data into information and knowledge which can help control, cost and maintain high quality of patient care. Moreover, data analysis using data mining will manage the variety of sources of data, the large volume of this data and also the velocity of the data in real time. The hidden patterns from data sets is the main objective of unsupervised learning as well as producing an inference from it (Basheer et al., 2019). Also, to introduce Healthcare analysts and practitioners to the new advancements in technology to effectively handle large and heterogeneous healthcare data.

On the other hand, predictive analytics is the process of using data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data (Conn, 2014). The goal is to go beyond knowing what has happened to provide a best assessment of what will happen in the future. Hence, predictive analytics in the health industry will enable the healthcare providers, organizations or any related authority to do prediction on diseases to avoid epidemic breaks out, cure ailment and ultimately avoid preventable deaths. In addition, the results from predictive analytics will increase the accuracy of diagnoses as it is important to help for strategic decision making. For

example, decision aids based on predictive analytics have shown to improve value-based clinical decision-making in preventing the readmission in the general inpatient setting (Mukherjee, 2019; Bates et al., 2014). Based on the preliminary study and literature review, this research only focused on two classifier techniques in data mining which are decision tree technique using J48 algorithm and rule induction technique using JRip algorithm. The selection of these two techniques because of the experiment result is much understandable since the rules generated are easy to understand and explained and many previous researches used both of them. A survey was conducted by Zand (2015) that used various techniques to predict breast cancer. Another experiment by Sahle (2016), to discover the factors that affect postnatal care visit in Ethiopia is conducted by using both decision tree technique and rule induction technique as these techniques support both numeric and nominal attributes and produce accurate and readable rules.

There is a large amount of data that is recorded and stored by the health center. The users are actively using Microsoft Excel. However, the traditional statistical methods are unable to analyse and process complex data. The main issues identified are fraud vulnerability, prone to error and time-consuming. Therefore, based on the literature and initial findings, data analytics techniques can be adapted to manage, monitor, process and integrate the big volume of patients' data (Lin et al., 2017). Cross Industry Standard Process for Data Mining or CRISP-DM is the methodology used and adopted throughout the experiment process. Experiments and data analysis will be done using Weka Tools. The techniques used are decision tree technique using J48 algorithm and rule induction technique using JRip algorithm. The model was developed by using the JRip algorithm in rule induction technique. A number of researches were done on predictive analytics on health data (Suhaimi, et al., 2019). The patient's data privacy of healthcare data must not be compromised. Therefore, the name of the organizations will not be mentioned. The objective of this research is to discover the common disease occurred among students based on the pattern diagnosis for the past five years. This study explores the suitable analytics technique to analyse the healthcare dataset and provide useful knowledge that assist to more accurate decision making for curing the ailments and to predict the future diseases that may occur.

Methodology

This research applied the Cross Industry Standard Process for Data Mining or commonly well-known with its acronyms, CRISP-DM is the methodology used and adopted throughout the experiment process. It is one of the methodologies used to develop predictive analytical models for data mining projects. This methodology consists of six stages which are:

Business Understanding

The first stage is understanding of the problem that needs to be solved. At this stage, data were collected using various methods such as interview and document review. For this research the data acquired will uncover the pattern of the patients' records from 2011 till 2015.

Data Understanding

After the collection of data, the data will be analysed. Basically, the steps involved in this stage are collecting the data, describing the data, exploring the data and verifying the quality of the data. The data analytics techniques used and applied are decision tree technique and Bayes theorem.

Data Preparation

Data will be polished and converted to the suitable format. Converting data to the tabular form, removing or improving the missing values and converting data to different types (for instance; the image is converted into form) if necessary.

Modelling

For this experiment, the researcher used a clustering technique. Clustering is the most common unsupervised learning technique that works to explore data in the data analysis process to find hidden patterns, corresponding to the objective of unsupervised learning. (Niaksu, 2015) and (Tarmizi *et al*., 2019).

Evaluation

The researcher evaluates the suitability of the model. The steps involved in this stage are to evaluate the result from the model created, review the process of developing the model and determine the next actions to be taken.

Deployment

The results will be used in the decision making in future based on the final report produced

Results and Discussion

Based on the experiments conducted, the findings are presented in the form of a diagnosis dataset. The datasets given also separated according to the years which were from 2011 until 2015. In the diagnosis dataset, it illustrated the program name and the college name of the patients. From there, we can understand that there was a relation between the college, program and disease that occurred. By using this data, we can discover the pattern of correlation between disease, gender, college and the program of the students.

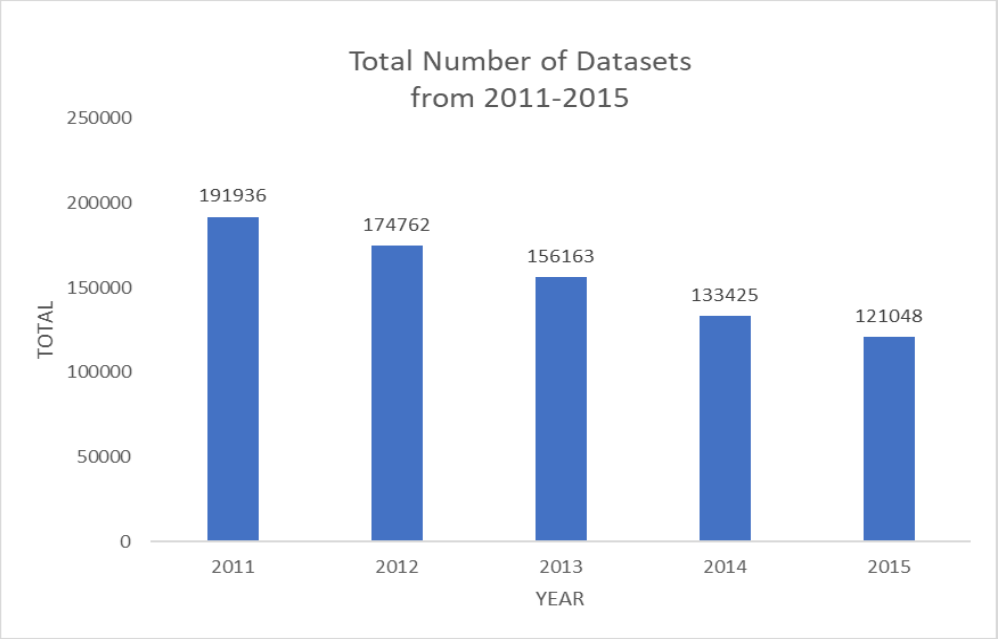

Fig.1: The total number of datasets collected for five years.

The initial instances in this dataset were 198, 855 and the attributes in this dataset were 16 attributes. By using the Weka Tools, data were filtered. After cleaning up the dataset,

after removing the noisy data, the cleaned data has reduced to 5898 records; it is about 3% of the total records. After the cleaning process, 192957 records were removed. It is accumulated to 97% of data removal. The dataset contains different types of diseases treated at the health center. Therefore, the scope of this experiment will only focus on the most common system, the "Diseases of the respiratory system" records. Based on initial readings and understanding, we found that there were four attributes that are most related to each other. They are chapter_name, block_name, division_description and subdivision_description. Since all the diseases were under the respiratory system, we excluded the chapter_name attribute from this experiment.

Based on the findings, the most common organ involved in the respiratory system was Acute Upper Respiratory Infection (URI) with 1805 records. There were 5 types of diseases listed under the respiratory system. They were:

1. Acute upper respiratory infections.
2. Chronic lower respiratory diseases.
3. Other diseases of the upper respiratory tract.
4. Influenza and pneumonia.
5. Other acute lower respiratory infections.

Acute Upper Respiratory Infections of Multiple and Unspecified Sites is also known as common cold/fever in general. It was diagnosed as the most common disease occurring among patients with the highest record of 1501. The second and third highest records were the Acute Pharyngitis with 183 records and Acute Tonsillitis with 107 records. The second part of this experiment is to discover the patterns of diagnosis that were conducted based on several hypotheses. For hypothesis 1, the experiment was conducted by selecting only 2 attributes. They are block_name and gender_name. Table 1 shows the result of the experiment. The results showed that the majority of the female students were classed into these 5 clusters.

a) Hypothesis 1: Most of the male students were diagnosed to have diseases related to the respiratory system

Table 1
Result According To Gender

Attributes	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
gender_name	Female (1065)	Male (803)	Female (41)	Female (38)	Male (37)
block_name	J00-J06 Acute upper respiratory infections	J00-J06 Acute upper respiratory infections	J40-J47 Chronic lower respiratory disease	J40-J47 Chronic lower respiratory disease	J30-J39 Other diseases of upper respiratory tract

Figure 2 shows the result from the first cycle of the experiment. We chose the number of clusters on the x-axis and the gender on the y-axis. The z-axis chosen was gender too. From 5 clusters, it showed that 3 clusters were dominated by the female patients. Therefore, hypothesis 1- most of the male students were diagnosed to have diseases related to the respiratory system cannot be proven since the result showed the female students were more

likely to have diseases related to the respiratory system. This result is very general; therefore, the research scope was narrowed down by making another hypothesis.

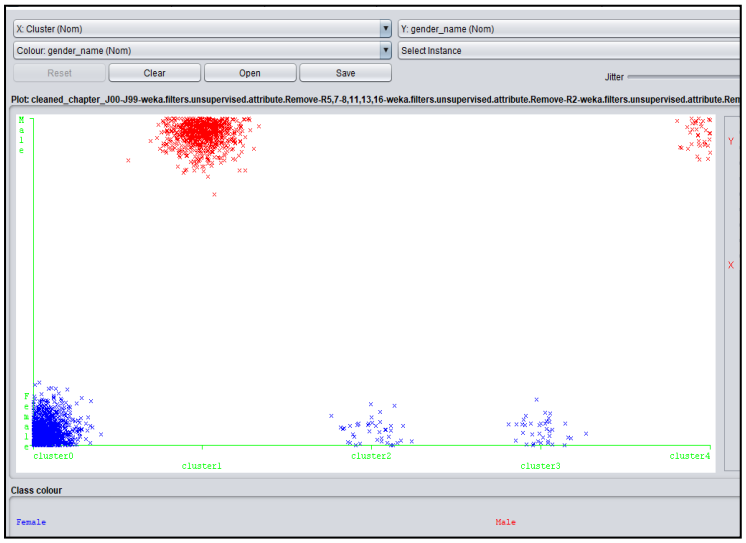

Fig.2: Cluster vs. gender

b) Hypothesis 2: Most of the male students were diagnosed to have disease that related to respiratory conditions due to smoke inhalation

The experiment was conducted using 4 attributes which were gender_name, block_name, division_description and subdivision_description. There were 4 main diagnoses related to J06.9 Acute Upper Respiratory Infection, Unspecified, {[Upper Respiratory: Disease, Acute, Infection NOS]}. Extension information found that J06.9 which is related to infection, respiratory and upper is actually related to the infection of an external agent. However, it was assumed that this dataset is not as detailed as it should be; under the subdivision_description, there should be another code that specified the symptom was related to the smoke inhalation (J68.2)

Table 2
Results based on 4 attributes

Attributes	C1	C2	C3	C4	C5
gender_name	F (1065)	M (803)	F (41)	M (38)	M (86)
block_name	J00-J06	J00-J06	J40-J47	J40-J47	J30-J39
division_description	J06	J06	J06	J06	J03
subdivision_description	J06.	J06.9	J06.9	J06.	J03. s

The clustering assignments graph displayed is based on the subdivision_description (x-axis), cluster (y-axis) and gender_name (z-axis) is shown in Figure 3 below. The blue dots represented the female patients and the red dots represented the male patients. We clicked at the blue dot, to discover the output of random instances in cluster 3. These instances were classified in cluster 3 as it tallied to the main result that displayed in Table 2.

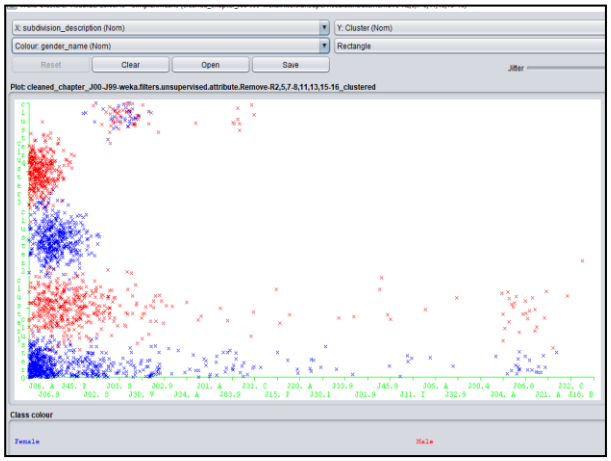

Fig. 3: Smoke inhalation

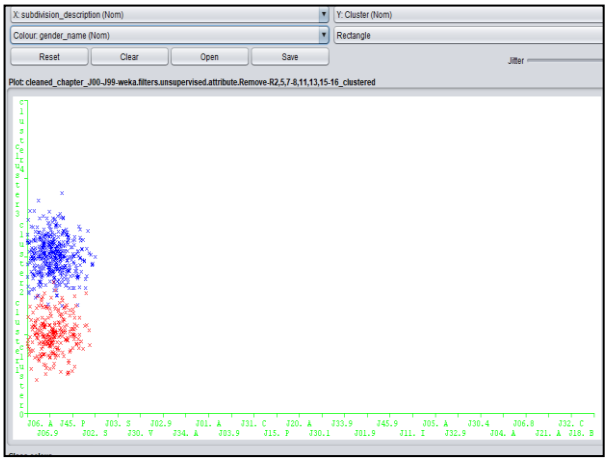

Fig. 4: Focus on J06.9 subdivision_description

The focus is on the subdivision_description = J06.9 (Figure 4). The clustering assignment's result is as below. The blue dots represented the female patients and the red dots represented the male patients. The blue dot showing output of random instances in cluster 3. These instances were classified in cluster 3 as it also tallied to the main result that displayed in Table 2. It was found that these instances were grouped into a cluster based on the similarity of gender_name, block_name, division_description and subdivision_description. However, there was a lack of interpretation as the cluster was labelled starting from 0, not 1 – the result shown was cluster 2 not cluster 3 and cluster 1 represented cluster 2. The comparison between two clusters (cluster 2 and cluster 3) was made. The instances classed into cluster 2 were 803 (male) and the instances classed into cluster 3 were only 41 instances (female). Hence, the hypothesis 2: most of the male students were diagnosed to have disease that related to respiratory condition due to smoke inhalation was proven after conducting this experiment.

Based on the experiment conducted, we have identified the techniques that are commonly used in order to develop predictive analytical models related to the health data. The data retrieved was health datasets related to the diagnosis that occurred from 2011 to 2015. However, only 3% of the collected data can be used to perform this research. Due to the missing attribute values in the input dataset and the confidential issue from the diagnosis type of datasets, the experiment focused on the most common diseases among students. Based on the predictive analytical model developed, we were able to evaluate the interestingness of the resulting pattern. We were able to identify the most common disease for the students over the past five years.

Conclusion

This research is about data-driven predictive analytics in healthcare. The result of this research is identifying the common disease and prevalence causes among students in using Weka. The results identified the highest probability disease occurred based on the history of patients' records. The predictive analytical model is developed based on the common symptoms that usually occurred among students. Based on the predictive result, we are able to understand patterns on raw data and interpret its results. The results will help the organization in general or the university specifically to identify the likelihood of future outcomes. It is very important to understand the probability of common diseases throughout

the year as the health center would be able to accommodate its resources accordingly. In addition, the potential of the predictive model from the data analytics would greatly help the health center to procure sufficient medical supplies and coordinate treatment for the residents in the most optimum way. However, this study has several limitations. One of the limitations is the condition of the historical data itself. Some of the data were incomplete and mismatched. Therefore, a full cycle experiment could not be conducted. This however can be fixed if mutual cooperation of the university health center with our research team should improve in the near foreseeable future. In addition, privacy, security and confidentiality concerns about patient data is also an issue that was raised as the research progressed. It is hoped new data insights can be acquired in the future to produce new insights and innovations in the field of data analytics.

Acknowledgements

Special thanks to Pusat Kesihatan Universiti Teknologi MARA (UiTM) for your active participation and inputs to this research. This research is fully supported by Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara (UiTM).

References

Abawajy, J. (2014), Comprehensive analysis of big data variety landscape, *International journal of parallel, emergent and distributed systems*, vol. 30, no. 1, 5-14, Retrieved December , 2019, from https://www.oracle.com/big-data/

Ariffin, M. A. M., Ishak, R., Ahmad, S. A., & Kasiran, Z. (2020). Network Traffic Profiling Using Data Mining Technique in *Campus Environment. International Journal*, 9(1.3).

Basheer, M. Y. I., Mutalib, S., Hamid, N. H. A., Abdul-Rahman, S., & Ab Malik, A. M. (2019). Predictive analytics of university student intake using supervised methods. *IAES International Journal of Artificial Intelligence,* 8(4), 367.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7), 1123-1131.

Conn, J. (2014). Predictive analytics tools help hospitals reduce preventable readmissions. *Modern healthcare*, 44(31), 16.

Fang, H., Zhang, Z., Wang, C. J., Daneshmand, M., Wang, C., & Wang, H. (2015). A survey of big data research. *IEEE Network*. Institute of Electrical and Electronics Engineers Inc.

G. Sahle (2016), Ethiopic maternal care data mining: discovering the factors that affect postnatal care visit in Ethiopia. *Health Information Science and Systems*, 4, 4.

Lin, Y. K., Chen, H., Brown, R. A., Li, S. H., & Yang, H. J. (2017). Healthcare predictive analytics for risk profiling in chronic care: A Bayesian multitask learning approach. *Mis Quarterly*, 41(2).

Liu, W., & Park, E. K. (2014). Big data as an e-health service. *In 2014 International Conference on Computing, Networking and Communications, ICNC 2014*, 982–988. IEEE

Mukherjee, S. (2019). Predictive Analytics and Predictive Modeling in Healthcare. Available at SSRN 3403900.

Mohsen, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I., Siddiqa, A., & Yaqoob, I. (2017). Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges. *IEEE Access,* 5, pp.5247–5261.

Niaksu, O. (2015). CRISP data mining methodology extension for medical domain. *Baltic Journal of Modern Computing*, 3(2), 92.

Noor, N. L. M., Aljunid, S. A., Noordin, N., & Teng, N. I. M. F. (2018). Predictive Analytics: The Application of J48 Algorithm on Grocery Data to Predict Obesity. *In 2018 IEEE Conference on Big Data and Analytics (ICBDA)*, 1-6. IEEE.

Zand, H. K. K. (2015). A comparative survey on data mining techniques for breast cancer diagnosis and prediction. *Indian Journal of Fundamental and Applied Life Sciences*, 5(S1), 4330-4339.

Zhang, Y., Ren, J., Liu, J., Xu, C., Guo, H., & Liu, Y. (2017). A survey on emerging computing paradigms for big data. *Chinese Journal of Electronics*, 26(1), 1-12.