Community Service Increasing Lecturer Competence Through Data Mining Training Using Rapidminer Tools in the Master of Public Health Study Program, Faculty of Health, Hang Tuah University, Pekanbaru

Eka Sabna[1], Azlina[2], Arif Arrafi[3]

Hang Tuah University, Pekanbaru, Indonesia[1,2,3]

Corresponding Author:
Eka Sabna
(es3jelita@htp.ac.id)

Received: October 18, 2024
Revised: November 07, 2024
Accepted: November 16, 2024
Published: December 05, 2024

ABSTRACT

Technological developments in the Industry 4.0 era open up enormous opportunities in data collection and processing. Currently, health data makes up around 30% of all global data, and by 2025, this figure will reach 36%. The ability to understand such segmented data can provide a major strategic advantage to medical organizations everywhere. This large amount of data can be carried out in research using various approaches, including using the Data Mining Approach. Data Mining can be applied to find knowledge patterns from patient profiles and health history data (patient history data). The knowledge gained can be used for analysis and decision making, including to predict the type of disease, determine the pattern of disease spread, and see the effectiveness of treatment.

Some Lecturers in the Master of Public Health Study Program do not know much about the basic concepts of Data Analysis using Data Mining concepts. This activity aims to provide knowledge about analyzing health data using a Data Mining Approach to lecturers in the Master of Public Health Study Program. The Data Mining technique discussed is the prediction of diabetes using the Decision Tree Algorithm. The data used was obtained from public data, namely Kaggle.

Keywords: *Diabetes, Data Mining, Prediction, Lecturers, Decision Tree*

1. INTRODUCTION

Currently, health data makes up around 30% of all global data, and by 2025, this figure will reach 36%. The ability to understand such segmented data can provide a major strategic advantage to medical organizations everywhere. Technological developments in the Industry 4.0 era open up enormous opportunities in data collection and processing. This large amount of data can be carried out in research using various approaches, including using the Data Mining Approach, which is one of the technological developments in the computer field (Fayyad et al., 1996) . Data Mining techniques are becoming increasingly popular and increasingly important, especially in the medical field. Data Mining can be defined as the process of searching for previously unseen patterns and trends in very large amounts of data (Jiawei, n.d.) (Oded & Lior, 2010). Data Mining can be applied in analyzing health data. Data Mining is a data processing method to look for hidden patterns in the data so that these patterns can be used as knowledge (Kumar Yadav, 2012)(Jiawei, n.d.). In general, the application of Data Mining in the health sector includes improving clinical decision making, increasing diagnostic accuracy, increasing treatment efficiency, avoiding dangerous drug and food interactions, and better customer relationships (4).

The tool used for this training is Rapidminer. RapidMiner is open source software. RapidMiner is a solution for analyzing data mining, text mining and predictive analysis. RapidMiner uses various descriptive and predictive techniques to provide insights to users so they can make the best decisions (5).

The Public Health Masters Study Program, Faculty of Health, Hang Tuah University, Pekanbaru, is one of the Study Programs at Hang Tuah University, Pekanbaru. Lecturers in the Master of Public Health Study Program have been carrying out the data analysis process using statistics and SPPS software. Lecturers do not know much about the basic concepts of Data Analysis using Data Mining concepts. This training provides another alternative in data analysis, namely the Data Mining approach and Rapidminer tools which are currently the choice of health sector researchers to carry out data analysis such as Prediction, Classification, Clustering and Data Association.

The Public Health Study Program is one of the Study Programs at Hang Tuah University Pekanbaru. Public Health Study Program lecturers do not know much about the basic concepts of Data Analysis using Data Science concepts, therefore training activities are proposed for Health Data Analysis using Data Science. This activity aims to provide knowledge about analyzing health data using Data Science to Lecturers in the Public Health Study Program.

The aim of this activity is to provide training on Data Science so that lecturers can analyze health data using a Data Science approach (Kumar Yadav, 2012). The benefit of this activity is to provide and increase lecturers' knowledge in analyzing health data so that lecturers in public health study programs can use Data Science as another alternative where up to now the data analysis process has only used statistical analysis.

2. METHODS

The implementation method for this PKM activity is planned for 8 months which includes several stages. The stages in implementing PKM are as follows:

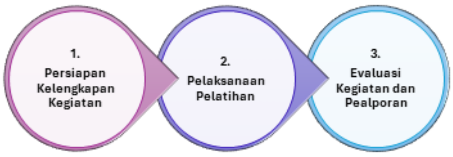

1.
Persiapan Kelengkapan Kegiatan
2.
Pelaksanaan Pelatihan
3.
Evaluasi Kegiatan dan Pealporan

Figure 1. Method of Implementing PKM Activities

Figure 1 shows the stages of PKm implementation which consists of 3 stages, namely:

1. **Preparation of activity equipment**

   In this activity there are several activities carried out including:

   a. Coordinating the Pkm Team with the Master of Public Health Study Program regarding training schedules and locations.
   b. Preparing training equipment, namely preparing the software used during training.
   c. Preparation of Materials and Modules for participants. The modules used in the training are made in the form of tutorials and theories to make it easier for participants to understand the material.

2. Implementation of Training

Training will be carried out where participants will be given knowledge about Data Mining. Training material to lecturers will be delivered following the method/steps in Data Mining, namely CRIPS-DM (Mauritsius & Binsar, 2020) (Hotz, 2024) . Figure 2 shows the stages of CRIPS-DM:

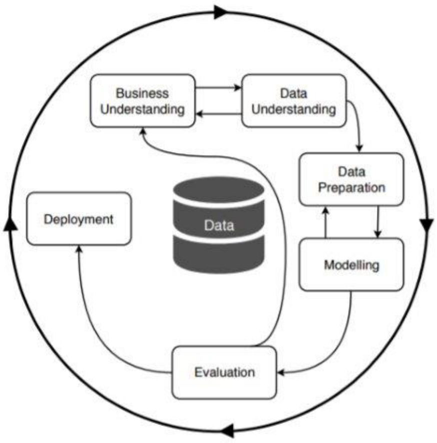


Figure 2. Steps for Implementing Data Mining Training

Details of the training implementation method follow the CRISP-DM flow as in Figure 2.

1) ***Business Understanding***
Activities carried out include: clearly determining overall goals and requirements, translating these goals and determining limitations in formulating Data Mining problems, and then preparing an initial strategy to achieve these goals.

2) ***Data Understanding***
This stage provides the analytical foundation for a study by creating a summary and identifying potential problems in the data.

3) ***Data Preparation***
Activities carried out include: selecting cases and parameters to be analyzed (Select Data), carrying out transformations on certain parameters (Transformation), and cleaning the data so that the data is ready for the modeling stage (Cleaning). The process at this stage uses the Rapidminer application.

4) ***Modeling***
At this stage, statistical and machine learning methods are used to determine the data mining techniques, data mining tools and data mining algorithms that will be applied. Then the next step is to apply the Data Mining techniques and algorithms to the data with the help of Rapidminer tools.

5) ***Evaluation***
Interpreting the results of Data Mining produced in the modeling process in the previous stage. Evaluation is carried out on the model applied in the previous stage with the aim that the determined model is in accordance with the objectives to be achieved. The process at this stage uses several evaluation techniques including Cross Validation, Confusion Matrix, RMSE MSE R2 and Davies Bouldin.

6) ***Deployment***
Using the generated model and generating reports.

3. Evaluation of Activities and Reporting

a. Evaluation

After the PKM activities are implemented, evaluation needs to be carried out by:

1) Measuring the effectiveness and efficiency of the training carried out.
2) Continuity of collaboration with partners will continue after PKM activities, such as participating in training and mentoring related to the health data analysis process.

b. Report

The final stage in this activity is compiling a report from the beginning of the activity to the evaluation stage. This report can be used as a reference.

3. RESULTS AND DISCUSSION

Training materials for lecturers are delivered according to the methods/steps in Data Mining. The case discussed is Prediction of Diabetes Patients (Zunaidi et al., 2020) (Hussein, 2020). The following are the stages of the data mining process using CRISP-DM :

a) Business Understanding

Problems:

- This data set comes from the National Institute of Diabetes and Digestive and Kidney Diseases. The goal of this data set is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the data set.
- Specifically, all patients here are women at least 21 years old. In the file (.csv) we can find several variables, some of which are independent (several medical predictor variables) and only one target dependent variable (Result).time.
- Objective: Find patterns from data sets of patients suffering from diabetes to predict diabetes.

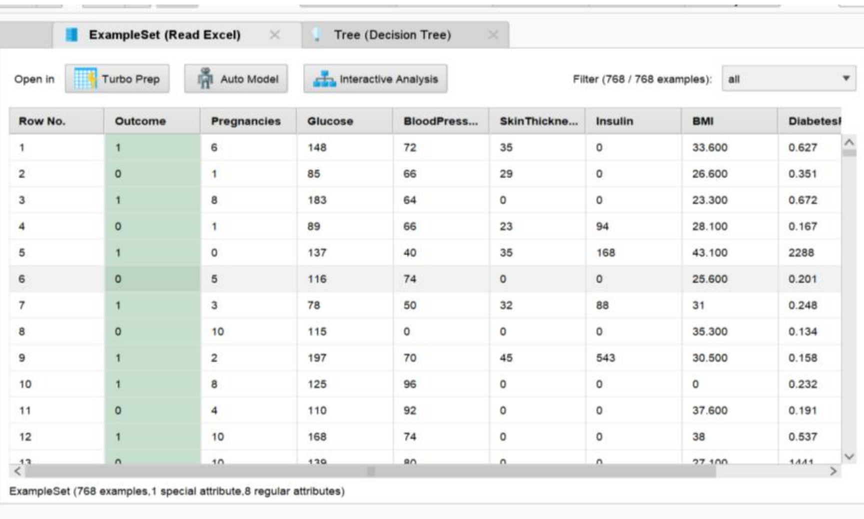
b) Data Understanding

Dataset: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset.
Information about dataset attributes :

- Pregnancies : To express the Number of pregnancies (To express the number of pregnancies)
- Glucose (Glucose): To express the Glucose level in blood (to express the Glucose level in the blood)
- BloodPressure (Blood Pressure): To express the Blood pressure measurement (to express the blood pressure measurement).
- SkinThickness: To express the thickness of the skin.
- Insulin: To express the insulin level in blood (to express the insulin level in the blood)
- BMI: To express the Body mass index (to express body mass index)
- DiabetesPedigreeFunction: To express the Diabetes percentage (to express the percentage of Diabetes)
- Age: To express the age (to express age)
- Outcome: To express the final result 1 is Yes and 0 is No (to express the final result)

c). Data Preparation

This stage ensures that the data is clean and there are no missing data, duplicate data and data anomalies.

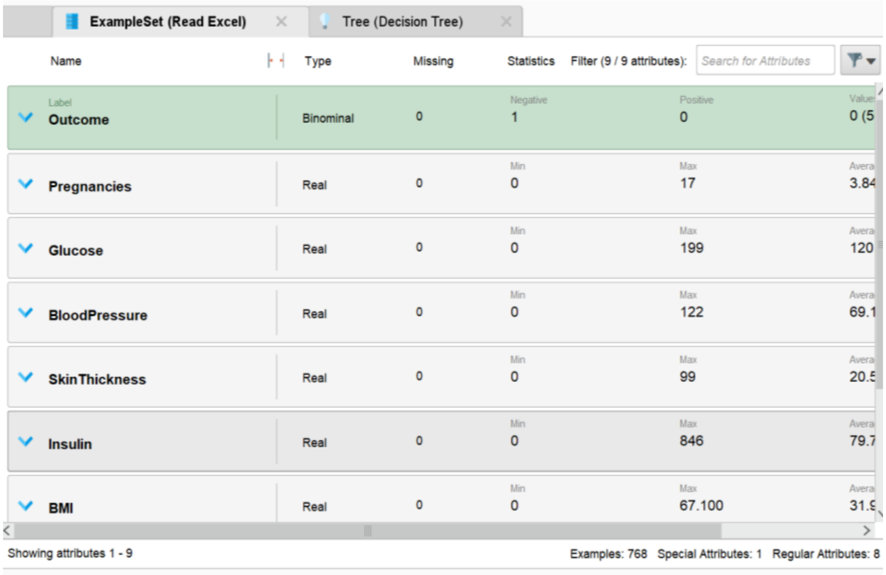


Row No.	Outcome	Pregnancies	Glucose	BloodPress...	SkinThickne...	Insulin	BMI	Diabetes!
1	1	6	148	72	35	0	33.600	0.627
2	0	1	85	66	29	0	26.600	0.351
3	1	8	183	64	0	0	23.300	0.672
4	0	1	89	66	23	94	28.100	0.167
5	1	0	137	40	35	168	43.100	2288
6	0	5	116	74	0	0	25.600	0.201
7	1	3	78	50	32	88	31	0.248
8	0	10	115	0	0	0	35.300	0.134
9	1	2	197	70	45	543	30.500	0.158
10	1	8	125	96	0	0	0	0.232
11	0	4	110	92	0	0	37.600	0.191
12	1	10	168	74	0	0	38	0.537
13	0	10	139	80	0	0	27.100	1441

ExampleSet (768 examples,1 special attribute,8 regular attributes)

Figure 3 . Data set

In figure 3 , there are 768 patient data with 9 attributes (1 target/label attribute and 8 regular attributes) . In figure 4 , missing does not contain missing values.



Name	Type	Missing			
Label Outcome	Binominal	0	Negative 1	Positive 0	Value 0 (5
Pregnancies	Real	0	Min 0	Max 17	Avera 3.84
Glucose	Real	0	Min 0	Max 199	Avera 120
BloodPressure	Real	0	Min 0	Max 122	Avera 69.1
SkinThickness	Real	0	Min 0	Max 99	Avera 20.5
Insulin	Real	0	Min 0	Max 846	Avera 79.7
BMI	Real	0	Min 0	Max 67.100	Avera 31.9

Showing attributes 1 - 9

Examples: 768   Special Attributes: 1   Regular Attributes: 8

Figure 4. There are no missing valuesd.

d). Modeling

The modeling used is Decisive Tree. Figure 4 Decision Tree modeling with Rapidminer.

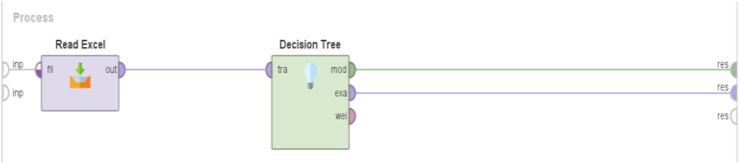


Figure 5. Prediction Model Decision Tree

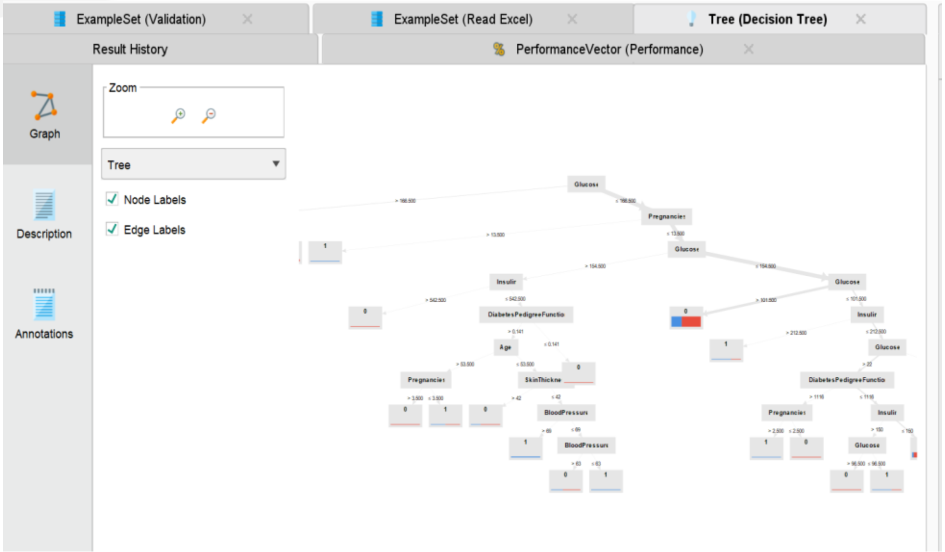


Figure 6. Results of the Decision Tree Algorithm process

e). Evaluation

This stage is to evaluate the predictive model. 10-Fold Cross-Validation is a model evaluation technique used to accurately measure the performance of machine learning models and prevent overfitting. This process involves dividing the dataset into 10 mutually exclusive parts (folds), where each part is used as test data in turn, while the other parts are used as training data (Husen et al., 2022) . 10-Fold Cross-Validation Process (Figure 7) :

1. Data Sharing:

   The dataset is divided into 10 subsets (fold). Each subset is the same size, if possible.

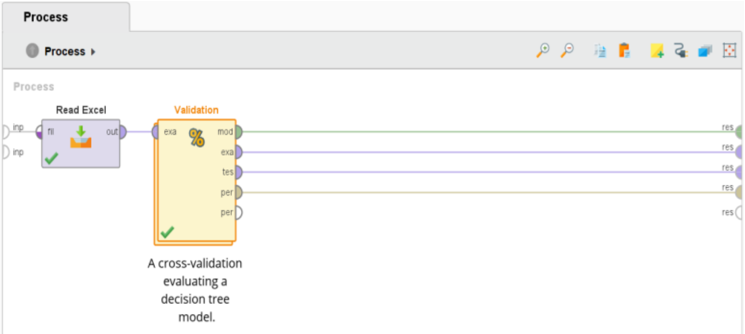
2. Training and Testing:

   The training and testing process was carried out 10 times. At each iteration:

   - One subset (fold) is used as test data.
   - The remaining 9 subsets are used to train the model.

3. Repetition:

   This process is repeated until each subset has been used as test data once. Thus, every data in the dataset will be tested.



A cross-validation evaluating a decision tree model.

Figure 7. Evaluation process with 10-Fold Cross-Validation

The following is some documentation during the activity. This activity was carried out by lecturers in the Public Health Postgraduate Study Program (Figure 8).

Figure 8. Documentation of activities during training

4. CONCLUSION

The implementation of this PkM activity was driven by several factors. First, the study program is very cooperative. Second, is support from the lecturers. Openness from lecturers (participants) to receive new knowledge. Suggestions for further activities are to collaborate with other parties so that knowledge about data mining can be applied more broadly to the health sector.

Thank-you note

Thank you to the Hang Tuah Pekanbaru Foundation through the Hang Tuah Pekanbaru Research and Community Service Institute for providing moral and material support so that this activity can be carried out well and smoothly.

REFERENCES

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37–37. https://doi.org/10.1609/AIMAG.V17I3.1230

Hotz, N. (2024). What is CRISP DM? - Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

Husen, D., Sandi, D., Bumbungan, S., Yogyakarta, U. A., Informatika, M. T., Mining, D., & Forest, R. (2022). Analisis Prediksi Kebakaran Hutan dengan Menggunakan Algoritma Random Forest Classifier. 16, 150–155.

Hussein, M. (2020). Prediksi Harga Minyak Dunia Dengan Metode Deep Learning | Hussein | Fountain of Informatics Journal. https://ejournal.unida.gontor.ac.id/index.php/FIJ/article/view/4446/pdf_60

Jiawei, H. (n.d.). Data mining concepts and techniques - 2006. Retrieved May 22, 2023, from https://elibrary.bsi.ac.id/readbook/221528/data-mining-concepts-and-techniques

Kumar Yadav, S. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification Saurabh Pal. World of Computer Science and Information Technology Journal (WCSIT), 2(2), 51–56.

Mauritsius, T., & Binsar, F. (2020). Cross-Industry Standard Process for Data Mining (CRISP-DM) – MMSI BINUS University. https://mmsi.binus.ac.id/2020/09/18/cross-industry-standard-process-for-data-mining-crisp-dm/

Oded, M., & Lior, R. (2010). Data Mining And Knowladge Discovery Handbook. In Journal of Chemical Information and Modeling (Vol. 53, Issue 9). https://doi.org/10.1017/CBO9781107415324.004

Zunaidi, M., Nasyuha, A. H., & Sinaga, S. M. (2020). Penerapan Data Mining Untuk Memprediksi Pertumbuhan Jumlah Penderita Human Immunodeficiency Virus (HIV) Menggunakan Metode Multiple Linier Regression (Studi Kasus Dinas Kesehatan Provinsi Sumatera Utara). Jurnal Teknologi Sistem Informasi Dan Sistem Komputer TGD, 3(1), 137–147. https://doi.org/10.53513/JSK.V3I1.205.