JOIV : Int. J. Inform. Visualization, 8(1) - March 2024 318-325

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Personalized Learning Models Using Decision Tree and Random Forest Algorithms in Telecommunication Company

Alexander Bryan Wiratman [a,*], Wella [a]

[a] *Information System Department, Universitas Multimedia Nusantara, Serpong, Tangerang, 15810, Indonesia*

Corresponding author: [*]alexander.bryan@student.umn.ac.id

***Abstract*— In response to the rising popularity of online training, this study addresses the crucial need for effective assessment methods at PT XYZ. The research focuses on developing a comprehensive solution through a data visualization dashboard and a machine learning model. The data visualization dashboard, created using Tableau, provides an interactive platform for exploring training data. It offers valuable insights into employees learning progress and needs, empowering them to monitor their advancement and identify areas for improvement effectively. Simultaneously, a machine learning model was developed using Python and Google Collab, employing decision trees and random forest algorithms. The model exhibited promising results with an accuracy rate of 69% for decision trees and 70% for random forests, indicating its proficiency in predicting skill groups. Furthermore, the study rigorously evaluated the dashboard and machine learning model using a 20% holdout dataset, affirming their effectiveness. The dashboard, deployed on a web server, ensures accessibility to all PT XYZ employees, enhancing user experience and engagement. Notably, the dashboard's user-friendly interface allows employees to actively participate in their learning journey, while the machine learning model generates personalized training recommendations based on their progress and needs. In summary, this research provides a practical and innovative solution to the challenge of online training assessment at PT XYZ. By combining data visualization techniques and machine learning algorithms, the developed tools significantly enhance the efficiency and effectiveness of training programs. These findings contribute valuable insights into online training assessment methodologies and pave the way for improved learning experiences in the digital age.**

***Keywords*— Data visualization; machine learning model; classification; employee training.**

Manuscript received 15 Jun. 2023; revised 26 Oct. 2023; accepted 7 Nov. 2023. Date of publication 31 Mar. 2024.

I. INTRODUCTION

The first outbreak of the Covid-19 pandemic in Wuhan, China, in December 2019, has spread worldwide, becoming a global pandemic that has impacted the healthcare, social, economic, and environmental sectors [1], [2]. Many countries have implemented lockdowns and social restrictions to minimize the spread of the virus, forcing people to adapt to new situations, such as the concept of work-from-home (WFH), to limit public mobility [3]–[5]. This pandemic has increased the utilization of technology to replace face-to-face activities, including learning, working, and even shopping. The WFH concept offers advantages such as flexibility and a better work-life balance, thus opening new opportunities to leverage online technologies in the workplace [6]. In Indonesia, the transition from a pandemic to an endemic status of COVID-19 has influenced government policies regarding restrictions on public activities, including changes in the working environment toward more freedom [7]. Companies like PT XYZ need to establish policies on adopting a work-from-office (WFO) or WFH approach.

PT XYZ is a telecommunications company with over 5000 employees that provides training programs as one of its policies to enhance employee productivity and performance and strengthen the company's culture [8]. Digitalization serves as PT XYZ's solution in managing employee training procedures, especially during the pandemic, by adopting online learning services from various providers such as Udemy, Percipio, and LinkedIn Learning [9]. In the past, PT XYZ implemented onsite training initiatives. However, due to the pandemic outbreak, the company was compelled to impose limitations on physical mobility. Consequently, the firm had to change its training methods by transitioning to online platforms to ensure the continuity of its training programs. Currently, PT XYZ continues to conduct training activities online in conditions in Indonesia that are getting better from a pandemic to an endemic, offering access to

online learning services divided into two types: assigned learning and self-learning [10]. To cater to the assigned learning type, organizations want technologies that can effectively facilitate the process of picking training courses that align with both the specific field of work and the individual preferences of employees, ensuring accuracy in the training selection process.

To achieve those goals, several data classifications can be used. Classification is a supervised machine learning method where the model tries to predict the correct input data label [11]. When doing classification, the model is fully trained using the training data, assessed using test data, and then utilized to make predictions on fresh, unused data [12]. Two widely employed techniques for data classification are decision trees and random forests [13]. Most decision trees are made up of two major procedures: the building (induction) and the classification (inference) procedures [14]. As decision trees use the "divide and conquer" method, they perform well if a few highly relevant attributes exist but less if many complex interactions are present [14].

The decision tree algorithm is associated with certain drawbacks. Firstly, it exhibits an unstable nature, which is a notable weakness. This means that even slight modifications in the input data can lead to substantial alterations in the structure of the decision tree. Secondly, the algorithm is less proficient in accurately forecasting outcomes based on continuous variables [15]. The Random Forest (RF) classifier exhibits lower sensitivity than other machine learning classifiers about the quality of training samples and overfitting. This is attributed to generating many decision trees by randomly selecting training samples and variables for splitting at each node within the trees [16]. The RF classifier is suitable for classifying hyperspectral data, where the curse of dimensionality and highly correlated data pose major challenges to other classification methodologies [16]. The Random Forest algorithm possesses certain limitations, primarily the challenge of interpretation and the necessity of appropriately tweaking the model to suit the data [17].

Data classification using machine learning algorithms, such as random forest and decision tree, within the context of developing a machine learning model using the CRISP-DM methodology is a crucial step in addressing the research problem. This study aims to create a personalized learning classification model in providing assigned learning topics for employees. Furthermore, the research includes data visualization to make dashboards that can be used for the company to make decisions from the information shown in the dashboard. Decision-making is one of the company's strategies to gain profits in a volatile and competitive economy [18].

II. Materials and Method

However, there is an issue regarding the creation of assigned learning due to the inability to track and analyze employee training activities as effectively as offline training. Therefore, there is a need to develop data visualization for employee training, which can map out their activities in both self-learning and assigned learning. Additionally, classification based on self-learning should be implemented to facilitate the provision of assigned learning that aligns with individual employee preferences (personalized learning). As one of the outcomes of this research topic, data classification has become an essential part of the field of artificial intelligence and has grown rapidly in recent years. Classification can be done for various purposes with different data types, such as images and text. An example of classification with images can be done to map and inventory a wetland area [19]. Classification with text data can also be done, such as in the banking world, to predict the level of fraud in credit card usage [20]. Classification with various data types can be done using machine learning and various algorithms that can be compared for their quality in finding the best algorithm for a machine learning model in data classification problems [21].

Various machine learning algorithms for classification, such as decision trees and random forests, are often used because they produce very good accuracy [20]–[22]. One study found that random forest had the best accuracy in researching the severity of road accidents compared to logistic regression and decision tree algorithms and analyzing predictors of rapid eye movement sleep behavior disorder in Parkinson's patients compared to decision tree algorithm [21], [22]. In another study, decision trees showed good accuracy even when compared to random forests, specifically in sentiment analysis on YouTube and predicting the operational efficiency of banks [23], [24]. Data classification and the development of machine learning models as part of the data mining process can be applied using the most popular method in data mining projects, namely CRISP-DM [25]. CRISP-DM has been de facto established in various data mining projects or research for over two decades [26]. CRISP-DM has successfully addressed cases such as classification and data visualization [4], [27]–[29]. CRISP-DM consists of six comprehensive stages, from problem identification to applying the built solution through model formation and evaluation [27]. CRISP-DM consists of six stages, which are depicted in Figure 1: business understanding, data understanding, data preparation, modelling, evaluation, and deployment [30].

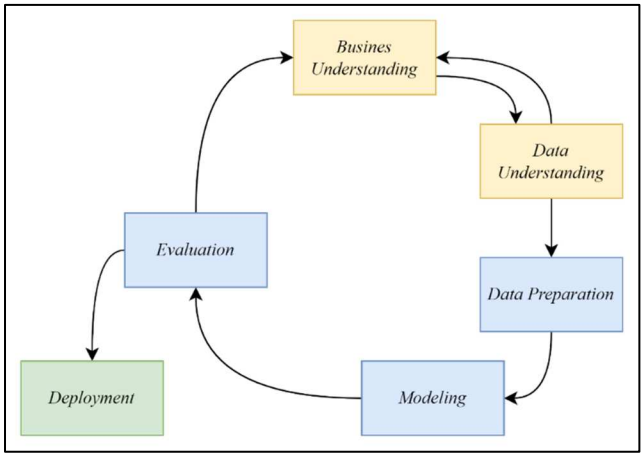

Fig. 1 Research Methodology CRISP-DM

III. Results and Discussion

This study applies data mining techniques, combining the development of machine learning models and data visualization [4]. The machine learning models utilized in this research are decision tree and random forest, implemented using the Python programming language and Google Collab tools. Data visualization uses the Tableau tool to recognize data patterns [30]. The data mining methodology employed in this study is CRISP-DM (CRoss Industry Standard for Data Mining), which offers a systematic approach to problem-solving where it can help support business decisions by defining a non-rigid six-phase process [31].

A. Business Understanding

On January 6, 2023, PT XYZ initiated information gathering through an online interview using Microsoft Teams to address issues related to employee training data. The interview with a Senior People Analytics and Data Intelligence Analyst revealed that PT XYZ, a telecommunications service provider in Indonesia with over 5,000 employees, had implemented online training due to the COVID-19 pandemic. The company utilized platforms like Udemy, Percepio, and Linkedin Learning for self-learning and assigned learning activities. A data mining project was undertaken to analyze and transform the training data into valuable insights, creating a data visualization dashboard and a personalized learning model. These developments aimed to assist PT XYZ in making informed decisions and providing tailored assigned learning to employees based on their preferences.

B. Data Understanding

The data from PT XYZ regarding employee training activities was gathered through an interview process with PT XYZ representatives. The data used consisted of employee training activities from the company's LinkedIn Learning platform, specifically from July 2022 to February 2023, as it marked the start of the platform's usage by PT XYZ and the beginning of the research. The data was in Microsoft Excel format (.xlsx) with a size of 7,898 KB and contained 49,838 rows, including the header and 29 columns. Many rows were due to the data displaying training activities for each employee on different dates and times for each accessed learning content. The contents of the first five rows of the data can be seen in Figure 2. The library used for data manipulation was straightforward, using Pandas to read the data in Microsoft Excel format (.xlsx).

Learning	NIK	Direktor	Dir_Sho	Band	Admins	Area	Age	Age Ran	YoS
42070	D543DF45	Planning a	PNT	1	HEAD OFFI	HEAD OFFI	21	≤ 25	0
41960	D543DF45	Planning a	PNT	1	HEAD OFFI	HEAD OFFI	21	≤ 25	0
41965	D543DF45	Planning a	PNT	1	HEAD OFFI	HEAD OFFI	21	≤ 25	0
41982	D543DF45	Planning a	PNT	1	HEAD OFFI	HEAD OFFI	21	≤ 25	0
42077	D543DF45	Planning a	PNT	1	HEAD OFFI	HEAD OFFI	21	≤ 25	0

YoS Ran	S/F	JobCat	Content Name	Content Provider	Content Type	Content ID	Hours Viewed	Percent	Started Date
≤ 2	Non S/F	Staff	A Day In The Life of a Data Scientist	LinkedIn	Course	2859039	1.0664	100%	03/09/2022
≤ 2	Non S/F	Staff	Ableton Live 11 Essential Training: T	LinkedIn	Course	2876317	9.2744	100%	10/09/2022
≤ 2	Non S/F	Staff	Accessibility for Web Design	LinkedIn	Course	606090	1.9481	100%	25/09/2022
≤ 2	Non S/F	Staff	Advance Your Data Engineering Skill	LinkedIn	Path	5a148d53498e47e9f9c70aaf		100%	11/07/2022
≤ 2	Non S/F	Staff	Advanced and Specialized Statistics	LinkedIn	Course	5034174	5.0722	100%	09/09/2022

Started Time	Latest View Date	Latest View Time	Completion Status	Total Assessments	Number of Assessments Completed	Assignment	PA H2 2022	Skills
20:30	09/04/2022	11:39	Y	0	0	Non-Assigned	P1	Career Management, Data Science
22:03	09/11/2022	18:12	Y	13	0	Non-Assigned	P1	Ableton Live
02:30	09/25/2022	16:41	Y	0	0	Non-Assigned	P1	Accessibility, COM
06:05	08/20/2022	14:47	Y			Non-Assigned	P1	Big Data, Data Engineering
03:55	09/09/2022	20:36	Y	8	0	Non-Assigned	P1	Stata, Statistical Data Analysis

Fig. 2 First Five Rows of the Data

C. Data Preparation

The acquired data will be cleaned using Python in Google Collab to handle duplicate and missing data. Duplicate data containing redundant information for employees will be filtered to retain only the first occurrence. Missing data will be identified and examined, with specific columns showing empty values. However, based on discussions with PT XYZ, it is determined that the empty data does not require modification as it does not impact calculations or data accuracy. In developing a machine learning model, particularly for classification using decision trees and random forest algorithms, further data processing is required to prepare the data for the classification process. Based on the data types of various columns or variables in the data, some variables, particularly the target variable to be used, "Skills," are of object data type. The "Skills" variable must be converted into numerical values using label encoding, simultaneously transforming it into a categorical variable for the classification process. Subsequently, the feature and target variables must be separated into two distinct datasets. Feature selection is conducted after the feature and target variables have been processed. Not all features in the data will be utilized, considering their relevance to the target variable for classification purposes. Some selected features include "Dir_Short," "Band," "Admin," "Area," "Age," "Yos," "S/F," and "JobCat." Following the feature selection, label encoding is applied to the selected features, as it is necessary for both decision tree and random forest algorithms during the classification process. Label encoding is performed on all selected features, particularly object data types. Once all the preparations related to features and targets are completed, the next step is to split the data into training and testing datasets, with 80% allocated for training and 20% for testing.

This research has resulted in two outputs: a data visualization dashboard and a machine learning model. Both outputs are part of the data mining process and have been processed using the CRISP-DM methodology, which includes business understanding, data understanding, data preparation, modeling, evaluation, and deployment [29]. The six stages of CRISP-DM can be applied to both research outputs simultaneously.

A. Modelling

The data used in creating a data visualization dashboard is the complete dataset without any modifications, similar to the

data used for building the machine learning model. The modeling phase for creating the data visualization dashboard begins with using Tableau, specifically Tableau Desktop Professional Edition version 2023.1.0 64-bit, for data visualization. There are no connections to other datasets as only one dataset is used, which includes all the variables required for data visualization.

The created data visualization dashboard, titled "Employee Training Demography," is displayed in Figures 3 and 4. The main view of the dashboard is shown in Figure 3 and includes a filter in the form of a "Date" parameter. This parameter allows the selection of desired month and year periods. The available options for the parameter/filter cover the data from July 2022 to February 2023.

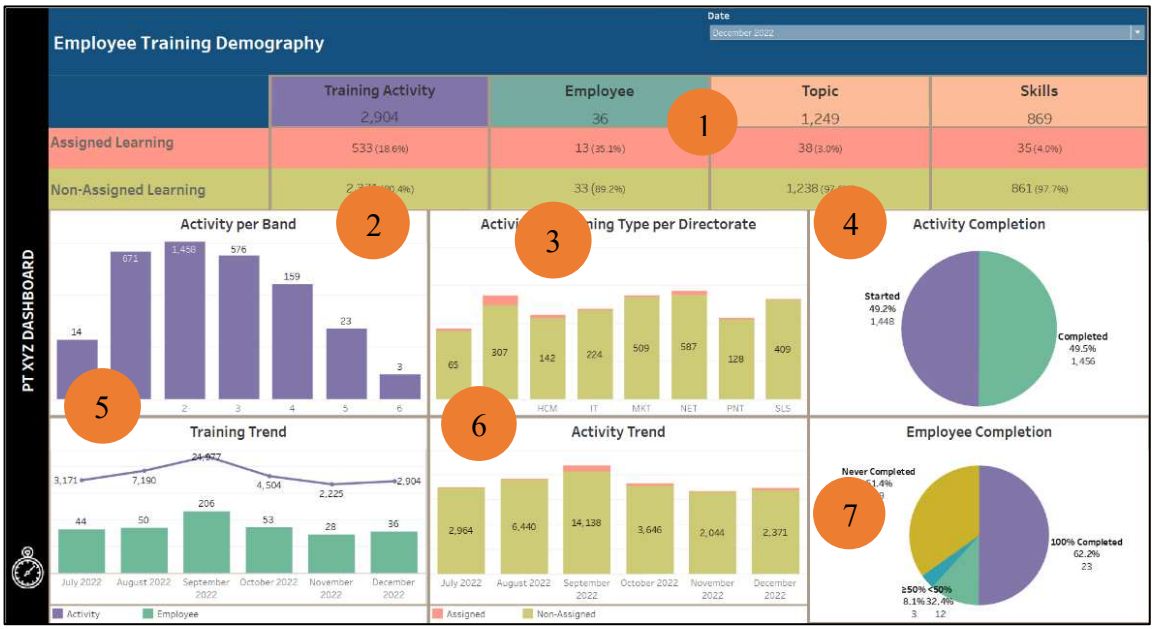

Fig. 3 Dashboard Employee Training Demography (1)

Fig. 4 Dashboard Employee Training Demography (2)

In the dashboard depicted in Figure 3, several visualizations can be described according to their numbered orange labels as follows:

1. Numeric information displays the total number of training activities, participating employees, accessed topics, and performed skills. These four pieces of information represent the combined totals for both assigned learning and non-assigned learning training types, along with the individual values for each training type, accompanied by their percentages of the total. Training activities for each selected period are calculated based on the number of Learning IDs. The number of employees is calculated based on the number of NIKs. The count of topics is based on the number of

Content Names, while the count of skills is based on the number of unique skills. The assigned learning and non-assigned learning categories are determined by the Assignment category in the data, which contains "Assigned" or "Non-Assigned."

2. The "Activity per Band" bar chart displays the number of learning activities for each band.
3. The "Activity by Learning Type per Directorate" bar chart illustrates learning activities divided into assigned and non-assigned learning for each directorate.
4. The "Activity Completion" pie chart shows the proportion of learning activities initiated or completed.
5. The "Training Trend" combined bar and line chart demonstrates the trend of learning activities compared to the number of employees. The displayed time range is the last 12 months based on the selected month and year filter.
6. The "Activity Trend" bar chart represents the trend of learning activities for each month. The displayed time range is also the last 12 months based on the selected month and year. It includes two categories: assigned learning and non-assigned learning, displayed in different colors.
7. The "Employee Completion" pie chart illustrates the proportion of employees' completion of learning activities, categorized as never completed, <50%, ⩾ 50%, and 100% completed.
8. In addition to the visualized data, a button in the dashboard's bottom-left corner acts as navigation to a more detailed dashboard related to time information, represented by a white clock shape.

The next dashboard, shown in Figure 4, generally has a similar layout to the previous dashboard. The difference lies in the button at the bottom-left corner, which now appears as an arrow and serves to return to the previous dashboard. The green-labelled visualizations in this dashboard are described as follows:

1. The "Employee" pie chart displays the proportion of employees engaging in learning activities during weekdays (Monday-Friday) and weekends (Saturday-Sunday).
2. The "Activity" pie chart illustrates. the proportion of learning activities conducted during weekdays and weekends.
3. The "Access Time by Employee" pie chart shows the distribution of employee access time for learning activities in four categories: before working hours (00:00-07:59), morning (08:00-11:59), afternoon (12:00-16:59), and after working hours (17:00-23:59).

Both dashboards depicted in Figures 3 and Figure 4 include dashboard actions that provide automatic filtering for all visualizations within the dashboard. For example, selecting a specific directorate in the "Activity by Learning Type per Directorate" visualization will filter the entire dashboard to display data only for that directorate. In the next dashboard, filters can be applied based on different time categories in the charts. For instance, selecting "Afternoon" in the "Access Time by Employee" pie chart will filter the dashboard to display data for the afternoon period. To reset the filters and view the dashboard without specific filters, the selected components within the charts can be clicked again.

The machine learning model uses decision trees and random forest algorithms in Google Collaboratory (Google Collab). The necessary libraries, such as Pandas and Sklearn, are used for decision trees and random forest models. Pandas facilitate data processing in Python, while Sklearn supports various aspects of building machine learning models, including label encoding, model creation, data splitting, and accuracy calculation.

To address the large raw data size (49,838 rows), a subset containing 40% of the data is randomly selected for model creation. This random subset selection is achieved using the Random library. The dashboard and machine learning model creation process is outlined, emphasizing the tools, libraries, and visualizations used for each step.

B. Evaluation

During the evaluation phase, the focus is on evaluating the machine learning model's performance in achieving the research problem's objectives. The data visualization dashboard moves on to the next phase directly as it does not involve the evaluation stage. The machine learning model built with the decision tree algorithm uses additional parameters, such as max_depth = 8, min_samples_split = 5, and min_samples_leaf = 2. These parameters are selected to prevent the decision tree from becoming too complex and to avoid oversampling. The accuracy of the decision tree model ranges from 9% to 10%, depending on the random subset used. However, with the same random subset, the subsequent machine learning model built with the random forest algorithm performs worse than the decision tree, with an accuracy of only 4% to 5%.

The low accuracy may be due to imbalanced classes in the 3,312 "Skills" variables targeted for classification. The number of skills employees learn is uneven, with thousands accessing certain skill-related learning. In contrast, other skills have only a few or just one employee accessing them. There are steps to address this issue, such as oversampling or under-sampling, to reduce the difference between skills for employees working on them. However, after trying these techniques, they did not yield good results regarding model accuracy like before.

The skills are grouped to address imbalanced classes caused by the 3,312 skills with significantly different numbers of employees accessing them. Grouping the skills allows multiple skills to be combined into one group, merging the number of employees accessing them. This approach is taken to avoid imbalanced classes. The skills are grouped into five categories based on keywords found in each skill. The selection of keywords is based on the keywords from skills with the highest number of employee accesses. To examine various frequently occurring keywords, a word cloud visualization using Python, specifically the Pandas, WordCloud, and matplotlib.pyplot libraries. The word cloud is generated based on each word in the "Skills" column or variable in the data, as implemented in the Python code. The resulting word cloud is displayed in Figure 5.

Fig. 5 Word Cloud Result

The keywords used to divide the skills into five groups are as follows:

TABLE I
WORD CLOUD RESULT

Group	Member
Group 1 (5.354 skills)	skills with the keyword "Management"
Group 2 (5.106 skills)	skills with the keyword "Management"
Group 3 (3.455 skills)	skills with the keyword "Microsoft"
Group 4 (2.381 skills)	skills with the keyword "Data"
Group 5 (33.541 skills)	skills with other keywords

These five groups have relatively balanced numbers, with more than 2,000 employees. The number of employee accesses for each group is displayed in Figure 4.20. It can be concluded that the group with the highest employee access is Group 5, with keywords not mentioned by other groups, followed by Group 1 with the "Management" skills keyword. The group with the fewest employee accesses is Group 4 with the "data" skills keyword. The selection of these keywords can be modified based on other relevant keywords according to the company's needs. In this study, these five groups will be used.

After dividing the groups, the next step is to retrain the machine learning models using these groups as the target. The accuracy of both the decision tree and random forest algorithms significantly improves, reaching 69% for the decision tree and 70% for a random forest. Based on the accuracy results, it can be concluded that the decision tree algorithm slightly outperforms the random forest algorithm in providing skill group recommendations with their keywords. However, the evaluation continues with calculating precision, recall, and F1 score. The evaluation results for both models are presented in Table 2.

TABLE II
EVALUATION RESULTS OF THE MODELS

	Decision Tree	Random Forest
Accuracy	69%	70%
Precision	0,35	0,25
Recall	0,23	0,22
F1 Score	0,22	0,20

Based on the accuracy, precision, recall, and F1 score results, it can be said that both machine learning models, the decision tree and random forest, have relatively balanced outcomes. The decision tree algorithm performs slightly better than the random forest algorithm in accuracy and precision but is slightly lower in recall and F1 score. Therefore, the machine learning model using the decision tree algorithm will be used as the solution for PT XYZ's problem.

C. Discussion

The research has yielded two external solutions to the problem: data visualization dashboards and machine learning models. As an integral component of the data mining project, developing the two external techniques employs the CRISP-DM methodology, which has undergone rigorous

examination in existing scholarly literature. This substantiates the widespread adoption of the data mining approach as the de facto standard in several mining projects by many authors worldwide [32]. Data visualization dashboards are created utilizing Tableau technologies, similar to those employed in prior studies, albeit on distinct subjects [33]–[36]. Tableau has demonstrated its capability to transform training data visualizations of PT XYZ into valuable insights that enhance decision-making processes. The algorithmic selection process for developing machine learning models aligns with the study objective of creating a classification system that facilitates personalized learning by recommending appropriate training activities. The chosen classification technique is derived from the decision tree and random forest algorithms, which have demonstrated high predictive accuracy and outperformed alternative classification algorithms [37], [38]. The findings of this investigation indicate that the random forest algorithms exhibit a minor superiority in terms of accuracy when compared to the decision trees, aligning with prior studies [39]. The precise outcomes of these advanced random forest algorithms exhibit a notable disparity compared to the findings of other studies [40]. It can be inferred that the previous research findings have been effectively incorporated into this study, particularly in selecting research methodologies, tools, and machine learning algorithms. Despite some conflicting findings, the results are generally consistent in supporting the superiority of the decision tree algorithm over the random forest algorithm in terms of accuracy.

IV. Conclusion

This research resulted in a data visualization dashboard and a machine learning model developed using PT XYZ's employee training data. The research followed the CRISP-DM methodology, which involved interviews to fulfill each stage. Previously, the analysis and information regarding PT XYZ's employee training activities were unavailable, but now they can be easily obtained using the Tableau dashboard created in this study. The Tableau dashboard also supports live data updates, enabling its long-term use by PT XYZ in collaboration with LinkedIn Learning as an online training provider. The machine learning model aims to predict groups of various skills with specific keywords to be assigned to employees based on their relevant criteria. Both machine learning models were compared regarding prediction quality using accuracy, precision, recall, and F1 score using the decision tree and random forest algorithms. The research concludes with the following statements based on the formulated problem: Two dashboards were successfully created using Tableau to visualize PT XYZ's online training activities. The information presented in the dashboards underwent manual data validation using raw data, and each visualization provides valuable insights for the company. The personalized learning machine learning model using the decision tree algorithm achieved good accuracy, precision, recall, and F1 score. The personalized learning machine learning model using the random forest algorithm also achieved good accuracy, precision, recall, and F1 score. Based on comparing accuracy, precision, recall, and F1 score, the decision tree algorithm outperformed the random forest algorithm for personalized learning.

Several recommendations can be applied to future research utilizing employee training data for data visualization dashboard creation and machine learning models, including: Incorporating a longer time of training data to provide more diverse information and insights in the dashboard. Exploring algorithms other than decision trees and random forests for classifying skills required for employee personalized learning. Assessing the effectiveness of the online training activity data visualization dashboard for PT XYZ employees after a specified duration involves a comprehensive analysis based on several key performance indicators (KPIs) and criteria defined by the company.

Acknowledgment

This work was supported by Universitas Multimedia Nusantara, Department of Information Systems, Faculty of Engineering, and Informatics.

References

[1] World Health Organization, “WHO Director-General’s opening remarks at the media briefing on COVID-19 - 11 March 2020,” https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

[2] M. Mofijur et al., “Impact of COVID-19 on the social, economic, environmental and energy domains: Lessons learnt from a global pandemic,” Sustainable Production and Consumption, vol. 26, pp. 343–359, Apr. 2021, doi: 10.1016/j.spc.2020.10.016.

[3] World Economic Outlook, “Global economy on firmer ground, but with divergent recoveries amid high uncertainty,” https://www.imf.org/en/Publications/WEO/Issues/2021/03/23/world-economic-outlook-april-2021.

[4] S. Navisa, Luqman Hakim, and Aulia Nabilah, “Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM,” Jurnal Sistem Cerdas, vol. 4, no. 2, pp. 114–125, Aug. 2021, doi:10.37396/jsc.v4i2.162.

[5] R. Rafiq, M. G. McNally, Y. Sarwar Uddin, and T. Ahmed, “Impact of working from home on activity-travel behavior during the COVID-19 Pandemic: An aggregate structural analysis,” Transportation Research Part A: Policy and Practice, vol. 159, pp. 35–54, May 2022, doi: 10.1016/j.tra.2022.03.003.

[6] M. J. Beck and D. A. Hensher, “Insights into the impact of COVID-19 on household travel and activities in Australia – The early days of easing restrictions,” Transport Policy, vol. 99, pp. 95–119, Dec. 2020, doi: 10.1016/j.tranpol.2020.08.004.

[7] Yefta Christopherus Asia Sanjaya and Rizal Setyo Nugroho, “Indonesia Disebut Sudah Endemi Covid-19, Ini Bedanya dengan Pandemi,” https://www.kompas.com/tren/read/2022/12/23/110913365/indonesia-disebut-sudah-endemi-covid-19-ini-bedanya-dengan-pandemi?page=all.

[8] Jatinder Kumar Jha, Jatin Pandey, and Biju Varkkey, “Examining the role of perceived investment in employees’ development on work-engagement of liquid knowledge workers: Moderating effects of psychological contract,” *Journal of Global Operations and Strategic Sourcing*, vol. 12, no. 2, Nov. 2018.

[9] Carl Dahlman, Sam Mealy, and Martin Wermelinger, “Harnessing The Digital Economy For Developing Countries,” 334, 2016.

[10] Kompas.com, “COVID-19 di Indonesia Mulai Berangsur-angsur Menjadi Endemi, Apa Artinya?,” https://nasional.kompas.com/read/2022/01/17/12000001/covid-19-di-indonesia-mulai-berangsur-angsur-menjadi-endemi-apa-artinya .

[11] P. C. Sen, M. Hajra, and M. Ghosh, “Supervised Classification Algorithms in Machine Learning: A Survey and Review,” Emerging Technology in Modelling and Graphics, pp. 99–111, Jul. 2019, doi:10.1007/978-981-13-7403-6_11.

[12] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study,” PLOS Medicine, vol. 15, no. 11, p. e1002683, Nov. 2018, doi:10.1371/journal.pmed.1002683.

[13] C. R. Dhivyaa, K. Sangeetha, M. Balamurugan, S. Amaran, T. Vetriselvi, and P. Johnpaul, "Skin lesion classification using decision trees and random forest algorithms," Journal of Ambient Intelligence and Humanized Computing, Nov. 2020, doi: 10.1007/s12652-020-02675-8.

[14] S. B. Kotsiantis, "Decision trees: a recent overview," Artificial Intelligence Review, vol. 39, no. 4, pp. 261–283, Jun. 2011, doi:10.1007/s10462-011-9272-4.

[15] M. Marudi, I. Ben-Gal, and G. Singer, "A decision tree-based method for ordinal classification problems," IISE Transactions, pp. 1–15, Jul. 2022, doi: 10.1080/24725854.2022.2081745.

[16] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 114, pp. 24–31, Apr. 2016, doi: 10.1016/j.isprsjprs.2016.01.011.

[17] R. Valavi, J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita, "Modelling species presence-only data with random forests," Ecography, vol. 44, no. 12, pp. 1731–1742, Oct. 2021, doi:10.1111/ecog.05615.

[18] "Job Security as a Mediating Variable between Innovative Leadership and Innovative Work Behavior among Employees," Journal of System and Management Sciences, vol. 13, no. 1, Feb. 2023, doi:10.33168/jsms.2023.0128.

[19] B. Carneiro da Rocha and R. Timoteo de Sousa Junior, "Identifying Bank Frauds Using CRISP-DM and Decision Trees," International Journal of Computer Science and Information Technology, vol. 2, no. 5, pp. 162–169, Oct. 2010, doi: 10.5121/ijcsit.2010.2512.

[20] N. P. Dileep, P. V. Sarma, R. Prasannachandran, V. Surendran, and M. M. Shaijumon, "Electrostatically Coupled Nanostructured Co(OH)2–MoS2 Heterostructures for Enhanced Alkaline Hydrogen Evolution," ACS Applied Nano Materials, vol. 4, no. 7, pp. 7206–7212, Jul. 2021, doi: 10.1021/acsanm.1c01163.

[21] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, and S. P. Ong, "A Critical Review of Machine Learning of Energy Materials," Advanced Energy Materials, vol. 10, no. 8, Jan. 2020, doi: 10.1002/aenm.201903242.

[22] W. Chong-Wen, L. Sha-Sha, and E. Xu, "Predictors of rapid eye movement sleep behavior disorder in patients with Parkinson's disease based on random forest and decision tree," PLOS ONE, vol. 17, no. 6, p. e0269392, Jun. 2022, doi: 10.1371/journal.pone.0269392.

[23] M. Aufar, R. Andreswari, and D. Pramesti, "Sentiment Analysis on Youtube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study," 2020 International Conference on Data Science and Its Applications (ICoDSA), Aug. 2020, doi:10.1109/icodsa50139.2020.9213078.

[24] P. Appiahene, Y. M. Missah, and U. Najim, "Predicting Bank Operational Efficiency Using Machine Learning Algorithm: Comparative Study of Decision Tree, Random Forest, and Neural Networks," Advances in Fuzzy Systems, vol. 2020, pp. 1–12, Jul. 2020, doi: 10.1155/2020/8581202.

[25] J. S. Saltz, "CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps," 2021 IEEE International Conference on Big Data (Big Data), Dec. 2021, doi: 10.1109/bigdata52589.2021.9671634.

[26] B. Carneiro da Rocha and R. Timoteo de Sousa Junior, "Identifying Bank Frauds Using CRISP-DM and Decision Trees," International Journal of Computer Science and Information Technology, vol. 2, no. 5, pp. 162–169, Oct. 2010, doi: 10.5121/ijcsit.2010.2512.

[27] Layth Almahadeen, Murat Akkaya, and Arif Sari, "Mining Student Data Using CRISP-DM Model," *International Journal of Computer Science and Information Security* , vol. 15, no. 2, pp. 305–316, Feb. 2017.

[28] Rudy Herteno, "Visualisasi Secara Spasial Cluster Kerusakan Sarana dan Prasarana Sekolah," *Journal Speed*, vol. 8, no. 2, pp. 61–68, 2016.

[29] Dita Munawwaroh, Arum, and H. Primandari, "Implementasi 28-Dm Model Menggunakan Metode Decision Tree Dengan Algoritma Cart Untuk Prediksi Lila Ibu Hamil Berpotensi Gizi Kurang," *Delta: Jurnal Ilmiah Pendidikan Matematika*, vol. 10, no. 2, pp. 367–380, 2022.

[30] H. A. Parhusip, S. Trihandaru, A. H. Heriadi, P. P. Santosa, and M. D. Puspasari, "Data Exploration Using Tableau and Principal Component Analysis," JOIV : International Journal on Informatics Visualization, vol. 6, no. 4, p. 911, Dec. 2022, doi: 10.30630/joiv.6.4.952.

[31] Z. N. I. Zailan, S. A. Mostafa, A. I. Abdulmaged, Z. Baharum, M. M. Jaber, and R. Hidayat, "Deep Learning Approach for Prediction of Brain Tumor from Small Number of MRI Images," JOIV : International Journal on Informatics Visualization, vol. 6, no. 2–2, p. 581, Aug. 2022, doi: 10.30630/joiv.6.2.987.

[32] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," Procedia Computer Science, vol. 181, pp. 526–534, 2021, doi:10.1016/j.procs.2021.01.199.

[33] D. Aryanti and J. Setiawan, "Visualisasi Data Penjualan dan Produksi PT Nitto Alam Indonesia Periode 2014-2018," Ultima InfoSys, vol. 9, no. 2, pp. 86–91, Mar. 2019, doi: 10.31937/si.v9i2.991.

[34] R. S. Oetama, T. T. Heng, and D. Tjahjana, "Sebuah Pola Cluster Geospatial Eksplorasi Kejahatan Narkoba di DKI Jakarta," Ultima InfoSys : Jurnal Ilmu Sistem Informasi, vol. 11, no. 1, pp. 57–62, Jul. 2020, doi: 10.31937/si.v9i1.1514.

[35] P. Afikah, I. R. Affandi, and F. N. Hasan, "Implementasi Business Intelligence Untuk Menganalisis Data Kasus Virus Corona di Indonesia Menggunakan Platform Tableau," Pseudocode, vol. 9, no. 1, pp. 25–32, Mar. 2022, doi: 10.33369/pseudocode.9.1.25-32.

[36] C. Goh, "Data Dashboarding in Accounting using Tableau," Journal of Economics and Business, vol. 6, no. 1, Mar. 2023, doi:10.31014/aior.1992.06.01.502.

[37] W. Budiaji, "Penerapan Reproducible Research pada RStudio dengan Bahasa R dan Paket Knitr," Khazanah Informatika : Jurnal Ilmu Komputer dan Informatika, vol. 5, no. 1, pp. 1–5, Jun. 2019, doi:10.23917/khif.v5i1.7202.

[38] A. Ghosh and R. Maiti, "Soil erosion susceptibility assessment using logistic regression, decision tree and random forest: study on the Mayurakshi river basin of Eastern India," Environmental Earth Sciences, vol. 80, no. 8, Apr. 2021, doi: 10.1007/s12665-021-09631-5.

[39] X. Zhou, P. Lu, Z. Zheng, D. Tolliver, and A. Keramati, "Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree," Reliability Engineering & System Safety, vol. 200, p. 106931, Aug. 2020, doi: 10.1016/j.ress.2020.106931.

[40] T. Prasandy, K. Nurkhasanah, M. P. Sari, and T. R. Fazry, "Perbandingan Hasil Penggunaan Metode Decision Tree Dan Random Tree Pada Data Training Aplikasi Pencarian Tukang," Ultima InfoSys : Jurnal Ilmu Sistem Informasi, vol. 10, no. 2, pp. 93–97, Jan. 2020, doi:10.31937/si.v10i2.1166.