INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 13, Issue 5, (2023) E-ISSN: 2222-6990

Suicide Prediction Using Machine Learning Techniques and Interactive Visualization of Suicide Information

Nur Afrisya Zairu Hisham[1], Nor Rashidah Paujah @ Ismail[2*], Suraya Masrom[3], Fadzilah Abdol Razak[4], Siti Khadijah Binti Baharin[5]

[1,3]Computing Sciences Studies, College of Computing, Informatics and Media, Universiti Teknologi MARA Perak Branch, Tapah Campus, Perak Malaysia, [2,4]Mathematical Sciences Studies, College of Computing, Informatics and Media, Universiti Teknologi MARA Perak Branch, Tapah Campus, Perak Malaysia, [5]Information Technology Department, Camfil Malaysia Sdn Bhd, Batu Gajah, Perak Malaysia

Corresponding Author Email: norra914@uitm.edu.my, 2019892326@uitm.edu.my

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v13-i5/16805 DOI:10.6007/IJARBSS/v13-i5/16805

Published Date: *25* May 2023

Abstract

Effective approaches to raising community awareness regarding suicide cases are critical to lowering the suicide rate over time. Many people do not commonly recognize suicide facts and rates since the method of dissemination is unappealing. Non-interactive and difficult methods of disseminating suicide information may lower suicide awareness, thus increasing the suicide rate. Therefore, an interactive web dashboard that contains suicide information and the prediction of the suicide rate has been developed by using Tableau Desktop software. The web dashboard of interactive visualization can attract users to read suicide information and learn more about suicide history. Apart from that, the focus of this study is to test the ability of machine learning to perform prediction by using the regression models of supervised machine learning. Three Machine Learning algorithms namely Random Forest Regressor, Decision Tree Regressor and Support Vector Regressor were used to predict the suicide rate. These three algorithms were compared to find the best model for this study. Random Forest Regressor outperformed the two machine learning algorithms with the highest R^2 and lowest prediction error.

Keywords: Suicide Rate, Suicide Prediction, Machine Learning, Supervised Learning, Interactive Visualization

Introduction

Globally, around 800,000 persons commit suicide each year, and there are more than 20 suicide attempts for every suicide committed (WHO, 2021). It is reported by WHO (2014), suicide has become one of the top causes of death worldwide, accounting for more victims

than malaria, HIV/AIDS, breast cancer, war and homicide. However, this number may be underestimated due to reporting issues (Katz et al., 2016). Moreover, due to its taboo nature and sensitivity, suicide deaths may not be reported, acknowledged, or mistakenly labeled as an accident or another cause of death (Bilsen, 2018).

It is crucial to identify the suicide risks and prevent them from happening to reduce the death rate caused by suicide over time. However, finding the actual risks and preventing suicide is not easy work since suicide is a rare outcome with a multifactorial cause. Even though it is difficult to prevent suicide by identifying its risks and stopping it from occurring, spreading the information about the suicide rate can help create awareness among society and make them more cautious and concerned about their surroundings. Many people do not widely know suicide information and its rate due to a few factors, such as how it is spread is not attractive. Therefore, an interesting method to deliver the information should be applied.

This study was conducted to predict the suicide rate and to visualize information on suicides interactively on a web dashboard. Machine learning prediction was implemented in this study because its approaches were able to examine a large number of variables simultaneously, such as to identify combinations of factors associated with suicidal thoughts and behaviors (Oppenheimer et al., 2021). The two main objectives of this study are:

a) To evaluate the performances of machine learning algorithms in predicting the suicide rate.
b) To develop a web dashboard with interactive visualization for suicide information and suicide rate prediction.

Literature Review

Suicide Prevention and Prediction

Suicide is a serious issue because it is one of the main causes of mortality worldwide and the second biggest cause of death for people under the age of 29 as reported by (WHO, 2014). In order to prevent suicide, it is crucial to identify the suicide risk factors contributing to suicidal behavior. According to a study conducted by Bilsen (2018), the most important risk factors for late school-age children and adolescents were: mental disorders, previous suicide attempts, specific personality characteristics, genetic loading, and family processes in combination with triggering psychosocial stressors, exposure to inspiring models and availability of means of committing suicide. Besides that, there are also other factors that directing an individual to commit suicide such as sexual harassment, workplace harassment, religious script enhancing self-sacrifice and also the portrayed death in movies by a heroic and famous artist (Gaur, 2019). There is also a potential risk factor that can be seen through suicidal behaviors such as alcohol dependence and drug use. A study conducted by Pillon et al (2019) has shown that people with alcohol dependence are between 2.6 and 3.7 times more likely to attempt suicide than non-alcohol.

In preventing suicide, it should be predictable first. Clinicians currently assess patients' risk of future suicide attempts using face-to-face interviews. A study conducted by Nock et al (2022) found that the face-to-face interviews method failed to predict which patients would end up dying from suicide and was unable to save them by using the information gained from the interview. A systematic review of 40 studies conducted by Luoma et al (2002) found that on average, 45% of suicide victims had contact with primary care providers within one month of suicide. Recent studies suggest that applying machine learning (ML) methods to electronic health records (EHRs) can improve clinicians' ability to identify patients at high risk of suicide (Kessler et al., 2020). A prognostic study conducted by Nock et al (2022) found that the

prediction of suicide attempts in one month and six months after a patient visited an emergency department was significantly improved using machine learning models applied to data from a clinician assessment, patient self-report, and electronic health records. A study conducted by Ryu (2018) demonstrated the effectiveness of a machine learning model for predicting people who have suicidal thoughts among the general population.

Machine Learning

Machine learning is a subfield of artificial intelligence that enables software applications to improve their accuracy in making predictions about the future without being specifically programmed to do so (Jan et al., 2022). The term machine learning was introduced by Arthur Samuel in 1959, where it is being identified as a field of study. To prepare a machine for use, it is first trained using a training set as an example, tested with additional data to evaluate its accuracy, and deemed ready for use if it is learned correctly (Harita et al., 2020).

Nowadays, machine learning is used variously in many applications, such as for the search engine in the web and robotics. Besides, it can be used for financial applications such as credit card fraud and identifies the risks (Harita et al., 2020). New routes for learning patterns of human behavior, identifying mental health and risk factors, personalizing and optimizing therapies, and developing predictions on disease progression can be done using machine learning techniques (Thieme et al., 2020).

Types of Machine Learning Techniques

In machine learning techniques, there are four learning techniques involved which are supervised learning, unsupervised learning, semi supervised learning, and reinforcement learning (Harita et al., 2020). In this study, supervised learning will be implemented.

Supervised Learning

Supervised learning is a type of machine learning that makes the model able to predict future outcomes after they are trained based on past data with the goal to produce a function that is approximated enough to be able to predict outputs for new inputs when introduced to them (Sen et al., 2020). Supervised learning can be categorized into two types: regression and classification. Regression involves predicting a continuous numerical output variable based on one or more input variables, while classification involves predicting a categorical output variable based on one or more input variables (Harita et al., 2020). The purpose of regression and classification is to develop a mathematical model in order to predict the dependent variable by analyzing a set of independent variables (Kowsher et al., 2022).

There are various types of algorithms can be applied to evaluate the prediction performance such as Linear Regression, Logistic Regression, Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naïve Bayes, XGBoost, and also K-Nearest Neighbors (KNN). Numerous researchers have discussed and used a variety of machine learning algorithms and datasets to make predictions in various fields. For example, Jain et al (2021) have used eight ML algorithms, namely Decision tree, Random Forest, Support Vector Machine, Naïve Bayes, Logistic Regression, XGBoost, Gradient Boosting Classifier and Artificial Neural Network to predict the mental health of an individual, utilizing a huge dataset (1429 individual's survey). Another example is a study conducted by Reddy et al (2018) to predict stress in working employees using six ML algorithms which are Logistic Regression, KNN Classifier, Decision Trees, Random Forest, Boosting and Bagging. Rahman et al (2022) have used eight algorithms

separately and found that Decision Tree and Random Forest provide better performance than others among all methods in predicting heart disease.
There will be only three algorithms performance of regression models evaluated and focused on this study which are Decision Tree, Random Forest and Support Vector Machine. Comparison between these algorithms helps to evaluate the performance of machine learning techniques algorithms in selecting the best model for this study. The algorithm with the most accurate result produced will be used to predict the suicide rate.

Methodology

Data Requirements

This study used a dataset from a well-known open data source, Kaggle, named "Rates Overview 1985 to 2016". This dataset was chosen because it meets the dataset requirements for the system that will be constructed. It consists of numerical and categorical variables, where the data type of the data target, also called the independent variable, is a continuous variable. This dataset can be used as the benchmark dataset for future works on machine learning performance comparison.
The dataset contains many cases that should be enough to predict and visualize the information, with 27,280 cases and 12 variables which are country, year, sex, age group, count of suicides, population, suicide rate, country-year composite key, HDI for year, GDP for year, GDP per capita, and generation (based on age grouping average). Even though this dataset had not been completely cleaned, it was still usable and straightforward to go through the data management procedures. Moreover, by using this global data obtained, this study provides a fundamental framework that can be replicated in the Malaysian context in the nearest future since Malaysian data does not yet exist.

Summary of Research Methodology

The systematic approach applied in this study is Machine Learning Lifecycle which is known as a cyclic process that is efficient in building machine learning systems (Yang et al., 2021). This lifecycle was selected because the project develops prediction models that use machine learning techniques to predict the suicide rate and visualize it on a web dashboard interactively. The lifecycle illustrated in Figure 1 comprises five main phases of this study.

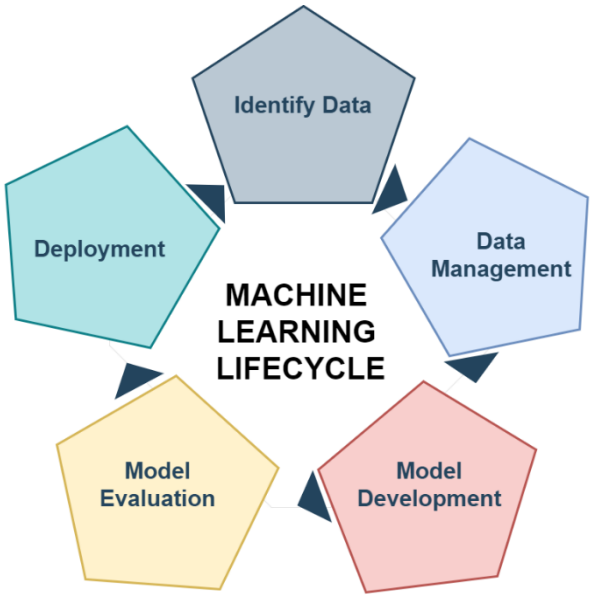

Figure 1. Project methodology using Machine Learning Lifecycle

The Overall Process of Development

In order to accomplish the purpose of this study, three main processes that consist of the five phases were conducted. Figure 2 visualizes processes associated with the phase carried out to achieve the goals.

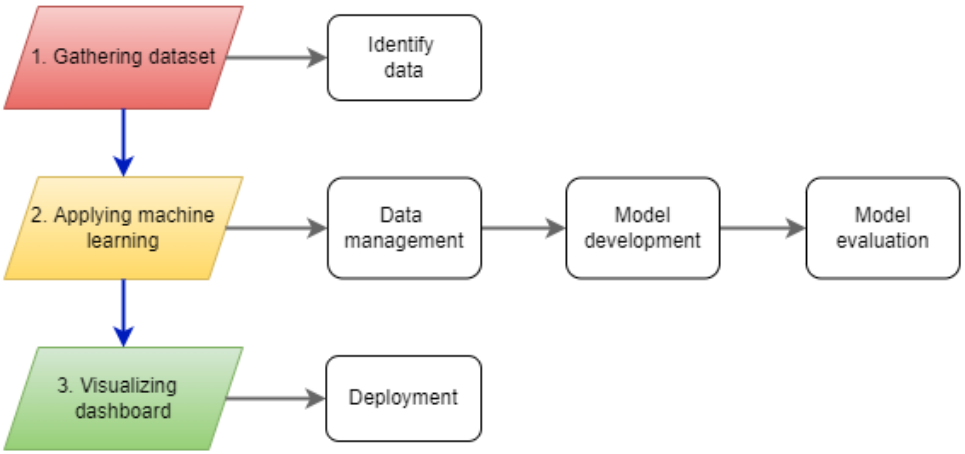

Figure 2: Visualizing the overall process

The main procedures involved are explained in the section below.

i) Gathering Dataset

To select and gather a suitable dataset for the study, numerous studies were done on suicide cases. As a result, it can be summarized that the most related variables or attributes that influenced the suicide rate are the age, generation, gender, country, population, and year of the incident occurred. After identifying the dataset, several processes were going through to obtain clean and usable data.

ii) Applying machine learning

The machine learning application has been implemented in Jupyter Notebook and Tableau Desktop, which involved three significant processes. To begin, data management has been performed with three sub-processes: data preparation, data wrangling and data analysis. Since the dataset prepared had missing values and unnecessary variables, the data wrangling process was performed to manage it. Then, three regression algorithms with their tuning parameters were applied to measure their performances.

Next, three sub-processes were involved in the model development: model selection, training and testing. As known, this study involved the continuous data target for prediction by using supervised machine learning algorithms. Thus, the regression model was selected to evaluate the prediction performance using Random Forest Regressor, Decision Tree Regressor, and Support Vector Regressor. After that, the model underwent the training and testing process by using the dataset that has been split into two subsets using the Sklearn train_test_split function with a ratio of 80% for training and 20% for testing data.

Then, the model evaluation is conducted where the performance of regression models is measured by using the calculated mean absolute error (MAE), mean squared error (MSE), r

squared (R^2 Score) and root mean square error (RMSE) as well as RMSE with cross-validation (CV).

iii) Visualizing the dashboard

Before proceeding with the deployment process in making predictions by implementing a machine learning application on Tableau, the Tableau Analytics Extension API must be installed. In this project, TabPy (the Tableau Python Server) has been installed through Anaconda Navigator and applied to the Tableau Desktop server. The last main process involved the deployment process, where users can interact with the model by selecting a filter to predict the suicide rate. A filter section is provided that allows users to select specific information to display from the original dataset. This includes country, year, age group, gender, number of populations, and the number of suicides with its rate. After ensuring the interactive dashboard can be used, and its features are well-functioning, the dashboard is published on Tableau Public to be easily accessible to the public.

System Architecture

Based on Figure 3, the system architecture is illustrated as a structural design for the overall process of this project development. To accomplish the objective of this project which is to apply and evaluate the performances of machine learning algorithms in predicting the suicide rate in society, Python is used as the programming language software. The processing data and prediction application can be fulfilled with the use of machine learning implementation. Lastly, the result of suicide rate prediction and the summarization of suicide information will be visualized interactively on a dashboard using the Tableau Desktop software and published on the Tableau Public website.

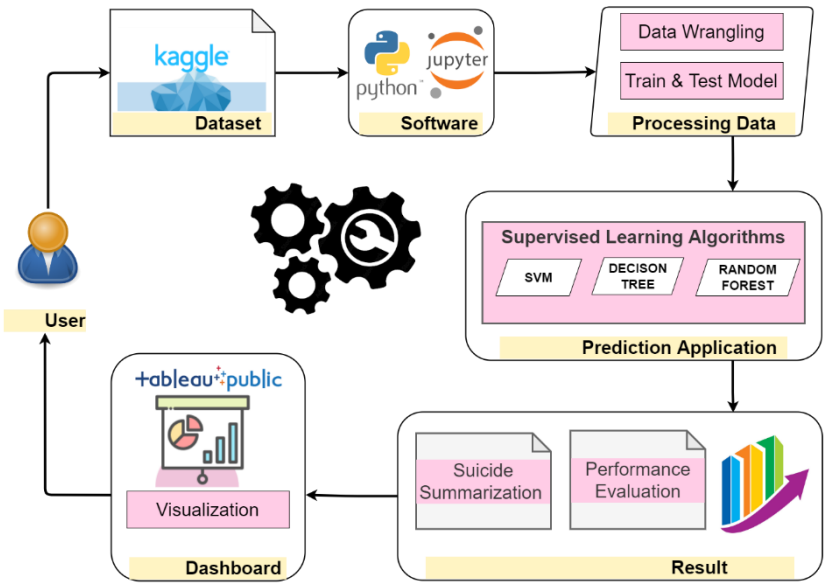

Figure 3. System architecture for suicide rate prediction and profiling suicide information project

Data Visualization

As an alternative to display suicide information and its rate prediction efficiently in increasing awareness on suicides, interactive visualization was used to present and deliver the information. Due to the complexity of data and ways of explaining the information, people

usually lose interest in looking up to this main problem. Therefore, using data visualization for managing and processing data is essential in representation because it is an important feature of the data communication process. Hence, it will be helpful to present complex and massive data information in a simple way where it can be easily interpreted (Tamayo et al., 2018). Furthermore, visualization is also known for playing a vital role in collecting, cleaning, analyzing, and sharing data that help in showing the result intuitively.

Data visualization helps users to read and understand presented information easily (Ko & Chang, 2017). The software tools and platform of the dashboard used is Tableau because it has many advantages. For example, Tableau can be connected with various data sources, which can present data visualization through charts, maps, dashboards, and stories through a simple interface. It is easy to create an interactive visualization expressing the desired format by dragging and dropping the data using this software. Moreover, it is known as software that users can use to explore and understand their data through the interactive visualization created (Ko & Chang, 2017).

Results and Discussion

Evaluation of Machine Learning Performance

There were five types of performance metrics being measured for each regression model, which were by using the calculated mean absolute error (MAE), mean squared error (MSE), r squared (R^2 Score) and root mean square error (RMSE) as well as RMSE with cross-validation (RMSE with CV). Details on performance metrics used are described in Table 1.

Table 1
Performance metrics used to evaluate the model performance

Performance Metric	Description
MAE	• Calculates the average size of errors in a set of predictions without considering their direction. • Calculate the distance between the actual value and the predicted value.
MSE	• Totaling the sum of error from the absolute error. • The lower MSE value, the higher the prediction accuracy.
R^2 Score	• Measure how well a model matches a particular dataset. • The higher value of R^2 Score, the better the model's fit.
RMSE	• The standard deviation of the prediction errors calculates how far data points from its regression line. • The lower the RMSE value, the better the performance produced.
RMSE with CV	• It is RMSE that applies the cross-validation for its dataset.

Table 2 below presents the result of performance metrics recorded in four decimal places. Based on the study performed, the best model with the highest values of R^2 Score as well as the lowest values of MAE, MSE, RMSE and RMSE with CV application is the Random Forest Regressor. Hence, this model is being selected to predict the suicide rate.

Table 2
Result of performance metrics for 3 regression algorithms

Performance Metric	Decision Tree	Random Forest	Support Vector Machine
MAE	1.8916	0.2783	8.2564
MSE	15.4319	2.1001	287.7145
R^2 Score	0.9533	0.9936	0.1286
RMSE	3.9283	1.4492	16.9622
RMSE with CV	6.3376	4.0012	21.9936

Interface Design of Interactive Dashboard

Four main dashboards were created, consisting of worldwide suicide profiling information, suicide rate prediction by user input, a summary of prediction, and evaluation results for three types of machine learning regression models.

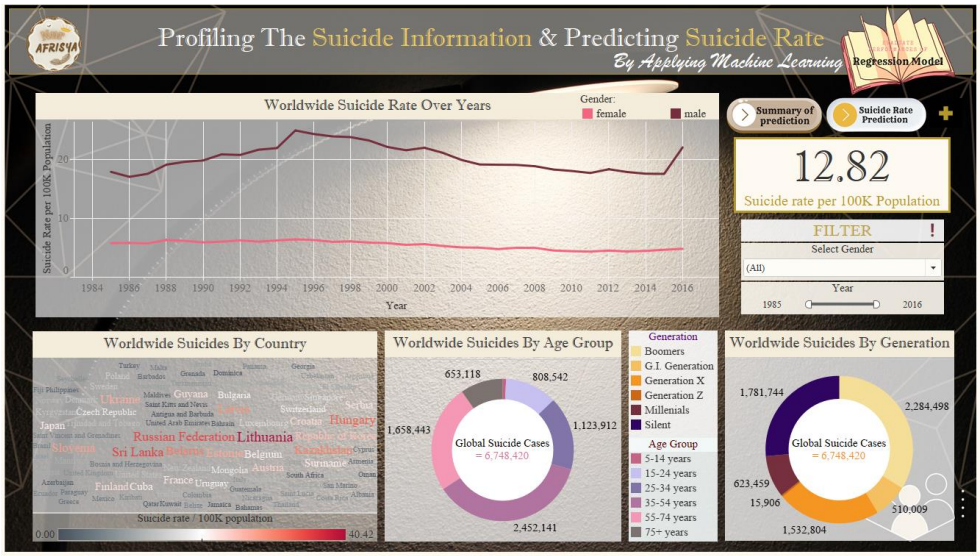

Figure 4. The main dashboard that profiling the worldwide suicide information

The dashboard shown in Figure 4 profiles the worldwide suicides information from 1985 until 2016. On that dashboard, the suicide rate displayed in the text block will be updated when the user applies a filter for the line graph that visualizes the worldwide suicide rate over the years. According to WHO (2021), the suicide rate is calculated as the number of suicides per 100,000 population as stated below:

$$Suicide\ rate = Number\ of\ suicides \times \frac{100,000}{Number\ of\ population}$$

Moreover, by applying the filter for gender or year, it will also update the other graph on the dashboard, such as the cloud graph, doughnut graph, and line graph.

For the next dashboard, as shown in Figure 5, there is a tree map graph that can be filtered by the user to display the specific information from the dataset used. Besides, a filter can be applied to predict the suicide rate with Random Forest Regressor model implementation.

From this prediction, many conclusions can be made about suicides. For example, users can conclude that as the number of suicides in a population increases, the suicide rate per 100,000 over that population will also increase.

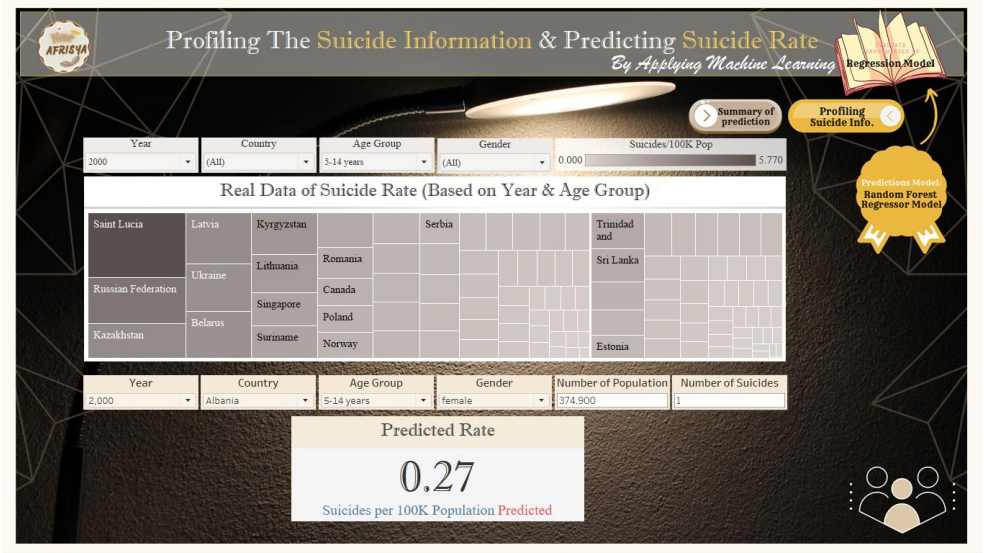

Figure 5. The dashboard that allows users to predict the suicides rate

The third dashboard in Figure 6 summarises the prediction made on Jupyter Notebook. The text table shows that the predicted rate does not have a huge difference from the original rate. Besides, a comparison between the first and second years of the suicide rate predicted is displayed to give more knowledge to the user. Furthermore, the comparison between worldwide average of suicide rate in two years and five years is being compared. This prediction is visualized by using the testing data.

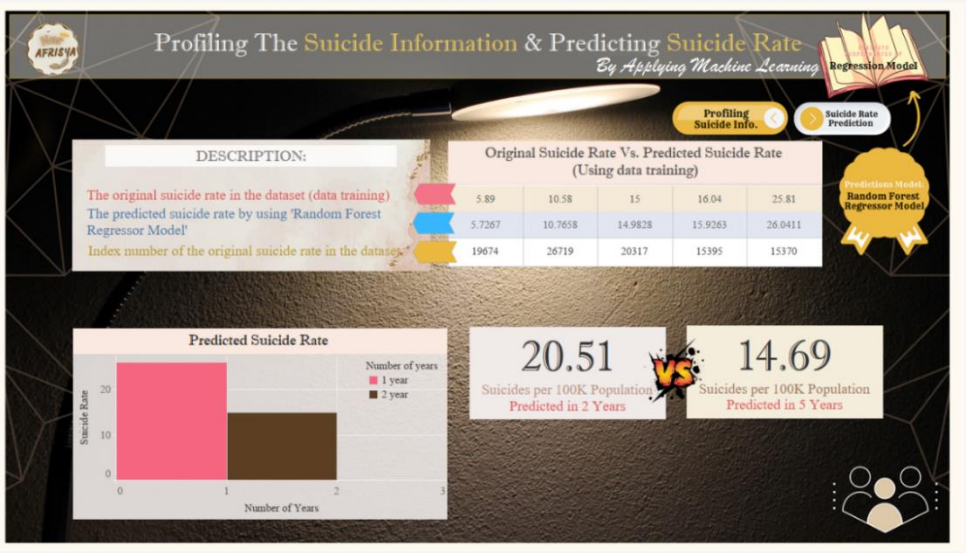

Figure 6. The dashboard that summarizes the suicide rate predicted

Lastly, a dashboard summarizes the comparison of the three types of regression models' performance evaluated in the Jupyter Notebook. From the dashboard, as shown in Figure 7, users can see and learn the measurement values used to evaluate each model and what is the reason for the best model being chosen. Besides, a doughnut graph that presents the model with its R-square score is provided. This can help users conclude the best model result easier

since R-square shows the value of which model is closest to the regression line and makes it most accurate if the value is closest to 1.

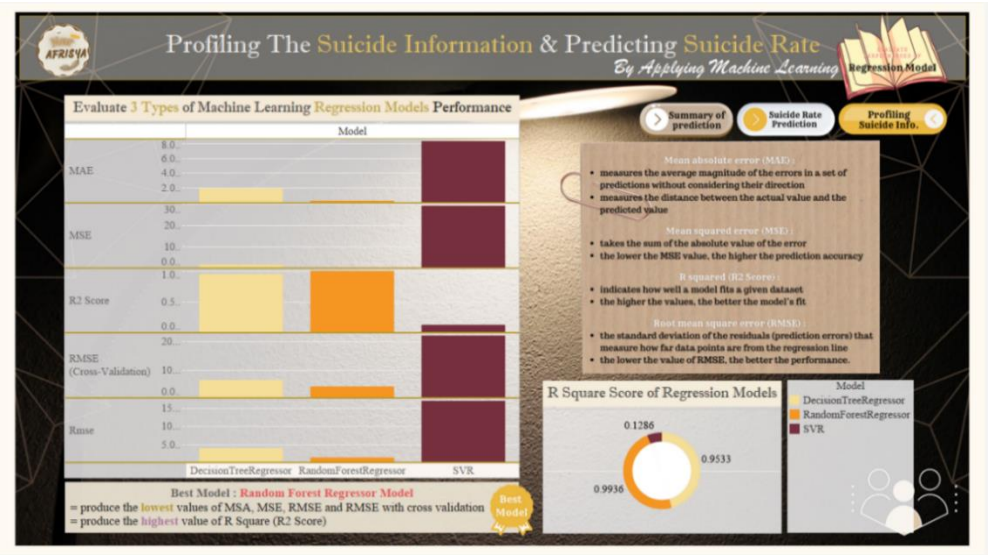

Figure 7. The dashboard that explains the evaluation result

Conclusion

This study had two objectives, which were to evaluate the performance of machine learning algorithms in predicting the suicide rate and to develop an interactive web dashboard for suicide information and suicide rate prediction. The study successfully predicted the suicide rate using the best machine learning algorithm, Random Forest, out of the three algorithms evaluated (Support Vector Machine, Decision Tree, Random Forest). The Random Forest algorithm had the least prediction error. The web dashboard was developed using Tableau Desktop and the Tableau Analytics Extension API, TabPy, to allow real-time prediction.

The study found that each variable used to forecast the suicide rate, such as age group, country, year, gender, generation group, population size, number of suicides, and suicide rate, was interdependent and had the potential to affect the global suicide rate. The study's findings suggest that an interactive web dashboard with machine learning algorithms can effectively predict and raise awareness about suicide, potentially aiding in efforts to reduce the global suicide rate.

Acknowledgments

The authors gratefully acknowledge the Universiti Teknologi MARA Perak Branch, Tapah Campus, for giving the authors the opportunity, support and facilities to accomplish this project.

References

Bilsen, J. (2018). Suicide and Youth: Risk factors. Frontiers in Psychiatry, 540. https://doi: 10.3389/fpsyt.2018.00540.

Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., ... & Pathak, J. (2019, May). Knowledge-Aware Assessment of Severity of Suicide Risk for Early Intervention. In *The World Wide Web Conference* (pp. 514-525), doi: 10.1145/3308558.3313698.

Harita, U., Kumar, V. U., Sudarsa, D., Krishna, G. R., Basha, C. Z., & Kumar, B. S. S. (2020, November). A Fundamental Study On Suicides And Rainfall Datasets Using Basic Machine Learning Algorithms. In *2020 4th International Conference on Electronics, Communication And Aerospace Technology* (ICECA) (pp. 1239-1243). IEEE.

Jain, T., A., Hada, P. S., Kumar, H., Verma, V. K., & Patni, A. (2021). Machine Learning Techniques for Prediction of Mental Health. *In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1606-1613). IEEE.

Jan, N., Maqsood, R., Nasir, A., Alhilal, M. S., Alabrah, A., & Al-Aidroos, N. (2022). A New Approach To Model Machine Learning By Using Complex Bipolar Intuitionistic Fuzzy Information. *Journal Of Function Spaces*, *2022*, 1-17.

Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2020). Suicide Prediction Models: A Critical Review of Recent Research with Recommendations For the Way Forward. *Molecular Psychiatry*, 25(1), 168-179.

Katz, C., Bolton, J., & Sareen, J. (2016). The Prevalence Rates of Suicide Are Likely Underestimated Worldwide: Why It Matters. *Social Psychiatry and Psychiatric Epidemiology*, 51, 125-127.

Kowsher, M., Tahabilder, A., & Murad, S. A. (2020, July). Impact-Learning: A Robust Machine Learning Algorithm. In *Proceedings of the 8th International Conference on Computer And Communications Management* (pp. 9-13).

Ko, I., & Chang, H. (2017). Interactive Visualization of Healthcare Data Using Tableau. *Healthcare Informatics Research*, *23*(4), 349-354.

Luoma, J. B., Martin, C. E., & Pearson, J. L. (2002). Contact With Mental Health and Primary Care Providers Before Suicide: A Review of The Evidence. *American Journal of Psychiatry*, 159(6), 909-916.

Nock, M. K., Millner, A. J., Ross, E. L., Kennedy, C. J., Al-Suwaidi, M., Barak-Corren, Y., ... & Kessler, R. C. (2022). Prediction Of Suicide Attempts Using Clinician Assessment, Patient Self-Report, And Electronic Health Records. *JAMA Network Open,* 5(1), e2144373-e2144373, doi:10.1001/jamanetworkopen.2021.44373.

Oppenheimer, C. W., Bertocci, M., Greenberg, T., Chase, H. W., Stiffler, R., Aslam, H. A., ... & Phillips, M. L. (2021). Informing The Study Of Suicidal Thoughts and Behaviors In Distressed Young Adults: The Use of a Machine Learning Approach To Identify Neuroimaging, Psychiatric, Behavioral, and Demographic Correlates. *Psychiatry Research: Neuroimaging*, 317, 111386.

Pillon, S. C., Vedana, K. G. G., Teixeira, J. A., Dos Santos, L. A., de Souza, R. M., Diehl, A., ... & Miasso, A. I. (2019). Depressive Symptoms And Factors Associated With Depression And Suicidal Behavior In Substances User In Treatment: Focus On Suicidal Behavior And Psychological Problems. *Archives Of Psychiatric Nursing*, *33*(1), 70-76, doi: 10.1016/j.apnu.2018.11.005.

Rahman, M. M., Rana, M. R., Alam, M. N. A., Khan, M. S. I., & Uddin, K. M. M. (2022). A Web-Based Heart Disease Prediction System Using Machine Learning Algorithms. *Network Biology*, *12*(2), 64-80.

Reddy, U. S., Thota, A. V., & Dharun, A. (2018, December). Machine learning techniques for stress prediction in working employees. In *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-4). IEEE.

Ryu, S., Lee, H., Lee, D. K., & Park, K. (2018). Use of a Machine Learning Algorithm to Predict Individuals with Suicide Ideation in the General Population. *Psychiatry Investigation*, 15(11), 1030.Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised

Classification Algorithms In Machine Learning: A Survey And Review. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018* (pp. 99-111). Springer Singapore.

Tamayo, J. L. R., Hernandez, M. B., & Gomez, H. G. (2018). Digital Data Visualization with Interactive And Virtual Reality Tools. Review Of Current State Of The Art And Proposal Of A Model. *ICONO 14, Revista de comunicación y tecnologías emergentes*, 16(2), 40-65.

Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *27*(5), 1-53.

World Health Organization. *Preventing Suicide: A Global Imperative*. World Health Organization, 2014.

World Health Organization. (2021). Suicide Worldwide in 2019: Global Health Estimates.

Yang, C., Wang, W., Zhang, Y., Zhang, Z., Shen, L., Li, Y., & See, J. (2021). MLife: A Lite Framework for Machine Learning Lifecycle Initialization. *Machine Learning*, 110, 2993-3013.