Malaysian Journal of Computing, 7 (2): 1067-1081, 2022

eISSN: 2600-8238

PREDICTION OF MALAYSIAN WOMEN DIVORCE USING MACHINE LEARNING TECHNIQUES

Nazim Aimran[1*], Adzhar Rambli[2], Asyraf Afthanorhan[3], Adzmel Mahmud[4], Azlin Sapri[5] and Airena Aireen[6]

[1,2] *Center of Statistical and Decision Science Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia.*

[3] *Faculty of Business and Management, Universiti Sultan Zainal Abidin, Kampung Gong Badak, Kuala Terengganu, Malaysia*

[4] *Faculty of Economics and Administration, Universiti of Malaya, Malaysia*

[5,6] *Population and Family Research Division, National Population and Family Development Board, Malaysia.*

[1*]nazim@tmsk.uitm.edu.my, [2]adzhar_rambli@tmsk.uitm.edu.my, [3]asyrafafthanorhan@unisza.edu.my, [4]adzmel@lppkn.gov.my, [5]azlin@lppkn.gov.my [6]nurairena@lppkn.gov.my

ABSTRACT

This paper discusses the performance of three machine learning techniques namely Decision Tree, Logistic Regression and Artificial Neural Network for predicting divorce among Malaysian women. Secondary data were obtained from the Fifth Malaysia Population and Family Survey (MPFS-5) conducted by the National Population and Family Development Board (LPPKN). The total number of instances in the dataset was 7,644 ever married Malaysian women aged 15 to 59 years old. Divorce is currently a serious problem among the Malaysian community due to various reasons. In 2019, the divorce rate in Malaysia rose by 12% from the previous year. During the first three months of the Movement Control Order (MCO), i.e. from March 18 to June 18, 2020, the Syariah Court of Malaysia recorded 6,569 divorce cases. Worse, a total of 90,766 divorce cases were recorded from January to October 2020. Six predictive models were used for comparison, namely Decision Tree (C5.0 and CHAID), Logistic Regression (Forward Stepwise and Backward Stepwise), and Artificial Neural Network (Multi-Layer Perceptron and Radial Basis Function). Among the six predictive methods, the Decision Tree model (C5.0) was found to be the best model in classifying divorce among Malaysian women. The accuracy of the C5.0 model was 77.96% followed by the Artificial Neural Network (Multi-Layer Perceptron) and Logistic Regression (Forward Stepwise) model (74.68% and 67.89%, respectively). The order of important predictors in predicting divorce among Malaysian women is the wives' employment status (0.1531) followed by the husbands' employment status (0.1396), type of marriage (0.1327), race/ethnicity (0.1327), distant relationship (0.1212), the wives' qualification level (0.1115), age group (0.1053) and religion (0.0998).

Keywords: *Artificial Neural Network, Data Mining, Decision Tree, Divorce, Logistic Regression, Malaysian Women.*

Received for review: 11-02-2022; Accepted: 06-07-2022; Published: 01-10-2022

DOI: 10.24191/mjoc.v7i2.17077

1. Introduction

Divorce occurs when a husband and wife are unable to solve problems in marriage and decide to separate to stop all the problems arising in their marriage. This phenomenon has increased in the ongoing decades and affected family and community cohesion (Mohamed and Alkhyeli, 2016). According to the National Population and Family Development Board (NPFDB) (2016), the first five years of marriage reported the most divorce cases in Malaysia, with a proportion of 37.3% women and 35.4% male. The median age of divorce in Malaysia is 37 for males and 34 for females (The Department of Statistics Malaysia [DOSM], 2019). DOSM (2020) also reported that the number of divorces in Malaysia rose by 12% in 2019 compared to 2018.

Malaysia's crude divorce rate has remained unchanged for three consecutive years from 2016 to 2018 (1.6% in 2016 but increased to 1.8% per 1,000 population in 2019). The Covid-19 outbreak Movement Control Order (MCO) has worsened the scenario. According to the Syariah Judiciary Department of Malaysia, from March 18 to June 18, 2020, the total number of divorce cases nationwide was 6,569 (Rahim, 2020). This statement is in line with Bosro (2020), who indicates that The Syariah Court of Malaysia recorded 90,766 divorce cases from January to October 2020. Out of this number, 10,557 (11.7%) cases involved couples aged 50 years and above.

There are several factors that can lead to divorce such as low education level (Raymo *et al.*, 2013; Raley *et al.*, 2015), age group (Jones, 2018; Fetzer, 2017), employment status (Sayer *et al.*, 2011; Steverman, 2016; Folke and Rickne, 2020), long distance relationship (Krapf, 2017), incompatibility, attitudes of the couples, unscrupulous couples, lack of understanding, religiosity, polygamy and family interference (Abdul Ghani *et al.*, 2017).

2. Related Works

Machine Learning (ML) is a form of artificial intelligence which is placed to transform the twenty-first century (Nichols *et al.*, 2018). It has gained popularity in many areas such as in medical (e.g., Weng *et al.*, 2017), fraud detection (e.g., Guzella & Caminhas, 2009), transportation (e.g., Borucka, 2020), energy consumption (e.g. Abdelkader *et al.*, 2020), machinery (Sharma *et al.*, 2019) as well as in divorce (e.g., Sharma *et al.*, 2021; Yontem *et al.*, 2019).

Yöntem *et al.* (2019) developed the divorce predictor scale in predicting divorce by using a set of 54 questions on a 5-point scale (0= Never, 1=Seldom, 2=Averagely, 3=Frequently, 4=Always) and found that ANN is the best technique to predict divorce with 98.82% accuracy followed by RBF (97.64%) and Random Forest (97.64%). On the other hand, Sharma *et al.* (2021) compared the accuracy six techniques namely Perceptron, Naïve Bayes (NB), K Nearest Neighbour (KNN), Decision Tree (DT), Support Vector Machine (SVM) and Logistic Regression (LR) in predicting divorce. They found that, for the 60-40 split, Perceptron outperforms the other techniques with 98.53% accuracy followed by NB (97.06%), KNN (97.06%), SVM (97.06%), LR (97.06%) and DT (95.59%).

However, we have not found any study to predict the divorce using the personal information of the samples, specifically in Malaysia. In machine learning, LR and ANN are the most popular model and share common roots in statistical pattern recognition (Dreiseitl & Ohno-Machado, 2002). We also include DT in the study to observe the performance of the supervised algorithm machine learning as done by other studies (e.g., Mudunuru & Skrzypek, 2020; Siddiqui *et al.*, 2020). Thus, this study investigates the performance of DT, LR and ANN to predict divorce using the personal information of the samples from the dataset obtained from NPFDB.

2.1 Dataset and Features

The dataset consisted of 7,644 ever married women, inclusive of married, widow, and divorced women. Since the interest of the study is in modeling divorce, we selected 7,226 cases only, that is, married women and divorced women. The remaining 418 cases were excluded as they were widows. There are 9 variables in the dataset: *age, race*, *religion, quali*, *dr*, *wes*, *tom*, *hes* and *ms*.

Table 1. Description of Features

age	Age group (1 = Generation Z, 2 = Generation Y, 3 = Generation X, 4 = Baby Boomer Generation)
race	Race/Ethnicity (1 = Malay, 2 = Other Bumiputras, 3 = Chinese, 4 = Indian, 5 = Others)
religion	Religion (1 = Islam, 2 = Buddha, 3 = Hindu, 4 = Kristian, 5 = Confucian/Taoism/ Religion of Chinese Tradition, 6 = Tribal Religion/'FOLK')
quali	The wive'’s qualification level (1 = No Schooling, 2 = Pre-School Education, 3 = Primary Education, 4 = Lower Secondary Education, 5 = Upper Secondary Education, 6 = Pre-University, 7 = Tertiary Education, 8 = Others)
dr	Does the person have distance relationship (0 = No, 1 = Yes)
wes	Wives’ employment status (0 = Not working, 1 = Working)
tom	Type of marriage (1 = Family Arrangement, 2 = Introduced, 3 = Own Through Social Interaction, 4 = Own Through the Internet, 5 = Others, 98 = No Information)
hes	Husbands’ employment status (1 = Employer, 2 = Government Worker, 3 = Private Worker, 4 = Self-Employed, 5 = Family Workers Without Salary, 6 = Not Working)
ms (target)	Marriage status (0 = Still Married, 1 = Divorced/Separated)

2.1 Machine learning techniques

In this study, we compared the performance of the DT, LR and ANN to discover the best technique for predicting divorce among Malaysian women.

2.1.1 Decision Tree

The Decision Tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes (Song and Lu, 2015).

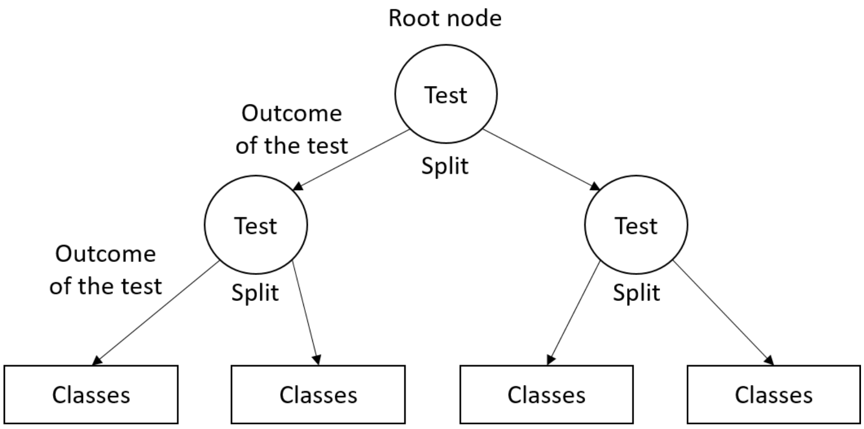

Figure 1. Decision Tree

During the construction of the tree, variable selection measures are used to choose the variable with the best partition. Finally, the decision tree divides the data into multiple segments, which are defined by the split rules for each step (Pandya and Pandya, 2015).

i. The C5.0

The C5.0 model works by building a top-down decision tree, starting from a root node and ending with a leaf (Buaton *et al.*, 2019). This algorithm is a supervised machine learning algorithm, which must be taught using data where the output values are known (Yu *et al.*, 2018).

The C5.0 is powerful in handling missing data and large amounts of input attributes and usually requires less training time to estimate the model. Compared to the other model types, the C5.0 is easier to understand because the rules derived from the model have a very simple and direct explanation (Pandya and Pandya, 2015).

ii. Chi-Square Automatic Interaction Detection (CHAID)

CHAID is a classification method that uses chi-square statistics (χ^2) to identify the best and optimal split decision tree (Lin and Fan, 2019). CHAID first checks the cross table between each input field and the targets by using the chi-square independence test. If more than one of these relationships is statistically important, CHAID will select the most important input field with the smallest *p*-value (Milanovic and Stamenkovic, 2016). If the input has more than two categories, these categories are compared and the categories with no difference in the results are grouped together. This is by adding a pair of the category that shows the smallest difference. When all the remaining categories are different at the specified test level, the category merging process will stop.

2.1.2 Logistic regression

Logistic regression is a statistical classification method suitable for modeling binary outcomes. While the outcome is binary (categorical), we can use both categorical or continuous variables as the explanatory variables (Sperandei, 2014). Logistic regression models the probability of an event as a function of other factors. These models are only able to state that there is a relationship between the explanatory and the outcome variables (Borucka, 2020). The target binary outcomes in this study refer to *ms* that takes on two values: 1 indicating divorce or 0 representing those who are still married. Generally, the logistic regression model is as below:

$$logit(p) = log\left(\frac{p}{1-p}\right) \text{ for which } p = P(Y = 1|X) \quad (1)$$

IBM Knowledge Centre (2021a) defines the forward stepwise method and backward stepwise method as follows:

i. Forward stepwise method

The forward stepwise method of field selection builds the equation from the simplest model possible. At each step, terms that have yet to be added to the model are evaluated, and if the best of those terms adds significantly to the predictive power of the model, it is added. Terms that are currently in the model are re-evaluated to determine if any of them can be removed without significantly detracting from the model. If so, they are removed. The process repeats, and other terms are added and/or removed. The final model is generated when no more terms can be added to improve the model and when no more terms can be removed without detracting from the model.

ii. Backward stepwise method

The backwards stepwise method is essentially the opposite of the forward stepwise method. In the former, the initial model contains all of the terms as predictors. In each step, the terms in the model are evaluated, and any term that can be removed without significantly detracting from the model is removed. Previously removed terms are re-evaluated to determine if the best of those terms adds significantly to the predictive power of the model. If so, it is added back into the model. The final model is generated when no more terms can be removed without significantly detracting from the model and when no more terms can be added to improve the model.

2.1.3 Artificial Neural Network (ANN)

Neural networks are used in many fields such as economy, financial institutions, and medical (Mossalam and Arafa, 2019). According to Kaur and Wasan (2006), ANN could generalize and learn from data, thus mimicking in some sense, the human ability to learn from experience.

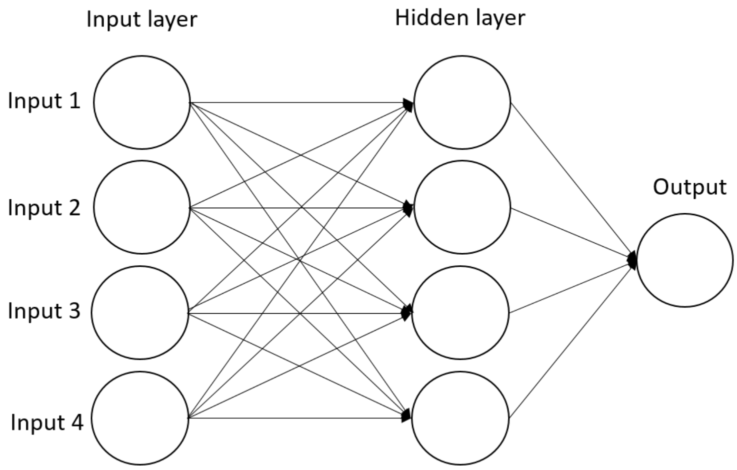

Figure 2. Neural Network

The structure of a typical neural network consists of (i) an input layer where data enters the network, (ii) a second layer (also known as the hidden layer), which comprises artificial neurons, each receiving multiple inputs from the input layer, and (iii) an output layer, which combines the results summarized by the artificial neurons (Fath *et al.*, 2020).

i. Multi-layer Perceptron

Multi-layer Perceptron (MLP) is a supplement of feed-forward neural network. The neurons in the MLP are trained with the backpropagation learning algorithm (Abirami and Chitra, 2020). It allows for more complex relationships at the possible cost of increasing the training and scoring time (IBM, 2022).

ii. Radial Basis Function

Similar to MLP, the Radial Basis Function (RBF) network is a feed-forward neural network composed of three layers: the input layer, the hidden layers, and the output layer (Fath *et al.*, 2020). Compared with MLP, RBF may have a shorter training and scoring time, but it may reduce the predictive ability.

2.2 Evaluation

2.2.1 Confusion Matrix

In this study, we used a confusion matrix to analyze the performance of a classification algorithm. This is one of the popular measures of classification performance and is widely used in classification studies (e.g., Kulkarni *et al.*, 2020; Al-jabery *et al.*, 2020). The confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data. A single prediction by a classifier can have four results, as displayed in the confusion matrix below:

Table 2. The Confusion Matrix Format

		Predicted		Total
		1	**0**	
Actual	**1**	TP	FN	TP + FN
	0	FP	TN	FP + TN
Total		TP + FP	FN + TN	TP + FP + FN +TN

According to Reddi and Eswar (2021), the measures derived from the confusion matrix are as follows:

- True positive (TP): The algorithm-predicted value is matched with the reality that news is fake. We can conclude that the algorithm has correctly classified and news shared in the social network is fake.
- False negative (FN): The predicted output is a false negative, where news is incorrectly classified that the news shared is negative, even though the news is genuine.
- True negative (TN): Predicted output is a true negative when the algorithm-predicted value is matched with the reality that news is genuine. We can conclude that the algorithm has been appropriately classified.
- False positive (FP): News is inaccurately classified that shared news is genuine news, even though it is fake news.
- Negative (N): A 0 value is used to represent a negative case, which means the news is genuine.
- Positive (P): A value of 1 is used to represent a positive case, which means the news is fake.

Table 3. The Formula of Accuracy Measure

Measure	Formula
Sensitivity	TP / (TP +FN)
Specificity	TN / (TN +FP)
Accuracy	(TP + TN) / (TP + FN + TN + FP)
Precision of True Class	TP / (TP +FP)

We determined the best model by identifying a model with the highest percentage in sensitivity, specificity, accuracy, and the precision of true class.

2.2.2 Receiver Operating Characteristics (ROC)

The ROC chart provides a means of comparison between the classification models. The chart shows False Positive Rate (1-Specificity) on the X-axis, a probability of target=1 when its true value is 0 against True Positive Rate (Sensitivity) on the Y-axis, and a probability of target=1 when its true value is 1. Ideally, the curve will climb quickly towards the top-left if the model correctly predicted the cases.

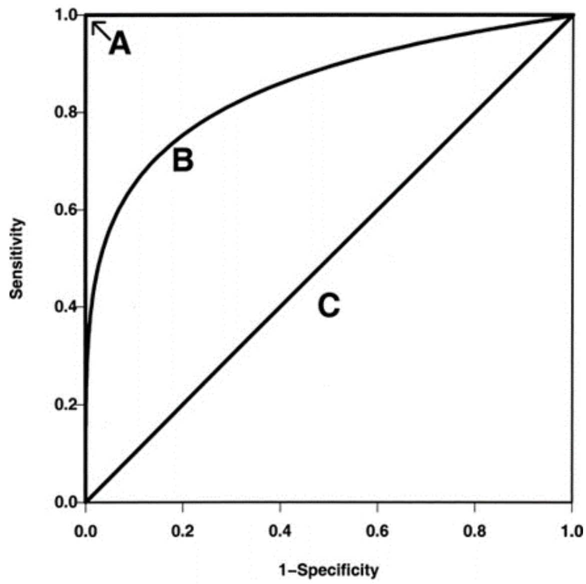

Figure 3. The ROC Chart

3. Implementation

The modeling was implemented by using IBM SPSS Statistics and IBM SPSS Modeler. As the reference, we used Cross Industry Standard Process for Data Mining (CRISP-DM) research framework as proposed by Rodrigues (2020). The framework consists of six major phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

Prior to the modeling, we reviewed the literature on the factors affecting divorce among Malaysians. The process undertaken included reviewing published manuscripts or reports, seeking clarification from the subject-matter experts such as from the ministries, universities and related agencies, and identifying suitable variables that can be used for the purpose of this study. This phase involved the qualitative evaluation of the variables in the instruments used. The significance of the variables included in this study was also validated by experts in the related field, as proposed by Rahlin *et al.* (2019).

Next, we identified *ms* of the Malaysian women as the target variable (or dependent variable) of the study. We also recognized the following predictor variables: *age, race, religion, quali, dr, wes, tom,* and *hes*. Other than that, we checked for the missing values and outliers in the dataset. From the analysis, this dataset contains no missing values and outliers. Lastly, we make sure that the proportion of the target variable is balance by using random oversampling (ROS) method.

4. Results

Table 4 shows that for the DT model, the C5.0 model gives better performance compared to the CHAID model in all the evaluation measures. In terms of accuracy, the C5.0 scored 77.62% compared to CHAID (71.12%). This finding indicates that the misclassification rate for the C5.0 is lower than CHAID's. By comparing the values of sensitivity and specificity for each model, it can be stated that both C5.0 and CHAID are significantly able to predict the event occurrence (status = divorced/separated) since the value of sensitivity for each model is higher than the specificity value.

Table 4. The Summary of Performance Measure for Decision Tree Model

Measures	C5.0	CHAID
Accuracy	77.62%	71.12%
Sensitivity	84.66%	76.85%
Specificity	70.54%	65.36%
Precision of True Class	0.7432	0.6908
ROC Area	0.845	0.766

Table 5 shows that for the LR model, the accuracy for the two logistic models is approximately the same; the forward stepwise method scored 67.93%, and the backward stepwise method scored 67.60%. This finding indicates that the misclassification rate for the forward stepwise is slightly lower than the backward stepwise. The sensitivity and specificity values indicate that the two models are significantly able to predict the event occurrence (status = divorced/separated) since the value of sensitivity for each model is higher than the specificity value.

Table 5. The Summary of Performance Measure for Logistic Regression Model

Measures	Forward Stepwise	Backward Stepwise
Accuracy	67.93%	67.60%
Sensitivity	67.73%	66.78%
Specificity	68.14%	68.43%
Precision of True Class	0.6818	0.6807
ROC Area	0.733	0.731

Table 6 shows that for the ANN model, the MLP model outperformed the RBF with an accuracy of 73.90% compared to Radial Basis Function (60.69%). This finding indicates that the misclassification rate for the MLP is better than the RBF. By comparing the values of sensitivity and specificity in both MLP and RBF, it can be concluded that both models are significantly able to predict the event occurrence (status = divorced/separated) since the value of sensitivity for each model is higher than the specificity value.

Table 6. The Summary of Performance for ANN Model

Measures	Multi-Layer Perceptron	Radial Basis Function
Accuracy	73.90%	60.69%
Sensitivity	78.48%	72.76%
Specificity	69.29%	48.56%
Precision of True Class	0.7199	0.5872
ROC Area	0.824	0.654

5. Discussion

The current study examined the performance of several data mining classifiers, namely the Decision Tree (The C5.0 and CHAID), Logistic Regression (forward stepwise and backward stepwise), and ANN (MLP and RBF). As stated, the following eight features of divorced women were studied in predicting the *Status* of the women: *age, race, religion quali, dr, wes, tom* and *hes*. All techniques were compared using accuracy, sensitivity, specificity, precision of true class, and ROC Area measures. The best technique in each model was selected, namely Decision Tree (The C5.0), Logistic Regression (Forward), and ANN (Multi-Layer Perceptron).

From our observation, all methods are significantly able to predict the event occurrence (status = divorced/separated). The evaluation based on the Decision Tree models revealed that the C5.0 was better at classifying the outcome compared to CHAID. Compared to the Logistic models, the performance of the forward stepwise method was slightly better than backward stepwise in all measures except in the specificity measure. On the other hand, MLP outperformed RBF in all measures in the ANN models.

Next, we compared the performance of the best method of Decision Tree (C5.0), Logistic Regression (forward), and ANN (MLP). Table 7 shows that Decision Tree (C5.0) outperformed Logistic Regression (forward) and ANN (MLP). The accuracy of Decision Tree (C5.0) was 77.96% followed by MLP (74.68%) and forward (67.89%). Decision Tree (C5.0) had the highest sensitivity value (85.33%) followed by ANN (MLP) (80.04%) and Logistic Regression (forward) (67.65%). The precision value for the C5.0 model is 0.7447. Therefore, it can be concluded that 74.47% of women predicted as divorced/separated women is true.

Table 7. The Performance Comparison of Best Techniques

Measures	**Decision Tree (C5.0)**	**Logistic Regression (Forwards)**	**ANN (Multi-Layer Perceptron)**
Accuracy	77.96%	67.89%	74.68%
Sensitivity	85.33%	67.65%	80.04%
Specificity	70.54%	68.14%	69.29%
Precision of True Class	0.7447	0.6814	0.7241
ROC Area	0.849	0.733	0.829

Based on the ROC graph shown in Figure 2, the C5.0 line was the farthest to the diagonal line, thus strengthening the conclusion that the model is the best model. Therefore, the C5.0 model was chosen as the best model for this study.

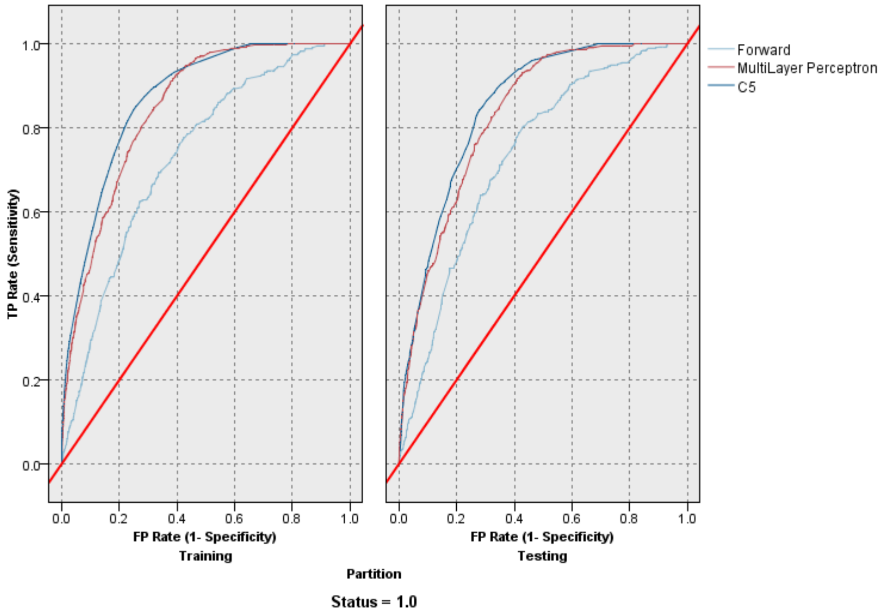

Figure 2. ROC Chart for Models Comparison

Table 8 shows the orders of the characteristics of Malaysian divorcees (women) based on the best model selected (The C5.0). Table 8 shows that all the predictors, except *religion*, have an importance value of more than 0.1. Order-wise, the *wes* appears to be the most important characteristic with importance value 0.1531 followed by *hes* (0.1396), *tom* (0.1327)*, race* (0.1327)*, dr* (0.1212)*, quali* (0.1115)*, age* (0.1053) *and religion* (0.0998).

Table 8. Predictor Importance of Best Models (The C5.0)

Predictors	Importance
wes	0.1531
hes	0.1396
tom	0.1327
race	0.1327
dr	0.1212
quali	0.1115
age	0.1053
religion	0.0998

Further analyses reveal that the tendency of a career woman to divorce is 3.44 times higher than unemployed woman. This is because, a career woman may be too focused with her career and it causes her to neglect her responsibilities as a wife and a mother. It can be concluded that the employment status of Malaysian women may be one of the root factors that gives rise to lack of communication and less time with family.

Next, we also found that Other Bumiputras wife has higher tendency to get divorce than Malay, Chinese and Indian races with 1.11 times, 1.48 times and 1.05 times higher significantly. We also found that 69.4% of Other Bumiputras women who get divorce is currently employed, in which we can relate this situation with the lack of communication and less time spent with the family.

To add, there is a significant association between the education level and employment status among the divorcees. The proportion of employed women gets higher as their level of education increases - No Schooling (58.3%), Pre-School Education (NA), Primary Education (66.0%), Lower Secondary Education (60.0%), Upper Secondary Education (78.3%), Pre-University (100.0%), Tertiary Education (90.0%), Others (NA).

On the flip side, the employment sector of the husband plays an important role in ensuring the stability of a relationship. From our analysis, woman with an unemployed husband has higher tendency to get divorce than woman with a government servant husband, private worker husband and self-employed husband with 1.82 times, 1.58 times and 1.88 times higher tendency significantly. We can relate this situation with the financial stability of a couple where Sayer *et al.* (2011) also found that woman with a jobless husband is more likely to divorce.

6. Acknowledgement

The authors would like to express their gratitude to the National Population and Family Development Board (LPPKN) for providing the data.

7. Disclosure Statement

In accordance with MJOC policy and my ethical obligation as a researcher, we are reporting that we have no financial and/or business interests in any party that may affect the research reported in the enclosed paper.

References

Abdelkader, E. M., Al-Sakkaf, A., & Ahmed, R. (2020). A Comprehensive Comparative Analysis of Machine Learning Models for Predicting Heating and Cooling Loads. *Decision Science Letters,* 9(3), 409-420.

Abdul Ghani, N., Norris, I. N. H., Abdullah, B., Ahmad, M. F., Zulkifli, N. I., & Hasan, M. R. (2017). Divorce Trends among the Malay Community in Perlis, Malaysia over a Ten-Year Period (2006 ~ 2015). *International Journal of Academic Research in Business and Social Sciences,* 7, 269-283.

Abirami, S. & Chitra, P. (2020). Energy-efficient edge based real-time healthcare support system. In *Advances in Computers*; Elsevier: Amsterdam, The Netherlands, 2020; Volume 117 (pp. 339–368).

Al-jabery, K. K., Obafemi-Ajayi, T., Olbricht, G. R. Wunsch II, D. C. (2020). 9 - Data Analysis and Machine Learning Tools in MATLAB and Python*, Academic Press*, (pp. 231-290).

Borucka, A. (2020). Logistic Regression in Modeling and Assessment of Transport Services. *Open Engineering,* 10(1), 26-34.

Buaton, R., Mawengkang, H., Zarlis, M., Effendi, S., Pardede, A. M. H., Maulita, Y., Fauzi A., Novvriyenni, N., Sihombing, A., & Lumbanbatu, K. (2019). Decision Tree Optimization in Data Mining with Support and Confidenc. *Journal of Physics: Conference Series, 1255(1)*. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1255/1/012056/pdf.

Department of Statistics Malaysia (2019). Marriage and Divorce Statistics, Malaysia, 2019 [Press Release]. *Retrieved from https://www.dosm.gov.my/v1/index.php?r=column/pdfPrev&id=d1BZVzBZYXVwOTBPdXhGVEJTQWl4dz09*

Department of Statistics Malaysia (2020). Marriage and Divorce Statistics, Malaysia, 2020 [Press Release]. *Retrieved from https://www.dosm.gov.my/v1/index.php?r=column/pdfPrev&id=QmZ1cE4xRFAvYWQ0R05hTk1rWm5KQT09*

Dreiseitl, S. & Ohno-Machado, L. (2003). Logistic regression and Artificial Neural Network Classification Models: A Methodology Review. *Journal of Biomedical Informatics*, 35 (5-6), (pp. 352-359).

Fath, A. H., Madanifar, F., & Abbasi, M. (2020). Implementation of multilayer perceptron (MLP) and radial basis function (RBF) neural networks to predict solution gas-oil ratio of crude oil systems. *Petroleum.* 6(1), 80–91.

Fetzer, M., (2017, April 12). *By the numbers: A breakdown of divorce by generation*. Avvo Stories. *https://stories.avvo.com/relationships/divorce/numbers-breakdown-divorce-generation.html*

Folke, O., & Rickne, J. (2020). All the Single Ladies: Job Promotions and the Durability of Marriage. *American Economic Journal: Applied Economics.* 12 (1), 260-87.

Guzella, T. S., & Caminhas, W. M. A review of machine learning approaches to spam filtering. *Expert Syst. Appl.* 36, 10206–10222 (2009).

IBM Knowledge Center. (2021a, Sept 20). Logistic Node Model Option. IBM Corporation. *https://www.ibm.com/docs/en/spss-modeler/18.1.0?topic=node-logistic-model-options*

IBM Knowledge Center (2021b, Sept 20). Predictor Importance. IBM Corporation. *https://www.ibm.com/docs/en/spss-modeler/18.1.0?topic=SS3RA7_18.1.0/modeler_mainhelp_client_ddita/clementine/idh_common_predictor_importance.html*

IBM Knowledge Centre. (2022, July 17). Basics (neural networks). IBM Corporation. *https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=networks-basics-neural*

Jones, G. (2018, August 9). Breaking Down Divorce by Generation. Goldberg Jones. *https://www.goldbergjones-wa.com/divorce/divorce-by-generation/#:~:text=As%20far%20as%20breaking%20down,at%20higher%20rates%20than%20ever*

Kaur, H., & Wasan, S. K. (2006). Empirical Study on Applications of Data Mining Techniques in Healthcare. *Journal of Computer Science*. 2(2), 194–200.

Krapf, S. (2017). Moving in or Breaking Up? The Role of Distance in the Development of Romantic Relationships. *European Journal of Population*. 34(3), 313–336.

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). 5-Foundations of Data Imbalance and Solutions for a Data Democracy. *Data Democracy, At the Nexus of Artificial Intelligence, Software Development and Knowledge Engineering*, ISBN: 978-0-12-818366-3, 83-106.

Kozak, J. (2018). Decision Tree and Ensemble Learning Based on Ant Colony Optimization. *Volume 781 of Studies in Computational Intelligence*. Springer.

Lin, C., & Fan, C. (2019). Evaluation of CART, CHAID, and QUEST Algorithms: A Case Study of Construction Defects in Taiwan, *Journal of Asian Architecture and Building Engineering*. 18(6), 539-553.

Milanovic, M., & Stamenkovic, M. (2016). CHAID Decision Tree: Methodological Frame and Application. *Economic Theme,* 54(4), 563-586.

Mohamed, R. I. A., & Alkhyeli, M. K. (2016). Early divorce prediction in Abu Dhabi. *Proceedings of The Spirit of Official Statistics: Partnership and Continuous Innovation.* 1 - 16, International Association for Official Statistics.

Mossalam, A., & Arafa, M. (2018). Using Artificial Neural Networks (ANN) In Projects Monitoring Dashboards' Formulation. *HBRC Journal*. 14(3), 385-392, DOI: 10.1016/j.hbrcj.2017.11.002

Mudunuru, V. R., & Skrzypek, L. A. (2020). A Comparison of Artificial Neural Network and Decision Trees with Logistic Regression as Classification Models for Breast Cancer Survival. *International Journal of Mathematical, Engineering and Management Science*. 5 (6), pp. 1170-1190.

National Population and Family Development Board (2016). Report on The Key Findings of the Fifth Malaysian Population and Family Survey (MPFS-5) 2014. Kuala Lumpur.

Nichols, J. A., Chan, H. W., & Baker, M. (2019). Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical reviews*. 11(1), 111–118. https://doi.org/10.1007/s12551-018-0449-9

Pandya, R., & Pandya, J. (2015). C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications.* 17(16), 18-21.

Rahlin, N. A., Awang, Z., Afthanorhan, A., & Aimran, N. (2019). Antecedents and Consequences of Employee Safety Climate in The Small Manufacturing Enterprises: Translation, Validation and Application of the Generic Safety Climate Questionnaire. *International Journal of Innovation, Creativity and Change.* 7(10), 307-328.

Raley, R. K., McClendon D. M., & Steidl, E. A. (2015). Credits & credentials: An in-depth analysis of the association between educational attainment and the risk of divorce. *Paper presented at PAA, May 1st, 2015.*

Raymo, J. M., Fukuda, S., & Iwasawa, M. (2013). Educational Differences in Divorce in Japan. *Demographic Research*. 28, 177-206.

Reddi, S., & Eswar, G.V. (2021). Chapter 9 - Fake news in social media recognition using Modified Long Short-Term Memory Network. *Intelligent Data-Centric Systems, Security in IoT Social Networks.* Academic Press, 205-227.

Rodrigues, I., (2020, February 17). *CRISP-DM methodology leader in data mining and big data*. Towards Data Science. *https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781*

Sayer, L. C., England, P., Allison, P. D., & Kangas, N. (2011). She Left, He Left: How Employment and Satisfaction Affect Women's and Men's Decision to Leave Marriages. *AJS; American Journal of Sociology.* 116(6), 1982-2018.

Sharma, A., Chudhey, A. S., & Singh, M. (2021*). Divorce case prediction using Machine learning algorithms. 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). Doi:10.1109/icais50930.2021.93958.*

Sharma, N., Chawla, V. K., & Ram, N. (2019). Comparison of Machine Learning Algorithms for the Automatic Programming of Computer Numerical Control Machine. *International Journal of Data and Network Science.* 4(1), 1-14.

Siddiqui, M. A., Khan, A. S., & Witjaksono, G. (2020). Classification of the Factors for Smoking Cessation using Logistic Regression, Decision Tree & Neural Networks. *AIP Conference Proceedings 2203, 020036 (2022). Doi: 10.1063/1.5142128.*

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*. *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044.

Sperandei S. (2014). Understanding Logistic Regression Analysis. *Biochemia medica*. 24(1), 12–18. https://doi.org/10.11613/BM.2014.003

Steverman, B. (2016, July 28). Don't Blame Divorce on Money. Ask: Did the Husband Have a Job?. Bloomberg. *https://www.bloomberg.com/news/articles/2016-07-28/don-t-blame-divorce-on-money-ask-did-the-husband-have-a-job*

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data*? PLoS One. 2017 Apr 4;12(4):e0174944.* DOI: 10.1371/journal.pone.0174944. PMID: 28376093; PMCID: PMC5380334.

Yöntem, M. K., Adem, K., İlhan, T., & Kılıçarslan, S. (2019). Divorce Prediction Using Correlation Based Feature Selection and Artificial Neural Networks. *Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi*, 9(1), 259-273.