FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

WID3006 - MACHINE LEARNING

PROJECT REPORT

LECTURER NAME: DR. AZNUL QALID BIN MD SABRI

Title: *Divorce Probability Prediction*

Name	Matric ID
AHMED AMAN IBRAHIM (LEADER)	17205436/1
LU YU (ASSOCIATE LEADER)	17206490/1
XIE XINYU	17200044/1
GUO YUWEI	17200505/1
MUHAMMAD ASYRAF BIN MUSTAFFA	17206591/2
SAAD HUMAYUN	17207002/1

TABLE OF CONTENTS

Section 1: Introduction to Problem

According to the Department of Statistics Malaysia, in the year 2020, the number of divorces in Malaysia increased 12.0 percent from 50,862 (2018) to 56,975 (2019). During the same time period, the number of divorces decreased by 1.2%. So even though marriages were decreasing in Malaysia, divorce was on the rise!

Divorces affect not only the lives of the couple involved but their children as well. Thus, a rising divorce rate is a prevalent problem in society. If couples were told beforehand that they are likely to get divorced, they would know to seek professional help. This could also help family counselors and therapists in problem formulation and coming up with an intervention plan.

Section 1.1 Hypothesis of the Problem

We believe there are things happening in a couple's life and their behaviour which are indicative of a divorce. Given these features, it is likely that we will be able to develop a model that can predict whether a couple is going to get divorced or not.

Section 1.2 Project Objectives

1.2.1 Obtain Dataset

Find a dataset or collect data about things happening in a couple's life and their behaviour. The labels on this data should be whether the couple got divorced or not.

1.2.2 Making a Model that Predicts Divorce Probability

Select a suitable machine learning model and train it on the dataset. Our model should predict divorces. The output should be the probability of divorce.

1.2.3 Identify Features Which are Predictors of Divorces

Perform statistical tests such as correlation to find suitable features of our model. Model can also be trained with and without certain features to see if they play an important role in predicting divorces.

1.2.4 Deployment of Ideas

Deploy the model on a website so that couples, counselors and therapists can use it to predict the probability of divorce. The format will be for the couple to fill in a questionnaire and when they submit, receive a probability value from the model.

Section 1.3 Literature Review and Similar Works Reading

1.3.1 Divorce Prediction Based on Relevant Feature Selection

Yöntem, M., Adem, K., İlhan, T., Kılıçarslan, S. (2019) conducted a study to find which features are predictive of divorce. The divorce prediction was conducted based on the treatment of the Gottmans using the Divorce Prediction Scale (DPS). A total of 174 participants, of which 84 (49%) were divorced, and 86 (51%) were married Couples. After the participants completed the "Personal Information Form" and "Divorce Prediction Scale", a multilayer perceptron neural network and C4.5 decision tree algorithm were used to study the success of DPS.

1.3.2 Machine Learning to Predict a Successful Marriage by personality

Rahayu, W. K. (2020) studies whether a successful marriage can be predicted when machine learning is applied to personality features. The research showed that machine learning techniques can be used to predict the quality of a relationship. General personality traits and relationship related personality traits were studied.

Section 2: Methodology

Section 2.1 Elaboration on Data

The dataset was collected by researchers in Turkey (Yöntem, Adem, İlhan, Kılıçarslan, 2019). It is in the form of a questionnaire which is given to the couple and the label is whether they got divorced or not. The questionnaire had 54 questions and asked questions such as: “I enjoy travelling with my spouse” or “even if I’m right, I stay silent so as not to hurt my spouse” etc.

Among the total 170 participants of the questionnaire, 84 (49%) were divorced and 86 (51%) were married couples. There were 84 males (49%) and 86 females (51%) in the study group. The ages of the participants ranged from 20 to 63 ($\bar{X}$ = 36.04, SD = 9.34). Of the participants, 74(43.5%) were married for love, and 96 (56.5%) were married in an arranged marriage. While 127 (74.7%) of the participants had children, 43 (25.3%) had no children.

Section 2.2 Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

The feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable y. In our case, we use Pearson’s correlation to check out the relationship between each feature and the target variable.

From our research, having the absolute value of Pearson’s correlation of each feature and the target variable **above 0.8** will bring a pretty good performance of selecting out the most strongly correlated features while retaining as many features as possible, hence we address 0.8 as our **correlation trade-off threshold**.

The raw data we retrieve from our source data file contains 170 rows, **54 columns**, however, after we filter out all the features that possess the correlation value less than 0.8, there are only **40 features** remain in the data set, given the variable name X_reduced, for the convenience of splitting train set and test set.

Table 1: Before feature selection (54 columns)

	Correlation
Q1	0.861324
Q2	0.820774
Q3	0.806709
Q4	0.819583
Q5	0.893180
Q6	0.420913
Q7	0.544835
Q8	0.869569
Q9	0.912368
Q10	0.834897
Q11	0.918386
Q12	0.868983
Q13	0.844743
Q14	0.864316
Q15	0.901220
Q16	0.886260
Q17	0.929346
Q18	0.923208
Q19	0.928627
Q20	0.907008
Q21	0.864519
Q22	0.825938
Q23	0.837504
Q24	0.839392
Q25	0.857052
Q26	0.872868
Q27	0.869788
Q28	0.846606
Q29	0.892954
Q30	0.874531
Q31	0.792607
Q32	0.829056
Q33	0.861328
Q34	0.835167
Q35	0.862624
Q36	0.886497
Q37	0.863597
Q38	0.883311
Q39	0.896180
Q40	0.938684
Q41	0.894356
Q42	0.739629
Q43	0.566242
Q44	0.847336
Q45	0.546450
Q46	0.443465
Q47	0.656409
Q48	0.619830
Q49	0.740704
Q50	0.755248
Q51	0.692681
Q52	0.651478
Q53	0.711176
Q54	0.806765

Table 2: After feature selection (40 columns)

	Correlation
Q1	0.861324
Q2	0.820774
Q3	0.806709
Q4	0.819583
Q5	0.893180
Q8	0.869569
Q9	0.912368
Q10	0.834897
Q11	0.918386
Q12	0.868983
Q13	0.844743
Q14	0.864316
Q15	0.901220
Q16	0.886260
Q17	0.929346
Q18	0.923208
Q19	0.928627
Q20	0.907008
Q21	0.864519
Q22	0.825938
Q23	0.837504
Q24	0.839392
Q25	0.857052
Q26	0.872868
Q27	0.869788
Q28	0.846606
Q29	0.892954
Q30	0.874531
Q32	0.829056
Q33	0.861328
Q34	0.835167
Q35	0.862624
Q36	0.886497
Q37	0.863597
Q38	0.883311
Q39	0.896180
Q40	0.938684
Q41	0.894356
Q44	0.847336
Q54	0.806765

Section 2.3 Model Design

We decided to go with a multilayer perceptron model. We used TensorFlow to customize our MLP model and coded it on Jupyter Notebook. Details regarding of the model selection procedures will be elaborated in the following chapters.

Section 2.4 Activation Function

We prefer to use **ReLU activation** in the hidden layers and only use **softmax activation** in the output layer. We used sigmoid as it is a binary classification task and we needed to output the probability value for the divorce.

Section 2.5 Loss Function

Our project decided to implement **binary Cross-Entropy Loss** as loss function, because **binary Cross-Entropy Loss** suits our binary classification model.

Section 2.6 Optimizer

We choose **Adam optimizer**, which is an extension to stochastic gradient descent (SGD), because it is an computationally efficient optimizer with learning rate=0.01 and no momentum.

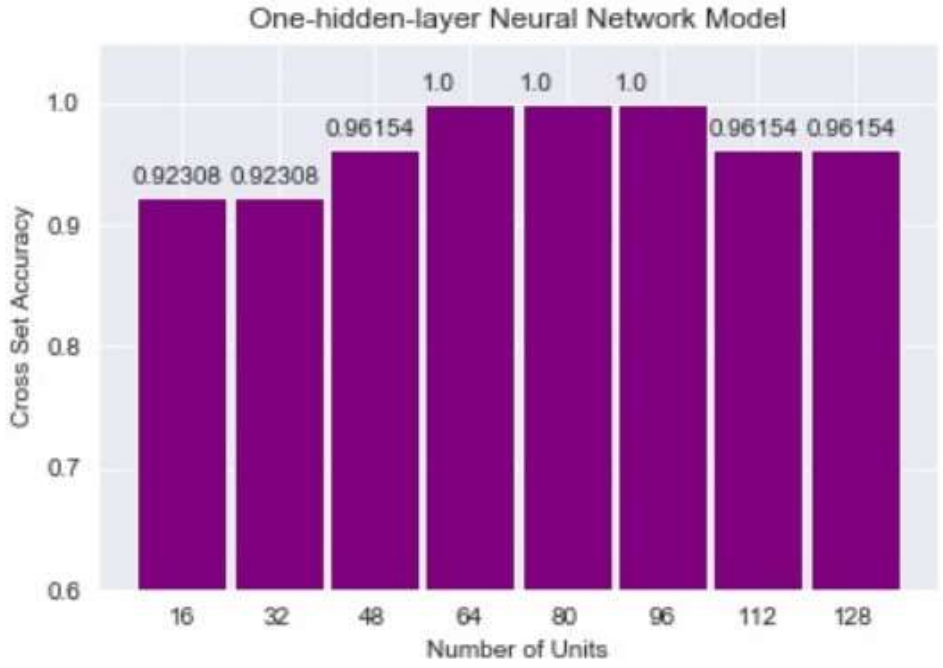
Section 2.7 Number of Units

Since the model we choose is a multi-layer perceptron, it is very important to decide the number of units in each layer. The number of units(neuron) of input layer and output layer is fixed because **the number of units of input layer is equal to the number of selected features**, and **the number of units of output layer is 1**, because our problem is a binary classification problem, so which means it is either 0 or 1. Therefore, we only focus on the number of units in the hidden layer. We decide to choose 64 **units in the hidden layer.**

The method to find the suitable number of units in the hidden layer is to **test the accuracy of the model for different sizes of units in the hidden layer.** We choose to test the 16, 32, 48, 64, 80, 96, 112, 128 units because these numbers are the number of units often selected in the neural network hidden layer, and some numbers are arranged in arithmetic series.

The result is shown in the below figure, it is clearly seen that when the units are 16, the accuracy has been pretty impressive, at about 92.3%. And with the number of units increasing (from 16 to 64), the accuracy also increases. And when the number of units is 64, the accuracy is amazingly standing at 100%. And when the units are bigger than 64, the accuracy keeps no change, at 100%. However, as the number of units in the hidden layer increases to 112, the prediction accuracy on cross validation set decreases, this actually indicates the excessive number units have caused model overfitting. Meanwhile, though 64, 80 and 96 units for one-hidden layer model all perform well in the cross validation set prediction, it is necessary to mention that the computation time is also increasing as the number of units increases. So in conclusion, we choose 64 units in the hidden layer because it reaches high accuracy and it is relatively computationally efficient.

Figure 2: Cross Validation Set Prediction Accuracy with One-hidden-layer Neural Network Model



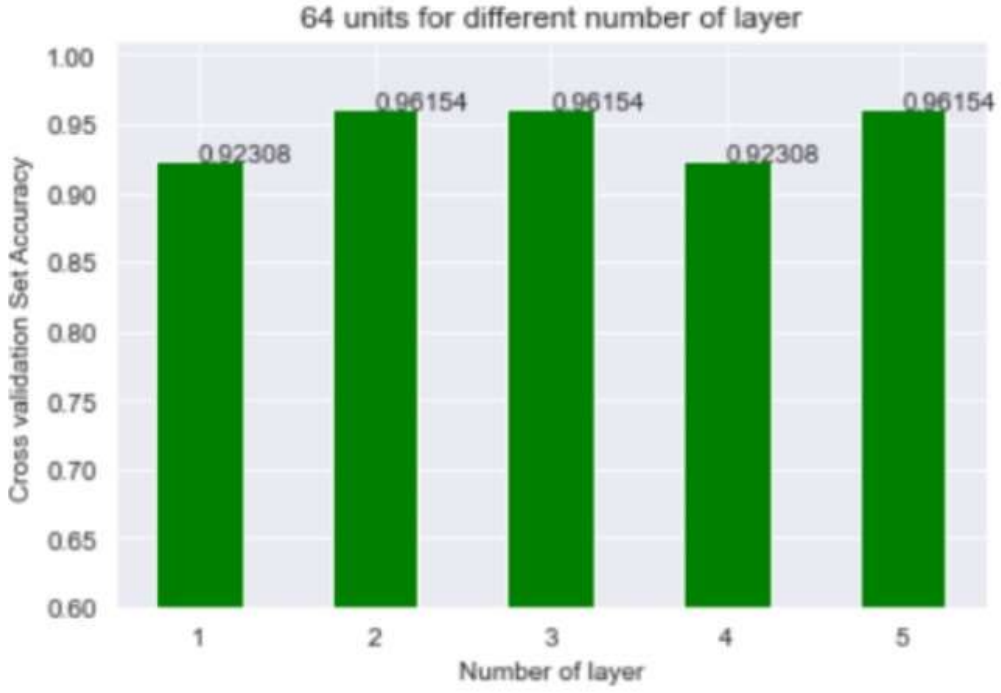
Section 2.8 Number of Hidden Layers

In this model, we choose only **one layer**, because one layer is the most suitable number of layers in this model.

Under normal circumstances, the more the number of layers, the higher the accuracy of the training model. Then more layers will also bring new problems such as overfitting and increased training time and cost. Therefore, it is particularly important to explore the number of layers suitable for a model. Hence, in our project, we set the units of each hidden layer to 64, and choose a different number of hidden layers, from 1 to 5.

From the figure shown below, we can clearly see that with the number of hidden layers increases, the accuracy doesn't increase too much. For example, for one hidden layer, the accuracy is 92.308%. and for the 5-hidden-layer model the accuracy is 96.154%. However, the computation time for the 5-hidden-layer model is exponential. As a result, from a balance point of view, a hidden layer is enough.

Figure 3: Cross Validation Set Prediction Accuracy with 64 Units for Different Number of Layers



Section 2.9 Regularization

Regularization is one of the most important concepts of neural networks. It is a technique to prevent the model from overfitting by adding extra information to it. This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model. It mainly regularizes or reduces the coefficient of features toward zero. In simple words, with a regularization technique, we reduce the magnitude of the features by keeping the same number of features.

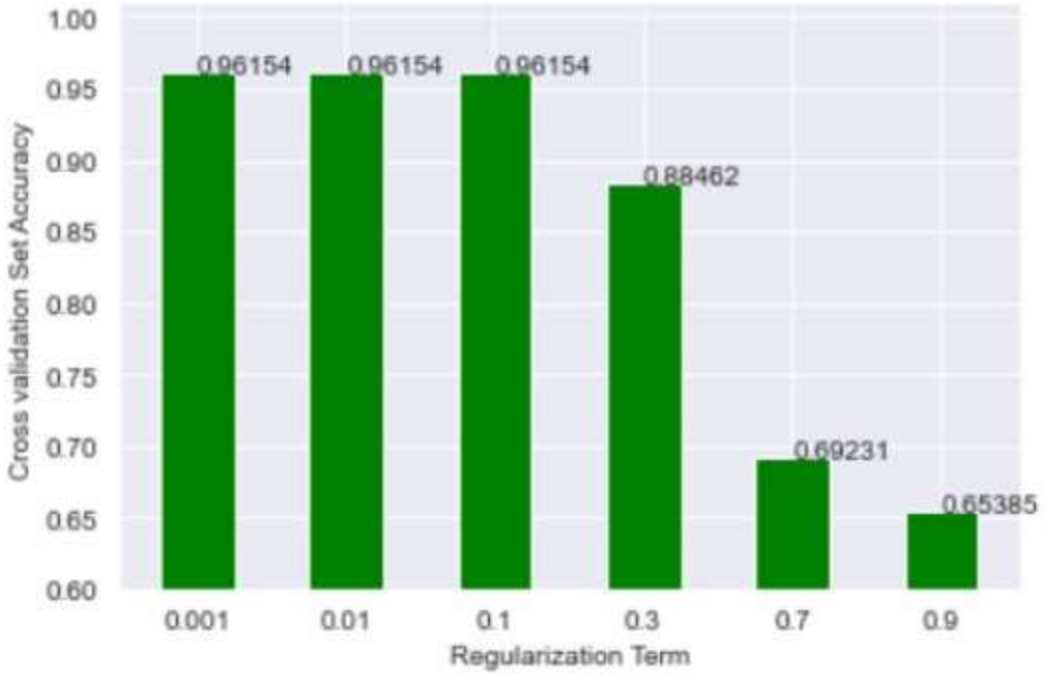
In this neural network model, **there is no need to use regularization**, there are a total of 3 reasons:

The first reason is that we have only one hidden layer with 64 units. For a model to be trained, training data and features are relatively small, so the probability of overfitting is small. However, the probability of underfitting may be relatively high. If you want to solve the problem of underfitting, regularization is not a suitable method. It is easier to deal with this problem by increasing the training library and features.

Secondly, we found that different regular terms will affect the training accuracy. By trying different regularization terms, it is found that the higher the regularization value, the lower the accuracy. As shown in the figure below, our regular term uses 0.001, 0.01, 0.1, 0.3, 0.7 and 0.9 respectively. We find that the cross validation set accuracy is continuously shrinking from the original 0.961 to 0.653 as the regular term increases.

Last but not least, our training epoch is quite small, only 5 epochs, so there are less data to be trained, it is less likely to have overfitting issues.

Figure 4: Cross Validation Set Prediction Accuracy with Different Regularization values



Section 3: Results and Discussions

Making predictions on the test set of 30 examples, we got 96% accuracy for our model. The confusion matrix for the predictions on the test set is seen below.

	Predicted 0	Predicted 1
Actual 0	16	1
Actual 1	0	8

Since one of the use cases of our model is for couples to know when to seek help if they are likely going to divorce, it is very important that we reduce the number of false negatives (Predicted 0 but Actual 1). As shown in the confusion matrix, we managed to obtain 0 false negatives.

We have used more performance metrics to evaluate our model. They are:

- **Precision:** ratio of correctly predicted positives to total predicted positives
- **Recall:** ratio of correctly predicted positives to total actual positives
- **F1 score:** weighted average of precision and recall (higher is better)

Precision: 0.88
Recall: 1.0
F1 score: 0.94

As we want our false negatives to be as low as possible, a recall as high as possible is preferred, even at the expense of precision.

Section 4: Suggestion for the Future Works

Since the dataset we used to train our model included only couples from Turkey, it might not generalize well for couples from other countries/regions. Different social and cultural values could well affect divorces. Therefore, it is important to collect data from couples in different regions. In consequence, our project team aims to create an online database to assemble diverse information from couples worldwide.

In addition, more features which could be indicative of divorces could also be explored and data collected about them. Currently, the dataset consists of subjective questions to the couple. Objective features such as the salaries, occupation and hours spent at work could also be explored to see if they are indicative of divorce probability. As a result, we group members decide to enhance the divorce factors study through an advanced and comprehensive social research.

Section 5: Appendix

Source Code:
https://github.com/politecat314/divorce-prediction/blob/master/Project_0603.ipynb

Dataset Referred:
https://drive.google.com/file/d/1bWkUQKOYvLS9Oy_dVGrsj8dKI1UCHJxS/view?usp=s haring

Video: https://drive.google.com/file/d/14KCC88nmMQV8SoCBC4qIA2ZvBsw22P9k/view?usp=sharing

Section 6: References

1. Yöntem, M., 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. Retrieved from https://dergipark.org.tr/tr/download/article-file/748448
2. Rahayu, W. K. (2020, February 6). *Machine Learning to Predict a Successful Marriage by Determining Some Important Variable*. RPubs. https://rpubs.com/wkania/Divorce_Preventor.
3. Klemz, J. (2018, July 31). How Dr. Gottman Can Predict Divorce with 94% Accuracy. Retrieved June 02, 2021, from https://reallifecounseling.us/predict-divorce-gottman/
4. Chen, P. (2020, February 20). Machine Learning for Predicting Divorce (Marriage Story for Nerds). Retrieved June 02, 2021, from https://peijin.medium.com/machine-learning-for-predicting-divorce-marriage-story-for-nerds-1033b69845f8 5. Scott, S., Rhoades, G., Stanley, S., Allen, E., & Markman, H. (2013, June). Reasons for Divorce and Recollections of Premarital Intervention: Implications for Improving Relationship Education. Retrieved June
   02, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4012696/
6. Chang, T., & Jones, G. (1990). Malay Divorce in Peninsular Malaysia: The Near-disappearance of an Institution. *Southeast Asian Journal of Social Science, 18*(2), 85-114. Retrieved June 2, 2021, from http://www.jstor.org/stable/24491673
7. Divorce statistics and facts: What affects divorce rates in the U.S.? (2020, August 12). Retrieved June 02, 2021, from https://www.wf-lawyers.com/divorce-statistics-and-facts/
8. Ortiz-Ospina, E., & Roser, M. (2020, July 25). Marriages and divorces. Retrieved June 02, 2021, from https://ourworldindata.org/marriages-and-divorces

9. More than 100,000 divorces in England and Wales last year. (2020, November 17). Retrieved June 02, 2021, from https://www.bbc.com/news/world-asia-china-54972762
10. Solicitors, B. (2020, January 21). Divorce and recent divorce rates: BSG Solicitors Lancaster and Preston. Retrieved June 02, 2021, from https://www.bsglaw.co.uk/news/2020/1/21/divorce-and-recent-divorce-rates