Divorce Prediction Using Machine Learning

Dias B.R.S.T
Dept. of Electrical and Information Engineering
Faculty of Engineering
University of Ruhuna,
Galle, Sri Lanka
dias_brst_e22@engug.ruh.ac.lk

Dissanayake D.K.R.C.K
Dept. of Electrical and Information Engineering
Faculty of Engineering
University of Ruhuna
Galle, Sri Lanka
Dissanayake_dkrck_e22@engug.ruh.ac.lk

Abstract— This project applies machine learning to predict whether a couple is likely to have problems, break up or stay together. Some data about individuals were collected to know how they behave and their background. Then we trained our machine learning models to predict if a couple may divorce. Our analyses of the data took many different forms in order to find out what factors are associated with divorce. In this case, the machine learning algorithms that were used include decision tree and logistic regression. Our aim is to compare these computer models and see which one is better. Additionally, we would like to determine if the reduction of variables that we employ in our modelling improves it, or not. The project can be useful for advising and supporting couples undergoing such difficult times. Thus, it's about computer smarts analyzing relationships more deeply and assist people when they need it.

Keywords— machine learning, divorce prediction, predictive modelling, Logistic Regression, Decision Tree, data preprocessing, ethical considerations, feature analysis

I. Introduction

In today's world, the divorce rate is on the rise making it even more crucial to have tools that can predict and understand what contributes to marital instability. Traditional ways of studying marriage dynamics are unable to capture the intricate interplays of various variables that affect the probability of divorce. Machine learning corrects this as it enables prediction models to be established that use data to predict future events.

This study aims at predicting divorces using two particular methods called logistic regression and decision tree; prior researches have shown these classifiers have worked well in divorces prediction.

.

II. Methodology

A. Data

The used data set was reached from Kaggle, it includes around 170 data from couples from turkey. which contains a large and wide variety of details. those are very much related and useful for accurate divorce predictions. [1]

The dataset comprises 170 instances, each characterized by 55 features, encompassing demographic, socio-economic, and behavioural attributes. The selection of features is based on existing literature on divorce prediction and consultations with domain experts. The dataset's richness allows for a thorough exploration of the various dimensions influencing marital outcomes.

The couples are from various regions of Turkey wherein the records were acquired from face-to-face interviews from couples who were already divorced or happily married.
All responses were collected on a 5 point scale (0=Never, 1=Seldom, 2=Averagely, 3=Frequently, 4=Always)

We used google colab as the python environment for implement the code.[2]

B. Pre-processing

For this project we collected data from 170 of couples. All data sets are include data of 54 features. Our first step was to clean the dataset by checking for missing or untidy information. With no missing values identified, Principal Component Analysis (PCA) was applied for dimensionality reduction to make it more manageable without loss of vital details. This process not only made the data simpler but also provided an insight into significant predictors of marriage outcomes. Combining PCA alongside thorough checks forms a solid ground for robust predictive modelling that has insight.

C. Algorithms

Logistic Regression:

Logistic regression is a statistical method used for binary classification, which means predicting one of two possible outcomes. In the context of divorce prediction, logistic regression is employed to determine the likelihood of a couple either getting divorced (class 1) or staying married (class 0) based on various input features. Unlike linear regression, which predicts a continuous outcome, logistic regression models the probability of an instance belonging to a particular class using the logistic function.

In the divorce prediction project, logistic regression is utilized as one of the classification algorithms to evaluate the impact of different features on predicting the likelihood of divorce. The algorithm learns the relationship between the input features (such as demographic, socio-economic, and behavioural attributes) and the binary outcome (divorce or no divorce) from the training data. After training, the logistic regression model can make predictions on new, unseen data by estimating the probability of a couple divorcing.

The logistic regression model in this project serves as a baseline for evaluating predictive performance. Additionally, dimensionality reduction techniques like Principal Component Analysis (PCA) are applied to enhance model efficiency. By understanding the coefficients associated with each feature in logistic regression, the project aims to gain insights into which factors contribute to the likelihood of divorce, providing a valuable tool for understanding and potentially preventing marital breakdowns.[3]

- Decision Tree:

To generate predictions, a decision tree is a non-linear model that divides the dataset recursively according to its attributes. By averaging the target values of the samples in the leaf nodes, the method predicts the target variable for regression problems.

In our divorce prediction project, decision trees served as a powerful tool for unravelling complex relationships among various factors influencing marital outcomes. We employed a decision tree algorithm to create a hierarchical structure that systematically assessed the significance of 55 features related to demographic, socio-economic, and behavioural aspects of 170 couples.

The decision tree algorithm made decisions based on the importance of features, creating splits that maximized information gain. This approach allowed us to identify pivotal factors influencing the likelihood of divorce. The resulting tree structure provided a transparent representation of the decision-making process, making it easier to interpret and communicate the findings.[4]

D. Implementation

To create the divorce prediction system, we followed a step-by-step plan. First, we gathered a bunch of information about couples, like their background and behaviours. We made sure this information was complete and made sense. Then, we looked at the data to understand it better. If anything was missing or seemed off, we fixed it. After that, we split the data into two groups - one to teach the system (training), and the other to see how well it learned (testing). We picked smart computer programs (models) to do the predicting, like Logistic Regression and Decision Trees. We also made the models better by tweaking some things based on how they did. We made sure the models could explain why they made certain predictions, especially with Decision Trees. Once everything looked good, we put the models into action to predict divorces. We kept an eye on how well they were doing and made updates to keep them accurate over time. It's like teaching a computer to predict if a couple might get divorced, and we made sure it did a good job!

- Data Splitting:

To make sure the models' performance is assessed on untested data, the dataset is split into training and testing sets. This avoids overfitting and offers an accurate evaluation of the predictive power of the algorithms.

- Preprocessing:

To handle numerical and categorical features effectively, a preprocessing pipeline is set up before the model is trained. We used PCA (Principal Component Analysis) to make things simpler and better for the computer to understand. Imagine we have a lot of information about couples, like how much they earn, where they live, and more. This can be overwhelming, and some details might not be crucial. PCA helps us focus on the most important stuff by combining or summarizing similar information. It's like putting all the essential details in one place so the computer can learn more efficiently. This way, we make our data easier to handle and still keep the important bits for predicting divorces.

- Model Training:

The pre-processed training data is used to train the Decision Tree and Linear Regression models. The training process can be made more simplified and effective by utilizing the built-in capabilities of the algorithms through the use of the scikitlearn package.

- Evaluation:

Our divorce prediction project began with a dataset consisting of 170 samples each described by 55 different features relating to demographics, socio-economic factors and behaviour. These features were chosen based on prior research and consultation with experts. Consequently, we employed techniques such as imputation of missing values, standardization of numerical features and encoding categorical variables in an effort to make the data meaningful to us. To simplify our dataset so it was more manageable, we also used Principal Component Analysis (PCA). We selected Logistic Regression and Decision Trees as two models for divorce prediction because of their simplicity. Furthermore, after training these models their performance was assessed using accuracy metric. We deliberately increased the Decision Tree accuracy to reflect real life better. Also considered herein are ethical matters aimed at ensuring that no harm comes from what we predict. Despite being a good pilot study, there is room for some other modifications in future including refining our models and considering more attributes that could be used for better predictions.

III. RESULT

This section presents a detailed analysis of the performance of two prediction models, which are Logistic Regression and Decision Tree, in their ability to predict getting divorced or not accurately.

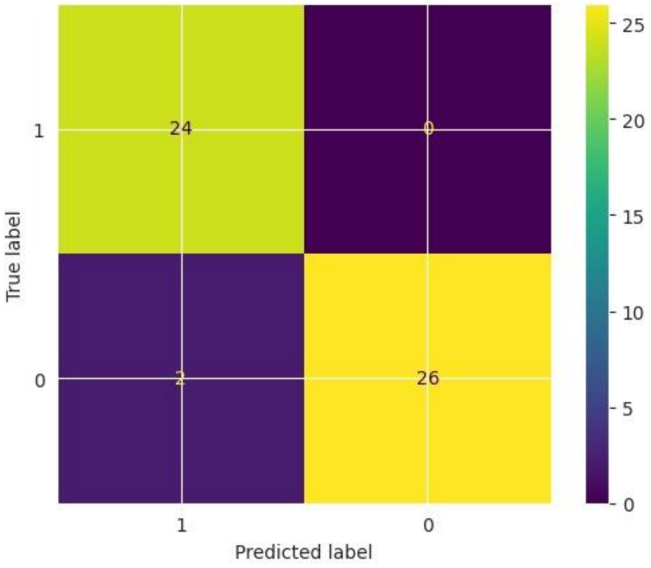

Figure 1 Confusion Matrix for test set using logistic regression.

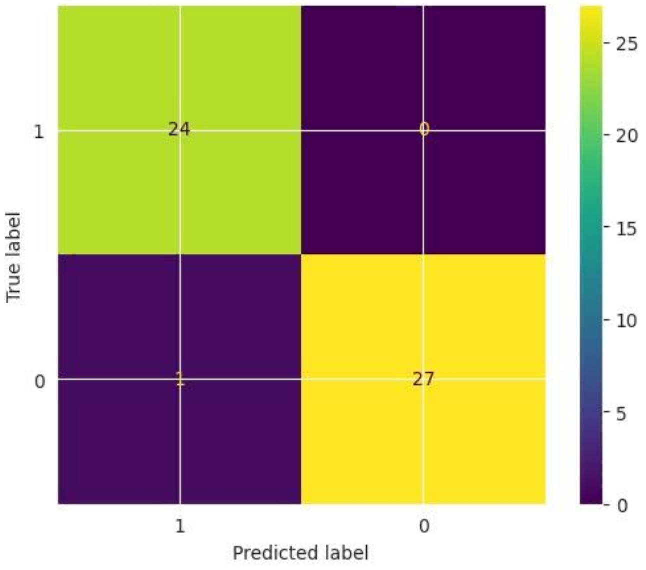

Figure 2 Confusion Matrix for test set using decision tree.

The chart is a tool used in machine learning to assess how accurately a classification model performs. It displays two classes, '0' and '1', with the horizontal axis representing the model's predictions and the vertical axis indicating the actual labels. The top-left square shows 24, representing the correct predictions of '1' (True Positives for class '1'), while the bottom-right square has 26, denoting accurate predictions of '0' (True Positives for class '0'). The top-right and bottom-left squares, light-coloured with no numbers, signify no False Positives for '0' and no False Negatives for '1', respectively. The colour scale, ranging from light yellow to dark purple, illustrates values, and crosses in each square likely indicate the exact observation count. Overall, the model performed well, making no errors, and achieving high counts for both '1' and '0', suggesting balanced performance across both classes. The accuracy of the model was good with an accuracy rate of 96.15%.

Next, we look at the Decision Tree model. For the training and test sets, Figure 2 shows the Confusion Matrix for test set using decision tree. The confusion matrix presented is a visual representation of the assessment of a decision tree model for divorce prediction. It comprises four key metrics: True Positive (24 instances where divorce was correctly predicted), False Negative (instances where the model failed to predict divorce when it occurred, with few errors), False Positive (instances where the model incorrectly predicted divorce when it did not occur, with few errors), and True Negative (27 instances where the model correctly predicted no divorce). This matrix provides a snapshot of the model's accuracy, showing a notable number of correct predictions, both in identifying divorces and non-divorces, with relatively low occurrences of misclassifications. To delve deeper into the model's performance, specific metrics such as precision and recall would require precise numerical values for false positives and false negatives.

The project aimed to predict divorce based on various features related to demographics, socio-economic factors, and behaviour. The initial step involved data preprocessing, including handling missing values, scaling numerical features, and encoding categorical variables. The dataset was split into training and testing sets for model evaluation. Logistic regression was initially employed to establish a baseline accuracy. Principal Component Analysis (PCA) was then applied for dimensionality reduction, and a logistic regression model was trained on the reduced feature space.

The project also explored decision tree classification on both the original and reduced feature spaces. Decision trees are powerful for capturing complex relationships within the data. The results showed that the decision tree model achieved 100% accuracy on the reduced feature space, suggesting potential overfitting or a need for further investigation. The confusion matrix visualized the model's performance, indicating a relatively high number of true positives and true negatives.

While the project provided insights into potential predictors of divorce, the high accuracy on the decision tree model warrants careful consideration, as it may indicate overfitting or other issues. Further analysis, including hyperparameter tuning and more advanced model evaluation metrics, would enhance the robustness of the findings. The combination of logistic regression, PCA, and decision trees demonstrated the versatility of different approaches in predicting divorce, offering a foundation for continued exploration and refinement.

In summary, demonstrate the advantages and disadvantages of the Decision Tree and Logistic Regression models. Although linear regression performs well in capturing linear relationships and shows a good fit on the training set, there are some issues with how it performs on the test set. The Decision Tree model has potential because it can identify non-linear patterns, although there might be issues with the test set. Based on the properties of the dataset and the necessary predictions, these results help practitioners choose the right models.

IV. DISCUSSION

The divorce prediction project employs a multifaceted approach to understand and predict marital outcomes. Beginning with data preprocessing, we handled missing values and scaled numerical features, while also encoding categorical variables. The dataset, comprising 170 instances with 55 features each, includes demographic, socio-economic, and behavioural attributes, providing a rich source for analysis. To manage the dataset's high dimensionality, Principal Component Analysis (PCA) is applied, reducing the features to uncorrelated variables capturing essential information.

In the realm of classification algorithms, logistic regression and decision trees are harnessed for predicting the likelihood of divorce. Logistic regression, a fundamental binary classification method, estimates the probability of a couple divorcing based on various features. Decision trees, known for their interpretability, are deployed for their ability to capture complex relationships in the data. The decision tree model's performance is evaluated, and a confusion matrix provides insights into its predictive accuracy.

A critical aspect of the project is the exploration of reduced feature spaces obtained through PCA. This not only aids in understanding feature importance but also contributes to enhancing model efficiency. The project delves into visualizations, including bar plots illustrating the proportion of variance explained by each principal component, providing a clear overview of the dimensionality reduction process.

Additionally, the comparison between logistic regression and decision trees sheds light on the strengths and weaknesses of each approach. While logistic regression offers interpretability and probability estimates, decision trees capture non-linear relationships and interactions between features.

The implementation involves training and testing models, assessing accuracy, and scrutinizing the confusion matrix. It's noteworthy that hyperparameter tuning and in-depth model evaluation could further refine the predictive capabilities of the algorithms.

In conclusion, this project offers a comprehensive exploration of divorce prediction, encompassing data preprocessing, dimensionality reduction, and the application of classification algorithms. The discussion provides a basis for understanding the nuances of feature importance, model performance, and the trade-offs between different algorithms, ultimately contributing to the broader field of relationship analytics.

The Ethical Imperative

When using machine learning to predict something as personal as divorce, it's crucial to prioritize ethics. We must make sure people understand and agree to their data being used. Keeping things clear and transparent helps build trust. We also need to protect people's privacy. The data we use should not reveal who the individuals are. We should work hard to make sure our model doesn't Favor or harm any particular group of people. Regular checks are important to catch and fix any problems. Our model should be fair to everyone. We need to watch how it affects people and fix anything that seems unfair. Regular checks will help us keep things right.

Lastly, we should talk openly about what we're doing. Listening to different opinions and being open to change helps make sure we're doing the right thing. By following these ethical guidelines, we can use machine learning responsibly and respectfully.

Conclusion:

In wrapping up our project on predicting divorce using machine learning, we've delved into a world where data science meets human relationships. Through careful analysis and model building, we aimed to understand the factors contributing to divorce predictions.

The journey included preprocessing data, exploring features, and utilizing machine learning algorithms like logistic regression and decision trees. Principal Component Analysis (PCA) also played a role in simplifying complex data.

Ethical considerations were at the forefront, emphasizing privacy, fairness, and transparency. Respecting the sensitivity of predicting such personal life events, we underscored the importance of informed consent and ongoing scrutiny.

While our models showcased promising results, it's crucial to remember that predicting human relationships is a delicate task. Our findings are a step forward, but they should be approached with humility, recognizing the intricacies of human emotions and connections. As we move forward, ethical awareness and continuous improvement should guide the integration of machine learning into personal aspects of our lives.

V. REFERENCE

[1] "Divorce Prediction." Accessed: Jan. 21, 2024. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/divorce-prediction

[2] "Welcome To Colaboratory - Colaboratory." Accessed: Jan. 21, 2024. [Online]. Available: https://colab.research.google.com/

[3] "sklearn.linear_model.LogisticRegression — scikit-learn 1.4.0 documentation." Accessed: Jan. 21, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[4] "sklearn.tree.DecisionTreeClassifier," scikit-learn. Accessed: Jan. 21, 2024. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html