**HDSC August ’22 Premiere Project Documentation: Divorce Prediction using ML.**

A project by team Kubernetes

Divorce is the process of terminating a marriage or marital union, dissolving the bonds of matrimony between a married couple under the rule of law of the particular country or state.

Sometimes divorce is the healthiest option for both people involved, but often divorce is preventable if both partners are willing to do the work to keep the marriage together. Although, most people are completely unaware of how much work it actually takes to keep a marriage together, and when they find out, they’re unwilling to do what it takes.

**1.1 Aim of the project:**

To investigate the provided divorce dataset, understand the trends and gather information on:

* How each question relates with divorce column.
* What decisions are popular across divorces.
* What decisions are popular across everyone.
* Correlation between questions.
* What’s the most agreeable question.
* Most important questions.

Later, deploy a machine learning model using the dataset that will predict the possibility of divorce between a couple.

**1.2 Flow Chart:**

This is the entire work flow of what has been planned and achieved under this project. Participating members of the team were divided into 5 sub-groups and had been assigned the task of Data wrangling Exploratory data analysis, Data findings, Modelling and validation and Model Deployment respectively.

*Figure 1: flowchart*

**1.3 Data Source:**

The dataset was sourced from Kaggle website: <https://www.kaggle.com/datasets/andrewmvd/divorce-prediction>

**1.4 Data Description:**

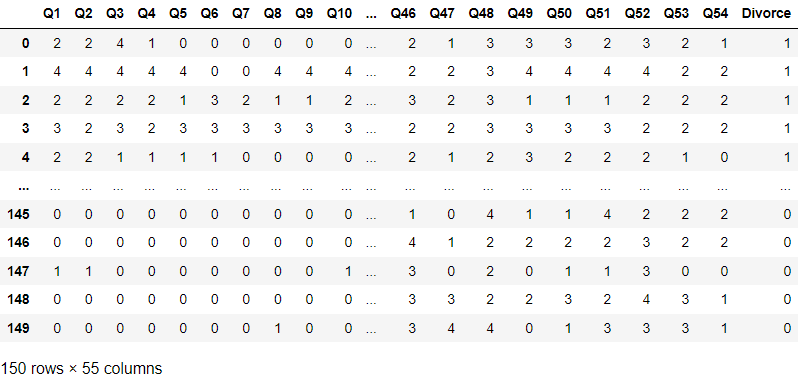
The Divorce Prediction dataset contains information about the answers to the 54 questions asked from 150 married people. Answers are on a 5-point scale (0 = Never, 1 = Rarely, 2 = Average, 3 = Often, 4 = Always). Each question had different probabilities of impact. Hence, we can use this dataset to predict whether a married couple will divorce or not.

The 54 variables or questions called data dictionary are:

1. If one of us apologizes when our discussion deteriorates, the discussion ends.
2. I know we can ignore our differences, even if things get hard sometimes.
3. When we need it, we can take our discussions with my spouse from the beginning and correct it.
4. When I discuss with my spouse, contacting him will eventually work.
5. The time I spent with my wife is special for us.
6. We don't have time at home as partners.
7. We are like two strangers who share the same environment at home rather than family.
8. I enjoy our holidays with my wife.
9. I enjoy traveling with my wife
10. Most of our goals are common to my spouse.
11. think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.
12. My spouse and I have similar values in terms of personal freedom.
13. My spouse and I have similar sense of entertainment.
14. Most of our goals for people (children, friends, etc.) are the same.
15. Our dreams with my spouse are similar and harmonious.
16. We're compatible with my spouse about what love should be.
17. We share the same views about being happy in our life with my spouse
18. My spouse and I have similar ideas about how marriage should be
19. My spouse and I have similar ideas about how roles should be in marriage
20. My spouse and I have similar values in trust.
21. I know exactly what my wife likes.
22. I know how my spouse wants to be taken care of when she/he sick.
23. I know my spouse's favourite food.
24. I can tell you what kind of stress my spouse is facing in her/his life.
25. I have knowledge of my spouse's inner world.
26. I know my spouse's basic anxieties.
27. I know what my spouse's current sources of stress are.
28. I know my spouse's hopes and wishes.
29. I know my spouse very well.
30. I know my spouse's friends and their social relationships.
31. I feel aggressive when I argue with my spouse.
32. When discussing with my spouse, I usually use expressions such as ‘you always’ or ‘you never’.
33. I can use negative statements about my spouse's personality during our discussions.
34. I can use offensive expressions during our discussions.
35. I can insult my spouse during our discussions.
36. I can be humiliating when we discussed.
37. My discussion with my spouse is not calm.
38. I hate my spouse's way of open a subject.
39. Our discussions often occur suddenly.
40. We're just starting a discussion before I know what's going on.
41. When I talk to my spouse about something, my calm suddenly breaks.
42. When I argue with my spouse, ı only go out and I don't say a word.
43. I mostly stay silent to calm the environment a little bit.
44. Sometimes I think it's good for me to leave home for a while.
45. I'd rather stay silent than discuss with my spouse.
46. Even if I'm right in the discussion, I stay silent to hurt my spouse.
47. When I discuss with my spouse, I stay silent because I am afraid of not being able to control my anger.
48. I feel right in our discussions.
49. I have nothing to do with what I've been accused of.
50. I'm not actually the one who's guilty about what I'm accused of.
51. I'm not the one who's wrong about problems at home.
52. I wouldn't hesitate to tell my spouse about her/his inadequacy.
53. When I discuss, I remind my spouse of her/his inadequacy.
54. I'm not afraid to tell my spouse about her/his incompetence.

**1.5 Data Preparation:**

1.5.1 Data wrangling

It was done completely using excel. The text was split into columns using comma as delimiter and removed duplicates using the remove duplicate icon on the tabs. Dataset was then split into two sections and then cleaned; duplicate cases were removed from the data.

*Figure 2. Divorce dataset*

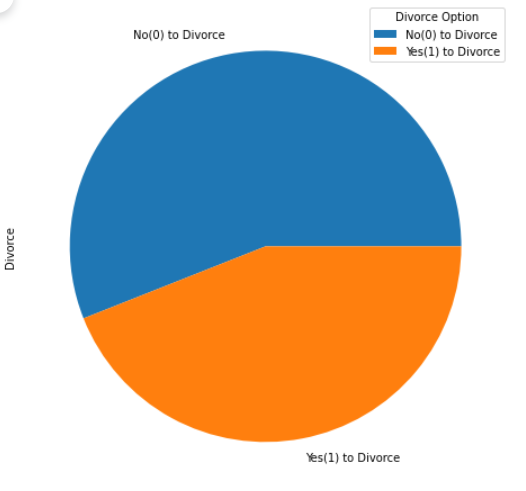
1.5.2 EDA - Exploratory data analysis

After plotting for the percentages of each choice for each question and popularity of choices for questions; the most agreeable questions included were:

'Q43','Q45','Q46','Q47','Q48','Q49','Q50','Q51','Q52','Q53'

Hence, the person be divorced or not is more likely to agree over these questions. where people choose silence over conflicts with their spouse despite the problem not being them, also they hesitate over communicating the inadequacies to their spouse. Hence loss of peace can be considered a No.1 factor for divorce.

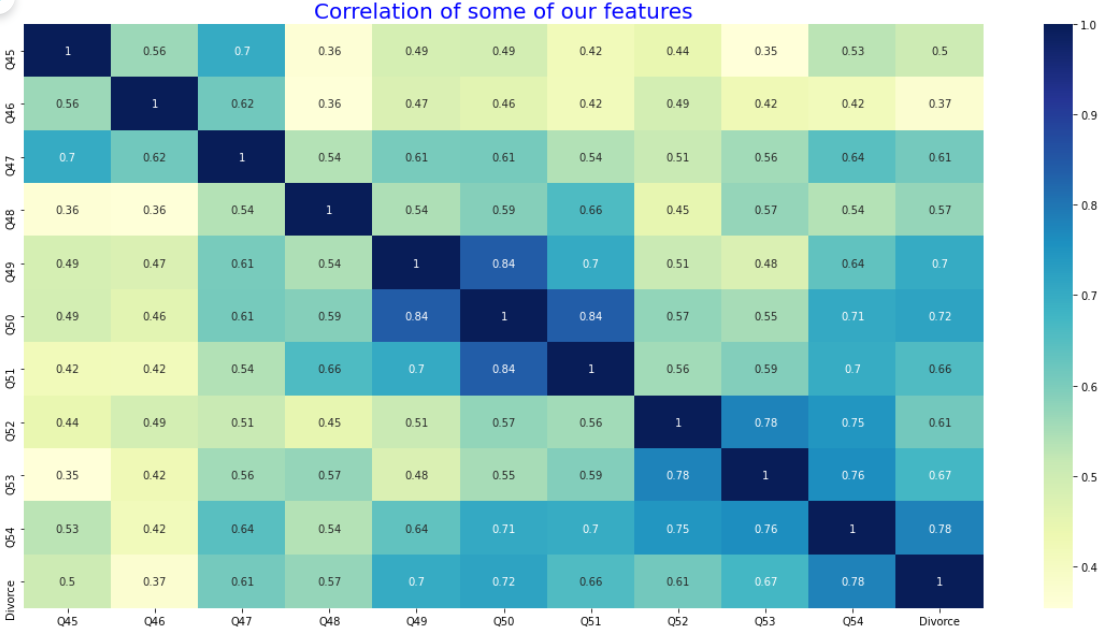
* Distribution of divorce:



*Figure 3: It can be seen from the pie chart, that there is a high rate [ 56%] for people who are not divorced while [ 44%] that of divorced people.*

* Correlation of features:

Positive correlation can be observed between some features of the dataset.

* Top 10 correlated questions with Divorce

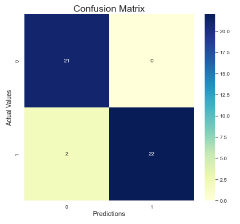
*Figure 4: Correlation plot between some features of the graph.*

* We can see that Q40 which had 47% of people choosing the “Never” option is the most correlated question with Divorce.
* From the observation it is seen that questions 'Q5', 'Q7', 'Q8', 'Q9', 'Q10', 'Q11', 'Q12', 'Q14', 'Q15', 'Q16', 'Q17', 'Q18', 'Q19', 'Q20', 'Q21', 'Q22', 'Q23', 'Q24', 'Q25', 'Q26', 'Q27', 'Q28', 'Q29', 'Q30', 'Q35', 'Q36', 'Q40' had only divorcees answer with “Frequently” and “Always”.
* We can infer that the most distribution of questions answered by married people are left skewed and divorcees is rightly skewed except for Q6 Q7 Q43 Q45 Q46 Q48 Q50 Q51 will check more on this?
* All of these questions ‘Q43','Q45','Q46','Q47','Q48','Q49','Q50','Q51','Q52','Q53' answers couldn't infer if participants were divorced explicitly although divorcees answers were more positive.

Based on the above plots:

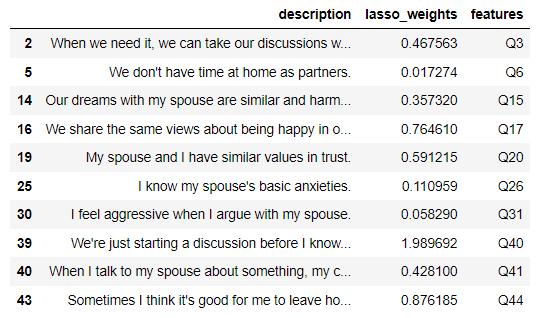
* We can infer that most people both married and divorced disagree with Q7, Q6 though everyone who agreed with the question 7 is divorced.
* Married couples’ answers varied across Q43, Q45, Q46.
* Q50 & Q51 more divorced people blamed the other spouse for problems while most married people said both parties are equally at fault.
* Q48 most married people also agreed that both parties are equal about who's right during discussions while divorced felt they were the right ones.

**1.6. Model training and evolution and validation:**

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables. With a classification problem, a base Logistic regression model was trained with 30% of the data was reserved for the testing, giving a recall value of 91.7% - which corresponds to no False Negatives (FN) and 2 False Positives (FP) and achieving a test accuracy of 95.6%.

*Figure 5: From our confusion matrix there are 2 false positives and no false negatives and we have an accuracy of 95.6%*

We performed feature selection using Lasso regularization (L1) to determine our features for prediction and we got 10 features whose weights were not reduced to 0.



*Figure 6: Feature Selection using L1 Regularization, only ten of our features were not reduced to 0*

Therefore, we re-trained and re-evaluated the model using features with non-zero weights. Figures depicted that the metrics of the regularised model didn’t change from our original model. This may be due to the cause of our dataset being linear.

Although, this resulted in the percentage of test accuracy being 95.6%, the goodness of fit of our model being 82.1% and the precision of our model is 100%.

**1.7 Model Deployment:**

Application was successfully built with a good model ready, Streamlit - an open-source app Framework, was used to deploy the model on Streamlit Cloud in order to make live predictions. Here is the link - <https://ababayato-divorce-prediction-app-app-rxf1jk.streamlitapp.com/>

**1.8 Conclusion:**

Against my better judgement more divorced people agreed with questions 10-30 probably means these where negative questions, thinking a positive answer meant a negative feeling; Otherwise, so there's a problem with the dataset considering positive answers to these questions state a good marriage and mainly divorcees had a positive answer to these questions.

**1.9 Future prospects:**

People can use this application and have a small test before any consultations. Hopefully this application can be developed further to support.

The most effective features and their values of significance obtained by applying correlation-based feature selection method on dataset, can be used by family counsellors and therapist to either provide their services remotely, or to devise a plan and prepare a case. Articles for better relationship management can also be suggested.

In the stages where there is nothing left to improve upon, divorce is inevitable. But, with more extensive work over this application can actually save a smiling family from parting ways.

**1.10 Contributing Members**

* Abhishek Man Tuladhar
* Adedamola Babayato
* Akindele Bola
* Damilare Adedoyin
* David Ekpo
* Flora Oladipupo
* Katherine David
* Oghenero Egbine
* Olayinka Tejumola
* Sakshi Jha
* Tobenna Okeke