Final Report: Relationship Outcome Modeling

By: Nick Ambrose and Killian Brait

Introduction

Marriage and divorce represent some of the most profound and complex aspects of human life. With global divorce rates on the rise, it is increasingly important to understand the factors contributing to this trend. The repercussions of divorce are not limited to the personal realm; they echo through social and economic landscapes, making this a critical area of study.

In this project, we leverage Data Science methods to delve into this intricate area. Our objective is to develop a predictive model that can estimate the likelihood of a couple's divorce based on their responses to specific questions. The ability to predict divorce carries significant implications. A successful prediction model can be used in couples therapy to identify areas of conflict and address them proactively. It can also contribute significantly to social science research by aiding in the understanding of societal trends and patterns.

Moreover, understanding the key predictors of divorce can yield practical benefits. This information can be used to help couples fortify their relationships by focusing on potential areas of conflict. It can also guide therapists to concentrate their therapy sessions on critical areas, enhancing the effectiveness of their interventions.

Our analysis is based on a dataset collected from a survey of Romanian couples. The dataset comprises answers to 54 questions related to various facets of the couples' relationships, such as communication, problem-solving, and compatibility. The responses are numerical, indicating the degree to which the respondent agrees or disagrees with the statement in the question. The target variable, "Divorce", is a binary variable indicating whether the couple got divorced (1) or not (0).

In the following sections, we describe the methods used for data preprocessing, correlation analysis, K-Means clustering, sentiment analysis, and K-Nearest Neighbors (KNN) modeling. We present the results of these analyses and discuss their implications. Finally, we summarize the main findings of the study and suggest directions for future research.

This comprehensive analysis allows us to shed light on the intricate factors contributing to divorce and provides valuable insights for therapists, social scientists, and couples alike.

Questions

1. Can we group couples into distinct clusters based on their survey responses and K Means clustering?
2. What are the defining characteristics of clusters in the data and what can we learn about healthy relationships from them?
3. How effective are the K-Nearest Neighbors (KNN) and K-Means models in predicting divorce, and how do our custom implementations compare with scikit-learn's implementations in terms of accuracy and execution time?
4. What are the key predictors of divorce, as indicated by the strongest correlations between features and the target variable?
5. Does a model built on this data extrapolate well to new data?
6. What practical applications and implications do the findings have for therapists, social scientists, and couples?

Methods

In our analysis, we employed a series of data preprocessing, exploratory, and predictive modeling techniques to gain insights into the dataset and develop a predictive model for divorce. Here is a detailed description of the methods used:

1. Data Preprocessing:

   The data preprocessing phase involved scaling the features to standardize their values. We used the standard scaler method, which transforms the features to have a mean of 0 and a variance of 1. This step was particularly important since we limited the number of features used in our KNN classifier from 54 to just 5, and we wanted to ensure all 5 features contributed equally to the model.

2. Correlation Analysis:

   To understand the relationships between the different features in our dataset, we conducted a correlation analysis. The correlation coefficient between each pair of features was calculated, resulting in a correlation matrix. This matrix was then visualized using a heatmap, which provided a color-coded representation of the relationships between the features. (results shown below)

3. K-Means Clustering:

   We used the K-Means clustering algorithm to identify clusters in the data. The optimal number of clusters was determined using the elbow method, which involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the "elbow" point where the decrease in WCSS becomes less pronounced. This allowed us to find the minimal number of clusters where the model still performs relatively well. Once the ideal number of clusters was identified, we trained SciKit Learn's and a custom KMeans model on the divorce dataset. We then analyzed the distribution of the clusters and calculated the mean values of the features for each cluster.

4. K-Nearest Neighbors (KNN) Modeling with Feature Reduction:

   We used the KNN algorithm to develop a predictive model for divorce. The number of features used in the model was optimized using five fold cross-validation, which involves partitioning the data into subsets, training the model on K - 1 subsets, and validating it on the remaining one. We implemented this with SciKit Learn's cross_val_score function. This process was repeated for different numbers of features to identify the optimal number. Once the ideal number of features was determined, we trained SciKit Learn's and our custom

KNN model on the divorce dataset to create a classifier based on the ideal number of features.

5. Sentiment Analysis:

   Sentiment analysis was performed on the text of the questions to gain insights into the emotional tone of the questions. This analysis was performed using textblob's natural language processing model's sentiment analysis functionality. However, this did not factor heavily in the findings of this report.

6. Correlation with Target Variable:

   We calculated the Pearson correlation coefficients between the features and the target variable to identify the questions that were most strongly related to divorce. The questions with the highest, middle, and lowest correlations were identified. The five highest questions were used to train the KNN model.

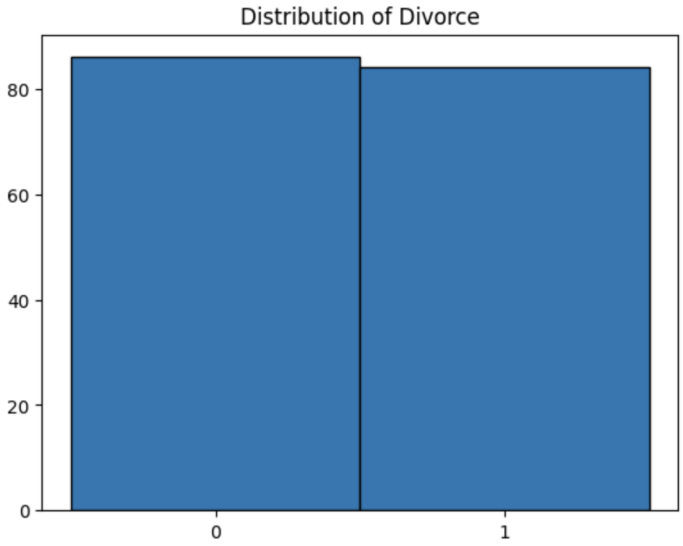
Each of these methods provided unique insights into the data and contributed to the development of our predictive model. The combination of these methods enabled us to thoroughly analyze the data and identify the key predictors of divorce.

Results/Discussion

The results of our analysis are organized according to the methods described in the previous section. The outcomes of each step in the analysis are presented below, along with some discussion of what they mean:

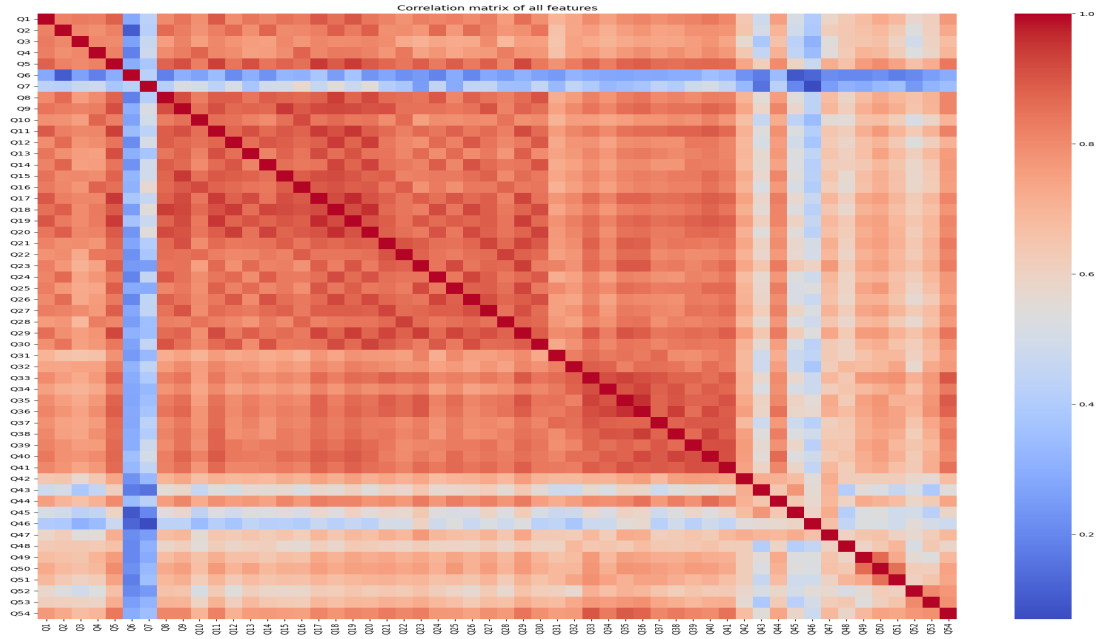
1. Initial Exploration:

   We began our analysis by exploring the distribution of the target variable, Divorce. This showed that we have a balanced dataset: there is a similar proportion of couples who got divorced versus those who did not. This means the model will not be biased towards one class



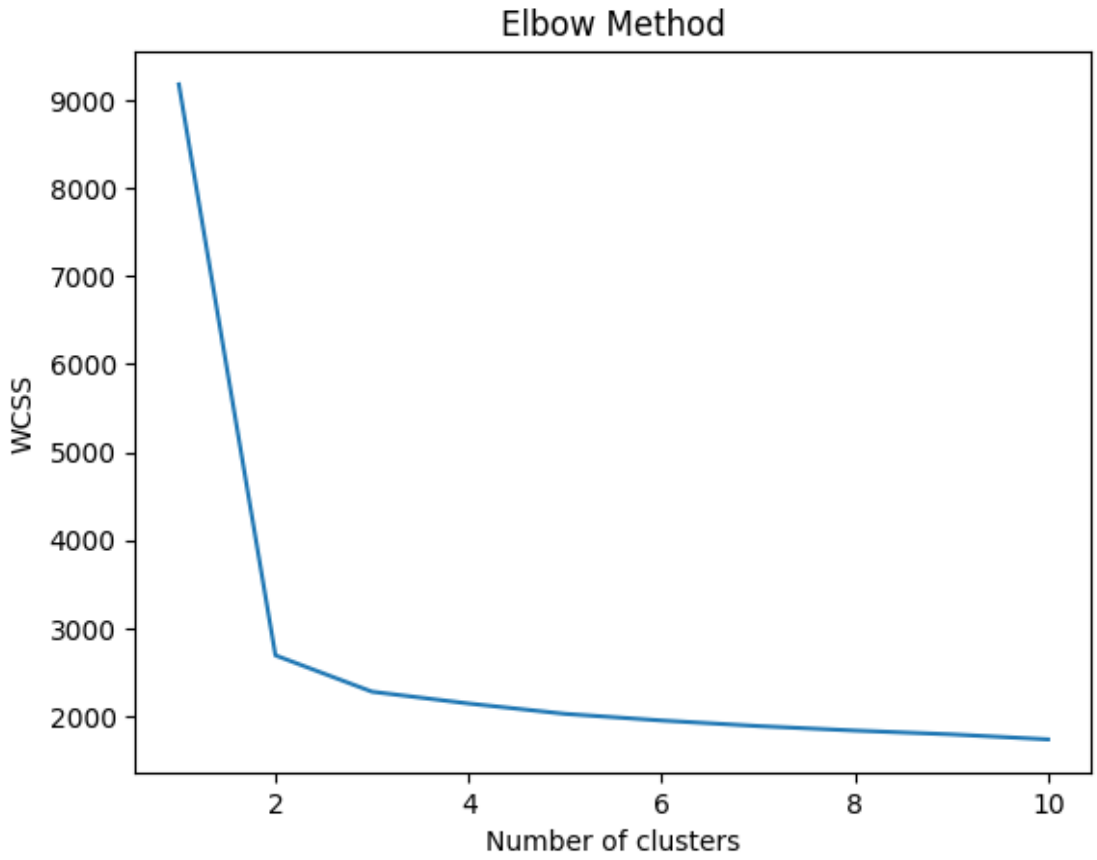
2. Correlation Analysis:

   The correlation heatmap gave us an overview of the relationships between the features in our dataset. This visualization revealed many features are highly correlated with each other, indicating we could reduce the number of features in our KNN classifier without losing much predictive power.

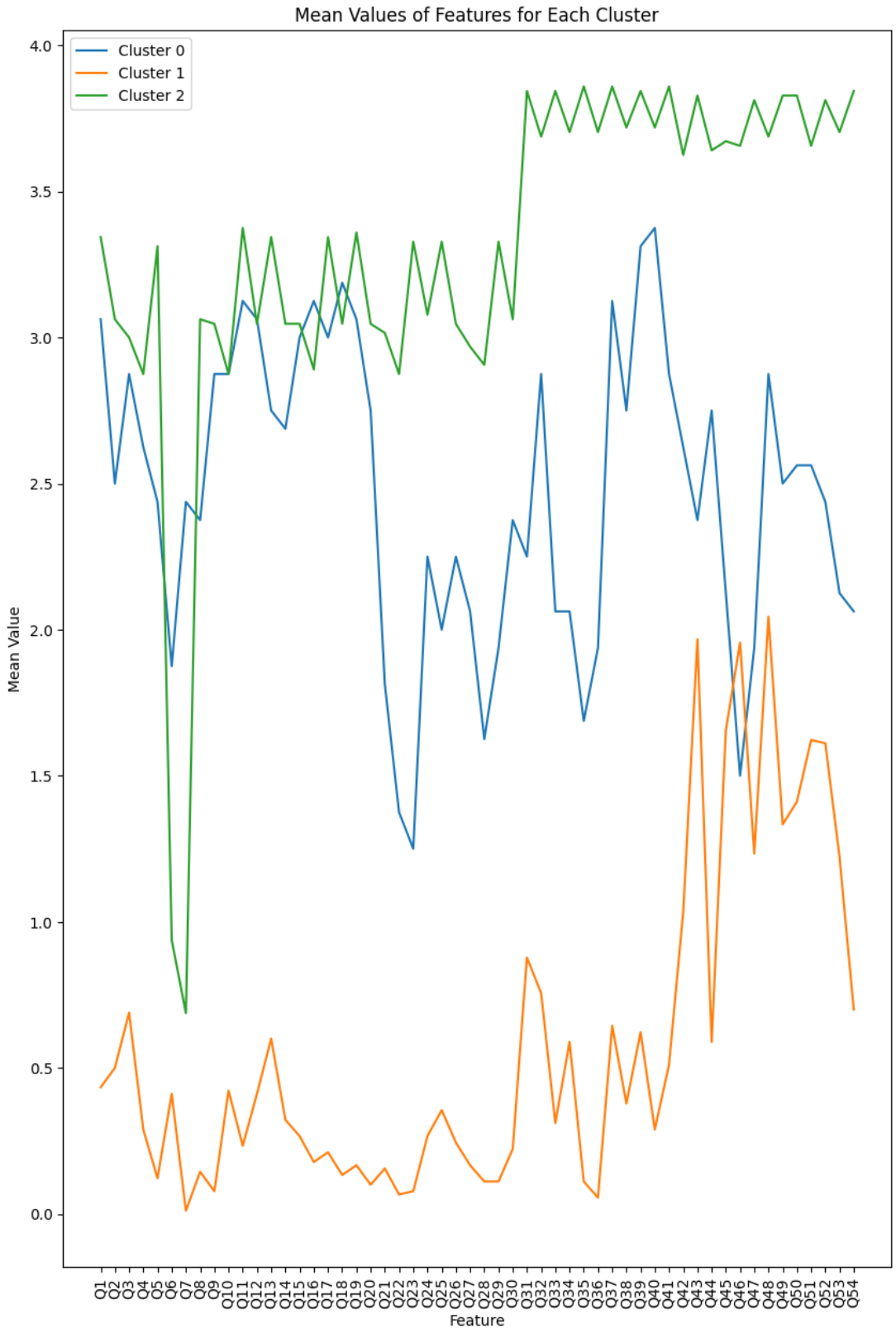


3. K-Means Clustering - Determining the Optimal Number of Clusters:

The elbow method indicated that the optimal number of clusters for our data was three. We weren't able to use our custom KMeans model for this because we did not implement a way to calculate the variance within clusters of the model. However, the SciKit Learn model allowed us to find the point at which the variance within the clusters was at a relatively low point, the "elbow" of a graph looking at the variance (WCSS) within each cluster.



We then ran our K-Means algorithm using the optimal number of clusters. Plotting the mean feature values for each cluster gave us informative results, as you can see below.

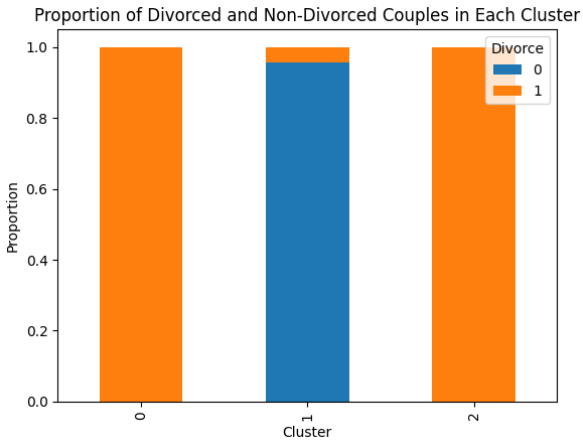

Mean Values of Features for Each Cluster
Cluster 0
Cluster 1
Cluster 2
Mean Value
Feature
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16 Q17 Q18 Q19 Q20 Q21 Q22 Q23 Q24 Q25 Q26 Q27 Q28 Q29 Q30 Q31 Q32 Q33 Q34 Q35 Q36 Q37 Q38 Q39 Q40 Q41 Q42 Q43 Q44 Q45 Q46 Q47 Q48 Q49 Q50 Q51 Q52 Q53 Q54

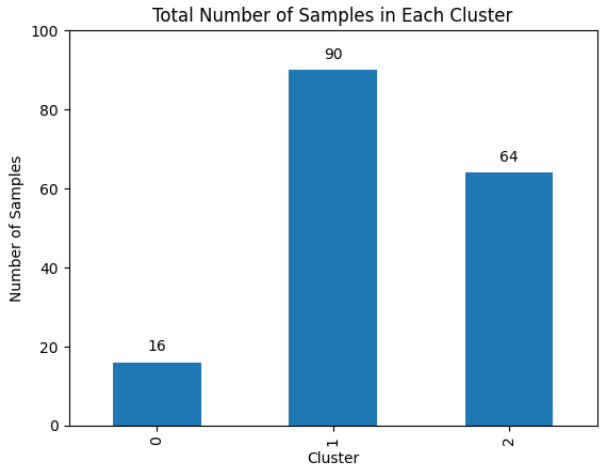
Here's how we interpreted this information:

- Cluster 2: This cluster has significantly higher averages than the other clusters for questions 21-54, showing that respondents in this cluster were far more likely to answer 3s or 4s on the questions, indicating that they "Frequently" or "Always" agreed with the question. Although, it should be noted that for the initial questions 1-20, this cluster had very similar averages to cluster 0.
- Cluster 1: This cluster had significantly lower averages for almost all the questions, showing that respondents in this cluster were far more likely to answer 0s or 1s on the questions, indicating that they "Never" or "Rarely" agreed with the question. Although, it should be noted that for question 46 this cluster had a higher average than cluster 0.
- Cluster 0: This cluster is the median group. They had averages that typically fell in between the other two clusters. In this cluster, respondents were likely to answer 1s, 2s, or 3s on the questions, indicating that they "Seldom", "Averagely", or "Frequently" agreed with the question. It should be noted that for questions 6, 7, 10, 12, 16, and 18 this cluster had a higher average than cluster 2, indicating for these questions this cluster had the highest average.

Without even looking at the questions to see what was asked, we can conclude that cluster 2 was the most likely to answer extremely in the affirmative, cluster 1 was most likely to answer extremely in the negative, and cluster 0 was the median group.

However, we still had to see how this was applicable to a lasting marriage versus one that ends in divorce.





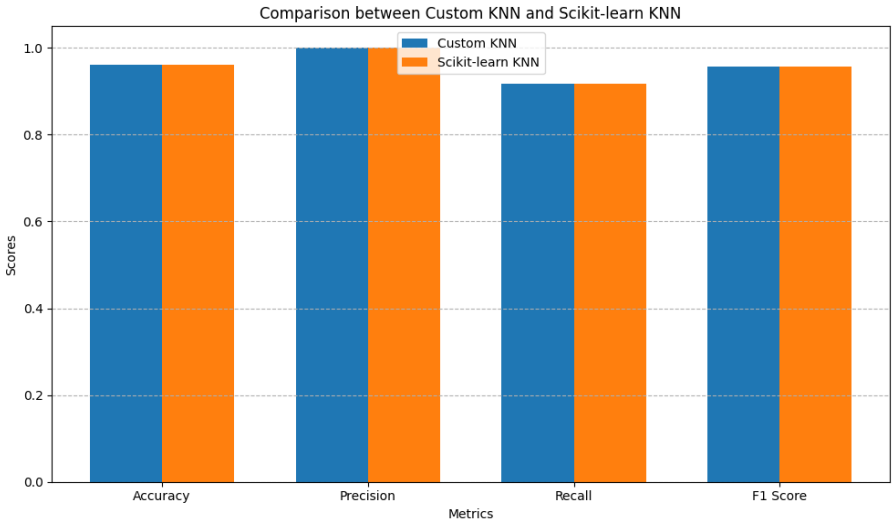
Our results show that both cluster's 0 and 2 have a 100% divorce rate. Looking back we can see that this may mean that if you answer the questions in the affirmative most of the time, then you are likely in a relationship that will end in divorce. However, this is based on averages, so there are likely some people that answered a few questions very differently than the rest of the group that this graph does not capture.

It's also important to take note of the value counts in each cluster, shown by the second graph above. The dataset is still balanced, even though two cluster's have a 100% divorce rate, they are smaller clusters. Taking this into account, we can likely ignore the predictive power of cluster 2 since it has only 16 samples, or 25% of cluster 0's size. Additionally, cluster 2 is our median group, so it's small and the median, meaning it's likely just a random group of samples that were outliers in the other two clusters.

This leads to a clear conclusion, if you answer the questions in the affirmative most of the time, then you are likely in a relationship that will end in divorce. However, if you answer the questions in the negative most of the time, then you are likely in a relationship that will not end in divorce. This is a very interesting result, and it's likely that this is a strong predictor of divorce.

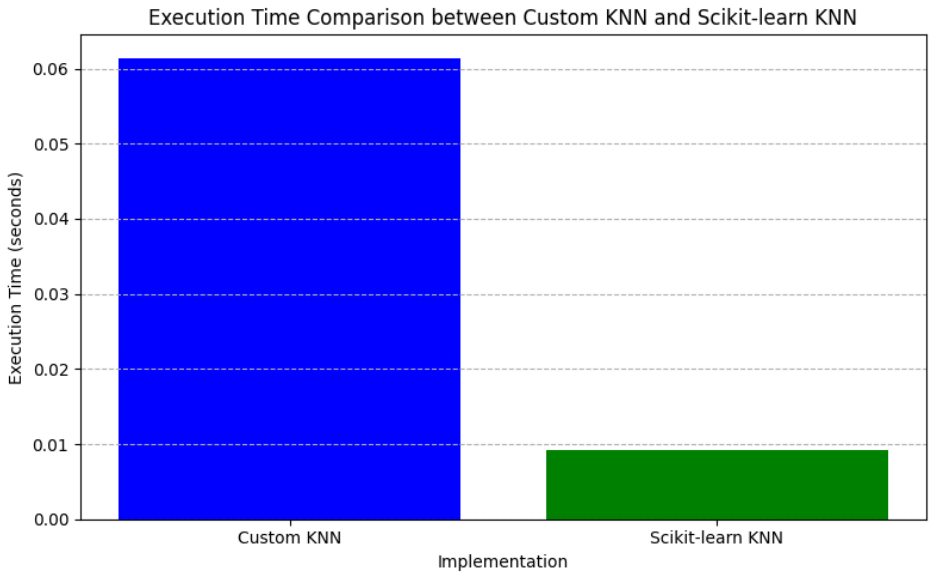
4. K-Nearest Neighbors (KNN) Modeling:

Both the SciKit-learn and from scratch KNN models with the optimal number of features achieved the same performance. The confusion matrix provided a detailed view of the model's performance, showing the number of true positive, true negative, false positive, and false negative predictions.




```
Confusion Matrix:
 [[27  0]
 [ 2 22]]
```

Let’s compare the speed at which each algorithm runs:



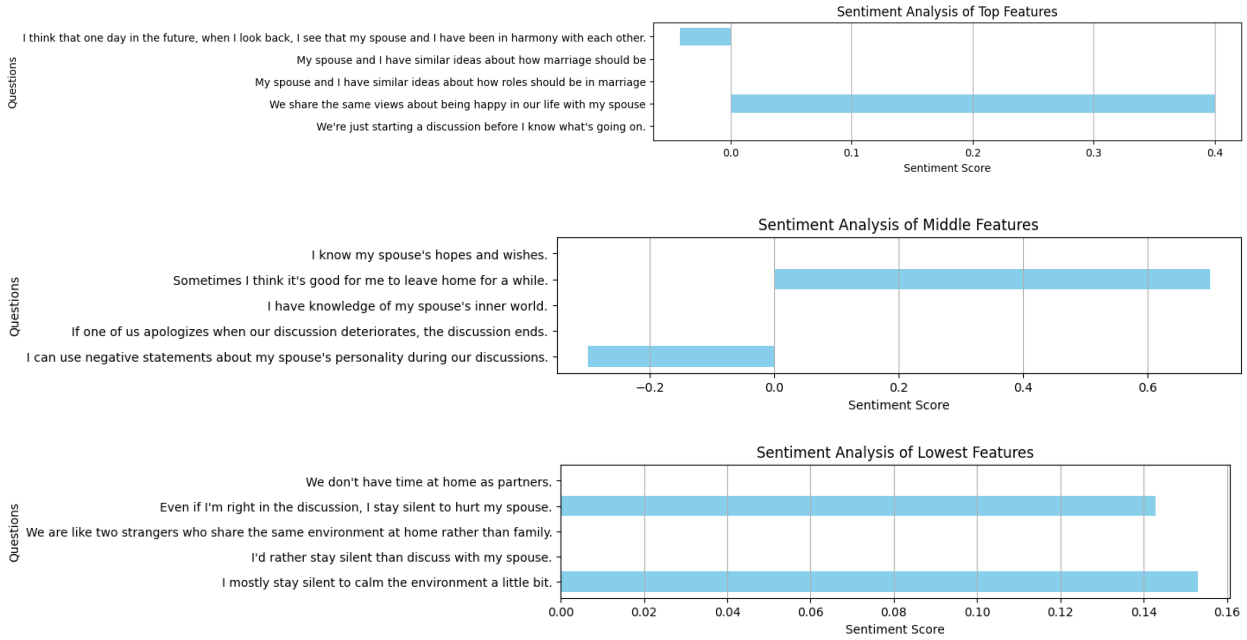
The bar chart clearly shows that scikit-learn's KNN implementation is faster than the custom implementation. This difference in execution time is not surprising, as scikit-learn is a highly optimized library that leverages efficient algorithms and data structures. For example, it leverages data structures like Ball Trees or KD Trees to quickly find the nearest neighbors, instead of computing pairwise distances between all points. For example, parts of scikit-learn are implemented in Cython, which runs much faster than interpreted Python code.

5. Correlation with Target Variable:

We identified the features (questions) that had the highest, middle, and lowest correlation with the target variable. The five questions with the highest, middle, and lowest correlations were singled out for sentiment analysis.

```
Top Features Questions:
        1. We're just starting a discussion before I know what's going on. (Correlation: 0.94)
        2. We share the same views about being happy in our life with my spouse (Correlation: 0.93)
        3. My spouse and I have similar ideas about how roles should be in marriage (Correlation: 0.93)
        4. My spouse and I have similar ideas about how marriage should be (Correlation: 0.92)
        5. I 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. (Correlation: 0.92)
--------------------------------------------------
Five Lowest Correlated Questions:
        50. I mostly stay silent to calm the environment a little bit. (Correlation: 0.57)
        51. I'd rather stay silent than discuss with my spouse. (Correlation: 0.55)
        52. We are like two strangers who share the same environment at home rather than family. (Correlation: 0.54)
        53. Even if I'm right in the discussion, I stay silent to hurt my spouse. (Correlation: 0.44)
        54. We don't have time at home as partners. (Correlation: 0.42)
--------------------------------------------------
Five Middle Correlated Questions:
        25. I can use negative statements about my spouse's personality during our discussions. (Correlation: 0.86)
        26. If one of us apologizes when our discussion deteriorates, the discussion ends. (Correlation: 0.86)
        27. I have knowledge of my spouse's inner world. (Correlation: 0.86)
        28. Sometimes I think it's good for me to leave home for a while. (Correlation: 0.85)
        29. I know my spouse's hopes and wishes. (Correlation: 0.85)
```

6. Sentiment Analysis:



As we can see, most of the questions have a sentiment score of 0.00, which means they are neutral. However, we think the sentiment analysis is placing too much weight on individual words since the sentences used in the sentiment analysis are rather short and concise. Even the ones that register some significant sentiment (>0.5) are neutral when you read them. We think this is a limitation of the sentiment analysis tool used and that it doesn't provide any useful information to bolster our results.

Overall, these results provided a multifaceted view of the data and the factors influencing divorce. The visualizations and statistics generated in our analysis shed light on the relationships between the features, the characteristics of the clusters, and the performance of the predictive models.

Conclusion

In this study, we explored the dynamics of marriage and divorce through a comprehensive analysis of survey responses from Romanian couples. Utilizing correlation, K-Means clustering, sentiment analysis, and predictive modeling, we uncovered significant insights into the factors influencing divorce.

Key findings include the identification of patterns in couples' relationships through correlation and clustering analyses, which may be useful to tailoring marriage therapy.

Specifically, the identification of a cluster that doesn't agree with the statements in the questions being far less likely to divorce, indicates that the statement in the questions reflect meaningful issues that are likely to plague a relationship and result in divorce if they exist.

Language barriers due to the translation of the questions from Romanian made it difficult to solidify any conclusions about the clusters. This was why rudimentary sentiment analysis was performed. However, this didn't provide any knowledge beyond what we'd already learned from looking at the clusters.

Nevertheless, clustering worked to create some great knowledge that would be worth further exploration. Specifically, are there some key things that don't happen in healthy relationships that almost inevitably lead to divorce? Some statements that people in a healthy relationship would never agree with? Based on our KMeans model, this report proposes there might be.

The K-Nearest Neighbors (KNN) model demonstrated good performance in predicting divorce, with an accuracy score of 96%. The confusion matrix further highlighted the model's ability to correctly classify divorced and non-divorced couples. This predictive capability is critical for identifying couples at risk of divorce and proactively addressing their issues.

These insights have practical applications for therapists and couples, providing guidance on areas to focus on for relationship strengthening. The identified key predictors of divorce can inform targeted interventions and help couples navigate their relationships more effectively.

The project also highlights the power of data science in understanding complex human phenomena like marriage and divorce. While we have achieved significant insights, opportunities for future research remain, including exploring other predictive models like random forest, increasing the dataset to train on more samples and more features such as the demographics of the couple, and an analysis of the questions themselves to better understand what the clusters we found illustrate.

In conclusion, our study not only develops a predictive tool for estimating the likelihood of divorce but also provides meaningful insights into the relationships' dynamics. By combining various data analysis techniques, we have created a valuable resource for social scientists, therapists, and couples, demonstrating the potential of data science to inform and enhance our understanding of human relationships.