JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

OPEN SOURCE AND SYSTEM PROGRAMMING

Project Report

Group Members:

Name	Enrollment no.
KHUSHI KALRA	9920103025
AVIRAL GUPTA	9920103021
DEVANSH CHUGH	9920103011

UNDER THE GUIDANCE OF:

Ms. Anuradha Gupta

Ms. Anubhuti Roda Mohindra

Ms. Chetna Gupta

Abstract

Divorce usually impacts the closest family members, over the years the divorce rate has increased dramatically, especially in the last two decades and worsening with the pandemic, where there has been a significant increase in the divorce rate in many countries of the world. In addition, we make use an automatic learning models (logistic regression) and 3 hybrid models based on voting criteria. Divorce is a major problem, especially in a context of confinement, where the rates of divorced couples have increased considerably, indirectly affecting the closest members of the family (such as children). Also, couples can lose a lot by going through a divorce process. This study can help them prevent these consequences. The prediction models in this study would help people decide whether to make the decision to marry or not, give them the opportunity based on compatibility to have a successful marriage.

Problem Statement

A divorce is a legal step taken by married people to end their marriage. It occurs after a couple decides to no longer live together as husband and wife. Globally, the divorce rate has more than doubled from 1970 until 2008, with divorces per 1,000 married people rising from 2.6 to 5.5. Divorce occurs at a rate of 16.9 per 1,000 married women.

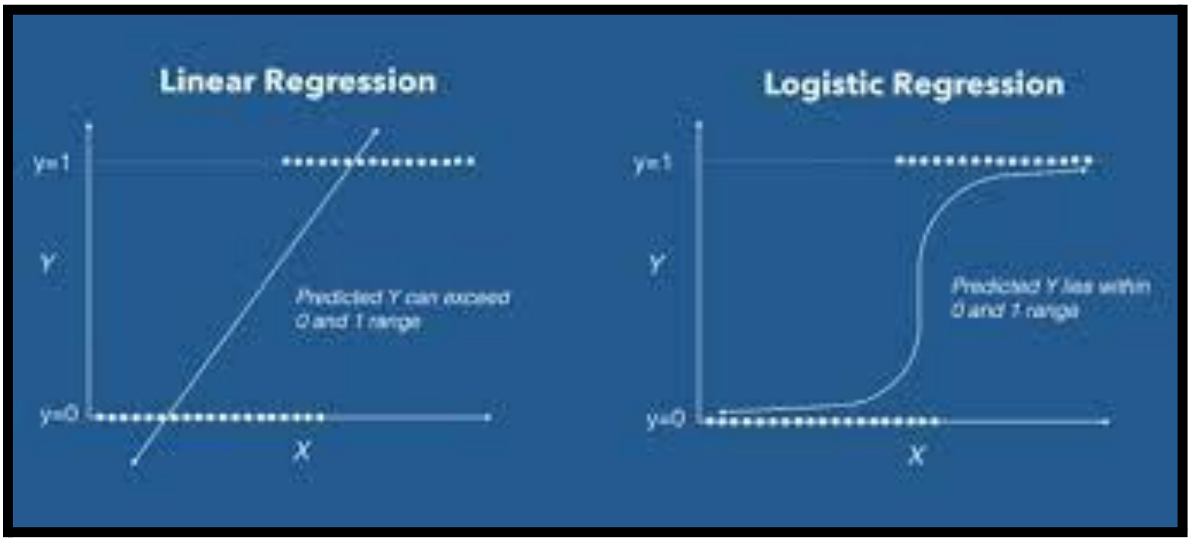
Given survey data from couples in Turkey, we are trying to predict if a given couple is divorced.We will use a logistic regression model to make our predictions. We will use principal component analysis to reduce the dimension of the data and show that the same results can be achieved with a smaller number of features, as well as to visualise the data.

Model used

- Model used

Logistic Regression :

Logistic model (or logit model) is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic mode.



Requirement Analysis

➢ **Software requirements**

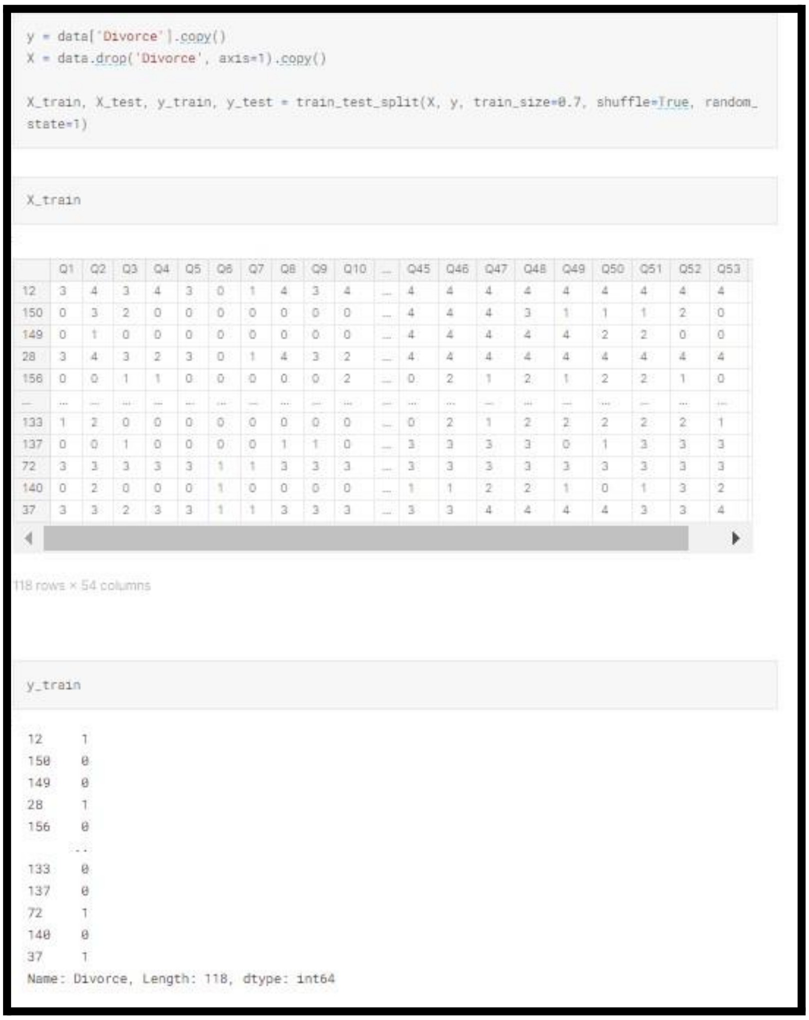
- Anaconda environment
- Jupyter Notebook
- Chrome browser

➢ **Hardware requirements**

- OS: Windows 7 or above (64-bit version)
- RAM: min 8 GB

- 2 GB min space
- Min 2 GB GPU

➢ **Language used: Python**



```
y = data['Divorce'].copy()
X = data.drop('Divorce', axis=1).copy()

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, shuffle=True, random_
state=1)
```

```
X_train
```

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q45	Q46	Q47	Q48	Q49	Q50	Q51	Q52	Q53
12	3	4	3	4	3	0	1	4	3	4	...	4	4	4	4	4	4	4	4	4
150	0	3	2	0	0	0	0	0	0	0	...	4	4	4	3	1	1	1	2	0
149	0	1	0	0	0	0	0	0	0	0	...	4	4	4	4	4	2	2	0	0
28	3	4	3	2	3	0	1	4	3	2	...	4	4	4	4	4	4	4	4	4
156	0	0	1	1	0	0	0	0	0	2	...	0	2	1	2	1	2	2	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
133	1	2	0	0	0	0	0	0	0	0	...	0	2	1	2	2	2	2	2	1
137	0	0	1	0	0	0	0	1	1	0	...	3	3	3	3	0	1	3	3	3
72	3	3	3	3	3	1	1	3	3	3	...	3	3	3	3	3	3	3	3	3
140	0	2	0	0	0	1	0	0	0	0	...	1	1	2	2	1	0	1	3	2
37	3	3	2	3	3	1	1	3	3	3	...	3	3	4	4	4	4	3	3	4

118 rows × 54 columns

```
y_train
```

```
12     1
150    0
149    0
28     1
156    0
      ..
133    0
137    0
72     1
140    0
37     1
Name: Divorce, Length: 118, dtype: int64
```

DETAILED DESIGN

The methodology for analyzing the divorces data collection, Training the model, Dimensionality Reduction and visualizes data.

1. DATA COLLECTION:

I have used the [Divorce Data Set from Kaggle]

This dataset contains data about 150 couples with their corresponding Divorce Predictors Scale variables (DPS) on the basis of Gottman couples therapy.

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.

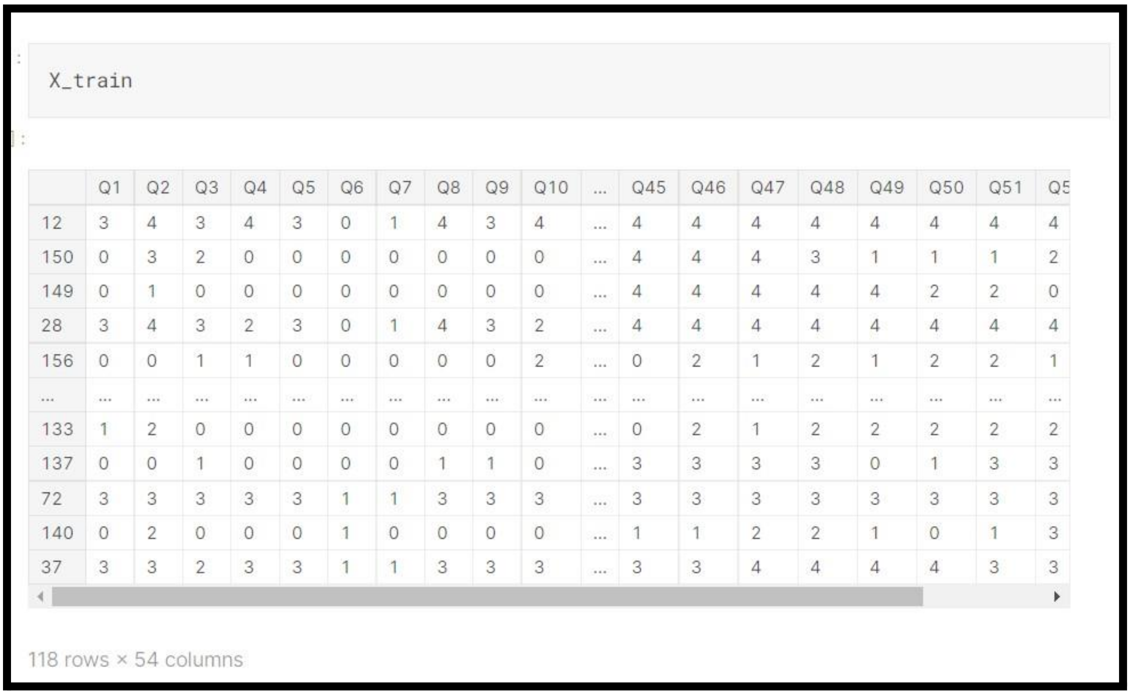
All responses were collected on a 5 point scale (0=Never, 1=Seldom, 2=Averagely, 3=Frequently, 4=Always).

2. TRAINING MODEL:

Training a model simply means learning (determining) good values for all the weights and the bias from labelled examples. In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimises loss; this process is called empirical risk minimization.

3.DIMENSIONALITY REDUCTION

Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.



```
X_train
```

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q45	Q46	Q47	Q48	Q49	Q50	Q51	Q5
12	3	4	3	4	3	0	1	4	3	4	...	4	4	4	4	4	4	4	4
150	0	3	2	0	0	0	0	0	0	0	...	4	4	4	3	1	1	1	2
149	0	1	0	0	0	0	0	0	0	0	...	4	4	4	4	4	2	2	0
28	3	4	3	2	3	0	1	4	3	2	...	4	4	4	4	4	4	4	4
156	0	0	1	1	0	0	0	0	0	2	...	0	2	1	2	1	2	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
133	1	2	0	0	0	0	0	0	0	0	...	0	2	1	2	2	2	2	2
137	0	0	1	0	0	0	0	1	1	0	...	3	3	3	3	0	1	3	3
72	3	3	3	3	3	1	1	3	3	3	...	3	3	3	3	3	3	3	3
140	0	2	0	0	0	1	0	0	0	0	...	1	1	2	2	1	0	1	3
37	3	3	2	3	3	1	1	3	3	3	...	3	3	4	4	4	4	3	3

118 rows × 54 columns

4.VISUALIZATION TECHNIQUE

Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualisations in Python. Matplotlib makes easy things easy and hard things possible.

- · Create publication quality plots.
- · Make interactive figures that can zoom, pan, update.
- · Customise visual style and layout.

- · Export to many file formats.
- · Embedded in JupyterLab and Graphical User Interfaces.
- · Use a rich array of third-party packages built on Matplotlib.

Seaborn

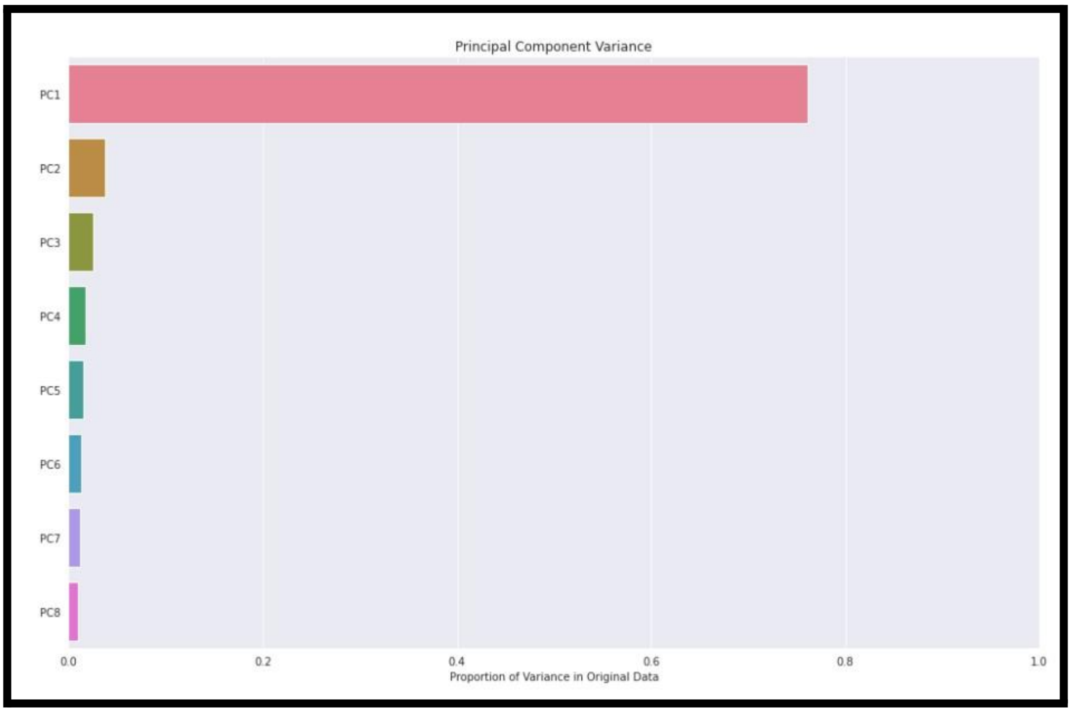
- Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.
- Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

Pandas

- pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

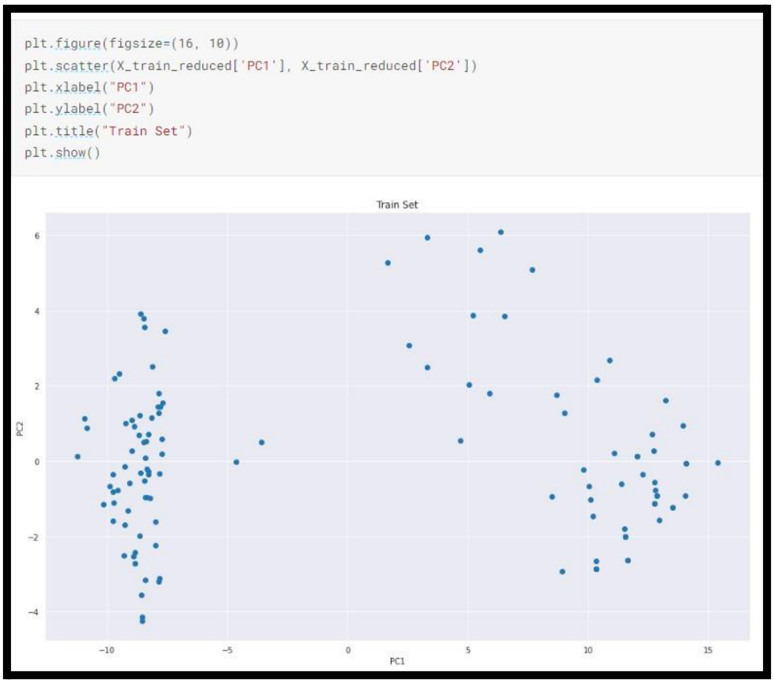
RESULT AND ANALYSIS

- Barplot of Proportion of Variance in Original Data

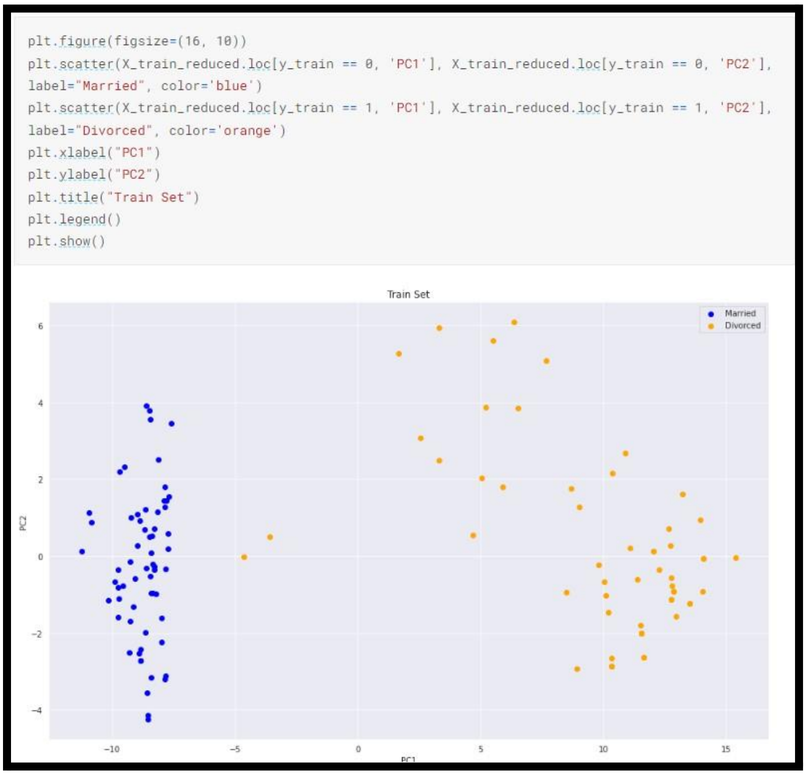


- SCATTERPLOTS

Training dataset

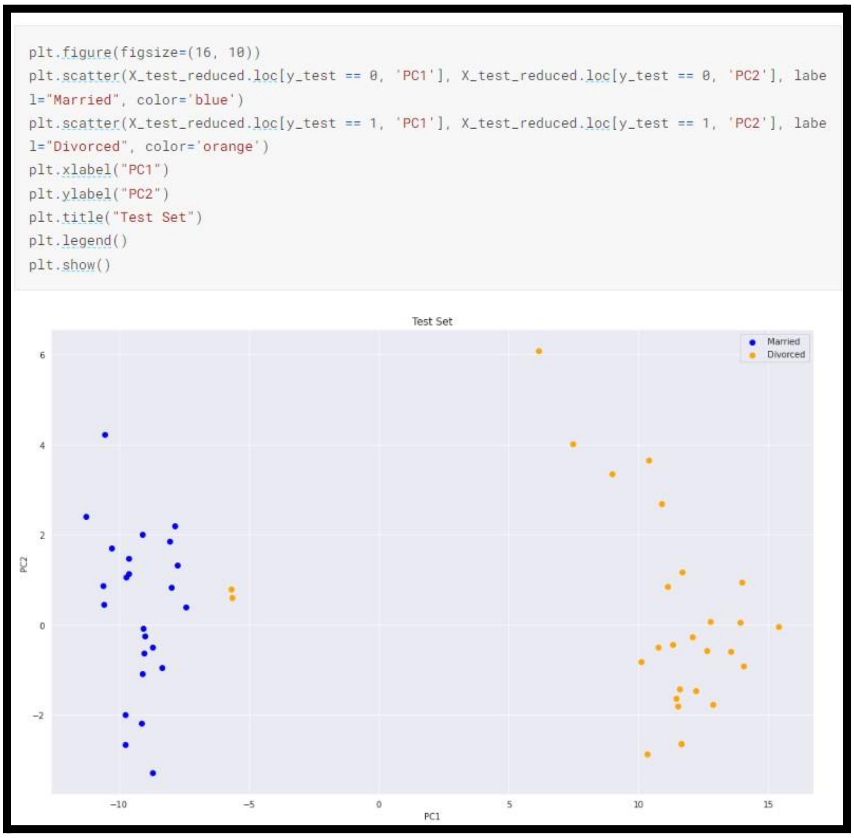


```
plt.figure(figsize=(16, 10))
plt.scatter(X_train_reduced['PC1'], X_train_reduced['PC2'])
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.title("Train Set")
plt.show()
```

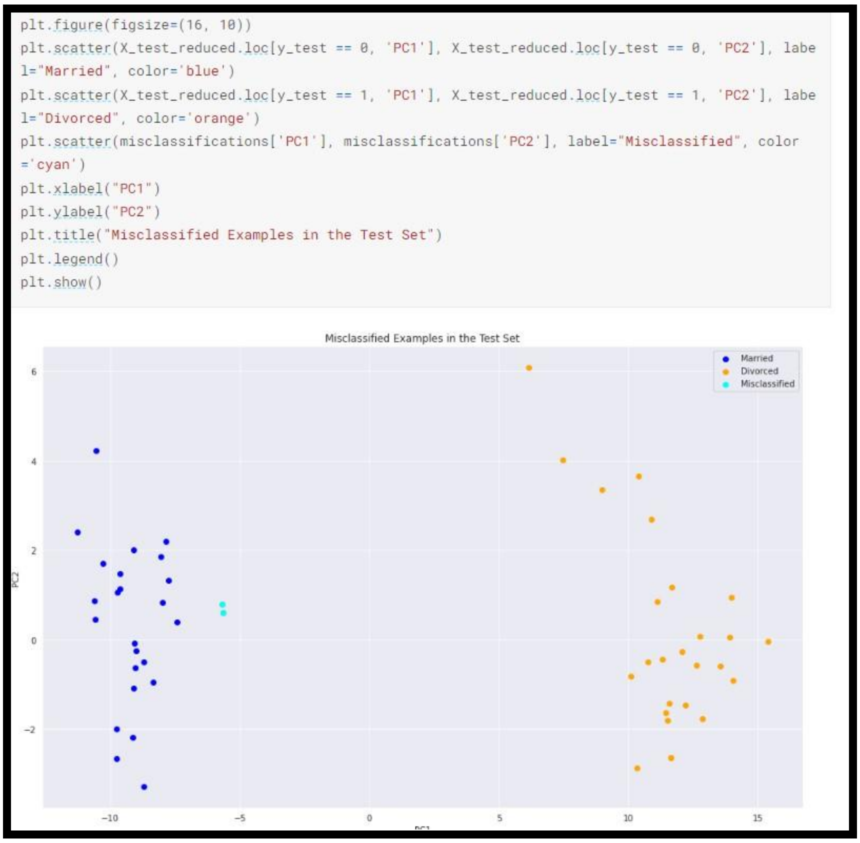


```
plt.figure(figsize=(16, 10))
plt.scatter(X_train_reduced.loc[y_train == 0, 'PC1'], X_train_reduced.loc[y_train == 0, 'PC2'],
label="Married", color='blue')
plt.scatter(X_train_reduced.loc[y_train == 1, 'PC1'], X_train_reduced.loc[y_train == 1, 'PC2'],
label="Divorced", color='orange')
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.title("Train Set")
plt.legend()
plt.show()
```

TESTING DATASET:

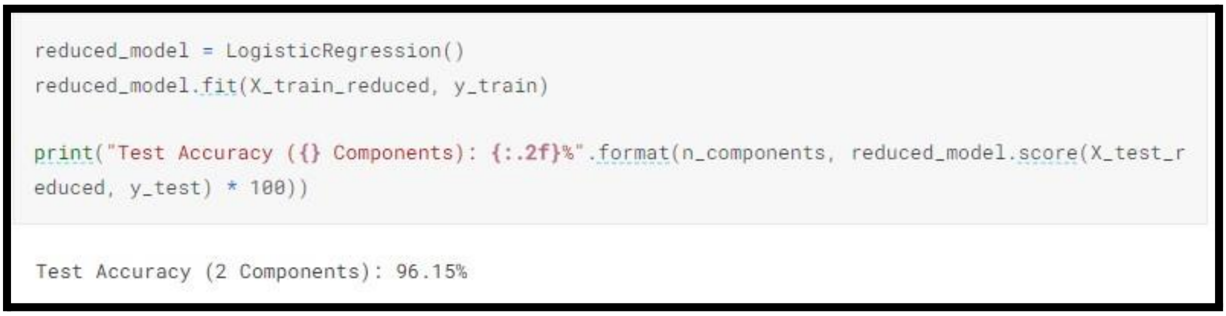


```
plt.figure(figsize=(16, 10))
plt.scatter(X_test_reduced.loc[y_test == 0, 'PC1'], X_test_reduced.loc[y_test == 0, 'PC2'], labe
l="Married", color='blue')
plt.scatter(X_test_reduced.loc[y_test == 1, 'PC1'], X_test_reduced.loc[y_test == 1, 'PC2'], labe
l="Divorced", color='orange')
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.title("Test Set")
plt.legend()
plt.show()
```



```
plt.figure(figsize=(16, 10))
plt.scatter(X_test_reduced.loc[y_test == 0, 'PC1'], X_test_reduced.loc[y_test == 0, 'PC2'], labe
l="Married", color='blue')
plt.scatter(X_test_reduced.loc[y_test == 1, 'PC1'], X_test_reduced.loc[y_test == 1, 'PC2'], labe
l="Divorced", color='orange')
plt.scatter(misclassifications['PC1'], misclassifications['PC2'], label="Misclassified", color
='cyan')
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.title("Misclassified Examples in the Test Set")
plt.legend()
plt.show()
```

Accuracy of model



```
reduced_model = LogisticRegression()
reduced_model.fit(X_train_reduced, y_train)

print("Test Accuracy ({} Components): {:.2f}%".format(n_components, reduced_model.score(X_test_r
educed, y_test) * 100))
```

```
Test Accuracy (2 Components): 96.15%
```

Conclusion

This project examines the fundamental reasons underlying the evolution of the divorce probability over the course of a marriage. Although learning about marital quality has been often proposed as an explanation for the divorce hazard, this mechanism finds limited support in the data. On the other hand, the divorce hazard can be fully explained by the assumption that the marital quality follows a random walk. In other terms, divorce can be fully explained by real changes in relationship quality. In this project we have concluded that the best performance of a model was obtained by using the 60/40 ratio of the training and test data.

The results were 0.9853 precision, 1.0 sensitivity and 0.9667 specificity.

Future Scope

In order to feed the dataset to retrain the models, in the future we plan to collect couple's data from different countries, evaluating their performance across the globe.The logistic regression has a lot of applications, but it lends itself strongly to analysing survey data and classifying subjects into 2 categories based on things such as Age, Income, Location, etc. This could be very useful in analysing political data, and media leanings.

References

1. Amiriparian, S., Awad, A., Gerczuk, M., Stappen, L., Baird, A., Ottl, S., & Schuller, B. (2019). Audio-based Recognition of Bipolar Disorder Utilising Capsule Networks. IEEE Xplore. https://doi.org/10.1109/ijcnn.2019.8852330
2. Betzig, L. (1989). Causes of Conjugal Dissolution: A Cross-cultural Study. Current Anthropology, 654-676.4
3. Carneiro, T., Medeiros, R., & Nepomuceno, T. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. IEEE, 9. https://doi.org/10.1109/ACCESS.2018.2874767
4. E. San Diego, "Divorce Statistics and Facts | what Affects Divorce Rates in the U.S.?" 2022, https://www.wf-lawyers.com/divorce-statistics-and-facts/.
   View at: Google Scholar

5. P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, “natural language processing: an introduction,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011.
   View at: Publisher Site | Google Scholar
6. N. Flores, S. Silva, C. Science, S. Silva, A. I. Group, and C. Science, “Machine learning model to predict the divorce OF a married couple,” *3C Tecnología_Glosas de innovación aplicadas a la pyme*, pp. 83–95, 2021.
   View at: Publisher Site | Google Scholar