**Assignment 7 – Project Proposal  
Project Team 1: Jared Anderson, Andrew Oveson, Vishal Yelisetti  
25 March 2025**

The dataset that we chose is the “Marriage Trends in India: Love vs. Arranged” dataset from Kaggle ([Dataset can be found here](https://www.kaggle.com/datasets/ak0212/marriage-trends-in-india-love-vs-arranged)). This dataset has 10,000 rows and 17 columns. The dataset has no missing values. The data has 3 continuous columns and 14 categorical columns. Our target variable is “Divorce\_Status”.

There is a class imbalance in our dataset for the “Divorce\_Status” column. This class imbalance is 9 to 1 (9000 “No” “Divorce\_Status” and 1000 “Yes” “Divorce\_Status”. We plan to try multiple approaches to address this mainly, oversampling, upweighting the minority class, or downsampling and upweighting the majority class.

The dataset as 3 numerical attributes and 14 categorical attributes. Although this is a higher amount of categorical attributes than ideal, 7 are Boolean values variants (Marriage Type, Gender, Caste Match, Rural/Urban, Spouse Working, Inter-Caste, and Inter-Religion), which would make one-hot-encoding fairly simple for the remaining categorical columns; Income Levels, Marital Satisfaction, Dowry Exchanged, and Parental Approval are all tri-option variants. Education Level (4 variants) and Religion (5 variants) are the remaining categorical columns. These should allow for a wide range of models to use.

This dataset appears to be pre-cleaned or otherwise intentionally prepared because there are no missing values across the entire dataset. In the dowry exchanged column, 10% of the data is “Not Disclosed” and the parental approval column has a “Partial” option for 20% of responses in addition to their Yes/No options. The numerical columns do not appear to have any substantial outliers.

The categorical columns are well-distributed for the most part; the highest irregularity is the divorce status column (our planned target), with a 90% False to 10% True distribution, as referenced above. The only other column that has a distribution beyond 70/30 is inter-religion, and four of the binary columns are within 60/40.

The numerical columns provide support to a well-sampled dataset. The years since marriage, children count, and age at marriage attribute histograms are all nearly flat across all buckets. The largest anomaly is in the age at marriage column which has end-buckets that are 50% taller than the others.

We can look at the data from a classification and regression angle. From a classification,

1) Can we predict whether a marriage will end in divorce based on factors like age at marriage, education level, caste match, religion, parental approval, urban/rural setting, dowry exchanged, marital satisfaction, and income level?

2) Can we classify whether a marriage received parental approval based on variables such as marriage type, age at marriage, education level, caste match, religion, urban/rural setting, dowry exchanged, and income level?

From a regression angle:

1) Can we predict the number of years since marriage based on factors like age at marriage, education level, caste match, religion, parental approval, urban/rural setting, dowry exchanged, marital satisfaction, and income level?

2) Can we predict the number of children a couple has based on variables such as marriage type, age at marriage, education level, caste match, religion, parental approval, urban/rural setting, dowry exchanged, marital satisfaction, and income level?

There are some data preparation techniques to be used as the samples are somewhat oversampled, we are planning on using SMOTE. We are also planning on using backward elimination to determine the features intended for use.

From a classification point of view, we are planning to use logistic regression for divorce status or dowry exchanges. We are also planning to use decision trees, Random Forest, SVM’s and Gradient Boost Machines. From a Regression Analysis, we are planning on using linear regression to determine simple metrics like years since marriage. Some other techniques that will be used are Decision Trees and several ensemble methods. We are also planning on using Lasso/Ridge regression to prevent overfitting the samples, tuning the hyperparameter to meet our training data needs.

**Dataset Link:** [**Marriage Trends in India: Love vs. Arranged | Kaggle**](https://www.kaggle.com/datasets/ak0212/marriage-trends-in-india-love-vs-arranged)

**Should we link this alternative dataset?**

Alternative Dataset: [**NFL Team Stats 2002 - Feb. 2025 (ESPN)**](https://www.kaggle.com/datasets/cviaxmiwnptr/nfl-team-stats-20022019-espn) **–** Potential things we could estimate with it: win/loss for the Vikings by game in 2025, betting odd for every game in the 2025 season (this would include win/loss and by how much, obviously), other options could include projecting the outcomes of all 2025 games and building expected end-of-season standings