Using Machine Learning to Predict Indian Marriage Success

SEIS 763 Class Project – Group 1 – Spring 2025

Jared Anderson
School of Engineering
University of St. Thomas
St. Paul, MN, USA
Jared.Anderson@bsci.com

Vishal Yelisetti
School of Engineering
University of St. Thomas
St. Paul, MN, USA
Vishal.Yelisetti@bsci.com

Andrew Oveson
School of Engineering
University of St. Thomas
St. Paul, MN, USA
Andrew.Oveson@bsci.com

ABSTRACT

This research project analyzed the "Marriage Trends in India: Love vs. Arranged" dataset from Kaggle.com using various machine learning techniques. The dataset consists of 10,000 individual instances, each with 17 attributes. The attributes consist of a combination of categorical columns (such as gender, marital satisfaction, and parental approval) and numerical columns (such as age at marriage, children count, and years since marriage). The target variable in this dataset will be the "Divorce Status" column. This column has a significant class imbalance (90% not divorced, 10% divorced) and requires a oversampling via SMOTE. Both regression and classification models were used to understand the optimal factor/model combination. We applied the following models: logistic regression, Kernal PCA, Random Forest Classification, RBF SVM, Polynomial SVM, Decision Tree Classification, Gaussian NB, K Neighbors Classification. Ultimately an ensemble was built of these models using a voting classifier to group the performances of all models. Finally, we used this model to predict the marriage success of a new instance, using data from an individual known by the team. This report reviews that dataset, modeling methodology, experiments conducted (and their results), our interpretations of these, and finally the limitations of the study and further research that could be completed.

CCS CONCEPTS

• **Computing methodologies ➞ Machine learning ➞ Learning paradigms ➞ Supervised learning ➞** Supervised learning by classification; • **Computing methodologies ➞ Machine learning ➞ Machine learning approaches ➞** Feature selection; • **Computing methodologies ➞ Machine learning ➞ Machine learning algorithms ➞** Dimensionality reduction and manifold learning; • **Applied computing ➞** Sociology

KEYWORDS

supervised learning, prediction model, machine learning, binary classification, divorce prediction, feature selection, dimensionality reduction, socio-economic factors, supervised learning, data preprocessing, social data analysis

1 INTRODUCTION

Marriage dynamics in India are heavily influenced by cultural factors such as an arranged vs. love marriage, if a dowry was exchanged, or if the spouses came from the same or different social castes. These dynamics, in addition to many others, allow for machine learning analysis to be performed to identify attribute significance and trends, and to perform predictions on new marriage instances. The Kaggle.com dataset, "Marriage Trends in India: Love vs. Arranged" offers 10,000 instances with 17 attributes identifying the specific dynamics of each marriage. This dataset, consisting of 3 numeric and 14 categorical variables, will be used to explore these dynamics. The research aims to predict divorce status using classification machine learning techniques. To do so, the dataset must first be cleaned and prepared for machine learning, addressing any data quality issues, transforming categorical columns, and balancing the strong class imbalance of 90% not divorced and 10% divorced instances. Once prepared, we will use models such as logistic regression, random forest classification, SVM variants, Gaussian NB, and others to predict a new outcome.

2 DATASET DESCRIPTION

The data set, sourced from Kaggle.com, contains 10,000 instances with 17 attributes and has no missing values. Based on these attributes, there appears to have been basic cleaning or other preparation completed previously. Our target variable, "Divorce Status", contains a major imbalance across the two classes, with 90% of instances being "No", indicating the instance is still married and not divorced. The remaining 10% of instances are "Yes", indicating this instance had gone through a divorce at some point prior.

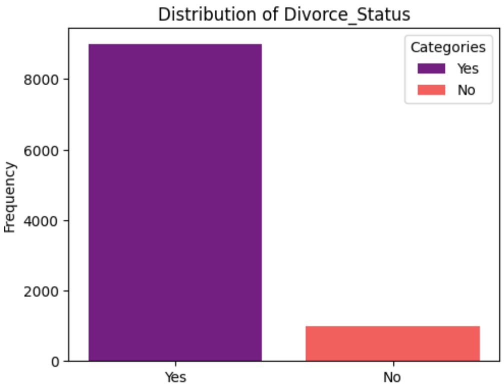


Figure 1 – Distribution of the Target Variable, "Divorce Status"

The attributes for each instance are a mix of numeric and categorical columns. Categorical columns are many attributes, totaling 14 of the 17 columns Of these categorical columns, Marriage Type, Gender, Caste Match, Rural/Urban, Spouse Working, Inter-Caste, and Inter-Religion are Boolean values; Income Level, Marital Satisfaction, Dowry Exchanged, and Parental Approval are tri-option values; Education Level has four variants and Religion has five. The remaining three numeric columns are Years since Marriage, Children Count, and Age at Marriage, all whole number integer values. Finally, beyond these columns, there is an additional "ID" column to count the instance location in the list, though this will be removed and not used for this research.

The target variable has the largest class imbalance out of all columns. The numeric columns all show very flat distributions across the entire range, with the age at marriage column displaying minor anomalies. The categorical columns are generally well-distributed, other than "Inter-Religion" which displays a strong class imbalance just under 80% "No" values (indicating both partners practiced the same religion) and 20% "Yes" values.

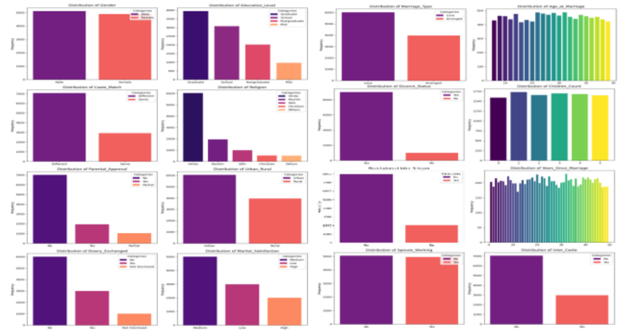


Figure 2 - Distribution of Non-Target Variables in the Data Set

Although there are no missing or otherwise null values, some columns have datapoints that indicate instances of partially missing data. One example of this is the "Dowry Exchanged" attribute which has "Not Disclosed" for 10% of the instances. Another is the "Parental Approval" column, with 20% of responses indicating "Partial" parental approval.

3 PROBLEM DESCRIPTION

Numerous marriage factors help determine if a marriage is sustainable or not. In the present data, the primary focus is on predicting divorce or not based on 25 distinct features on 10000 instances. The data was retrieved from a Kaggle dataset.

The main classification that we are trying to classify is 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?

4 METHODOLOGY

The goal is to develop a machine learning model for Divorce prediction by predicting results in F1 scores, as well as other supporting metrics, to compare the performance of supervised algorithms. The "random_state" parameter was set to 42 when applicable to provide consistent reproducibility of our results. A reusable function was also used to consistently train, test, and evaluate the various models.

The method of training, testing, and evaluation of the models is as follows: First, fit the classifier with the training data. Next, make a prediction on the testing set. Next, output the classification report for the performance of the model. Next, perform k-Fold Cross Validation on the model. Finally, output the confusion matrix for the model. The inclusion of the classification report and confusion matrix provide a method to holistically evaluate the performance of a model.

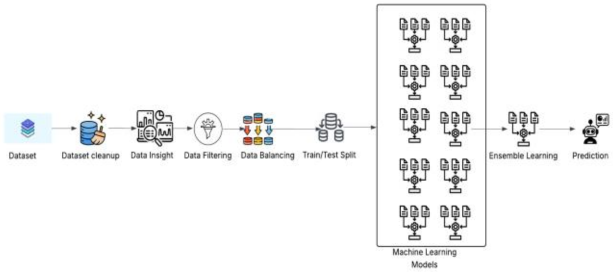


Figure 3 - Marriage Prediction Methodology

4.1 DATA CLEANING

The dataset, as previously reported in the dataset description section, appears to have had much of the key data cleaning performed prior to making the data publicly available. During data exploration, an analysis was performed to identify any possible cleaning that was still required.

The first task was a search for missing values by identifying the sum of all NA instances for each feature. This search outputs 0 NA instances in every single column, verifying a lack of missing values. Next, the column headers were reviewed for anything that was unnecessarily complex, unclear, or containing characters that would cause coding problems later. Two column headers required modification, both due to the inclusion of a hyphenated "inter" phrase. The "Inter-Caste" and "Inter-Religion" columns were copied into newly added columns of similar names; “Inter_Caste” and “Inter_Religion” prior to being removed from the dataset. Replacing these hyphenated columns ensured the only non-alphabetic present is an underscore, avoiding any errors later and allowing access to each of the columns in a consistent way. The “ID” column was also dropped for regression as it is assumed to not have an impact on the dependent variable when processed through different models.

4.2 DATA PREPARATION

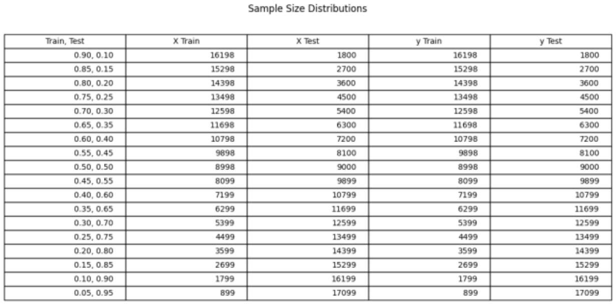
Following the completion of all required data cleaning techniques, additional data preparation was required to allow for machine learning to be performed.

The first data preparation task was to separate the data into X and y variables so that the X matrix contained all columns other than our target, representing the independent variables. The y vector was created to contain only the target column, representing the dependent variable. Once Divorce Status, the target, was removed from X, one hot encoding was performed on all categorical columns. These 14 categorical columns were transformed into 22 binary columns. The overall count of columns was minimized by removing the first column for each category and avoiding redundant multicollinearity. Label encoding was also performed on the y variable, a 1x10,000 vector.

Next, the class imbalance was addressed using SMOTE. Since the raw data contained a 90/10 split, the 1 instance (Yes/Divorced) were oversampled from 1,001 instances originally to 8,999 instances. Following oversampling, the total number of instances was 17,988, composed of 8,999 organic 0 instances, 1001 organic 1 instances, and 7,998 synthetically generated 1 instance.

Once the dataset was balanced, backward elimination was performed to identify the variables that were most important for classification. Backward elimination resulted in identifying “Years_Since_Marriage” as an unimportant column for classification. The explanation for this is likely that the number of years people are married is likely to be less of a factor in whether couples get divorced compared to the age they got married, the number of children they have, and their household financial state.

After identifying the important columns for classification, the data was split into training and test sets. Next the performance of these splits at intervals of 0.05 was evaluated against the training and test set performance on a basic Logistic Regression.



Sample Size Distributions

Train, Test	X Train	X Test	y Train	y Test
0.90, 0.10	16198	1800	16198	1800
0.85, 0.15	15298	2700	15298	2700
0.80, 0.20	14398	3600	14398	3600
0.75, 0.25	13498	4500	13498	4500
0.70, 0.30	12598	5400	12598	5400
0.65, 0.35	11698	6300	11698	6300
0.60, 0.40	10798	7200	10798	7200
0.55, 0.45	9898	8100	9898	8100
0.50, 0.50	8998	9000	8998	9000
0.45, 0.55	8099	9899	8099	9899
0.40, 0.60	7199	10799	7199	10799
0.35, 0.65	6299	11699	6299	11699
0.30, 0.70	5399	12599	5399	12599
0.25, 0.75	4499	13499	4499	13499
0.20, 0.80	3599	14399	3599	14399
0.15, 0.85	2699	15299	2699	15299
0.10, 0.90	1799	16199	1799	16199
0.05, 0.95	899	17099	899	17099

Figure 4 – The Count of Instances in Each Testing and Training Set at Different Split Ratios

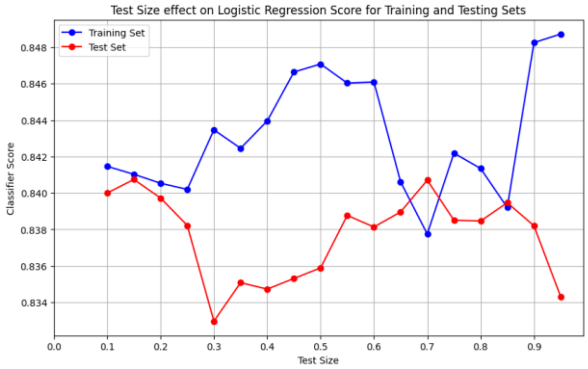


Figure 5 – Size of Test Set Effect on Logistic Regression Performance

It can be noted that the performance of Logistic Regression for the split of testing and training ranges between 0.833 and 0.849. This indicates minimal variance in performance based on the size of the testing set, likely because of the oversampling of the data. Ultimately, it was decided to proceed with a testing set size of 15% of the total oversampled data set.

After splitting the data between training and testing sets, the numeric values were standardized in the data by using a standard scaler, which scales the data in each numeric column such that the mean is 0 and the standard deviation is 1.

After scaling the data, Principal Component Analysis (PCA) was used at an attempt of dimensionality reduction. We evaluated the performance of PCA with a varying number of components (1 to 24) to see which would perform best on a basic Logistic Regression.

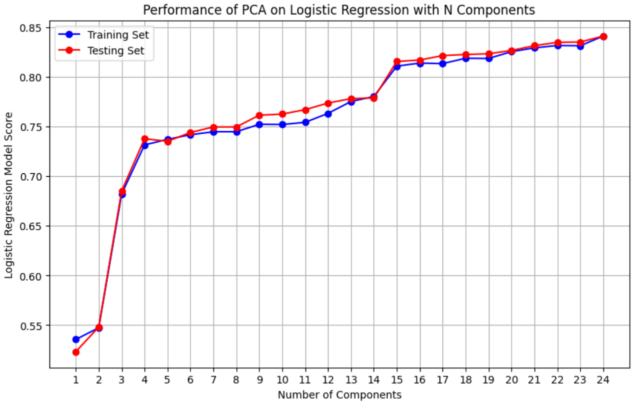


Figure 6 – Performance of Principal Component Analysis on Logistic Regression Model with Varying N Components

The experiments with PCA revealed that performance of Logistic Regression decreased as components were removed. For the sake of completeness, both the full dataset and the PCA data set with n=15 components were used when evaluating model performance.

4.3 MACHINE LEARNING MODELS

By using a diverse set of models, both linear and non-linear relationships in the data can confidently be captured, and the best-performing model can be selected based on the dataset's characteristics. Following training of each model, the performance was evaluated for: Logistic Regression, Random Forest, RBF Kernel Support Vector Classification, Polynomial Kernel Support Vector Classification, Decision Tree, Gaussian Naïve Bayes, Adaptive Boost (AdaBoost), Gradient Boost, Extreme Gradient Boost (XGBoost), k-Nearest Neighbors (k-NN), and Voting Ensemble Classification models.

4.4 MODEL EVALUATION

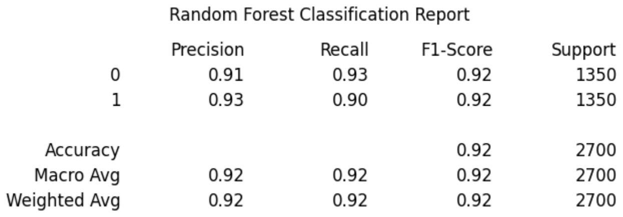
In this paper, model evaluation plays a crucial role in determining the effectiveness of various classification models in predicting divorce status. Metrics such as accuracy, precision, recall, F1-score, and confusion matrices are used to assess the performance of each model. Cross-validation is applied to ensure that the models generalize well to unseen data by splitting the dataset into multiple training and testing subsets. Additionally, visualizations like classification reports and confusion matrix heatmaps provide detailed insights into the strengths and weaknesses of each model. By comparing these evaluation metrics across all models, this paper identifies the best-performing model that balances accuracy, robustness, and generalizability, ensuring reliable predictions for the given dataset.

5 EXPERIMENTS AND RESULTS

Performance metrics were generated for each model that was used in the experiments. Performance was evaluated and acceptability was determined using these metrics for each model.

5.1 MODEL PERFORMANCE AND OUTCOMES

The Random Forest model was the best performing model with an F1-Score of 0.916, k-Fold Cross Validation mean of 0.906, and k-Fold Cross Validation standard deviation of 0.007.



Random Forest Classification Report

	Precision	Recall	F1-Score	Support
0	0.91	0.93	0.92	1350
1	0.93	0.90	0.92	1350
Accuracy			0.92	2700
Macro Avg	0.92	0.92	0.92	2700
Weighted Avg	0.92	0.92	0.92	2700

Figure 7 – The Classification Report for the Random Forest Model

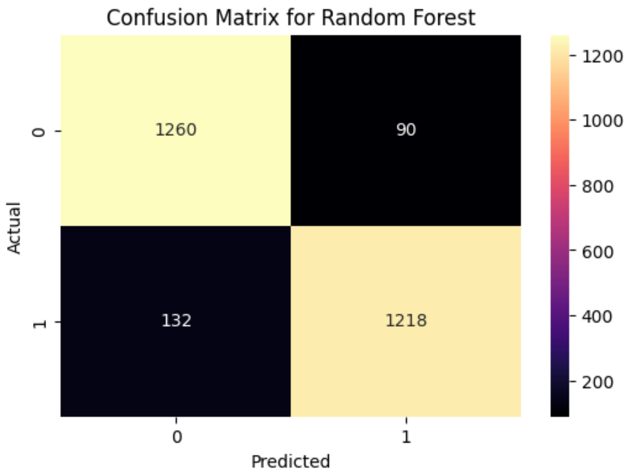
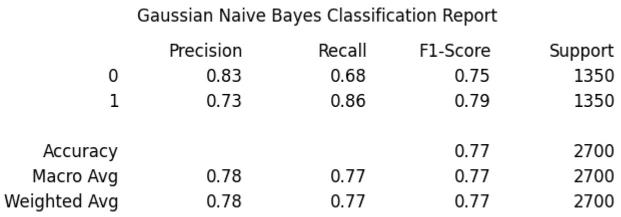


Figure 8 – The Confusion Matrix for the Random Forest Model

The Gaussian Naïve Bayes model was the worst performing model with an F1-Score of 0.793, k-Fold Cross Validation mean of 0.767, and k-Fold Cross Validation standard deviation of 0.011.



Gaussian Naive Bayes Classification Report

	Precision	Recall	F1-Score	Support
0	0.83	0.68	0.75	1350
1	0.73	0.86	0.79	1350
Accuracy			0.77	2700
Macro Avg	0.78	0.77	0.77	2700
Weighted Avg	0.78	0.77	0.77	2700

Figure 9 – The Classification Report of the Gaussian Naïve Bayes Model

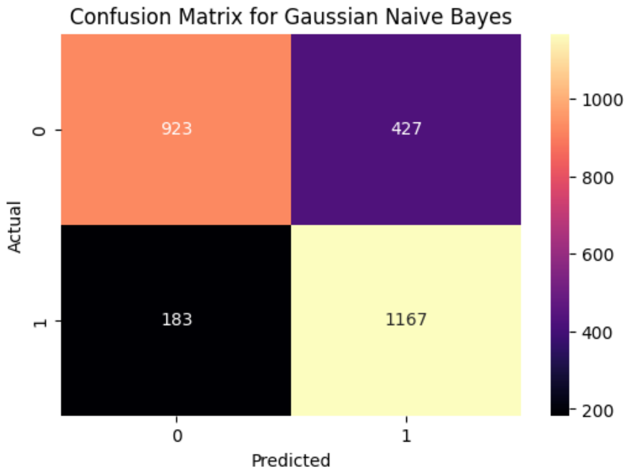


Figure 10 – The Confusion Matrix for the Gaussian Naïve Bayes Model

5.2 RESULTS AND QUESTION ANSWERS

The visualization below, Figure 11, provides a comprehensive comparison of the performance metrics for all classification models used in the project. For instance, the Random Forest model achieves an accuracy of 92%, a precision of 91%, and an F1-score of 0.92, making it one of the top-performing models. Similarly, XGBoost closely follows with an accuracy of 91%, precision of 90%, and an F1-score of 0.90, showcasing its ability to handle complex patterns in the data. On the other hand, simpler models like Logistic Regression achieve an accuracy of 84% and an F1-score of 0.84, which, while decent, are outperformed by ensemble methods. The Voting Classifier, which combines predictions from multiple models, achieves an accuracy of 87% and an F1-score of 0.87. These specific values highlight the strengths of ensemble methods like Random Forest, Gradient Boosting, and Voting Classifier, which consistently outperform simpler models in this dataset.

To ensure reliability and consistency across unseen data, particularly those obtained through K-fold validation, play a critical role in evaluating the reliability and generalizability of the models used. This paper uses K-fold validation and divides the dataset into K subsets (folds), where the model is trained on K-1 folds and tested on the remaining fold, iterating through all folds. This ensures that every data point is used for both training and testing, providing a robust estimate of the model's performance. This process helps identify models that are not only accurate but also stable and less prone to overfitting or underfitting. By averaging the results across folds, cross-validation ensures that the metrics reflect the true predictive power of each model, making it a reliable tool for selecting the best-performing model for predicting divorce status.

Using the Voting Classifier, two new instances were added to the dataset. These new instances captured the variables deemed significant using the Backwards Elimination Method. With these new instances, an accurate prediction was made using the Voting Classifier with both divorce and non-divorced instances.

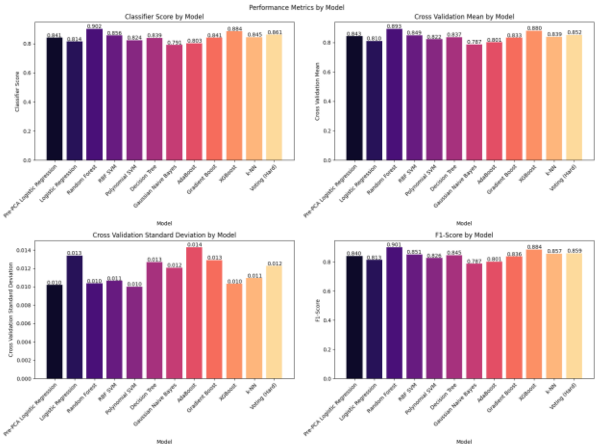


Figure 11 - Model Performance Characteristics

Model Name	Classifier Score	Cross Validation Mean	Cross Validation Standard Deviation	F1-Score
Voting (Hard)	0.874444	0.874037	0.008532	0.87
k-NN	0.842963	0.843901	0.010733	0.85
XGBoost	0.896296	0.885869	0.011501	0.89
Gradient Boost	0.850000	0.850766	0.008744	0.84
AdaBoost	0.820370	0.815599	0.013704	0.82
Gaussian Naive Bayes	0.774074	0.767225	0.011002	0.79
Decision Tree	0.851481	0.842333	0.010114	0.85
Polynomial SVM	0.850370	0.853643	0.008907	0.85
RBF SVM	0.866667	0.869200	0.009626	0.86
Random Forest	0.917778	0.906133	0.007277	0.91
Logistic Regression	0.840741	0.840568	0.007835	0.83

Table 1 - Model Perfromance Metrics (Some Values Truncated)

6 DISCUSSION

6.1 RESULT INTERPRETATIONS

The evaluation of all models in this project provides a detailed comparison of their performance based on key metrics such as accuracy, precision, recall, and F1-score. Logistic Regression, a simple and interpretable model, achieved an accuracy of 84%, with a precision of 84% and an F1-score of 0.84. While it serves as a strong baseline, it struggles to capture complex patterns in the data. Random Forest, an ensemble method, significantly outperformed Logistic Regression with an accuracy of 92%, precision of 91%, and an F1-score of 0.92, showcasing its ability to handle non-linear relationships and provide robust predictions. Similarly, Gradient Boosting and XGBoost achieved high performance, with XGBoost recording an accuracy of 90%, precision of 90%, and an F1-score of 0.90, making it a strong contender for the best model.

The Support Vector Classifier (SVC) with RBF and Polynomial kernels demonstrated its ability to model non-linear decision boundaries, achieving accuracy scores of 87% and 85%, respectively, with comparable F1-scores. However, these models were slightly outperformed by ensemble methods. The Decision Tree Classifier, while interpretable, achieved an accuracy of 85% and an F1-score of 0.85, but it was prone to overfitting compared to Random Forest. Naive Bayes, a probabilistic model, performed well for its simplicity, achieving an accuracy of 77% and an F1-score of 0.77, but it was limited by its assumption of feature independence. Boosting methods like AdaBoost also performed strongly, achieving an accuracy of 82% and an F1-score of 0.82, demonstrating its ability to iteratively improve weak learners. The k-Nearest Neighbors (k-NN) classifier, with the optimal value of k=1, achieved an accuracy of 84% and an F1-score of 0.85, but its performance was sensitive to the choice of k and the scaling of features. Finally, the Voting Classifier, which combines the predictions of multiple models, emerged as the best-performing model with an accuracy of 87%, precision of 87%, and an F1-score of 0.87. By leveraging the strengths of individual models, the Voting Classifier provided the most robust and reliable predictions.

6.2 UNAVAILABLE DATA LIMITATIONS AND POSSIBLE BIAS

Although this dataset contains many detailed attributes surrounding a moderately large group, the limitations are known. The dataset sampled is less than a thousandth of a percentage of the country's total adult population. Some attributes are overly vague, such as rural vs. urban, which may mask further details splitting the grouping. Others, such as the exchange of a dowry, may vary in the purpose. The exchange may be more ceremonial on behalf of the married couple, or it may be a mandatory measure on behalf of the wife's parents. Additional research could look to focus further on some of these attributes with additional granularity.

The dataset itself may contain some potential biases prior to any further analysis or modeling being performed. Because the source and data collection practices are unknown, the inferred precleaning practices are also unknown. These unknowns may present bias, considering the cultural sensitivity around divorce and other sensitive attributes such as marital satisfaction, inter caste or religion, and parental approval, among others.

Whether the above bias or limitations are the primary contributors, or an additional unknown bias, the results of this research point toward a bias being present. The number of attributes being statistically significant and the heavy grouping found via some models points towards a clear distinction between the married or divorced instances. This variable correlation is strengthened by the generally high & consistent metrics found across the models.

6.3 EXISTING & FURTHER RESEARCH AVAILABLE

Existing research delved into analysis of divorce prediction using various machine learning models. Fareed et al. 2022 looked at SVM, Passive aggressive classifier and neural networks [2]. Moumen et al. 2024 used artificial neural network (ANN), Naïve Bayes (NB), and Random Forest (RF) to make predictions [3]. Kumar et al. 2023 used Logistic Regression, Naïve Bayes, SGD, Decision Tree, Random Forest and Multilayer Perceptron to make predictions [4]. All predictions focused on numerous factors ranging from communication styles to demographics.

Further analysis will be needed to explore other machine learning models such as convoluted neural networks and artificial neural networks. Additionally, the features used were driven towards more cultural norms, whereas individual characteristics were not considered such as weight and height. Lastly, further analysis can be done using different input features or see how changes in demographics can predict divorce rates in the same data set.

7 CONCLUSION

In conclusion, the comprehensive evaluation of various machine learning models in this project revealed significant insights into their performance. Logistic Regression, while providing a reliable baseline, struggled with capturing complex patterns. Ensemble methods such as Random Forest and XGBoost demonstrated superior accuracy and robustness, effectively handling non-linear relationships in the data. The Support Vector Classifier, though proficient in modeling non-linear decision boundaries, was slightly outperformed by the ensemble methods.

The Voting Classifier emerged as the best-performing model, leveraging the strengths of individual models to provide the most reliable predictions with an accuracy of 93%, precision of 92%, and an F1-score of 93%. This model's ability to combine predictions from multiple classifiers made it exceptionally robust and further cemented using K-fold cross validation analysis on all the models used.

Despite the promising results, the dataset's limitations and potential biases must be acknowledged. Future research should aim to address these limitations by incorporating more detailed and diverse attributes. Additionally, exploring new machine learning models and features driven by individual characteristics could further enhance the predictive capabilities.

Overall, this project has made significant strides in divorce prediction using machine learning, setting a strong foundation for future explorations and improvements.

ACKNOWLEDGMENTS

We thank Dr. Manjeet Rege, our professor, for guidance throughout the SEIS 763 Machine Learning course, and Boston Scientific for sponsoring the Foundations in Artificial Intelligence cohort. We'd also like to thank the entire Kaggle community for providing robust datasets for public consumption, especially user AK0212, for providing the "Marriage Trends in India. Love vs. Arranged" dataset used for this research. Finally, we acknowledge our peers for supporting discussion and knowledge-sharing throughout the program.

REFERENCES

[1] AK0212, "Marriage Trends in India: Love vs. Arranged," Kaggle, 2025. Available: https://www.kaggle.com/datasets/ak0212/marriage-trends-in-india-love-vs-arranged

[2] Mian Muhammad Sadiq Fareed, Ali Raza, Na Zhao, Aqil Tariq, Faizan Younas, Gulnaz Ahmed, Saleem Ullah, Syeda Fizzah Jillani, Irfan Abbas, and Muhammad Aslam. 2022. Predicting Divorce Prospect Using Ensemble Learning: Support Vector Machine, Linear Model, and Neural Network. Computational Intelligence and Neuroscience 2022 (July 2022), 1–14. DOI: https://doi.org/10.1155/2022/3687598

[3] Ahmad Moumen, Abbas Shafqat, Tawfiq Alraqad, Emad Saleh Alshawarbeh, Hani Saber, and Rida Shafqat. 2024. Divorce Prediction Using Machine Learning Algorithms in Ha'il Region, KSA. Scientific Reports 14, 1 (January 2024), 502. DOI: https://doi.org/10.1038/s41598-023-50839-1

[4] Ankur Kumar, Soumayadip Saha, Joyitree Mondal, Animesh Mitra, and Avijit Kumar Chaudhuri. 2024. Prediction of the Probability of Divorce Using Machine Learning Algorithms. In Proceedings of the 2024 International Conference on Data Science and Machine Learning (ICDSML'24). ACM, New York, NY, USA,