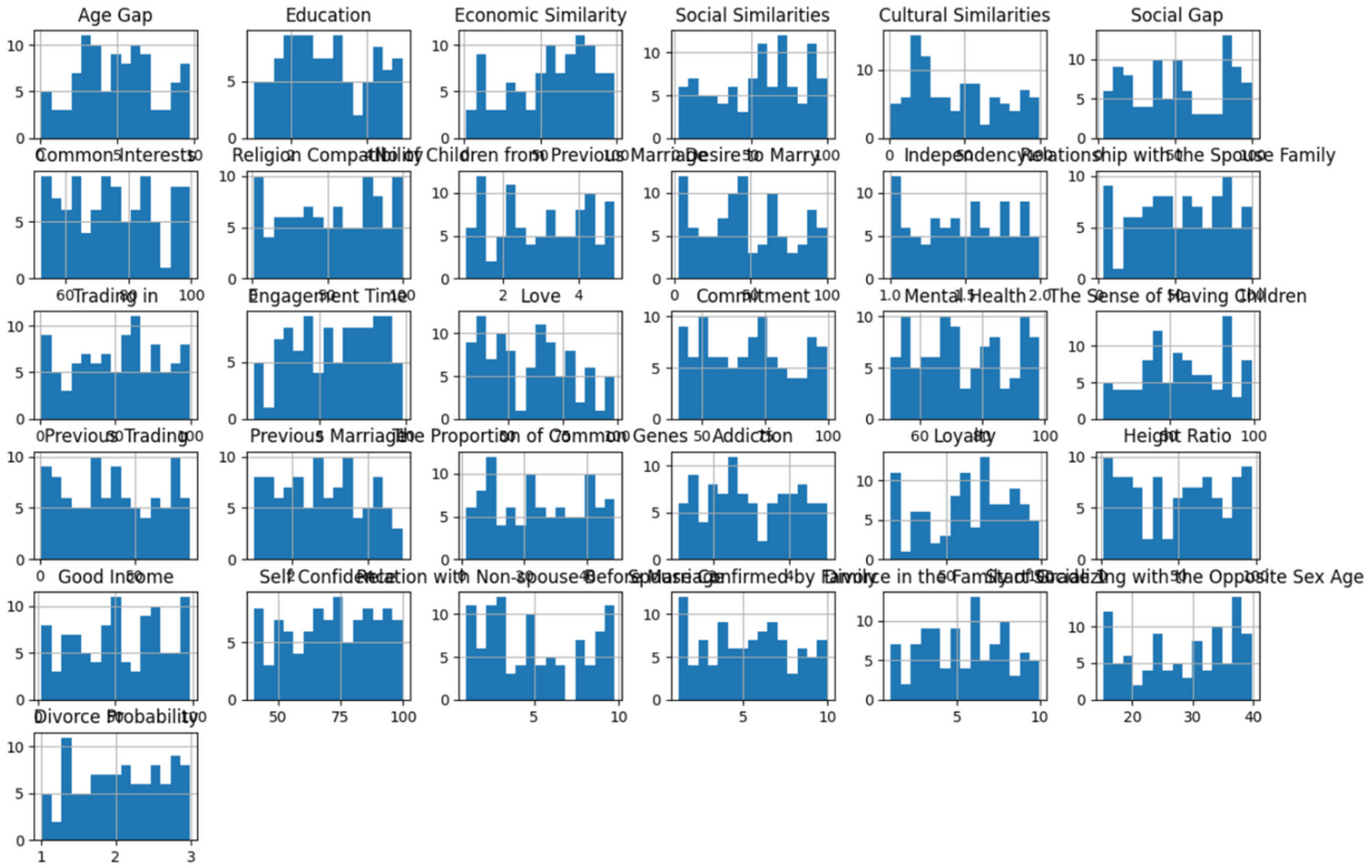
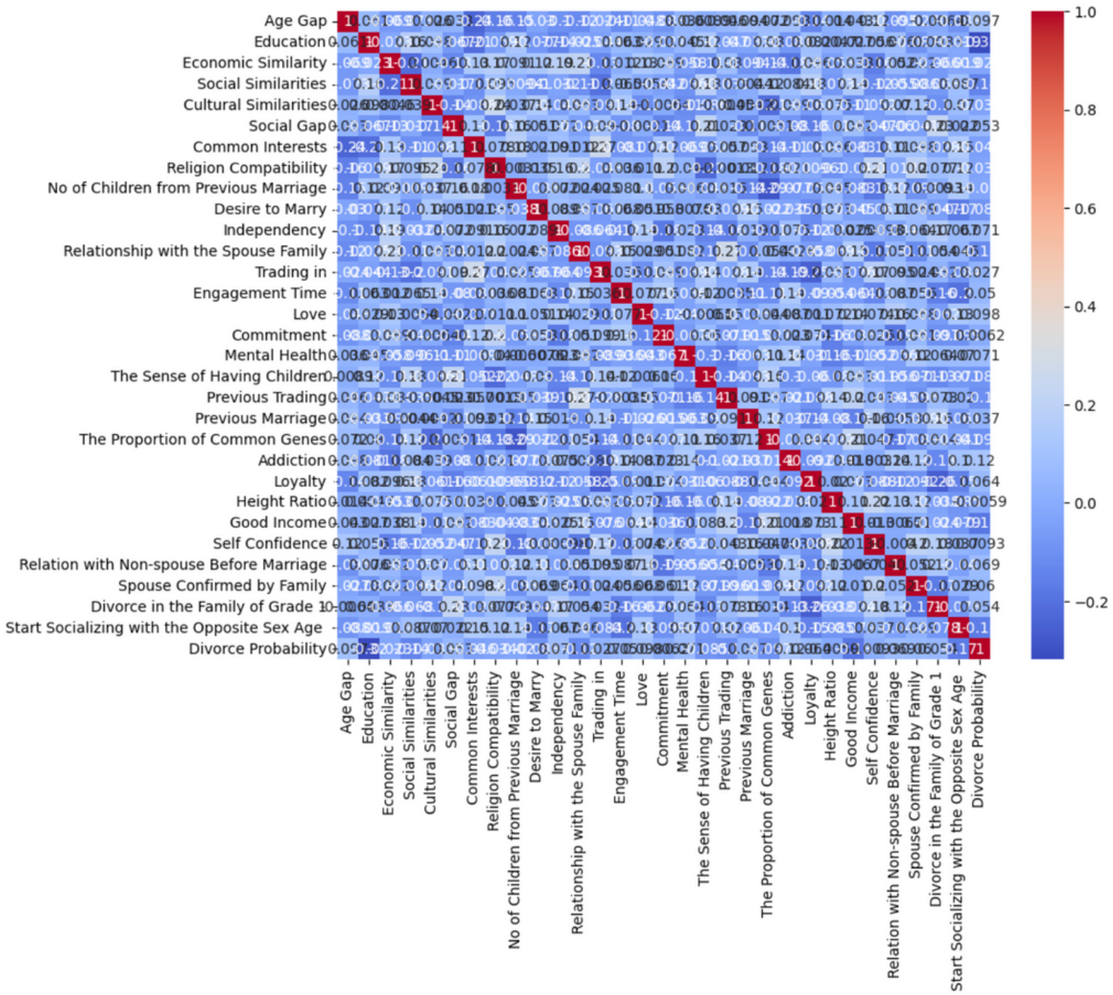
# Predicting Divorces using machine learning techniques

sai sindhuja upadrashta- 01
rohil kothapally-02
anthony taylor-01

A dataset titled "Divorce Survey," which includes a variety of parameters that might be used to predict divorce. These parameters include 'Age Gap', 'Education', 'Economic Similarity', 'Social Similarities', and several others.
Python Library Imports: . The libraries imported include:

- - `pandas` for data manipulation and analysis.
- - `SimpleImputer` and `StandardScaler` from `sklearn` for handling missing data and feature scaling.
- - `train_test_split` from `sklearn.model_selection` for dividing the data into training and testing sets.
- - `RandomForestRegressor` from `sklearn.ensemble` which indicates that a machine learning model is being used, likely for regression analysis.
- - Other utilities such as `mean_squared_error`, and `r2_score` for evaluating model performance.


Age Gap
Education
Economic Similarity
Social Similarities
Cultural Similarities
Social Gap
Common Interests
Religion
Compatibility
Children from Previous Marriage
Desire to Marry
Independency
Relationship with the Spouse Family
Trading in
Engagement Time
Love
Commitment
Mental Health
The Sense of Having Children
Previous Trading
Previous Marriage
The Proportion of Common Genes
Addiction
Loyalty
Height Ratio
Good Income
Self Confidence
Relation with Non-spouse Before Marriage
Spouse Confirmed by Family
Divorce in the Family
Standardizing with the Opposite Sex Age
Divorce Probability


Age Gap
Education
Economic Similarity
Social Similarities
Cultural Similarities
Social Gap
Common Interests
Religion Compatibility
No of Children from Previous Marriage
Desire to Marry
Independency
Relationship with the Spouse Family
Trading in
Engagement Time
Love
Commitment
Mental Health
The Sense of Having Children
Previous Trading
Previous Marriage
The Proportion of Common Genes
Addiction
Loyalty
Height Ratio
Good Income
Self Confidence
Relation with Non-spouse Before Marriage
Spouse Confirmed by Family
Divorce in the Family of Grade 1
Start Socializing with the Opposite Sex Age
Divorce Probability
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

# LINEAR REGRESSION

- Linear Regression is initialized and trained using the **LinearRegression()** class from the **sklearn.linear_model** library. The model is fit with the training data (**X_train** and **y_train**).
- Predictions are made on both training and testing sets (**X_train**, **X_test**) using the trained Linear Regression model.
- The performance of the model is evaluated using metrics such as mean squared error (MSE) and R-squared. These metrics are calculated for both the training and testing datasets to assess the accuracy and goodness of fit of the Linear Regression model.
- Results indicate the model's performance, with specific values provided for MSE and R-squared on both training and testing sets. The evaluation helps in understanding the model's effectiveness in predicting and its capability to generalize to new, unseen data.

# RANDOM FOREST REGRESSION MODEL

- The Random Forest Regressor is initialized with a specific number of estimators and a random state set for reproducibility.
- The model is trained on the training dataset, where it learns to predict the target variable from the features provided.
- Predictions are made on both the training set to evaluate the fit of the model and on the testing set to assess its generalization capabilities.
- The performance of the model is evaluated using metrics such as mean squared error (MSE) and R-squared ($R^2$), which provide insights into the accuracy and the proportion of variance in the dependent variable explained by the model.
- The results are analyzed to determine the effectiveness of the Random Forest model in the context of the data and project objectives, considering factors like overfitting and the ability of the model to generalize to new data.

# EVALUATION OF MODEL

- The results reveal that the model explains roughly 81.4% of the variance in the training set but shows a significant decrease in performance on the test set, where it accounts for only about 14.7% of the variance. This substantial discrepancy suggests that the model may be overfitting to the idiosyncrasies and noise within the training data rather than capturing underlying, generalizable patterns across the broader dataset.

- To improve the model's ability to perform well on new, unseen data, it is advisable to implement stronger regularization methods that help mitigate overfitting. Additionally, adjusting the model’s hyperparameters to simplify the model or utilizing approaches like cross-validation could enhance parameter estimation, leading to more stable and consistent performance across different data samples.

# IMPROVE THE MODEL

- Employed Random Search Cross-Validation to optimize model hyperparameters by exploring a grid of values and selecting combinations randomly.
- The tuning process was conducted iteratively for a set number of repetitions to ensure thorough exploration.
- Aimed to identify the best combination of hyperparameters that maximized performance on the validation set.
- Post-tuning, the model displayed increased sensitivity, though this led to decreased precision and specificity, slightly lowering the AUC from 0.80 to 0.78.
- Highlighted the trade-offs involved in model optimization, where improving one aspect of performance might adversely affect others, underscoring the complexities in achieving a balanced model.

- Utilize Random Search Cross-Validation to optimize hyperparameters like learning rate, max depth, min samples leaf, min samples split, and number of estimators.
- Implement grid search within cross-validation to explore predefined hyperparameter spaces.
- Use Negative Mean Squared Error (MSE) as the evaluation metric to gauge model accuracy.
- Ensure model reliability and stability across different datasets through statistical validation via cross-validation.
- Transition to more advanced algorithms such as Gradient Boosting or LightGBM based on outcomes of hyperparameter tuning for potential performance improvements.

# LIGHTGBM algorithm

- Implementation of GridSearchCV to explore a wide range of hyperparameters and identify the optimal settings for the models, focusing on parameters such as learning rate, tree depth, and number of estimators.
- The use of a scoring metric based on negative mean squared error to rank the performance of each parameter combination, facilitating a quantitative comparison across different sets of hyperparameters.
- Detailed results from the GridSearchCV showing the best parameters found and their corresponding performance scores, illustrating how different configurations impact model accuracy.
- Adjustments to the models' hyperparameters based on the GridSearchCV outcomes, with emphasis on how changes in parameters like max_depth, min_samples_split, and n_estimators affect the model's ability to generalize to new data.
- A discussion on the trade-offs between model complexity and performance, highlighting how increasing n_estimators or reducing the learning_rate can lead to more precise but computationally expensive models.
- Comparative analysis of model performance before and after hyperparameter tuning, using metrics such as R-squared and mean squared error, to demonstrate the effectiveness of the tuning process in enhancing model predictive power.

# CONCLUSION

- This project demonstrated the feasibility of using machine learning to predict marital stability with a high degree of accuracy. The insights gained from feature importance analysis can be particularly useful for marital counselors and relationship experts in identifying areas of concern before they escalate into potential grounds for divorce.
- Future work could explore deeper ensemble techniques and newer algorithms like neural networks for potential improvements in prediction accuracy. Additionally, incorporating more granular data, such as personal attitudes towards marriage and daily interaction patterns, could further enhance the model's predictive power.
- Overall, the project underscores the significant potential of machine learning in social science applications, providing valuable insights that can aid individuals and professionals in making informed decisions about marital relationships.

# Thanks