**Using Statistical Learning and Big Data to Predict Relationship Status**

*A Case Study of Middlebury College Students*

Quinlan McGaugh ’22 and Otis Milliken ‘24

Middlebury College

**Abstract**

**Introduction**

* **Intro and Motivation**

If you have taken a tour at Middlebury College, odds are high that you heard the tour guide boast an interesting statistic regarding marriage rates of Middlebury Students. Something like, “You know, something like 60 percent of Middlebury students marry other Middlebury students![[1]](#endnote-1)” Although some questions have been raised, and these relationships are a key piece of college life. During Covid-19, relationships were

Motivated by this interesting phenomenon, as well as series of polls like the Middlebury Campus’ *Zeitgeist 4.0* and Midd MarriagePact, the authors aimed to create a tool for predicting relationship status based on answers to a survey created by Middlebury sociology professor Peggy Nelson in 2011.

* **Why is the problem important?**

By using random forests, we’re exploring the determinants of relationship status at Middlebury College. There are various applications to student life planning and administration,

* Identify determinants of relationship status at Middlebury
* We can assess whether determinants of relationships have changed before and after Covid-19
  + Built using data collected in 2020, tested using data from 2022.
* If relationships are truly something Middlebury prides itself on, then they can use this information
* Entertainment
  + In an entertainment environment, a user can input their responses and see whether Middlebury students like them tend to be in relationships or not
* **Novel contribution to the field of AI**

Adds to a relatively indecisive breadth of literature that discusses optimization techniques of random forests.

Section II will discuss a brief literature review of random forests and decision trees. Section III will formalize our problem statement. Section IV will provide a full description of our methodology and results. Section V offers conclusions and discussion.

**Related Work**

**Problem Statement**

**Methods and Results**

*I. Data*

* One-Hot encoding of variables

*II. Statistical-Based Learning using Decision Trees and Random Forests*

*II.A Naïve Decision Tree*

*II.B Random Forests*

*III. Results*

*III.A Naïve Decision Tree*

*III.B Random Forest Gini Technique*

*III.C Random Forest Entropy*

*III.D Comparison of Techniques*

**Discussion and Conclusion**

**Bibliography**

**End Notes**

1. <https://middleburymagazine.com/features/bye-bye-love/> [↑](#endnote-ref-1)