# Machiavellian Expression and Relational Classification

***Midterm Project Report for DATA 1030, Fall 2021***

***https://github.com/ktoleary13/mach_rel_classification.git***

Katie O'Leary

## I. Introduction

Within personality studies, "The Dark Triad" describes sub-clinical yet offensive traits that are represented within populations globally. These personality traits are non-pathological but are colloquially negative descriptors: narcissism, psychopathy, and Machiavellianism. Machiavellianism, the manipulative personality, describes a person's perception of others and willingness to neglect others for personal gain. Here, a psychometric dataset containing Machiavellian and personality surveys will be used to build a classification model predicting relationship status. The relationship-status target variable describes a participant as married, never married, or previously married. With a classification model based on manipulative tendencies, relationship outcomes could be predicted and/or avoided. At a fundamental level, the ability of individuals to form meaningful relationships impacts their lifelong flourishing [1]. Thus, if a person expresses substantial Machiavellianism, they may inhibit their own ability to form a successful primary partnership.

Containing seventy thousand data points, this dataset houses one hundred and five original features. The features are a Machiavellian test questions (MACH IV), a personality assessment (TIPI), a vocabulary test, and demographic variables. Each of these is well documented in a separate text file provided by the "Open Psychometrics" website. Though there was no difference in dataset content, the dataset was accessed through Kaggle [2], where a single public project had been published. While this previous project attempted to use TensorFlow and machine learning, it is incomplete with no substantial outcomes. Additionally, this dataset was also used in Confirmatory Factor Analysis as a part of a thesis submitted to Australian National University [3]. While validating the data rigor and usability, the work does not inform the proposed machine learning model here but delves into the psychological research and refines definitions of Machiavellianism through factor analysis.

## II. Exploratory Data Analysis

During data exploration, the data was tailored for the project goals. For example, because relationship status is the target variable, datapoints with improper ages were immediately filtered out so that twenty-six thousand points remained [4]. Additionally, raw question needed to be scored

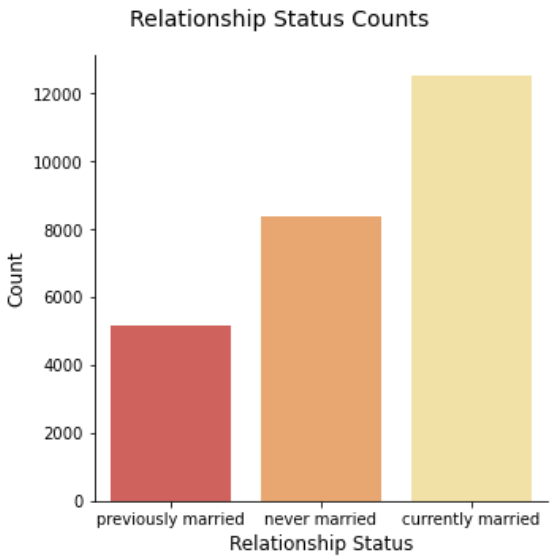


*Figure 1. Bar chart of relationship status target variable. Indicates that currently married individuals are the most represented.*

internally before exploration, and some features were winnowed due to repetitive or irrelevant information. However, once features were prepared, forty-seven variables displayed several trends in the dataset became noticeable.

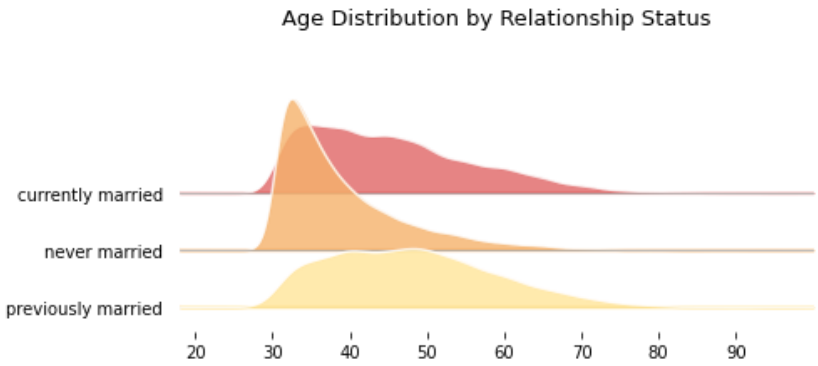


*Figure 2. Ridge plot of each relationship status versus the age of those individuals. Currently married and never married individuals skew young, which reflects the role of age in life partnerships.*

First, the proportion of datapoints in the target variable indicate that currently married individuals are the most represented (Figure 1). As demonstrated in Figure 2, the never married category skews younger than the others categories despite controlling for average age of marriage. While it is unsurprising that increased age increases opportunity and likelihood for marriage/previous marriage, the influence of age should be noted. Aside from age, considering the direct relationship between MACH IV test scores and relationship status is distinct, but not obviously so (Figure 3). Distributions reflect the population averages, but the means and quartile ranges differ more substantially (Figure 3). Further, these differences may expand in conjunction with personality test results [5]. For example, Figure 4 shows a distinction in MACH IV scores with signifiers of self-assurance.

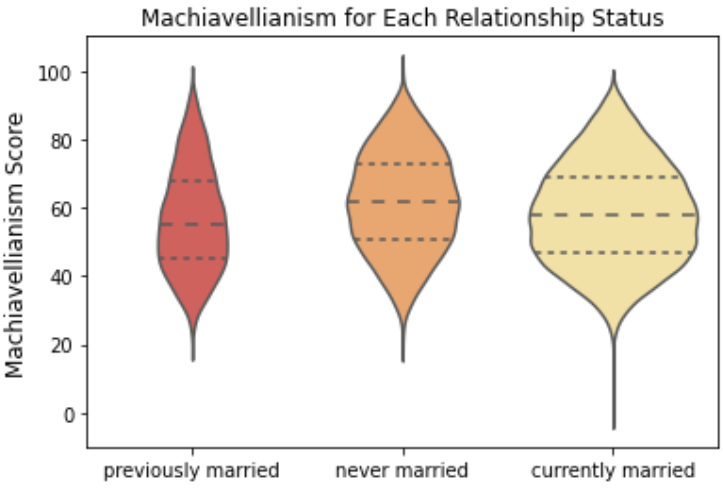


*Figure 3. Violin plot with quartile ranges of MACH IV scores for each relationship status. Never married individuals show highest means and quartile ranges.*

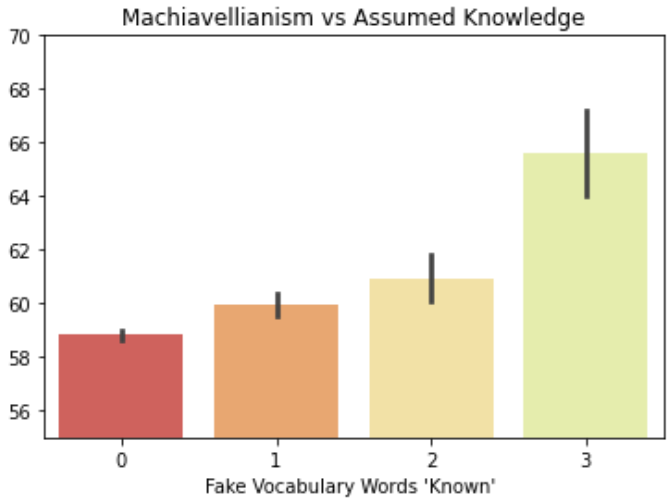


*Figure 4 Bar chart of MACH IV scores for number of fake voacbulary words known. The more words 'known', the higher the MACH IV score.*

## III. Data Processing

The dataset was split and processed according to the classes represented in the target variable. With no apparent group structure, it is neither IID nor a time-series dataset. Rather, each data point represents a single set of responses independent of all other points and

independent of the time the test was taken. Thus, the approach to splitting was based on the fraction of each represented class in the target variable. While the least represented class is still over fifteen percent, a stratified K-fold split with a shuffle of each class before sampling was implemented. This accounts for the variation represented in the features of each class and minimizes the chance of overrepresenting young ages in any one split.

Encoding and preprocessing the forty-seven features was simplified by the nature of the dataset. All MACH IV and TIPI questions were presented on a Likert-scale, which is a widely used ordinal encoder of survey responses. Thus, all features that contained disagreement to agreement scores[1] and education were encoded with OrdinalEncoder. For demographic variables and vocab survey responses, these categorical variables could not be assigned an ordinance without an unfair judgement on the survey response. For example, religion, handedness, orientation, vocabulary, etc.[2] have no ordinance and were therefore encoded with OneHotEncoder. Finally, the aggregate MACH IV score (20-100 scale) and age of individuals were encoded with MinMaxEncoder due to their natural boundaries. Then, all time-oriented variables[3], as continuous and theoretically boundless variables, were encoded with StandardScalar.

[1] MACH IV Questions ('Q1A', 'Q2A', 'Q3A', 'Q4A', 'Q5A', 'Q6A', 'Q7A', 'Q8A', 'Q9A', 'Q10A','Q11A', 'Q12A', 'Q13A', 'Q14A', 'Q15A', 'Q16A', 'Q17A', 'Q18A', 'Q19A', 'Q20A') , TIPI Scores ('extraver', 'agreeable', 'conscient', 'neuroticism', 'openness'), and Education (‘education’)

[2] Demographic Variables: ('race', 'voted','familysize', 'major','urban', 'gender', 'engnat', 'orientation','hand', 'religion', Vocabulary responses: ('voc_fake', 'voc_conf')

[3] Time Variables: (‘testelapse’, ‘PITtime’, NITtime’, ‘PVHtime’,’CVHtime’)

IV. References

[1]:Kern, Margaret L., et al. “Lifelong Pathways to Longevity: Personality, Relationships, Flourishing, and Health.” *Wiley Online Library*, John Wiley & Sons, Ltd, 16 Oct. 2013,

[2] Greenwell, Lucas. “Machivallianism Test Responses.” *Kaggle*, 28 May 2020, https://www.kaggle.com/lucasgreenwell/machivallianism-test-responses.

[3] Monaghan, Conal, “Two-Dimensional Machiavellianism: Conceptualisation, Measurement, and Well-Being”. *Thesis to Australian National University* 2019/01/0 https://doi.org/10.25911/5d149b879161d

[4] “Median Age at First Marriage: 1890 to Present.” *Census.gov*, United States Census Bureau , 2020, https://www.census.gov/content/dam/Census/library/visualizations/time-series/demo/families-and-households/ms-2.pdf.

[5] Paulhus, Delroy L, and Kevin M Williams. “The Dark Triad of Personality: Narcissism, Machiavellianism, and Psychopathy.” *Journal of Research in Personality*, Academic Press, 19 Nov. 2002,