Siddharth Tiwari
Mr. Meyer
CS Seminar: Machine Learning
28 May 2021

Using PCA and Naive Bayes to Observe Trends in "Machiavellianism" Amongst Different Demographics

1. Introduction

In psychology, Machiavellianism is a personality trait where a "person so focused on their own interests will manipulate, deceive, and exploit others to achieve their goals." (Jones 2009) To measure this affinity for manipulation and indifference to morality, psychologists devised the *Mach-IV* test, a twenty question personality survey that is used to measure the extent at which the "Machiavellian Construct" influences an individual's actions. (Christie & Geis 1970) This questionnaire, along with others, was administered by the Open-Source Psychometrics Project in an aim to understand how the strength of a population's Machiavellian trait varies between different populations and personalities. (Open Psychology Data 2019) In this project, I examine the relation of this trait with different demographics variables, such as gender, race, religion, marital status, etc.

Keeping this as the main task of my project, I observe the effect of principal component analysis (PCA) on the accuracy and runtime of classification by two different algorithms. This dataset contains a large number of instances and features. Since the runtime for these classification algorithms is influenced by the amount of data utilized by each one, I examine the effect of dimensionality reduction on accuracy and runtime. To classify these data, I utilize the Gaussian Naive-Bayes (NB) and k-Nearest Neighbors (KNN) algorithms, which, at the

mathematical level, are two entirely different algorithms. The NB algorithm assumes that there is no relation between any of the inputted features; (Shanbhag & Rao 2003) conversely, the KNN algorithm rests on similarities between the different features of the data, designating clumps of data as a "class." (Altman 1992)

2. Materials and Methods

Below, I detail the features of the data and methods used to conduct analyses.

2.1 Dataset

This dataset is published by the Open-Source Psychometrics Project and consists of 73,489 respondents and 105 metrics respectively. This version of the dataset was published on 26 March 2019. The following list contains descriptions of the features contained in this dataset:

- 20 Question *Mach-IV* Questionnaire.
    - Consists of 20 questions used to measure the presence of the "Machiavellian Construct" in the survey respondent.
    - Three values are recorded for each question (ex. Q1):
        - The user's answer (ex. feature name: Q1A)
        - The position of the item in the survey (ex. feature name: Q1I)
        - The time spent on the question in milliseconds (ex. feature name: Q1E)
- Ten Item Personality Inventory
    - Ten-Item Personality Inventory was used to briefly capture the respondent's personality traits
    - Features are labeled as "TIP" followed by the question number (ex. TIP1)
- Definition Validity Checklist
    - This questionnaire was admitted to measure the vocabulary/critical thinking of the respondent. These features were excluded completely from classifications.

- Demographic Variables
    - The following information was catalogued for each respondent (and is utilized in the dataset):
        - education: "How much education have you completed?", 1=Less than high school, 2=High school, 3=University degree, 4=Graduate degree
        - urban: "What type of area did you live when you were a child?", 1=Rural (country side), 2=Suburban, 3=Urban (town, city)
        - gender: "What is your gender?", 1=Male, 2=Female, 3=Other
        - religion: "What is your religion?", 1=Agnostic, 2=Atheist, 3=Buddhist, 4=Christian (Catholic), 5=Christian (Mormon), 6=Christian (Protestant), 7=Christian (Other), 8=Hindu, 9=Jewish, 10=Muslim, 11=Sikh, 12=Other
        - orientation: "What is your sexual orientation?", 1=Heterosexual, 2=Bisexual, 3=Homosexual, 4=Asexual, 5=Other
        - race : "What is your race?", 10=Asian, 20=Arab, 30=Black, 40=Indigenous Australian, 50=Native American, 60=White, 70=Other
        - married: "What is your marital status?", 1=Never married, 2=Currently married, 3=Previously married
    - To see other information that was recorded for each individual, visit the codebook (attached in the zip file). The demographic variables included above are explored in the "Data Visualization" section of this project.
- Other variables:
    - The following features were included in this dataset but were excluded from classifications: country, screenw, screenh

For additional information on each of these features, refer to “codebook.txt” in the zipped file.

The following figure contains distributions of the demographic variables in the dataset. As demonstrated by this figure, this dataset includes a wide variety of individuals from different demographics:

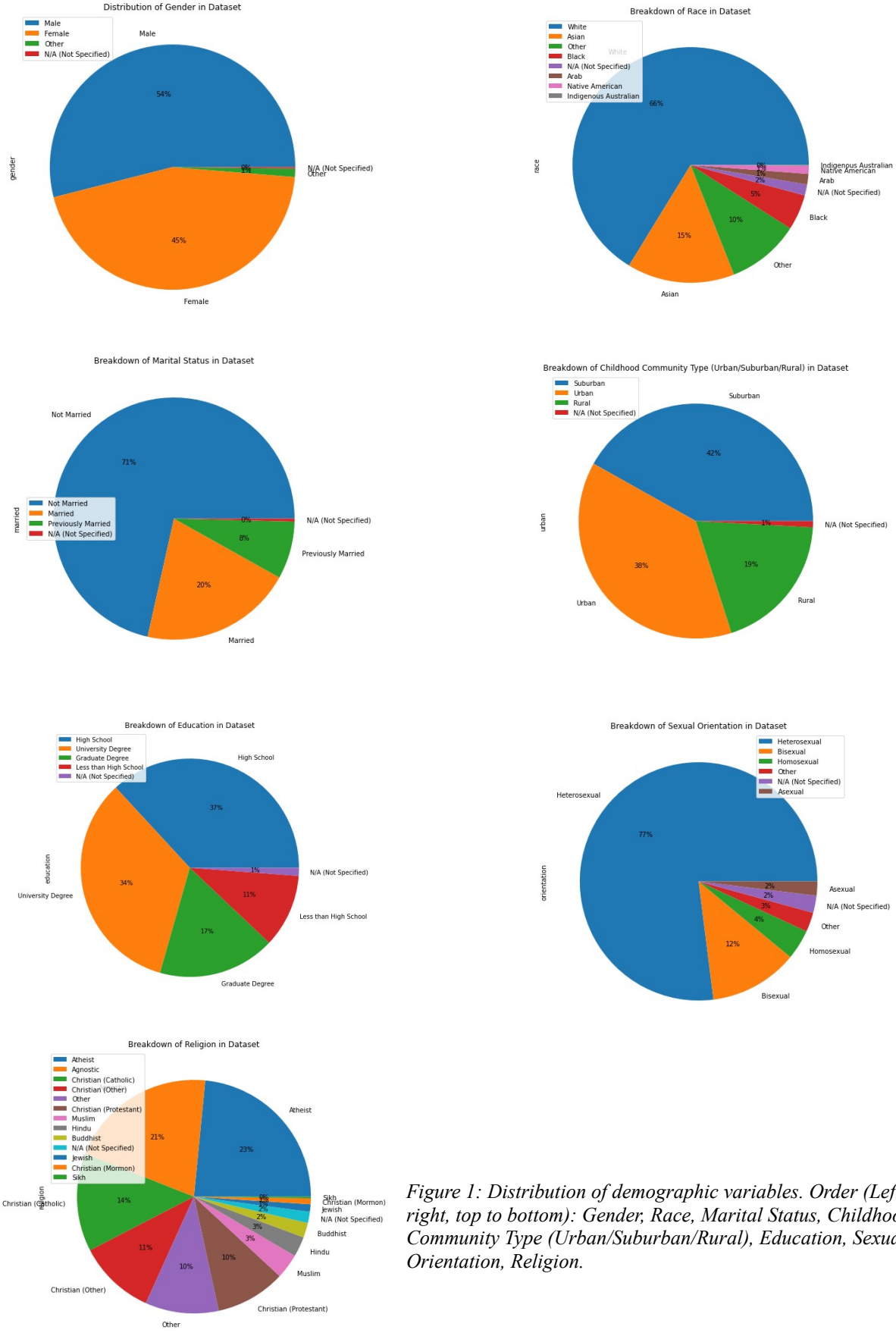

Figure 1: Distribution of demographic variables. Order (Left to right, top to bottom): Gender, Race, Marital Status, Childhood Community Type (Urban/Suburban/Rural), Education, Sexual Orientation, Religion.

2.2 Methods

I use the Mach-IV Features and the Ten Item Personality Inventory (Gosling 2003) as the independent variables (70 features) in my classifications and the demographic variables included in the distributions above (7 features). After removing rows with "NaN" and outlier values, I will commence the analysis by performing a general PCA on the data. This general PCA will output the number of principal components and the corresponding variance ratio. To demonstrate the effect of dimensionality reduction on accuracy and runtime, I will pick five principal component values that have different variance ratios. From here, I will use the five principal component values to conduct PCAs as well as NB and kNN classifications on the data. Since there are 7 different features that must be classified, 14 different classifications will be performed on the Mach-IV and TIP data. After conducting the analyses, the runtime for the PCA and all 14 analyses will be stored. Since this number will vary due to processor efficiency, this runtime value will be an average of five iterations. The average accuracy for each classification will be stored for each principal component as well.

3. Results.

Upon conducting a general PCA on the Mach-IV and TIP data, the following ratios were uncovered (in Figure 2):

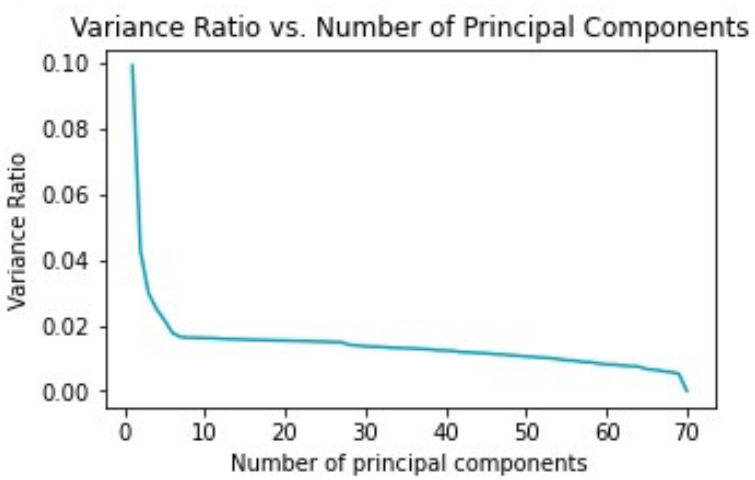

Figure 2: Variance ratio with respect to number of principal components.

Upon obtaining these results, I chose 5 principal component values to use in subsequent analyses: 1, 2, 6, 40, and 70 (all) principal components. These principal components, with respect to their variance ratios, are depicted graphically in Figure 3.

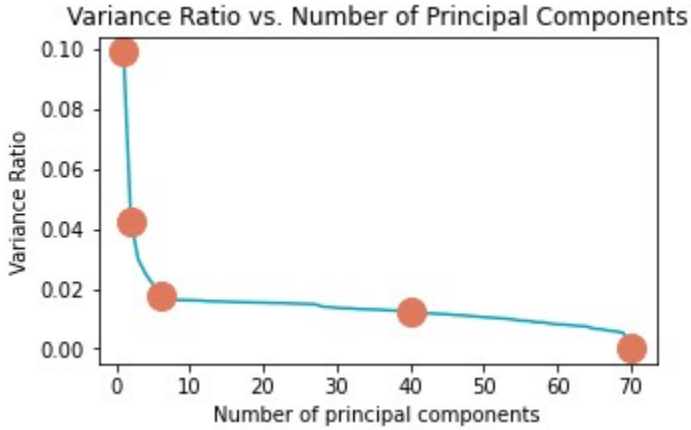

Figure 3: Variance ratio with respect to number of principal components as well as selected principal components for analysis. As depicted in this graph, the variance ratios for these components are relatively distinct.

After choosing these 5 principal components, I used them to perform dimensionality reduction and classifications for each dependent feature using the NB and KNN algorithm. The following results were obtained from these analyses (Table 1):

Accuracy scores and runtime for classifications by number of principal components.

n of principal components	Variance Ratio	Average NB accuracy score	Average KNN accuracy score	Runtime (in seconds)
1	0.0992324	0.5079	0.4577	7.52
2	0.04267516	0.5068	0.4617	8.28
6	0.01765231	0.5179	0.4696	15.88
40	0.0123412	0.4955	0.4733	987.21
70	6.0098e-33	0.1466	0.4769	1550.99

Table 1: Variance ratios, accuracy scores and runtime for classifications by number of principal components. NOTE: If you run these scripts, the accuracy scores for the NB and KNN algorithms might vary slightly from the accuracies presented in the table and the runtime will vary based on the power of your processor, any background processes running on your computer, etc.

4. Conclusion

Upon obtaining the accuracy scores and runtime for classifications by number of principal components, it is clear that dimensionality reduction is effective at reducing runtimes while preserving accuracy. In fact, the KNN accuracy score for 1 principal component is only 0.0192 or 1.92% less than the accuracy score for all features of the dataset. For this reason, performing a PCA can preserve the variance of the features in a dataset, meaning that classification accuracy is maintained between principal components.

The increased NB accuracy for 1 principal component seems unusual, as less variance is preserved with lower numbers of principal components. This phenomena can be explained by PCA's ability to reduce unnecessary variance or noise, especially in datasets with extremely interdependent features. (Bailey 2012) Since PCA can eliminate this noise, we see that 1 principal component's accuracy score is 0.3613 units higher than 70. In this way, PCA's reduction in variance can even boost accuracy results, especially in classifiers, such as NB, that work best with independent features.

In addition to the preservation of accuracy, the runtime of each analysis decreased with the number of principal components. As shown in Table 1, using 1 principal component reduces the runtime of the 14 classifications by 1543.48 seconds or 99.5%! Although a reduction of 26

minutes does not seem like a notable accomplishment, with larger amounts of data, this reduction in runtime is drastic. If this dataset contained millions of instances rather than several tens of thousands, this reduction in runtime would be closer to hours or days. For this reason, utilizing PCA to preserve classification accuracy and reduce runtime is successful; it can be used in cases where runtime is limited.

5. Bibliography

Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression".The American Statistician. 46 (3): 175–185.

Bailey, S. (2012). Principal Component Analysis with Noisy and/or Missing Data. Publications of the Astronomical Society of the Pacific, 124(919), 1015-1023. doi:10.1086/668105

Christie, R. & Geis, F. (1970) "Studies in Machiavellianism". NY: Academic Press.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A Very Brief Measure of the Big Five Personality Domains. Journal of Research in Personality, 37, 504-528.

Jones, D. N., & Paulhus, D. L. (2009). Machiavellianism. In M. R. Leary & R. H. Hoyle (Eds.), Handbook of individual differences in social behavior (pp. 93–108). New York, NY, US: The Guilford Press.

Open Psychology Data: Raw Data from online personality tests. (2019, March 21). Open-Source Psychometrics Project. https://openpsychometrics.org/_rawdata/.

Shanbhag, D. N., & Rao, C. R. (2003). Handbook of statistics. Elsevier.