Predicting benefit of psychedelic integration

Deborah Young

DSC 630 T301: Predictive Analytics

Term Project

Introduction:

Getting the most out of life is a noble goal for any person, but, in an age of disconnection, what avenues are available to us for improving our experience? Psychedelic integration in lifestyle is not a new development, but the effort to integrate them for legal medical practice in the US is gaining momentum. (*NBC News*, 2017)

The current-day exploration by medical professionals to utilize these substances for improving quality of life holds promising outcomes (MAPS Bulletin, 2022)

- Mental Health: Medical professionals are increasingly looking at psychedelics to treat conditions like PTSD, depression, and anxiety. Clinical trials using substances like psilocybin and MDMA have shown promising results.
- Spiritual Growth: For those seeking spiritual enlightenment or a deeper connection with themselves, guided psychedelic experiences can foster these connections.

The dataset I have chosen to utilize is from *An Exploration of Naturalistic Psychedelic Use*.

- The data was collected for the author's Master thesis "investigating the long-term effects of psychedelic use." (*An Exploration of Naturalistic Psychedelic Use*, 2023)
- "Data was collected using an online survey (N = 603) and covers a range of factors related to psychedelic use. These include questions related to demographics, drug use history (frequency and intensity of use), purposes for using psychedelics, long-term effects, as well as challenging, negative, and positive psychedelic experiences." (*An Exploration of Naturalistic Psychedelic Use*, 2023)

I am attempting to support the conclusions made by the author – that psychedelics are a beneficial integration for many people – and provide a machine learning framework that can hopefully be used in the future to predict a positive outcome for individuals based on their characteristics.

Predictive Analysis Steps:

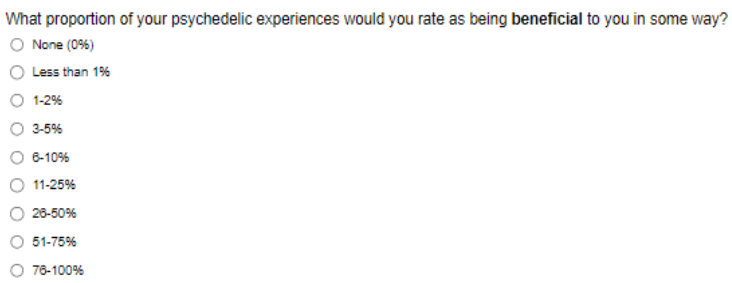
Data Exploration and Cleaning:

- Any duplicate rows are identified and removed.
- Columns with more than 5% missing values are removed to reduce dimensionality.

Creation of a DataFrame for Analysis:

- Find and select columns with categorical data by selecting columns with repeating values
- Review and remove non-ordinal variables

Modeling for All Classes of Target Variable:



Data Cleaning and Preparation:

- A value of "None (0%)" in the target column "Q88" is replaced with "Less than 1%" to avoid classes with only one value.
- The target variable (with all classes included) is separated and converted to a factor, while the remaining data is stored as feature data.
- Predictor columns are converted to ordered factors.

Feature Engineering & Selection:

- One hot encoding is applied to the feature data to create dummy variables.
- Final data is constructed, combining the target variable with the encoded features.
- Missing values are identified and omitted
- An attempt to perform Recursive Feature Elimination (RFE) was made but found to be too computationally intensive.
- A correlation matrix is used to identify and remove highly correlated features.

Modeling and Evaluation:

- *Support Vector Machine (SVM)*: Model is trained and evaluated with a confusion matrix.
  - Accuracy: 73%
- Random Forest model: rained and evaluated with a confusion matrix.
  - Accuracy: 75%

Modeling for specified (binary) class of Target Variable:

In the given text, the process of binary classification modeling is being described, specifically focusing on predicting whether a variable "Q88" is in the 76-100% range. Here's a summary of the process:

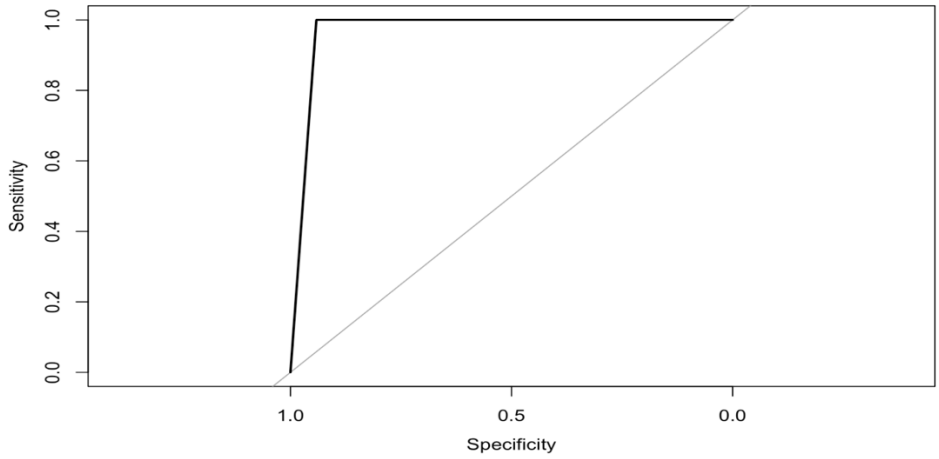
Feature Engineering & Selection:

- Creation of a binary target variable that captures whether "Q88" falls within the "76-100%" range.
- One hot encoding is applied to the feature data to create dummy variables.
- Final data is constructed, combining the target variable with the encoded features.
- Missing values are identified and omitted
- An attempt to perform Recursive Feature Elimination (RFE) was made but found to be too computationally intensive.
- A correlation matrix is used to identify and remove highly correlated features.

Model Building and Evaluation:

- *SVM Model*: An SVM model was trained, evaluated, with a confusion matrix.
  - Accuracy: 71%
- *KNN Model:* A k-Nearest Neighbors model was trained with k=5, and its accuracy and ROC curve were evaluated.
  - Accuracy: 71%
- *Random Forest Model:* A Random Forest model was trained, predictions were made, and an ROC curve was plotted.
  - Accuracy: 98.32%
  - Sensitivity: 94.12% (true positive rate)
  - Specificity: 100% (true negative rate)
  - Positive Predictive Value: 100% (precision)
  - Negative Predictive Value: 97.70%
  - Balanced Accuracy: 97.06%

- The Receiver Operating Characteristic (ROC) curve is again evaluated, with an Area Under the Curve (AUC) score of 0.9706.



Conclusion:

This result gives me hope that, in future implementation of psychedelic use for mental health, a machine learning model may be able to predict the likelihood of a positive outcome for an individual (or raise caution for those that may not benefit). The potential capability to predict benefit and/or the lasting impact on a person's health will be extremely useful in personalized medicine. The accuracy of the Random Forest model demonstrates that using information based on an individual's lifestyle, mindset, health status, and experiences, we may be able to predict the benefit of integrating psychedelic use. This information can support the use of psychedelics for spiritual growth, inform decriminalization and legalization for naturalistic users, and make a path for possible inclusion of psychedelics in the medical field.

References:

A Short Introduction to the caret Package. (n.d.). Retrieved August 9, 2023, from https://cran.r-project.org/web/packages/caret/vignettes/caret.html

An Exploration of Naturalistic Psychedelic use. (2023). figshare. https://doi.org/10.6084/m9.figshare.21708044.v1

Bento, C. (2018, December 3). *K-Means in Real Life: Clustering Workout Sessions*. Medium. https://towardsdatascience.com/k-means-in-real-life-clustering-workout-sessions-119946f9e8dd

Bhalla, D. (n.d.). Random Forest in R. *ListenData*. Retrieved August 9, 2023, from https://www.listendata.com/2014/11/random-forest-with-r.html

Gandhi, R. (2018, July 5). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Medium. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

KNN Algorithm | KNN In R | KNN Algorithm Example. (n.d.). Retrieved August 9, 2023, from https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-programming/

MAPS Bulletin 2022: Vol 32, No. 2. (2022, November 3). Multidisciplinary Association for Psychedelic Studies - MAPS. https://maps.org/news/bulletin/maps-bulletin-2022-vol-32-no-2/

NBC News: Will Psychedelic Therapy Transform Mental Health Care? (2017, September 29). Multidisciplinary Association for Psychedelic Studies - MAPS. https://maps.org/news/media/will-psychedelic-therapy-transform-mental-health-care/

Top 5 Predictive Analytics Models and Algorithms - insightsoftware. (n.d.). Retrieved June 18, 2023, from https://insightsoftware.com/blog/top-5-predictive-analytics-models-and-algorithms/

Understanding K-means Clustering with Examples. (2014, July 25). *Edureka*. https://www.edureka.co/blog/k-means-clustering/