Machine learning Project : Addiction Test Report

Code:

https://github.com/amineidel1/ADDICTEST

Realized By:

EL MOUTTAKI Adnan
IDELHAJ Amine
ZAHRAN Anouar

Supervised By:

CAZABET Rémy

18/12/2023

Table Of Contents

1. Introduction

This project, aligning with research efforts aimed at addressing complex issues, aims to simulate reality through a predictive system for potential addictions.

ADDICTEST is a solution designed to tackle emerging challenges in the healthcare sector and meet the needs of university students requiring addiction prevention. It predicts the risk of developing an addiction to various harmful substances by leveraging advanced machine learning models.

The project tasks are divided among the team as follows:

- **Anwar Zahran:** Data Cleaning & Data Encoding - Web App Back-End Development
- **Adnan El Mouttaki:** General Data Analysis - Web App Front-End Development
- **Amine Idelhaj:** Drug Use Prediction & Graph - Linking the models with the App

2. Data Cleaning

a. Initial Data Overview:

The database contains records for 1885 people. For each person, 12 attributes are known:

- **Personality measures:** NEO-FFI-R (neuroticism, extraversion, openness to experience, friendliness and conscientiousness), BIS-11 (impulsivity) and ImpSS (sensation seeking)
- **Personal information:** level of education, age, gender, country of residence and ethnic origin.
  In addition, participants were asked about their use of 18 legal and illegal drugs.
- **Legal drugs:** alcohol, benzodiazepine, nicotine, caffeine, chocolate.
- **Illegal drugs:** amphetamines, amyl nitrite, cannabis, cocaine, crack, ecstasy, heroin, ketamine, legalh, LSD, methadone, Mushroom, volatile substance abuse.
- **Fictitious drug:** Semeron - introduced to identify overclaimers.
  For each drug, they had to select one of the following answers: never used, used it more than a decade ago or over the course of a decade, more than a decade ago or in the last decade, year, month, week or day. **(See Table 1 in Annexe)**

b. Data filtering

Our data cleaning process involved removing entries related to a fictitious drug, ensuring that the analysis focuses on genuine drug consumption data.
As so 8 records were identified as unwanted, and subsequently they were removed, which keeps our dataset with 1876 records.**(See Table 1 in Annexe)**

3. Data Encoding

- The 'Gender' column has been encoded, with 'M' mapped to 1 and 'F' mapped to 0, facilitating numerical representation for gender.
- Ordinal features such as 'Age,' 'Education,' and various drug consumption categories (e.g., 'Alcohol,' 'Cannabis'), ordinal encoding has been performed, which involves mapping categorical values to their corresponding positions in predefined orderings, ensuring a meaningful numerical representation.
- Nominal features like 'Country' and 'Ethnicity' have been encoded using categorical encoding, assigning each category a unique integer.

➤The resulting dataset has undergone a type conversion to ensure uniformity, with all features now represented as **float64**.

➤This encoding process is crucial for our machine learning algorithms, facilitating the exploration and analysis of relationships within the data.**(See Table 2 in Annexe)**

4. General data analysis

a. Network Analysis of Drug Consumption Habits:

One notable aspect of the analysis is the application of network analysis to represent similarities in drug consumption habits among individuals.
Each person is depicted as a node, and edges connect individuals with similar drug consumption patterns.
The calculation of similarity utilized the Hamming distance for categorical data, providing a robust metric for assessing similarities in drug use.

b. Visualization of Drug Consumption Network:

The resulting network graph was visualized using the NetworkX library and Matplotlib. This graphical representation allowed for a comprehensive view of clusters and connections within the dataset, emphasizing individuals with significant similarities in drug consumption behaviors.
Adjusting the similarity threshold provides flexibility in exploring different levels of connections.**(See Graph 1 in Annexe)**

c. Psychological Correlations: **(See Graph 10 in Annexe)**

The analysis delved into the psychological traits of individuals, revealing intriguing correlations between personality scores and drug consumption.
Notably, the Sensation Seeking (SS) score emerged as a key factor, displaying positive correlations with various legal and illegal drug consumption.
This finding suggests that SS could be a robust personality indicator for predicting drug use tendencies.**(See Graph 6 and 9 in Annexe)**

d. Insights from Legal Drug Consumption:

The analysis of legal drug consumption, particularly alcohol and nicotine, uncovered correlations with demographic and psychological factors.
For instance, alcohol consumption exhibited a positive correlation with educational levels, indicating that individuals with higher education tend to consume alcohol more frequently.**(See Graph 3, 4 and 5 in Annexe)**

e. In-Depth Analysis of Illegal Drug Consumption:

The analysis delved into specific illegal drugs, such as cannabis, ecstasy, and cocaine. The findings highlighted gender and age disparities in drug consumption patterns. Notably, the frequency of cannabis use was negatively correlated with age, indicating higher consumption among younger individuals.**(See Graph 11 in Annexe)**

f. Psychological Traits and Drug Use:

Correlations between psychological traits and drug consumption were thoroughly explored.
The SS score exhibited consistent positive correlations with drug use across various substances, suggesting its potential as a predictive factor for susceptibility to drug consumption.**(See Graph 10, 19 and 20 in Annexe)**

g. Final Reflections:

This comprehensive analysis provides valuable insights into the complex interplay between personal characteristics, psychological traits, and drug consumption habits. The identified patterns and correlations can contribute to a deeper understanding of risk factors for substance abuse, plus the inclusion of network analysis adds a novel dimension, uncovering hidden connections among individuals with similar drug consumption behaviors.

5. Drug use prediction

a. Feature Engineering:

Low correlation features, including 'Age', 'Gender', 'Education', and 'AScore', were identified and subsequently removed from the dataset.
The elimination of these features was executed across the entire dataset and entailed updating the list of personal information attributes accordingly.
Additionally, categorical drug usage labels were encoded into a binary format, facilitating the subsequent classification tasks.

b. Model Prediction:

A comprehensive analysis employing 4 distinct classification models was conducted to predict drug usage based on personal information.

- **Naive Bayes:**
  A Gaussian Naive Bayes model was applied to predict drug usage for each substance.

Subsequent evaluation metrics, encompassing the Accuracy Score and F1 Score, were computed and reported for each drug category.**(See Figure 1 in Annexe)**

- **Random Forest Classifier:**
  The Random Forest Classifier was employed to predict drug usage, and its performance was assessed using established metrics.**(See Figure 2 in Annexe)**
- **SVM Classifier:**
  The Support Vector Machine (SVM) Classifier was utilized for drug usage prediction, with subsequent evaluation metrics presented for each substance.**(See Figure 3 in Annexe)**
- **Logistic Regression:**
  Logistic Regression served as the final classification model, and performance metrics were reported to gauge its efficacy in predicting drug usage.
  A grid search, leveraging the GridSearchCV method, was conducted to identify optimal hyperparameters for the Logistic Regression model.**(See Figure 4 and 5 in Annexe)**

c. Conclusion:

In summary, the application of multiple classification models, enabled a comprehensive analysis of drug usage prediction.
The evaluation metrics provided insights into the efficacy of each model, with the optimized Logistic Regression model standing out as the best estimator.

6. Graph

Step 1: Calculate Similarity Between Individuals

In the initial stage of our analysis, we focus on determining the similarity between various individuals. This is achieved by employing a method that transforms differences in their attributes into a quantifiable similarity score. These scores are pivotal in understanding the degree of resemblance between each pair of individuals in our study.

Step 2: Create a Graph

Using these calculated similarities, we then construct a graph. This graph is a visual and analytical representation of our data, where individuals are represented as nodes. Connections, or edges, between these nodes are established based on their similarity scores. We set a predefined threshold for similarity, ensuring that only significantly similar individuals are connected in our graph. This threshold is adjustable and can be tailored according to the specific needs of the analysis.**(See the code)**

Step 3: Visualize the Graph

The final step involves visualizing this graph. This visualization is not only crucial for a better aesthetic understanding of our data but also provides insights into

the complex network of relationships between the individuals based on their similarity. The layout and design of this graph are chosen to best reflect the underlying structure and connections within the data.**(See Graph 1 in Annexe)**

Additional Analysis with Gephi:

To enhance our analysis, we also utilize Gephi, an advanced tool for network visualization and analysis. Gephi offers a range of features that allow for deeper exploration and understanding of network data. Its capabilities in producing detailed and interactive visualizations complement our initial analysis, providing a more comprehensive view of the relationships and patterns present in our data. In this approach, each person is represented as a node, and edges are added between individuals who exhibit a high similarity in their drug consumption habits.**(See Gephi file)**

7. Implementation Web-Application (With Django)

Our Django web application is structured into three main pages, providing a streamlined and intuitive user experience:

Home Page:

- The Home page serves as the landing page of our application. It provides users with an overview of the project, including its purpose, the methodologies used, and a brief guide on how to navigate the application.
- This page sets the stage for the user journey, ensuring that users are well-informed before they proceed to the simulation.

Simulation Page:

- On the Simulation page, users can interact with the models.
- Users are prompted to enter relevant data or select specific criteria that the models will use to predict drug consumption patterns.
- This page is designed to be user-friendly, guiding the user through the necessary steps to provide the input required for accurate predictions.

Results Page:

- After the user submits their data, the application processes this input using the appropriate pickle-loaded models.
- The Results page then displays the predictions of drug usage patterns based on the user's input.
- This page is crucial for providing personalized insights and interpretations of the predictive analysis, offering users a clear understanding of the results.

Integration of Machine Learning Models:

- The integration of machine learning models into the Django application is a key feature of our project.
- When a user submits their data on the Simulation page, the Django backend retrieves the appropriate pickled model and uses it to make predictions.
- This seamless integration ensures that the predictions are not only accurate but also generated in real-time, providing immediate feedback to the user on the Results page.

Visualization: