Project Report Summary: Health Prediction from Smoking and Drinking Habits

This project focuses on predicting health outcomes based on individuals' smoking and drinking habits, by body signalsemploying a multi-faceted approach by data cleaning, outlier detection, exploratory data visualization, and machine learning classification.

Data Cleaning

The initial phase involved data cleaning to ensure the dataset's integrity and reliability. Handling missing values and addressing any inconsistencies set the foundation for subsequent analyses.

Outlier Detection:

Outliers, potentially skewing the predictive model, were identified and managed using robust outlier detection techniques. This step aimed to enhance the model's accuracy by mitigating the impact of anomalous data points.

Data Visualization:

Exploratory data visualization was conducted to glean insights into the distribution of health-related variables. Density plots were particularly informative, providing a nuanced understanding of the relationships between various features and health indicators. Comparative visualizations highlighted distinctions in smoking and drinking patterns across genders.

Machine Learning Classification:

Utilizing the scikit-learn library, an AdaBoostClassifier was employed to partition the data into training and testing sets. Subsequently, a LightGBM classifier was implemented for health prediction, offering a sophisticated solution for its efficiency and speed.

Confusion Matrix Analysis:

The predictive model's performance was assessed using a confusion matrix. Notably, for individuals who had never smoked, the model exhibited an impressive 85% accuracy. This outcome underscores the model's efficacy in identifying health patterns associated with non-smokers.

Conclusion:

In conclusion, this project showcases a comprehensive methodology encompassing data preparation, exploratory analysis, and machine learning classification. The achieved accuracy of 85% for non-smokers is promising, indicating the potential of the model in predicting health outcomes based on lifestyle choices. Future work may involve model refinement, feature engineering, and a deeper investigation into the interplay of additional factors. The project contributes to the growing body of knowledge at the intersection of data science and health, offering practical insights for personalized health interventions and public health strategies.