Project Objective

Introduction

The primary objective of this project was to predict alcohol consumption levels based on several features such as sleep time, income, emotional support, and gender. This prediction can assist in understanding behavioral patterns and providing insights into public health efforts.

Dataset Overview

The dataset comes from the Behavioral Risk Factor Surveillance System (BRFSS), which includes responses from adults about health-related behaviors. In this project, we focus on predicting alcohol consumption, represented by the variable DRNK3GE5. The features used in the analysis include:

SLEPTIM1: Hours of sleep
INCOME3: Income level
SEXVAR: Gender
EMTSUPRT: Emotional support status

Target Variable

The target variable, DRNK3GE5, represents alcohol consumption, measured on a scale of 1 to 76. The goal was to use the available features to predict this value.

Preprocessing

Data Cleaning

Handling Missing Values:

- The target variable DRNK3GE5 had no missing values.
- Missing values in SLEPTIM1 were imputed using the median.
- The EMTSUPRT variable had a significant amount of missing data (22,433 missing values). For simplicity, the "missing" category was created.
- INCOME3 had one missing value, which was imputed using the median of the column.

Feature Engineering:

- Binning: The SLEPTIM1 variable was categorized into bins to make the data easier to interpret and model.
- Encoding: The categorical feature SEXVAR was one-hot encoded to facilitate its use in machine learning models.

Feature Scaling:

- Numerical features like SLEPTIM1 and INCOME3 were scaled using StandardScaler to ensure equal weight across features in the models.

Methodology

Model Selection

Three different models to predict alcohol consumption:

- Linear Regression: A simple regression model to examine the linear relationship between features and the target variable.

- Random Forest Regressor: A powerful ensemble model that aggregates predictions from multiple decision trees.

- Gradient Boosting Regressor: Another ensemble method that builds decision trees sequentially, optimizing for errors made by previous trees.

Data Splitting

- The dataset was split into a training set (80%) and a test set (20%) using train_test_split from sklearn.

Results

Model Evaluation

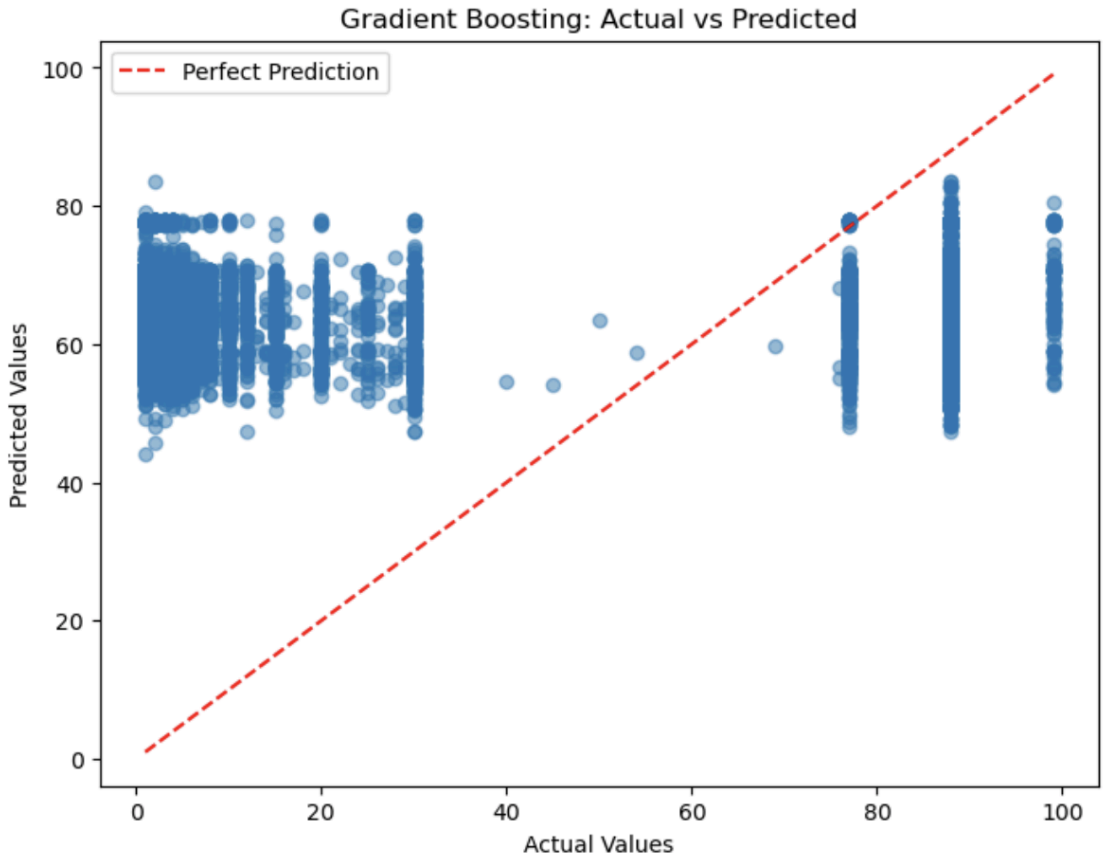
- The models were evaluated using Mean Squared Error (MSE) and R-squared (R2) metrics to determine their predictive accuracy.

Model	MSE	R2
Linear Regression	1356.43	0.014
Random Forest	1361.41	0.010
Gradient Boosting	1349.33	0.019

- Gradient Boosting provided the best performance with the lowest MSE and the highest R2.
- The models have relatively low R2 values, indicating that the features used do not fully explain the variability in alcohol consumption.

Visualizing Model Performance

For each model, the actual vs. predicted values were plotted to visually inspect the model's performance. Below is an example of the **Actual vs. Predicted** plot for the **Gradient Boosting** model.



Discussion

Insights

- Predicting Alcohol Consumption: The models were able to predict alcohol consumption with moderate accuracy. However, none of the models achieved high R2 values, suggesting that additional features or more sophisticated techniques may be required for a more accurate prediction.
- Importance of Features: While features like sleep time and income were included, other variables such as mental health or social influences might be crucial for better predictions.

Limitations

- Missing Data: A significant portion of the data was missing for EMTSUPRT, which may have impacted model performance.
- Feature Limitations: The features used in this analysis were somewhat limited in explaining alcohol consumption, which might explain the low R2 values.

Conclusion

Summary of Findings

- Best Performing Model: The Gradient Boosting Regressor achieved the lowest MSE and the highest R2, making it the best performing model.
- Model Improvement: The models could be improved with additional features, such as more detailed information on mental health, stress levels, or social support.
- Future Work: Future analyses could explore alternative models, better feature engineering, and more detailed data on the factors influencing alcohol consumption.

Recommendations

- Explore additional data sources or collect more granular data to improve prediction accuracy.
- Apply hyperparameter tuning for the models to see if performance can be enhanced further.

References

1. **BRFSS (Behavioral Risk Factor Surveillance System)**: [https://www.cdc.gov/brfss/annual_data/annual_2023.html].
2. **Sklearn Documentation**:[https://scikit-learn.org/stable/supervised_learning.html].