PREDICTING HEALTH HABITS

DRINKING BEHAVIOR PREDICTION

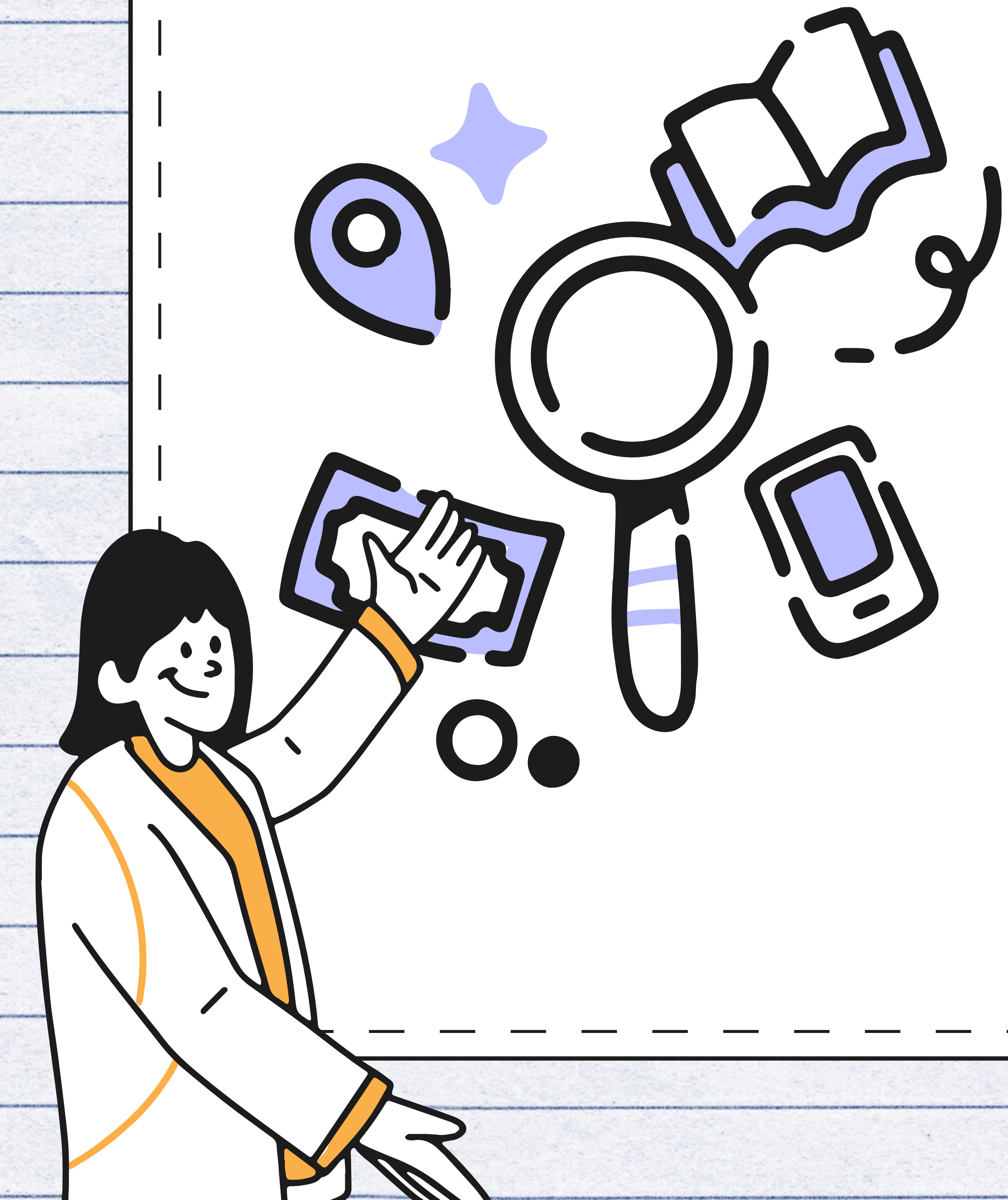
GROUP MEMBERS

01 Mohit Rajpurohit

02 Devanshu

01

PROBLEM STATEMENT



Introduction:

- In today’s health-driven world, understanding drinking habits is essential for better healthcare recommendations and lifestyle interventions.

Problem Statement:

- This project aims to predict whether an individual is a drinker based on medical test results and personal health data, using machine learning techniques.

02

DATASET DESCRIPTION

Source:

- Kaggle

Size:

- 50,000 rows
- 24 columns

Target Variable:

- is_drinker (0: Non-Drinker, 1: Drinker)

Sample Features:

- Age, Gender, Blood Pressure, Liver Enzymes, Cholesterol, etc.

03

DATA PREPROCESSING

Preprocessing Steps:

- Imported dataset and renamed columns for clarity
- Checked for null values and duplicates (none found)
- Removed ineffective columns after analysis

Feature Engineering:

- Added new features:
  - Liver_Enzyme_Ratio (ALT/AST)
  - Anemia_Indicator (based on hemoglobin levels)

Other Key Steps:

- Outlier treatment (IQR Method)
- Label Encoding (Categorical variables)
- Train-Test Split (80-20)
- Data Standardization (Standard Scaler)

04

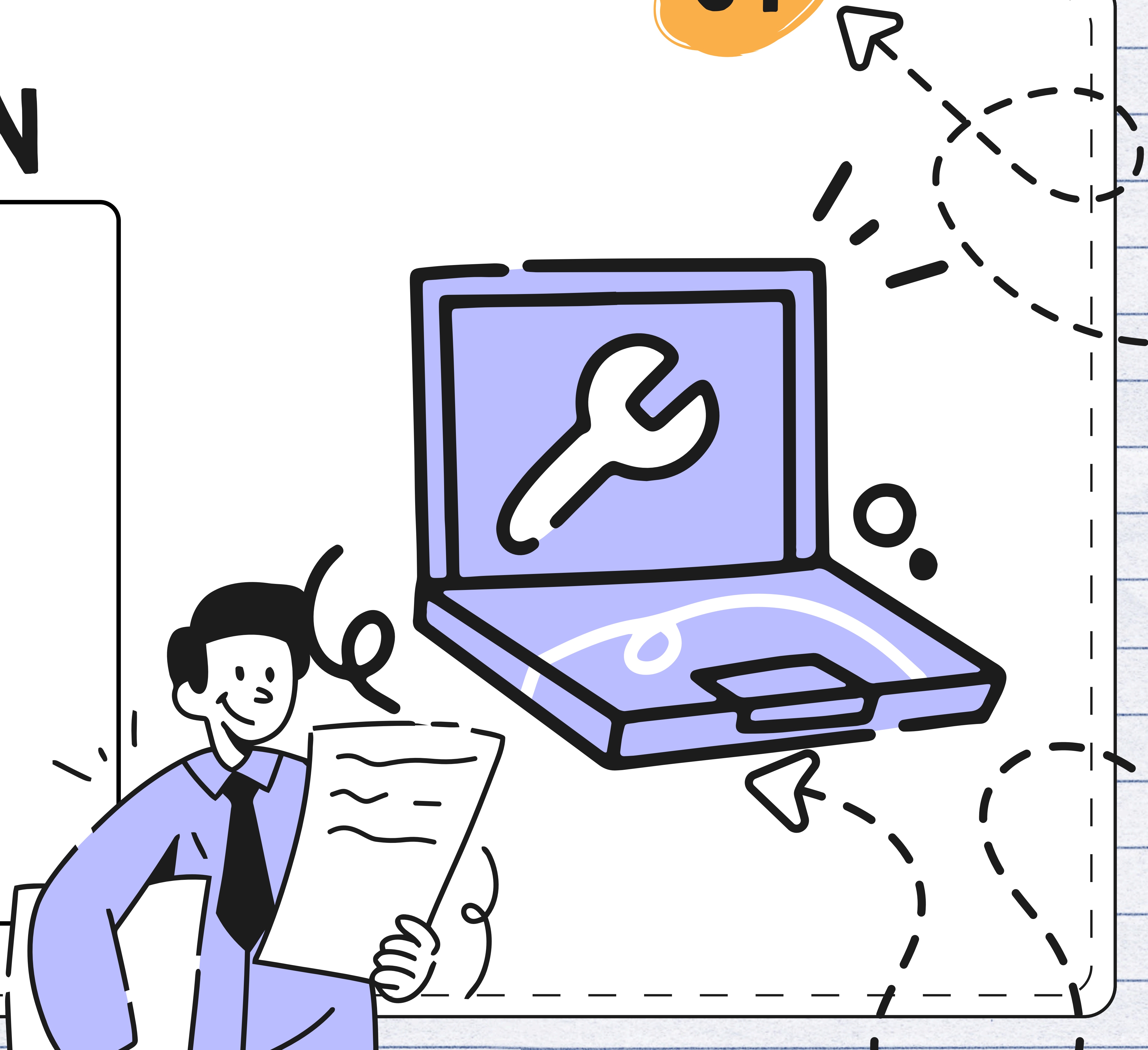
MODEL SELECTION

Models Used:

- Logistic Regression
- Random Forest Classifier
- XGBoost
- Gradient Boosting
- Stacking Classifier (ensemble)

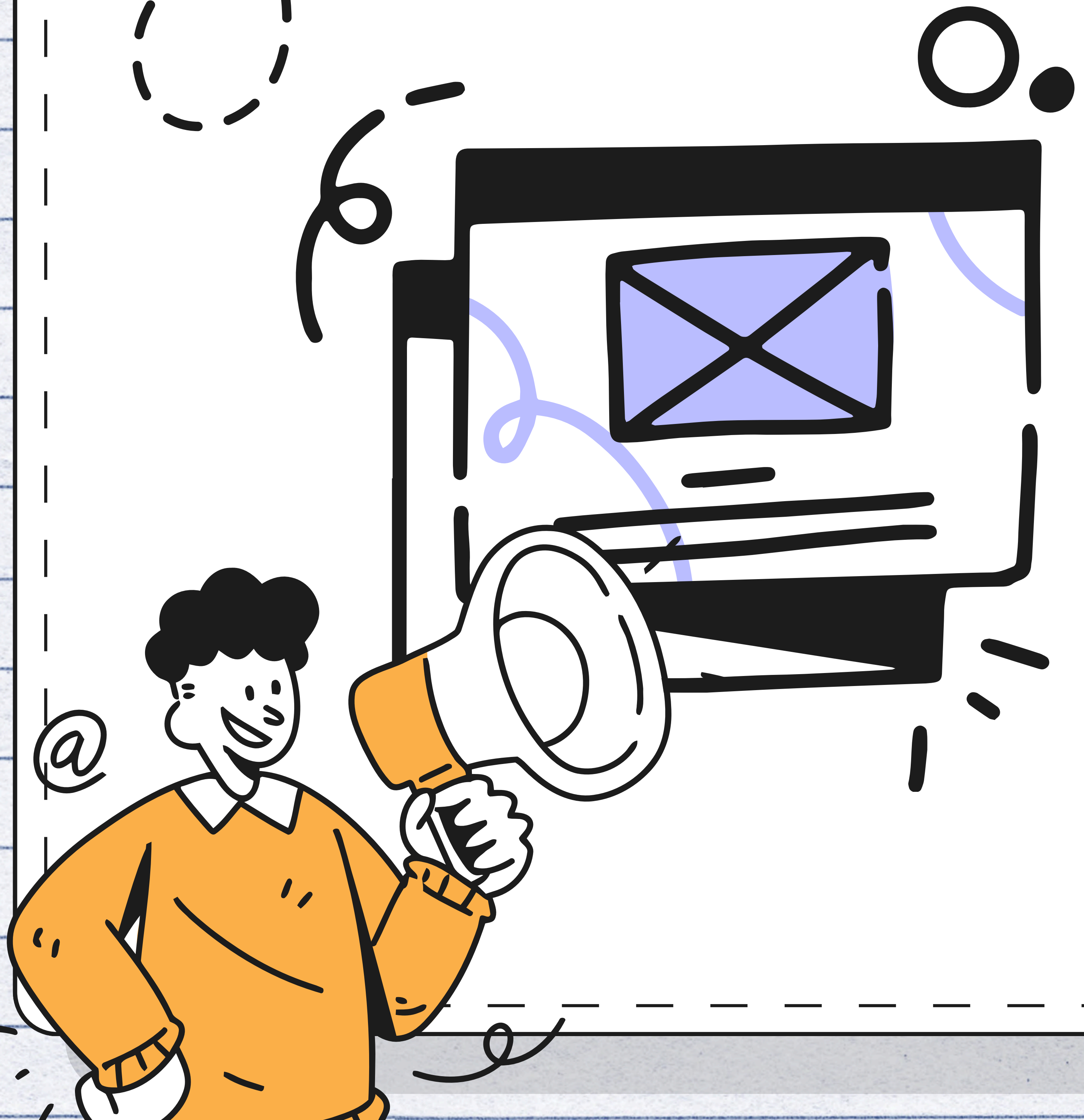
Training Strategy:

- Each model trained using standardized data
- Evaluated using accuracy and classification report



05

HYPERPARAMETER TUNING



Technique:

- RandomizedSearchCV

Purpose:

- To find the best set of hyperparameters to optimize model performance.

Example Tuned Parameters:

- Random Forest (n_estimators, max_depth)
- XGBoost (learning_rate, subsample)
- Gradient Boosting (n_estimators, max_depth)

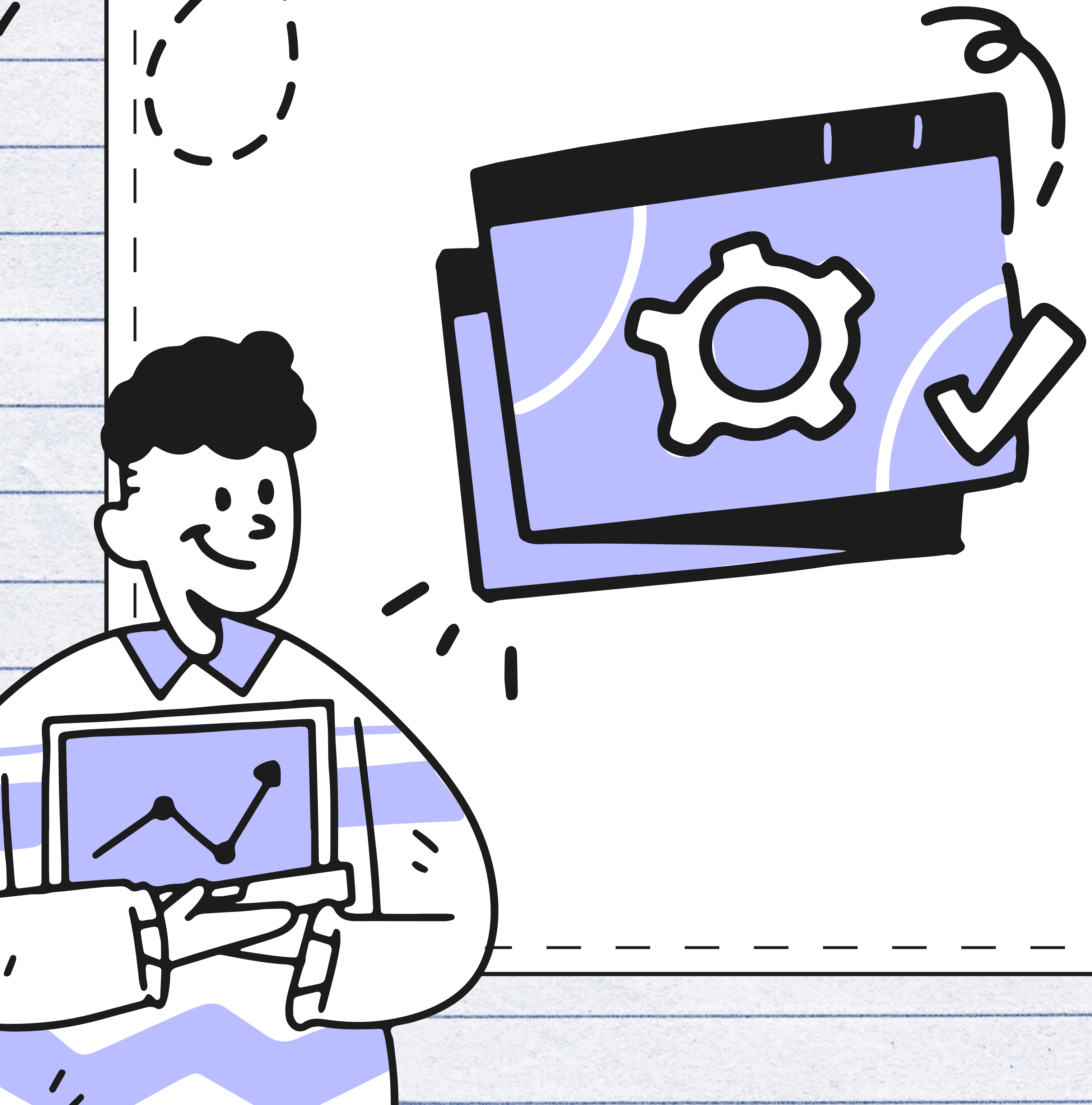
06

MODEL COMPARISON

Model	Train Accuracy	Test Accuracy
Logistic Regression	74.12%	72.01%
Random Forest	76.85%	73.25%
XGBoost	76.00%	73.10%
Gradient Boost	75.72%	72.90%
Stacking (Best)	76.30%	73.79%

07

EVALUATION



Train-Test Split:

- 80% Training
- 20% Testing

Validation:

- Cross-validation used to ensure model generalization

Model Submission:

- Final predictions generated using the best (stacked) model

Presentation:

- Clean and efficient workflow documented through Jupyter Notebook

08

INSIGHTS

Insights:

- Feature engineering enhanced model performance.
- Ensemble learning (Stacking) outperformed individual models by combining their strengths.

Conclusion:

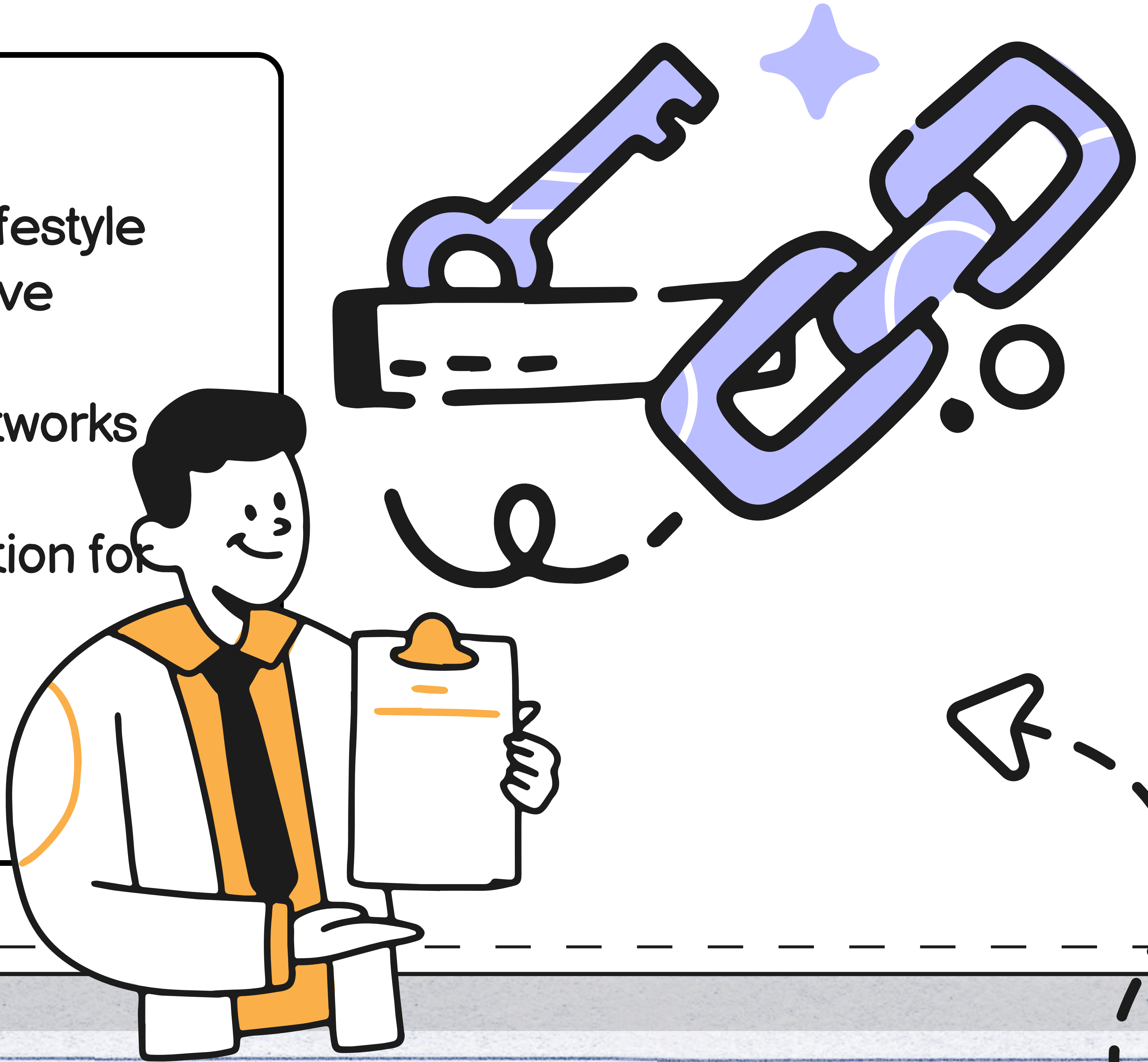
- Machine learning models can effectively predict drinking behavior based on medical data with good accuracy.
- Proper data preprocessing and model tuning are critical for maximizing performance.

9

FUTURE SCOPE

Future Scope:

- **Feature Enrichment:** Adding more lifestyle or psychological factors could improve prediction accuracy.
- **Deep Learning:** Exploring neural networks for capturing complex patterns.
- **Deployment:** Building a web application for real-time prediction.
- **Explainability:** Using SHAP values to understand feature impact better.



GITHUB LINK

Mohit: Link
Devanshu: Link

GRATEFUL FOR YOUR ATTENTION!!