Predicting Student Performance

An Analytical Approach

Introduction

- Overview of the project
- Importance of predicting student performance
- Brief summary of the dataset

Problem Description

Objective: Predict the students' final grades (G3)

Prediction can help in several ways:

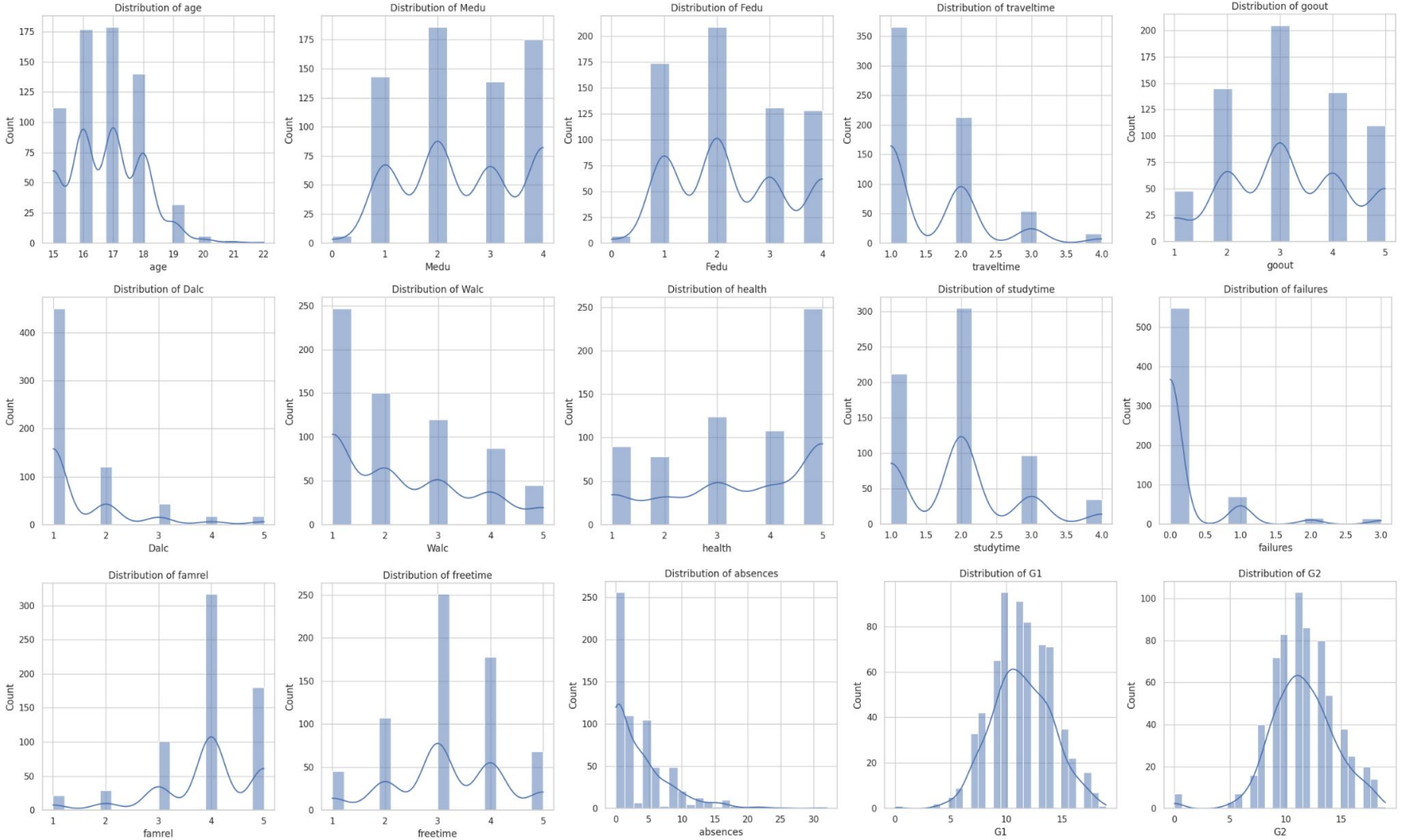
- Identifying At-Risk Students
- Resource Allocation
- Personalized Education Plans
- Policy Making
- Parental Engagement

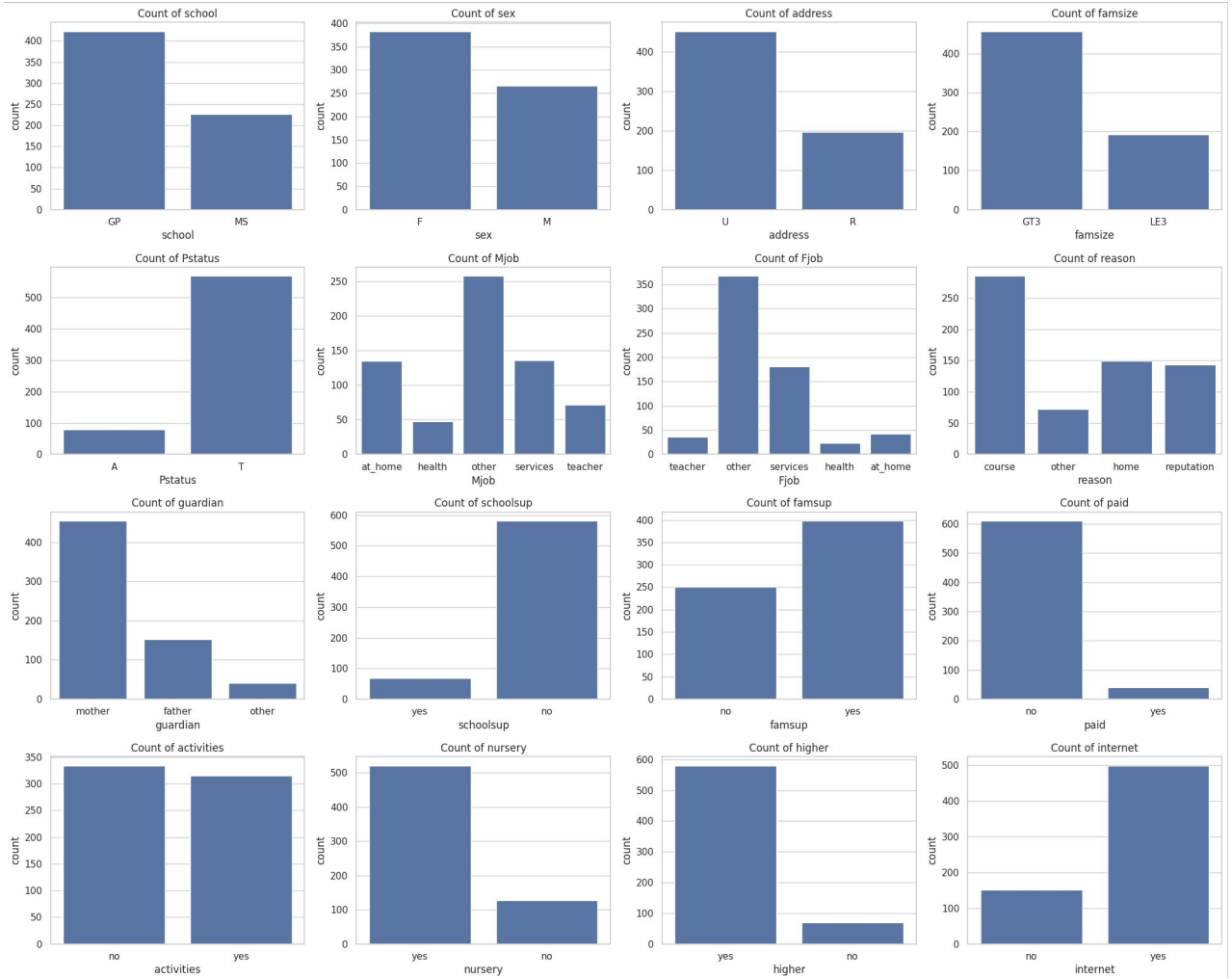
Dataset Overview

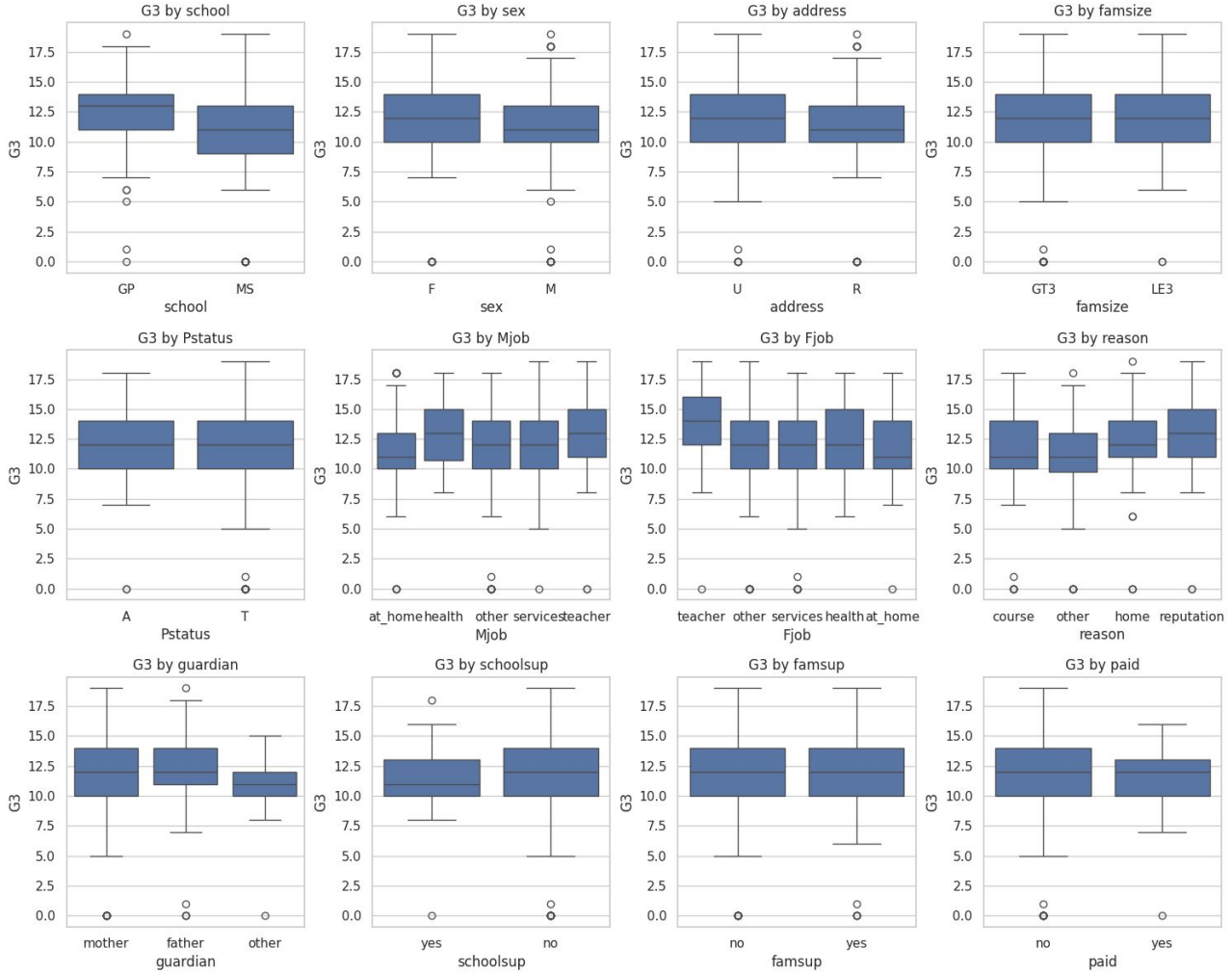
- **Demographics**: school, sex, age, address
- **Family Background**: famsize, Pstatus, Medu, Fedu, Mjob, Fjob, guardian
- **Educational Support**: traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic
- **Personal Information**: famrel, freetime, goout, Dalc, Walc, health, absences
- **Grades**: G1, G2, G3

EDA Procedure

- Visualize the Distribution of Numerical Features
- Visualize Categorical Features
- Feature Relationships with Target Variable


Distribution of age
Count
age
Distribution of Medu
Medu
Distribution of Fedu
Fedu
Distribution of traveltime
traveltime
Distribution of goout
goout
Distribution of Dalc
Dalc
Distribution of Walc
Walc
Distribution of health
health
Distribution of studytime
studytime
Distribution of failures
failures
Distribution of famrel
famrel
Distribution of freetime
freetime
Distribution of absences
absences
Distribution of G1
G1
Distribution of G2
G2


Count of school
count
GP
MS
school
Count of sex
F
M
sex
Count of address
U
R
address
Count of famsize
GT3
LE3
famsize
Count of Pstatus
A
T
Pstatus
Count of Mjob
at_home
health
other
services
teacher
Mjob
Count of Fjob
teacher
other
services
health
at_home
Fjob
Count of reason
course
other
home
reputation
reason
Count of guardian
mother
father
other
guardian
Count of schoolsup
yes
no
schoolsup
Count of famsup
no
yes
famsup
Count of paid
no
yes
paid
Count of activities
no
yes
activities
Count of nursery
yes
no
nursery
Count of higher
yes
no
higher
Count of internet
no
yes
internet


G3 by school
G3
school
GP
MS
G3 by sex
sex
F
M
G3 by address
address
U
R
G3 by famsize
famsize
GT3
LE3
G3 by Pstatus
Pstatus
A
T
G3 by Mjob
Mjob
at_home
health
other
services
teacher
G3 by Fjob
Fjob
teacher
other
services
health
at_home
G3 by reason
reason
course
other
home
reputation
G3 by guardian
guardian
mother
father
other
G3 by schoolsup
schoolsup
yes
no
G3 by famsup
famsup
no
yes
G3 by paid
paid
no
yes

Introduce the StudentPerformanceAnalyzer Class

- **Purpose**: Analyzes and predicts student performance.
- **Initialization**: Loads dataset and models.
- **Preprocessing**: Encodes data, splits into train/test, standardizes features.
- **Models**: Includes various classifiers (Logistic Regression, Decision Tree, SVM, etc.).
- **Training**: Trains models and evaluates metrics.
- **Evaluation**: Compares model performance.
- **Feature Importance**: Plots top features for selected models.

Model Selection

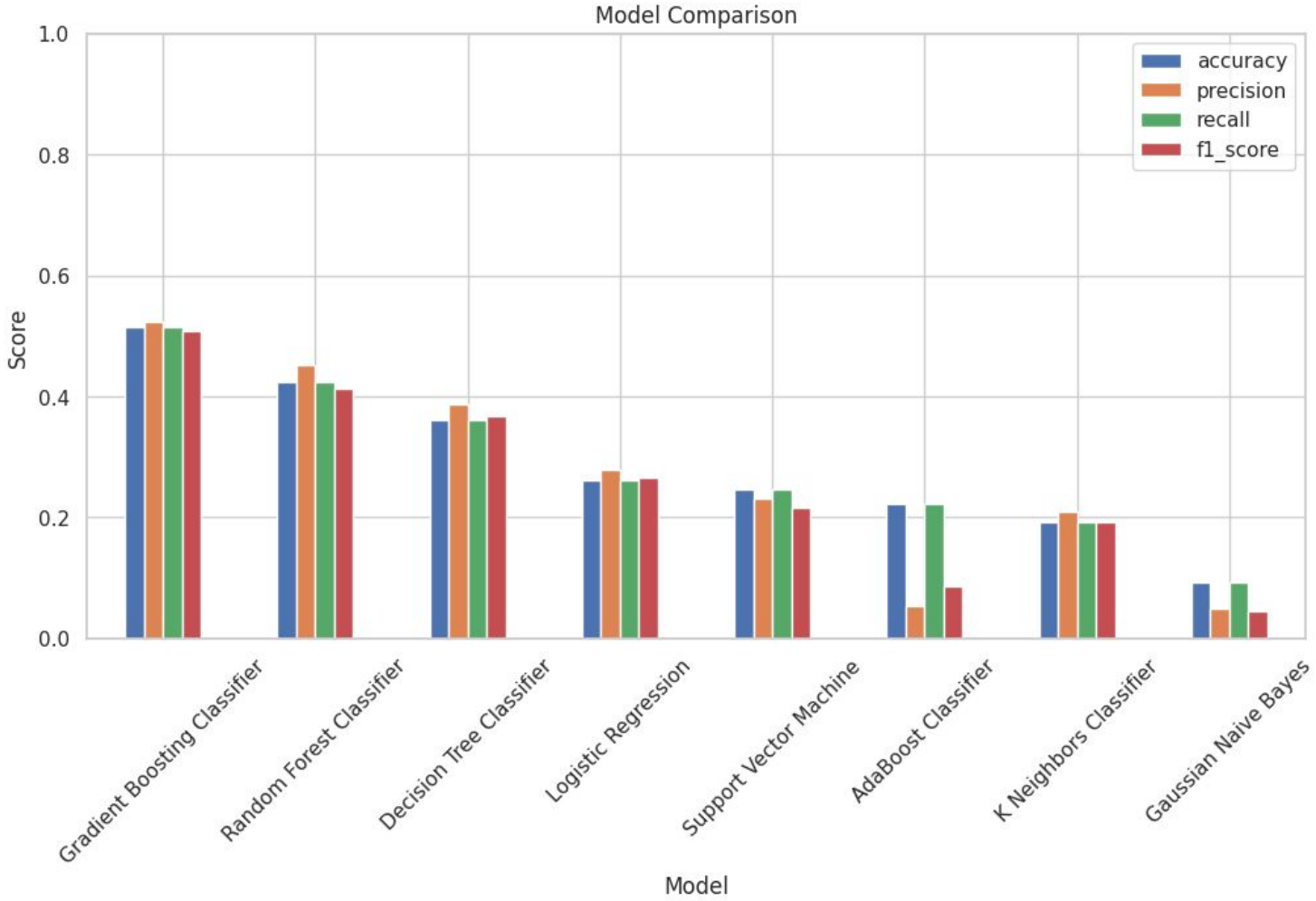
1. Logistic Regression
2. Decision Tree Classifier
3. Support Vector Machine
4. Random Forest Classifier
5. AdaBoost Classifier
6. Gradient Boosting Classifier
7. K Neighbors Classifier
8. Gaussian Naive Bayes

Model Training

- Train each model on training data
- Evaluate models on testing data
- Store evaluation metrics (accuracy, precision, recall, F1 score)

Model Evaluation

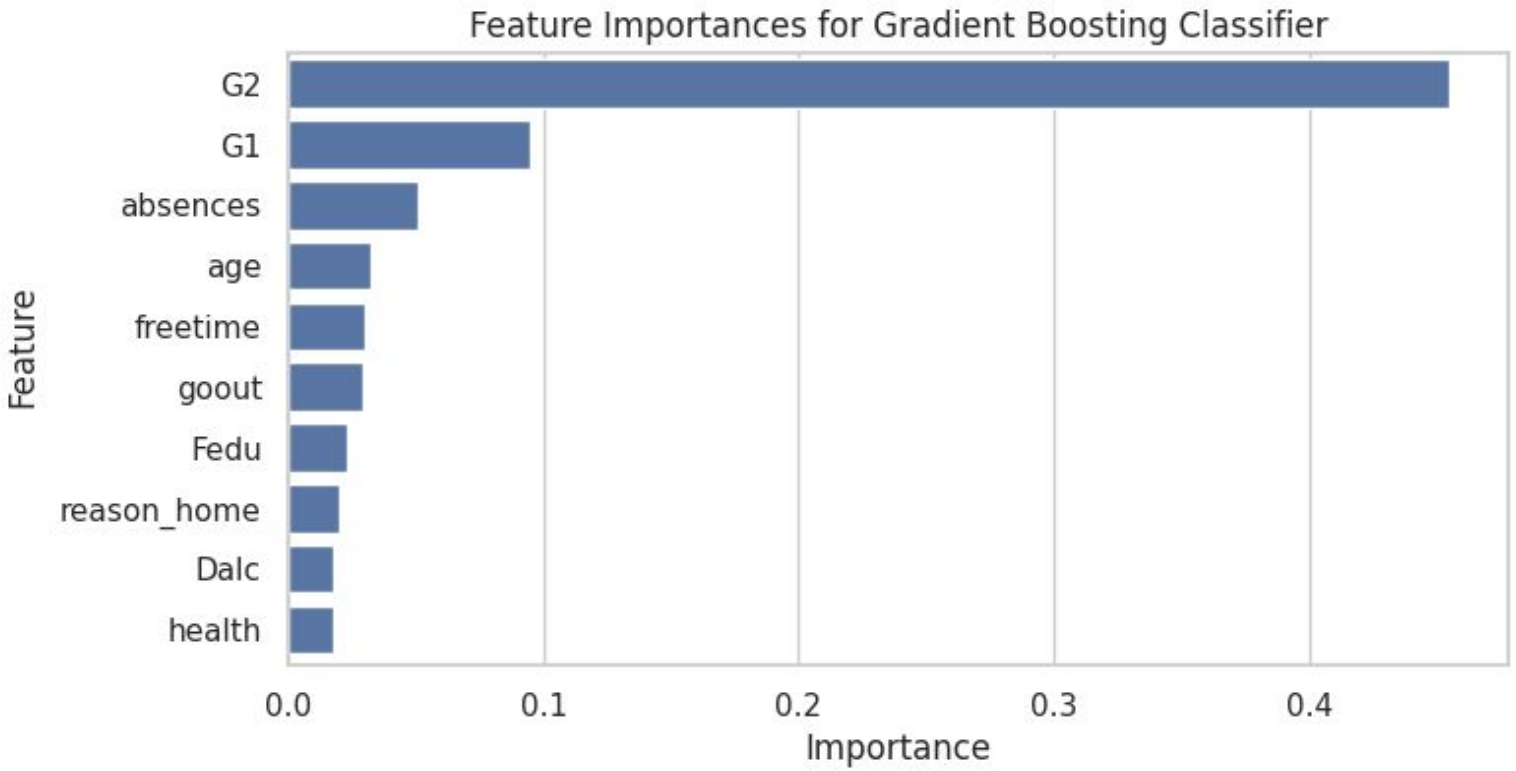
- Compare models based on evaluation metrics
- Highlight best performing models


Model Comparison
accuracy
precision
recall
f1_score
Score
1.0
0.8
0.6
0.4
0.2
0.0
Gradient Boosting Classifier
Random Forest Classifier
Decision Tree Classifier
Logistic Regression
Support Vector Machine
AdaBoost Classifier
K Neighbors Classifier
Gaussian Naive Bayes
Model

Model Comparison Table

Model	Accuracy	Precision	Recall	F1 Score
Gradient Boosting Classifier	0.515385	0.522385	0.515385	0.508159
Random Forest Classifier	0.446154	0.462195	0.446154	0.420462
Decision Tree Classifier	0.384615	0.411134	0.384615	0.388698
Logistic Regression	0.261538	0.279706	0.261538	0.266291
Support Vector Machine	0.246154	0.231989	0.246154	0.216996
AdaBoost Classifier	0.223077	0.053197	0.223077	0.085577
K Neighbors Classifier	0.192308	0.210256	0.192308	0.192286
Gaussian Naive Bayes	0.092308	0.050318	0.092308	0.045445

Feature Importance



Conclusion: Model Performance

Gradient Boosting Classifier:

Achieved the highest accuracy (**0.515385**) among all models.

Random Forest Classifier:

Performed moderately well with an accuracy of (**0.446154**).

Decision Tree Classifier:

Showed a lower accuracy (**0.384615**).

Other Models:

Other models showed lower performance.

Conclusion: Feature Importance

Gradient Boosting Classifier:

- Top features: G2, G1, absences, age, free time.
- G2 and G1 are the most important, showing continuous assessment is crucial.

Random Forest Classifier and Decision Tree Classifier:

- Similar top features: G1, G2, absences, free time, Fedu, Medu.
- Emphasize past performance, engagement, and family background.

Insights and Recommendations

- Importance of Continuous Assessment
- Attendance and Engagement
- Family Background
- Holistic Approach

References

- Project:
  GitHub Repository
- Dataset:
  Kaggle: Student Performance Data Set
- Additional Analysis:
  Kaggle Notebook: Annual Grades Average 5-Level Classification