Analyzing & Anticipating Student Performance to Shape Success Strategies

Shivam Dwivedi, Harsh Bhardwaj, Kumar Gurusahai, Mehul Singh

Group Number: 39

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY
DELHI

Motivation

Factors affecting student performance in schools?

1. Sex
2. Age
3. Family Income
4. Relationship status
5. **Study Hours**

6. Father Occupation
7. Mothers Occupation
8. Motivation of School
9. Aspiration

Objective:

1. Identify the codependence of various factors on student grades
2. Propose a suitable number of hours of study for the student
3. Aim to identify the factors negatively impacting the students performance
4. Give solutions to better student learning

Literature review

- ## Predicting Student Performance by Data Mining

by Kabakchieva (2013),

explored various data mining techniques to predict student performance and demonstrated the potential of predictive modeling in educational settings.

- ## Educational Data Mining: A Survey from 1995 to 2005

by Romero and Ventura (2007),

provided a comprehensive survey on understanding student learning behavior using a Deep mining based approach.

- ## Mining Education Data to Improve Student Retention

by Shahiri et al. (2015),

showcased the application of data mining in identifying at-risk students and creating methods that would help them better perform in their academic endeavour.

Dataset description

- Overview

  1. Dataset comprises files: student–mat.csv (Math course)
  2. Contains socio–demographic and academic information of 650 + 349 students.
  3. **Aimed at understanding the impact of various factors on student performance. (Link)**

- Parameters

  1. School: Student's school ('GP' or 'MS').
  2. Sex: Student's sex ('F' or 'M').
  3. Age: Student's age (numeric: 15 to 22).
  4. Address: Home address type ('U' – urban or 'R' – rural).
  5. Famsize: Family size ('LE3' – <= 3 or 'GT3' – > 3).
  6. Pstatus: Parent's cohabitation status ('T' – living together or 'A' – apart).
  7. Studytime: Weekly study time (1 –2 hours, 2 to 5 hours, 5 to 10 hours, or >10 hours).
  8. Failures: Number of past class failures (numeric).
  9. Schoolsup: Extra educational support (binary: yes or no).
  10. Higher: Wants to take higher education (binary: yes or no).
  11. Internet: Internet access at home (binary: yes or no).
  12. G1, G2, G3: Grades for three periods (numeric: 0 to 20).
  13. Go_out: Ability to go out with friends (numeric: 1 to 5)
  14. Free_time: How many hours of free time does the student have.(numeric: 1–5)

Extended Dataset description

- ## Overview

  1. Dataset from a higher education institution covering various undergraduate degrees.
  2. Contains enrollment information, academic paths, demographics, socio-economic factors.
  3. Contains information of 4424
  4. Captures academic performance across two semesters.

- ## Parameters

  1. Marital Status: Integer encoding of marital status, with options ranging from single to legally separated
  2. Application Mode: Categorizes the mode of application to the institution, from general contingent to special cases like international students or transfers.
  3. Application Order: Integer indicating the preference of the course chosen by the student, where 0 is the first choice.
  4. Course: Identifies the undergraduate degree program the student is enrolled in, from Biofuel Production Technologies to Nursing.
  5. Daytime/Evening Attendance: Binary feature indicating whether the student is attending daytime (1) or evening (0) classes.
  6. Previous Qualification: Encodes the level of education a student had before enrollment, such as secondary education or higher degrees.
  7. Previous Qualification (Grade): A continuous feature representing the grade of the student's previous qualification, scaled from 0 to 200.
  8. Nationality: Encodes the student's nationality, with options ranging from Portuguese to Colombian.
  9. Mother's Qualification: Integer representing the education level of the student's mother, from secondary education to higher degrees.
  10. Father's Qualification: Similar to the mother's qualification, this feature encodes the education level of the student's father.

Visualization (1)

Correlation with Studytime
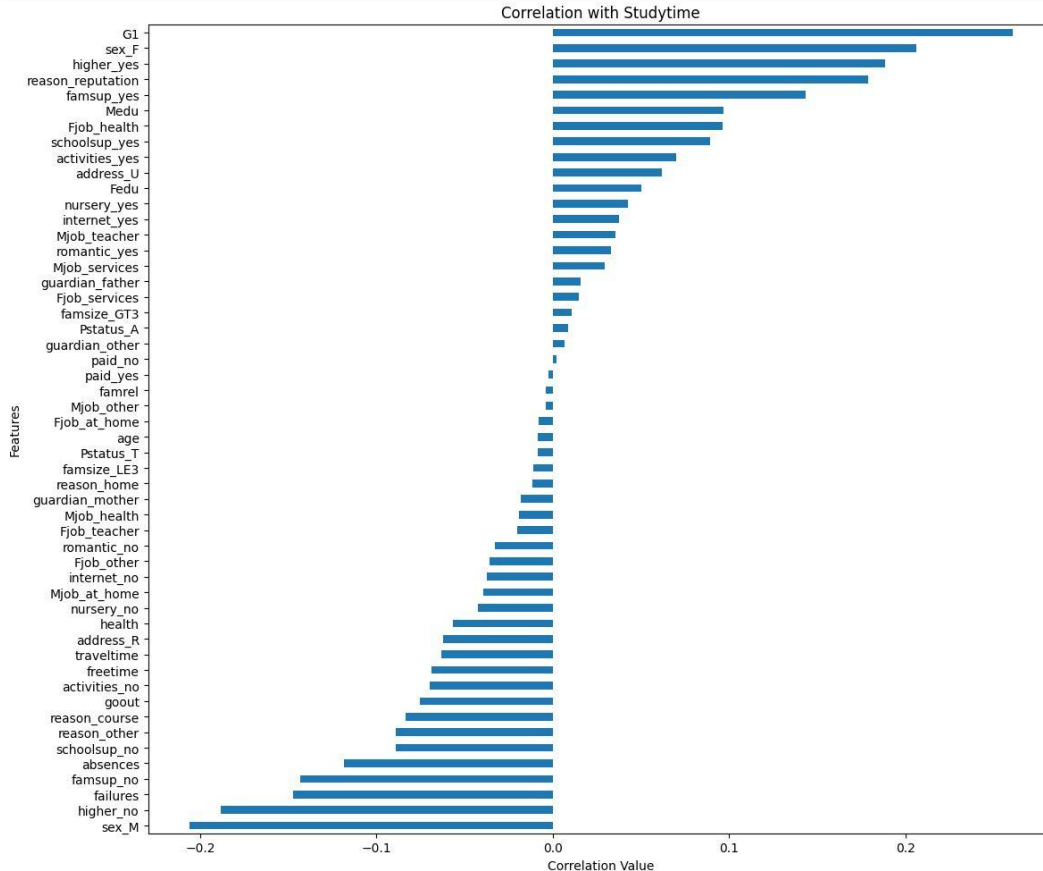
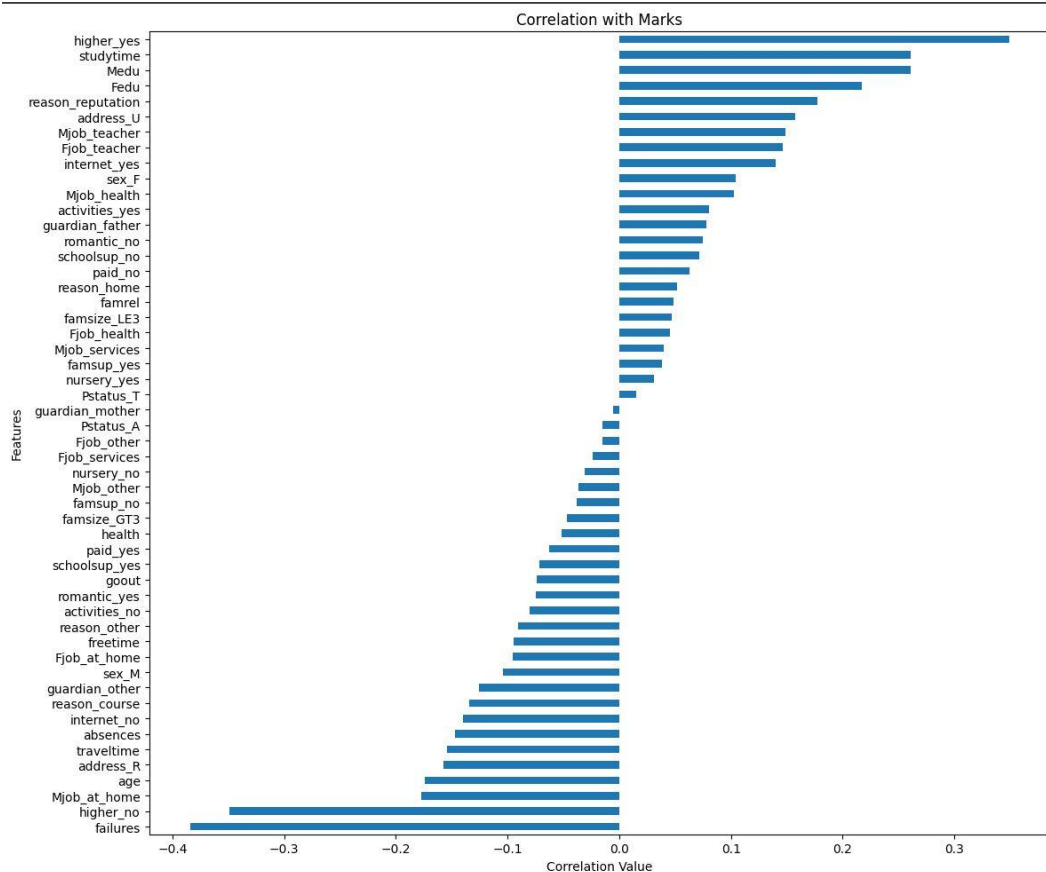Correlation of **Study Time** to other parameters:

Intuitive:

1. Grades
2. Internet
3. Aspiration
4. Mothers Edu.
5. Family Support

Incoherent:

1. Sex
2. Fathers Job
3. Activities
4. Relationship
   Yes

Visualization (2)
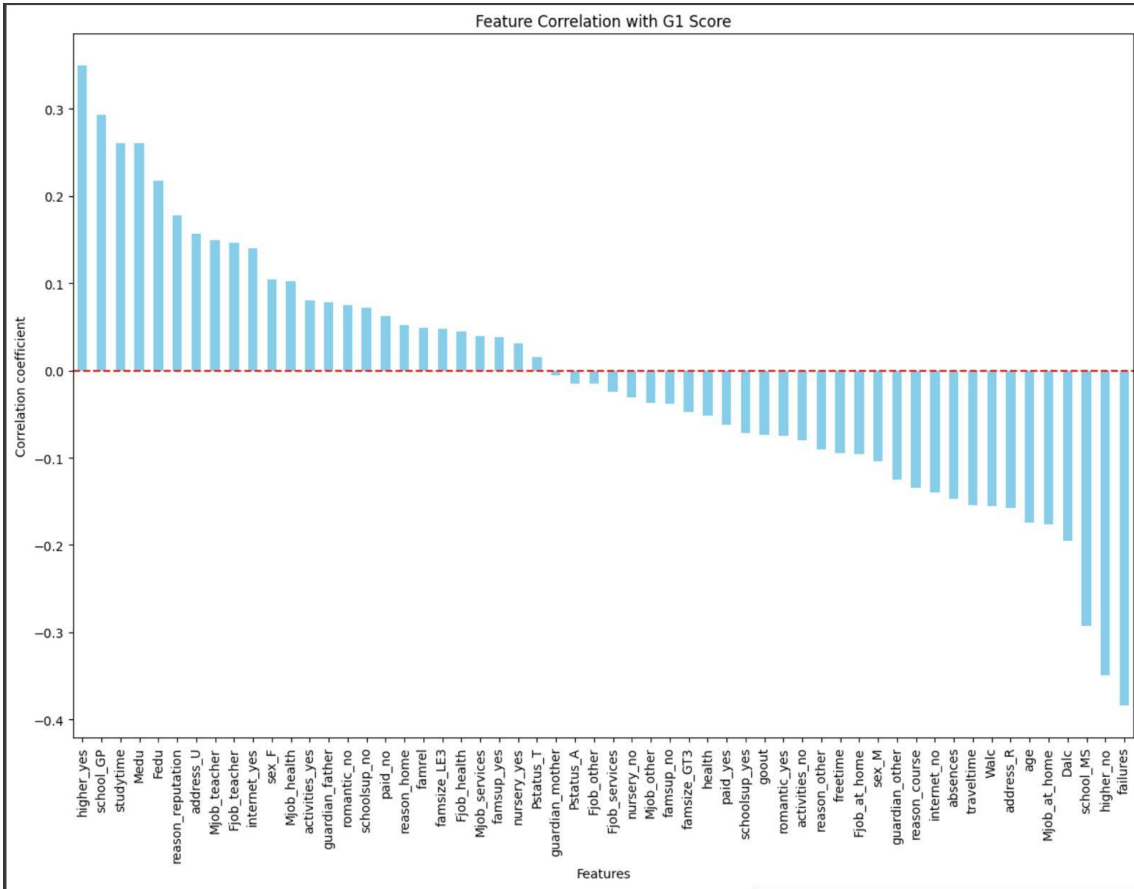

Correlation of **Marks** to other Parameters:

Intuitive:

1. Aspiration
2. Studytime
3. Fathers Edu.
4. Mothers Edu.
5. Internet

Incoherent:

1. Sex
2. Address
3. Teacher Jobs
4. Relationship No

Visualization (3)

Feature Correlation with G1 Score

Correlation of **G1** to other Parameters:

Intuitive:

1. Aspiration
2. Studytime
3. School
4. Mothers Edu.
5. Fathers Eu

Incoherent:

1. Sex
2. Teacher Jobs
3. Address
4. Relationship No

Pre–Processing + Methodology

- ## Removal of Irrelevant Columns

Columns such as: alcohol consumption: Dalc - Daily Alcohol Consumption, Walc - Weekly Alcohol Consumption

- ## One Hot Encoding (OHE)

OHE : Mapped categorical (string) values to numeric data in order to be able to include the data in the model.

- ## Analysis of important information

Categorical understanding of what to use:

Using the visualizations, we decided on which data should be involved in the actual model, based on its significance on its correlation to Marks and Studytime

Methodology

- ## Objective

Give a final analysis on the amount of hours a student should be studying to achieve their desired grade.

- ## Marks and Study Hours

Merged the categories of study hours and grades to increase the understanding of the processed data

- ## Determine the Weights

Determining the weights for each category

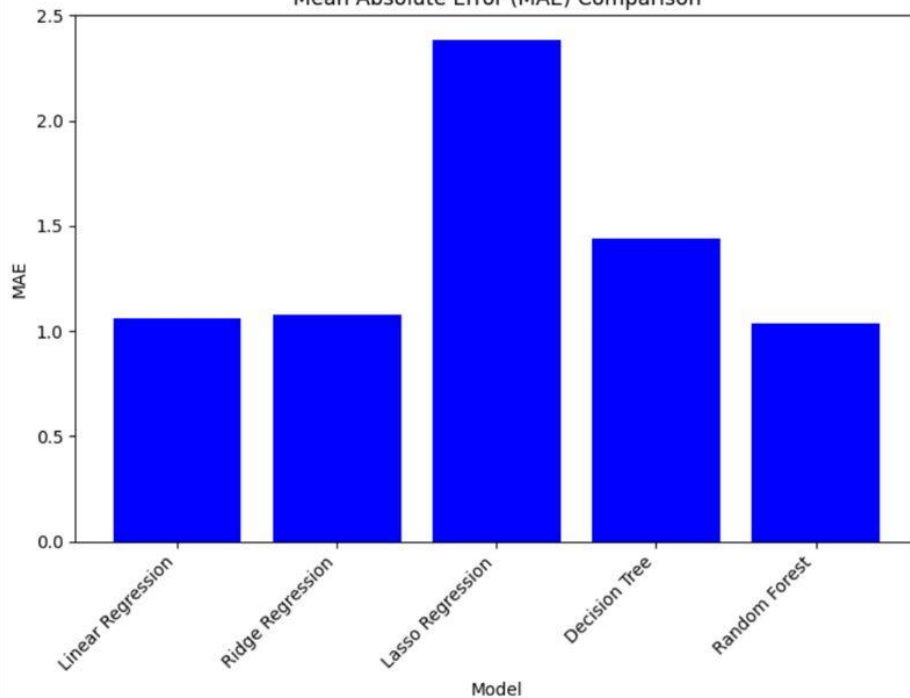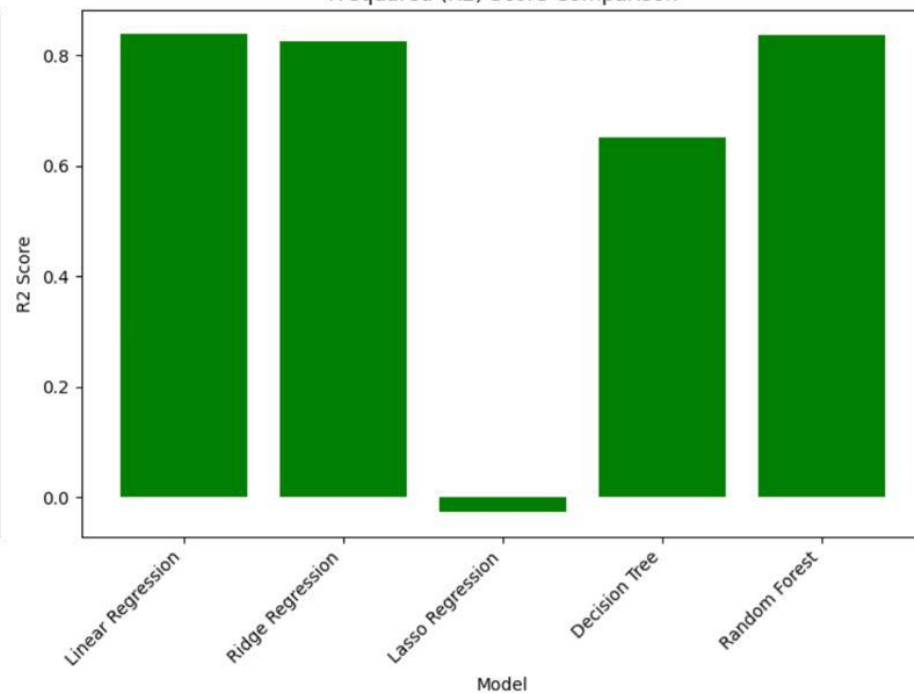We determined the weights of each of the previous categories based on the visualization of the data and the intrinsic correlation values.

Results (Ran the mode)
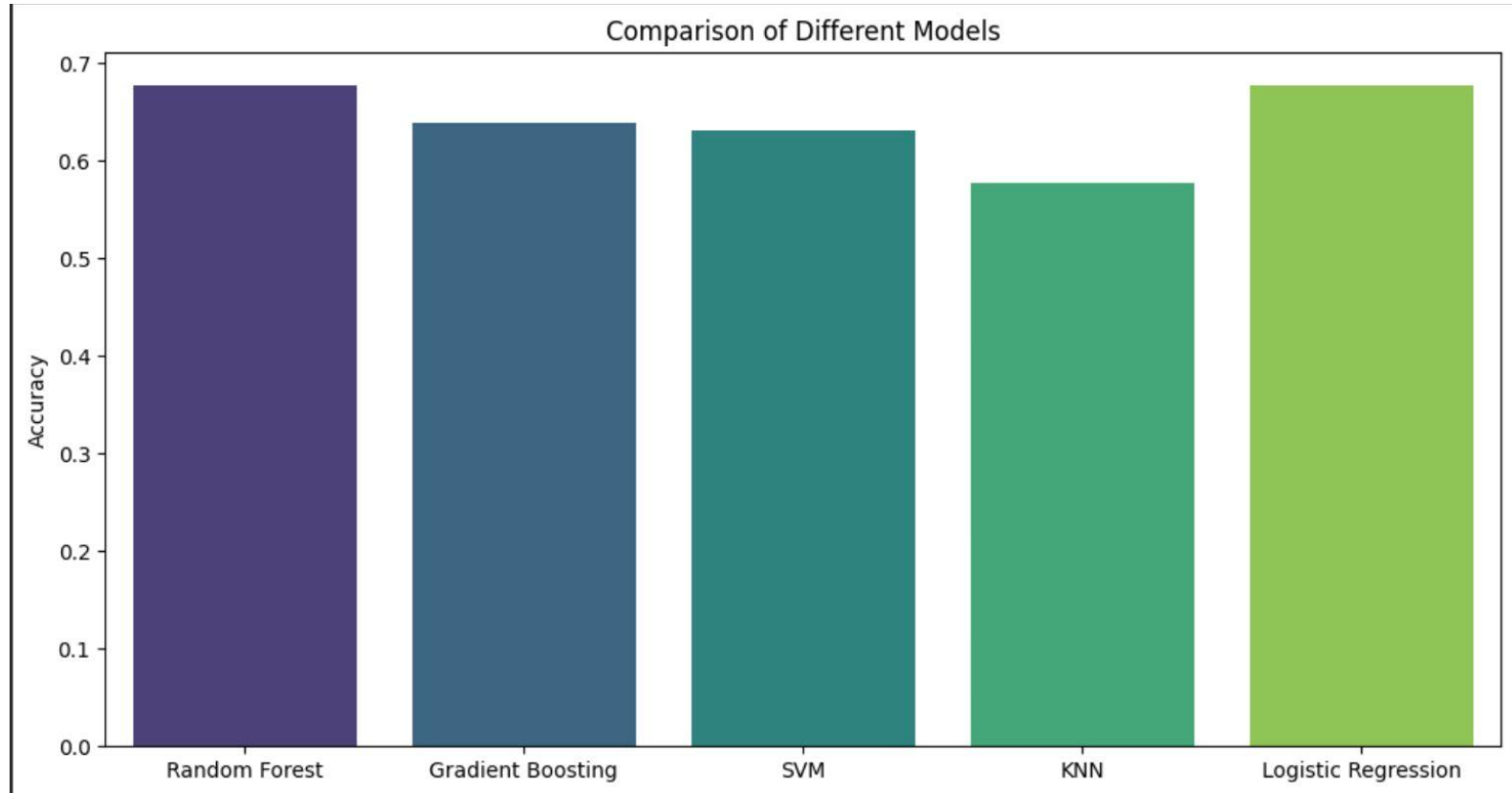
Comparison of Different Models

Results cont.

Speedbumps

1. Realizing that we can not verify the study increase time that we are predicting:

    a. We decided to come up with a new strategy:

        i. Predict Marks

        ii. Predict Study Time

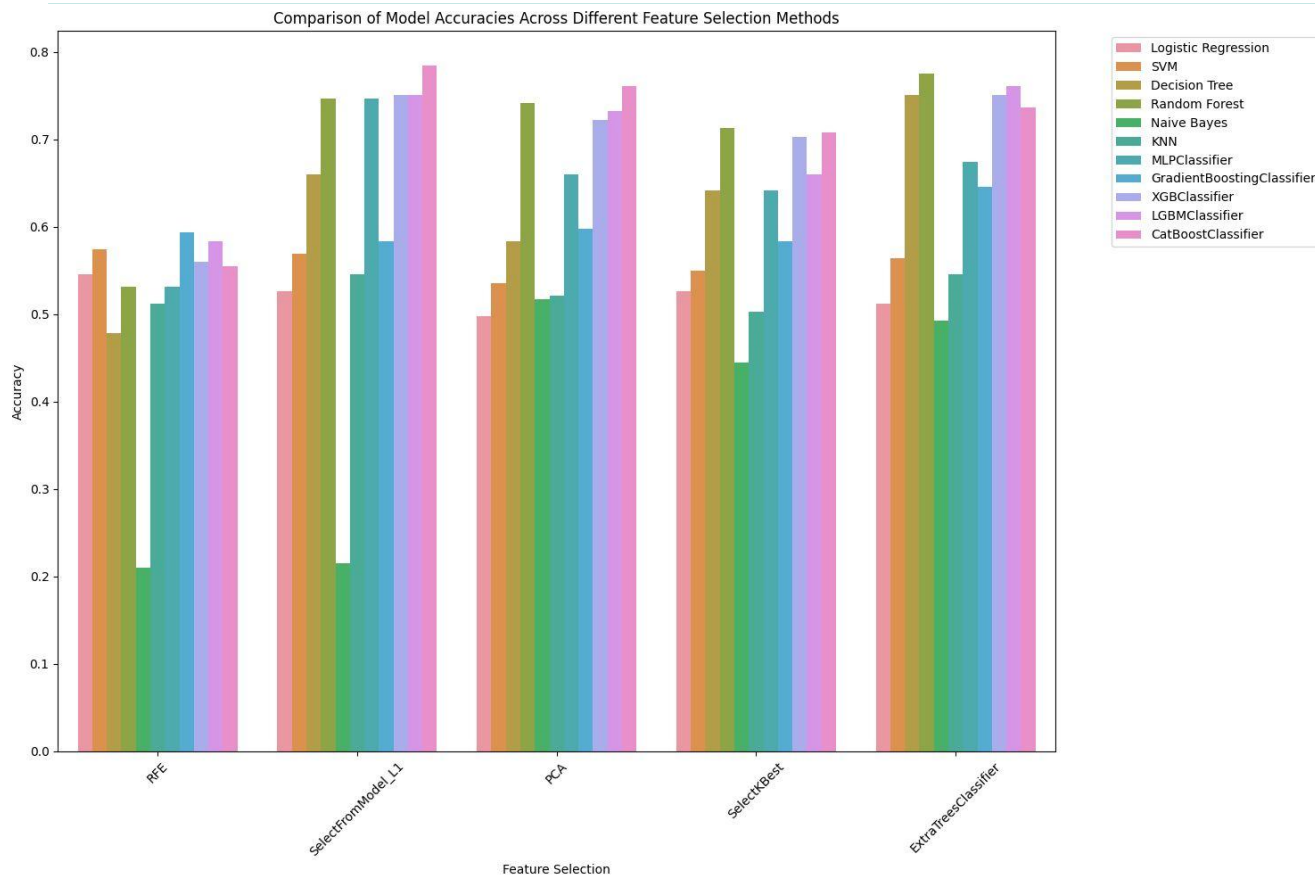   iii. Hide both and then learn the relationship between the two for further prediction
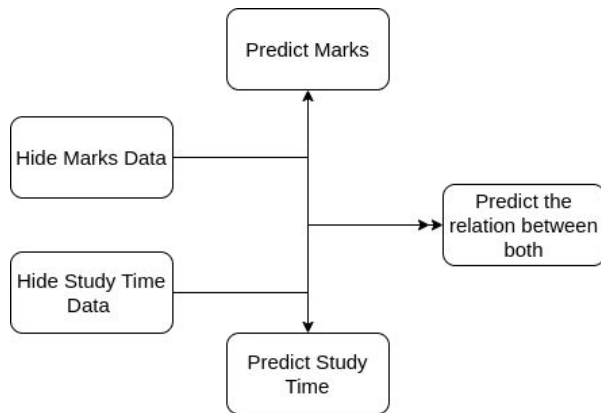
Mark Prediction

Comparison of Model Accuracies - Marks
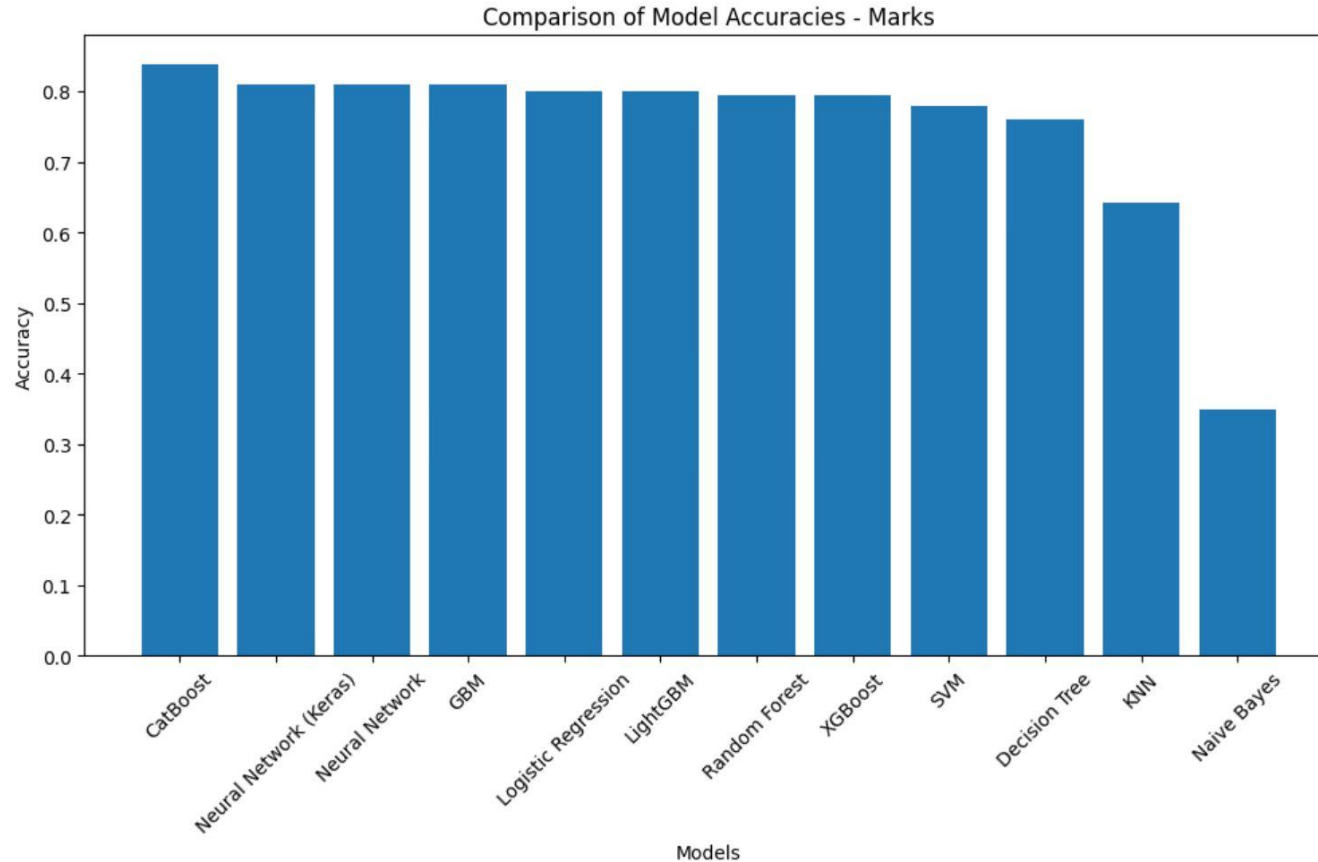
* Mark prediction ~ 82% accuracy

Study Time Prediction w/o grades

Comparison of Model Accuracies Across Different Feature Selection Methods

* Study time prediction ~ 78% acc

Grade Prediction (On a larger data set)
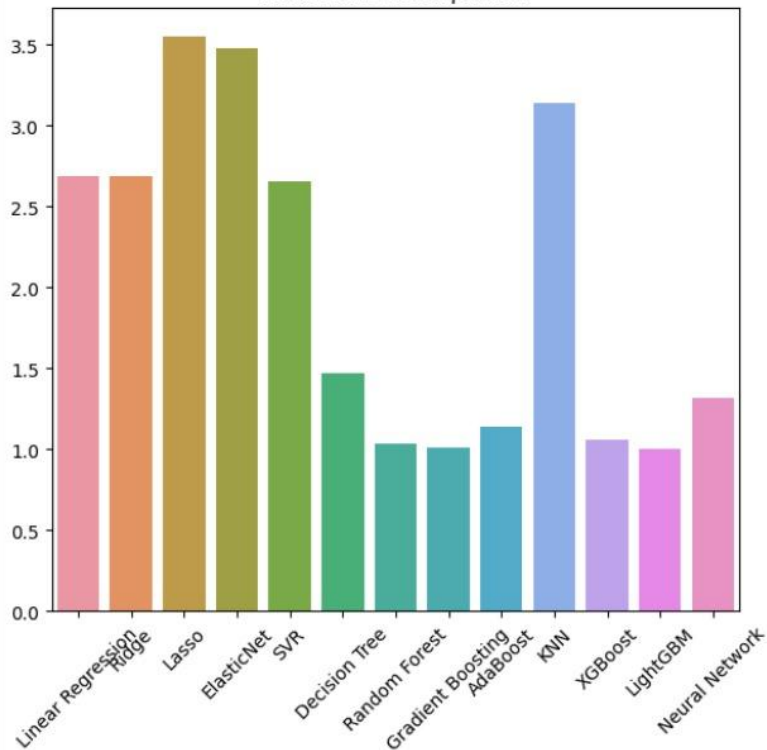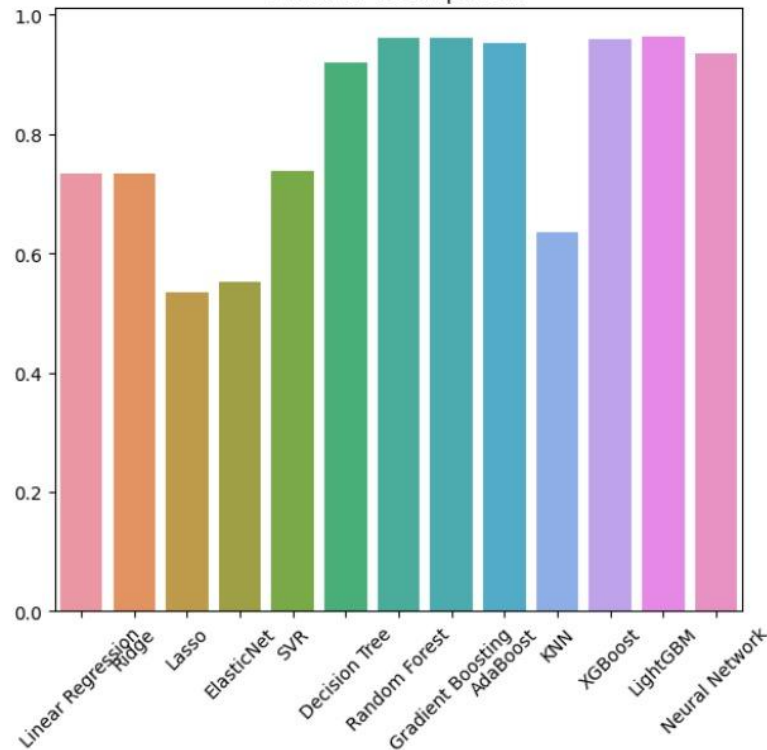
* ~ 95 % acc

Edge Cases

In case prediction fails or returns absurd values:

1. If the model suggests reducing study time to increase marks

    Return ('maintain current study routine')

2. If the model suggests increasing study time to obtuse amounts

    max(studyTime) = 5 hrs a day

    Return (max(studyTime))

User Interface

GradeAce Survey

What is your gender?
- Female
- Male

How old are you?

What is your home address type?
- Urban
- Rural

What is your family size?
- Less or equal to 3
- Greater than 3

Do your parents live together?
- Yes
- No

What is the highest level of education achieved by your mother?
None

What is the highest level of education achieved by your father?
None

What is your mother's job?
Teacher

Color Scheme selection: 'IIITD' color scheme to reduce stress.

A easy to identify simple logo. That indicates academic career growth

IOS Application

Further Works and Contribution

Work Distribution and Future Works:

Model Analysis and Training: Shivam and Harsh

Front end Application and Tuning: Mehul and Kumar

Documentation: Whole Team

Future Works:

Application Interface, Training on a larger
Group of student.