Linear Regression Project

By

Dhruv Sharma, Richard Clinger, Prabhash Venkat Paila, & Brian Hert

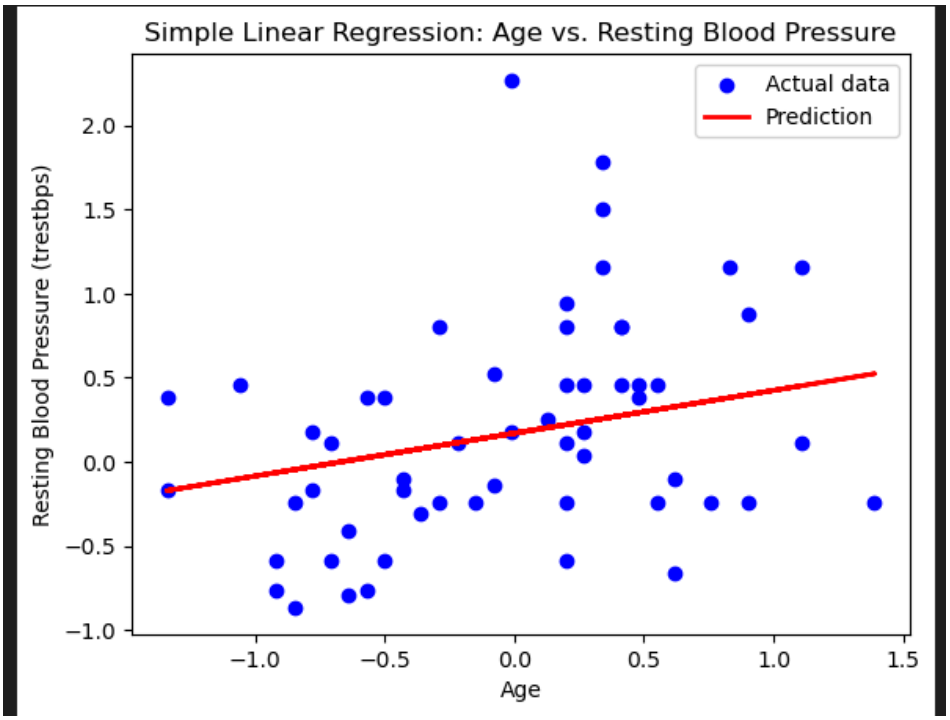
Part A

For part A, we applied Linear Regression on the dataset we preprocessed in our last project, in order to analyze the relationship of various features in the dataset and our target attribute (resting blood pressure). We used a simple Linear Regression Model as well as Multiple, Polynomial, Ridge, Lasso and Decision Tree Regression to test and predict our data.

Simple Linear Regression Model

In our simple Linear Regression Model, we used age as our independent variable and the resting blood pressure as our dependent variable. We then split the dataset into an 80-20 train test split, and finally used mean squared error (MSE) and R-squared to see how accurate our model is. Below represents our results for the Simple Linear Regression Model:

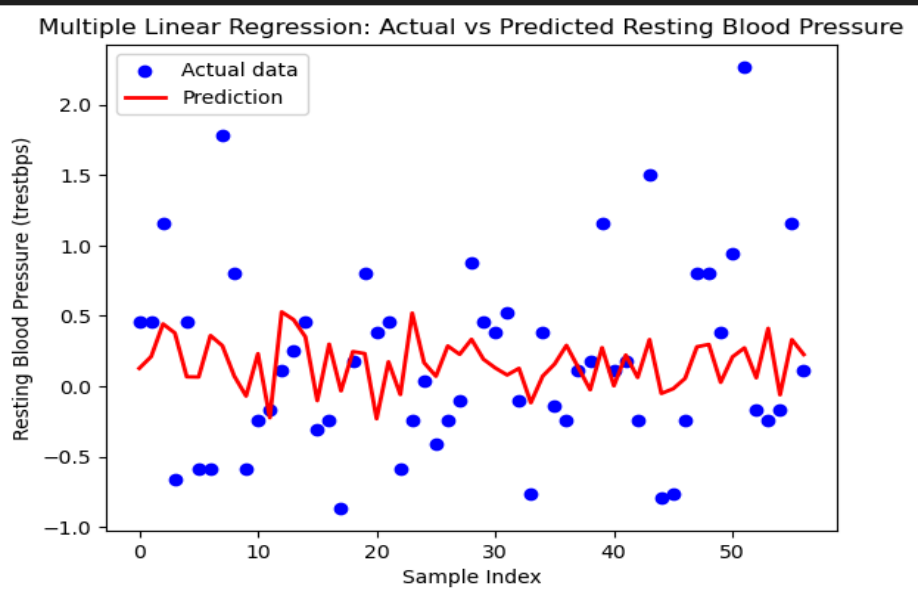
```
Simple Linear Regression
Mean Squared Error: 0.38073677984264553
R-squared: 0.1132068206849367
```



Multiple Linear Regression Model

In our multiple Linear regression model, we used age, chol and thalach as predictors for our trestbps (resting blood pressure). We used MSE and R-squared to help evaluate our model and the images below represent our findings:

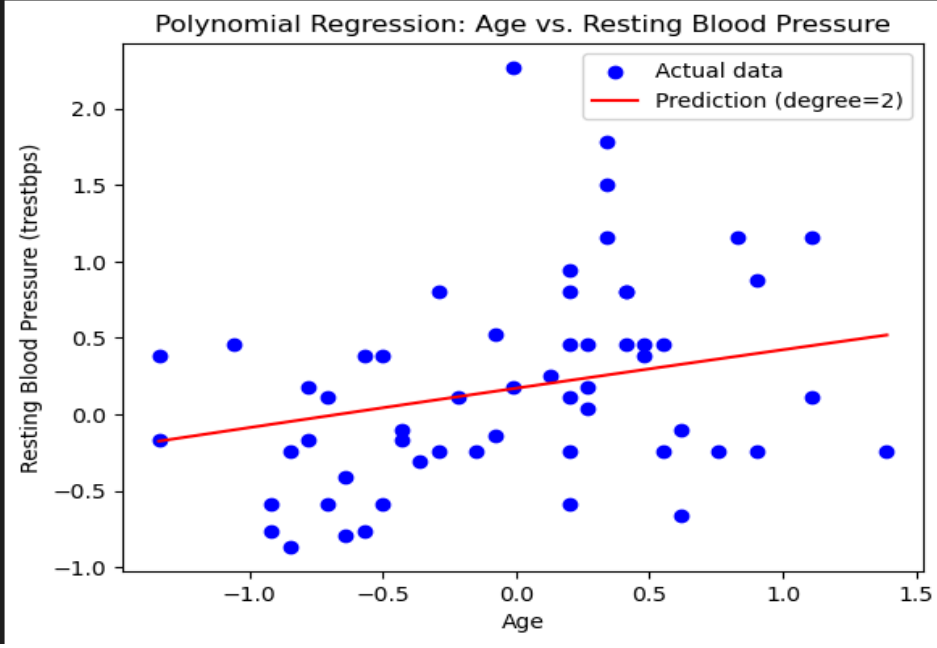
```
Multiple Linear Regression
Mean Squared Error: 0.37662225071217614
R-squared: 0.1227901773821839
```



Polynomial Regression

For the Polynomial Regression model, we used a degree of 2 to capture non-linear relationships between the age and trestbps. We used MSE and R-squared to evaluate our model, and below is the result:

```
Polynomial Regression
Mean Squared Error: 0.3804892180413344
R-squared: 0.11378342932503804
```



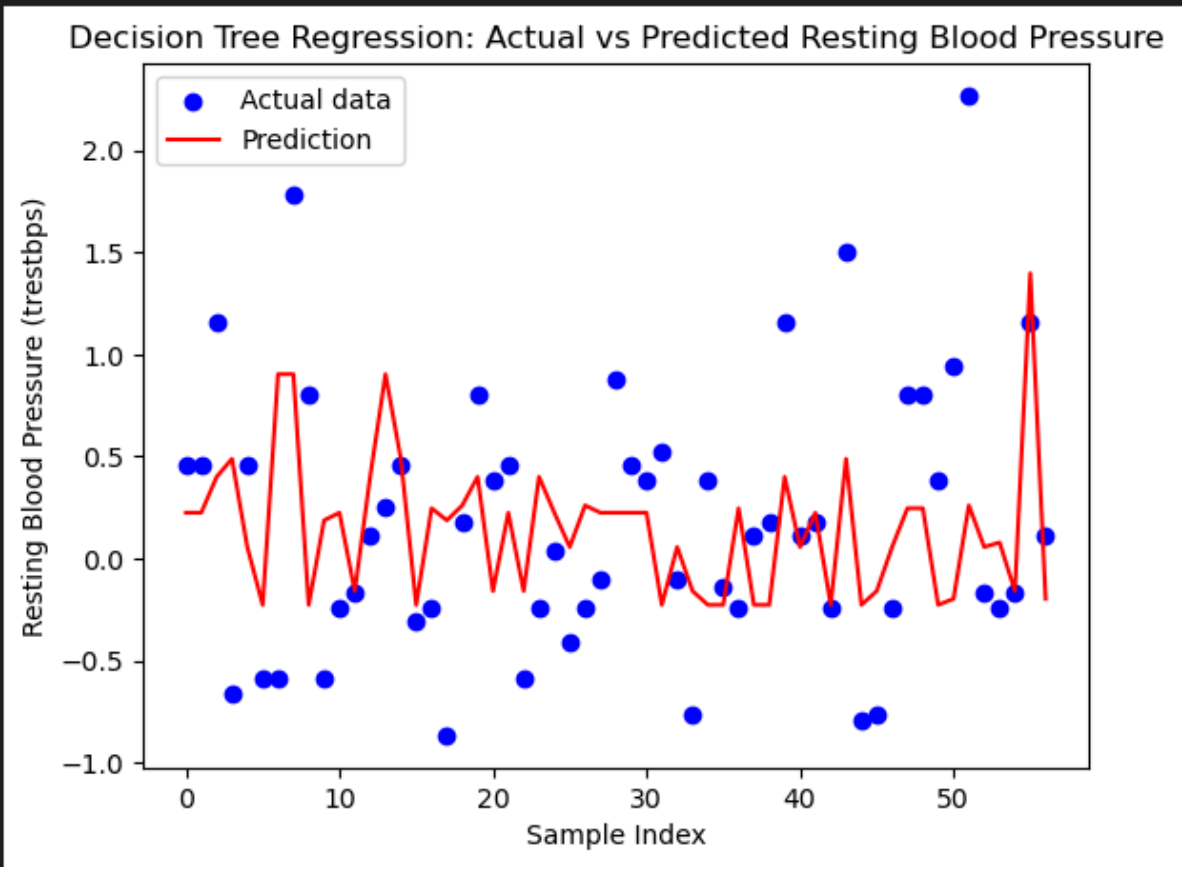
Ridge, Lasso and Decision Trees Models:

Ridge Regression added a regularization term to reduce overfitting by penalizing large coefficients, and Lasso regression adds a penalty that can shrink coefficients to zero, selecting features effectively. We used a decision tree regressor as an alternative to linear models for a non linear fit. We used MSE and R-squared to evaluate the models, and below are our findings.

```
Ridge Regression
Mean Squared Error: 0.3771117477541704
R-squared: 0.12165006520727184
```

```
Lasso Regression
Mean Squared Error: 0.4293824126299884
R-squared: -9.616879012686042e-05
```

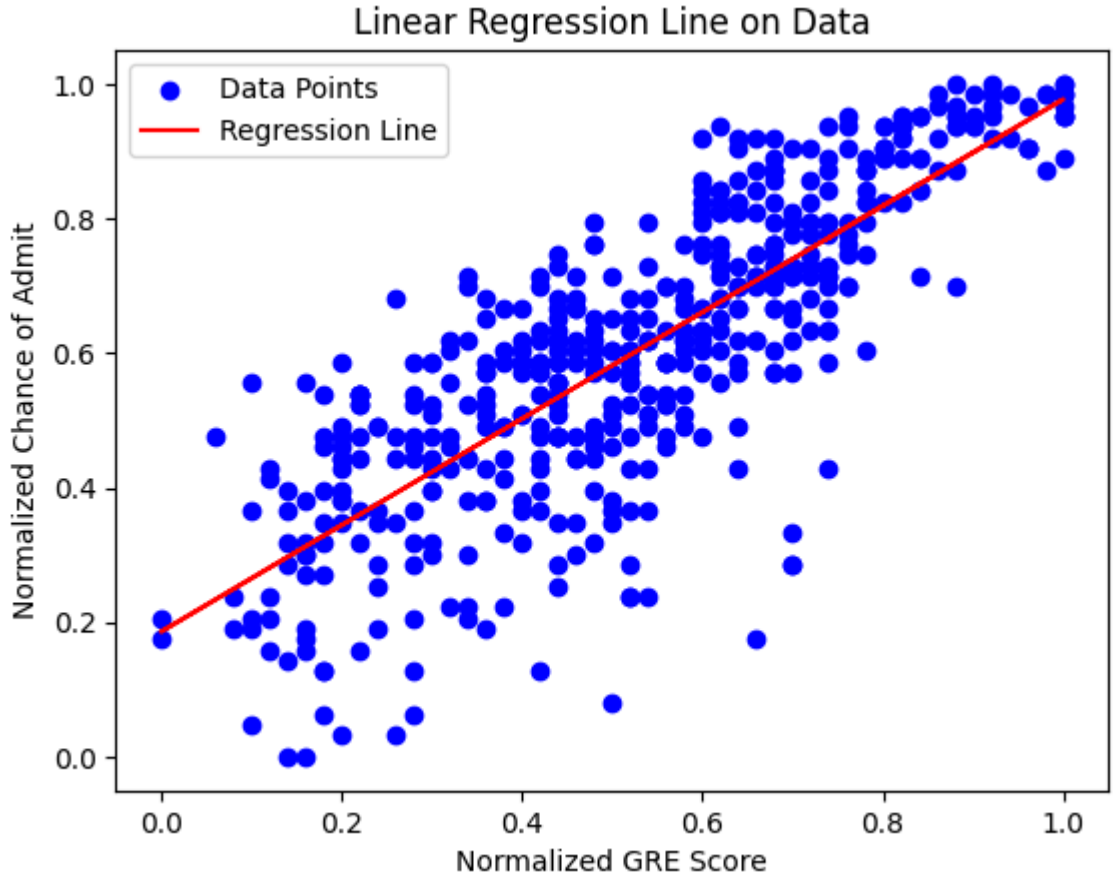
```
Decision Tree Regression
Mean Squared Error: 0.3885725703560488
R-squared: 0.0949560869767313
```



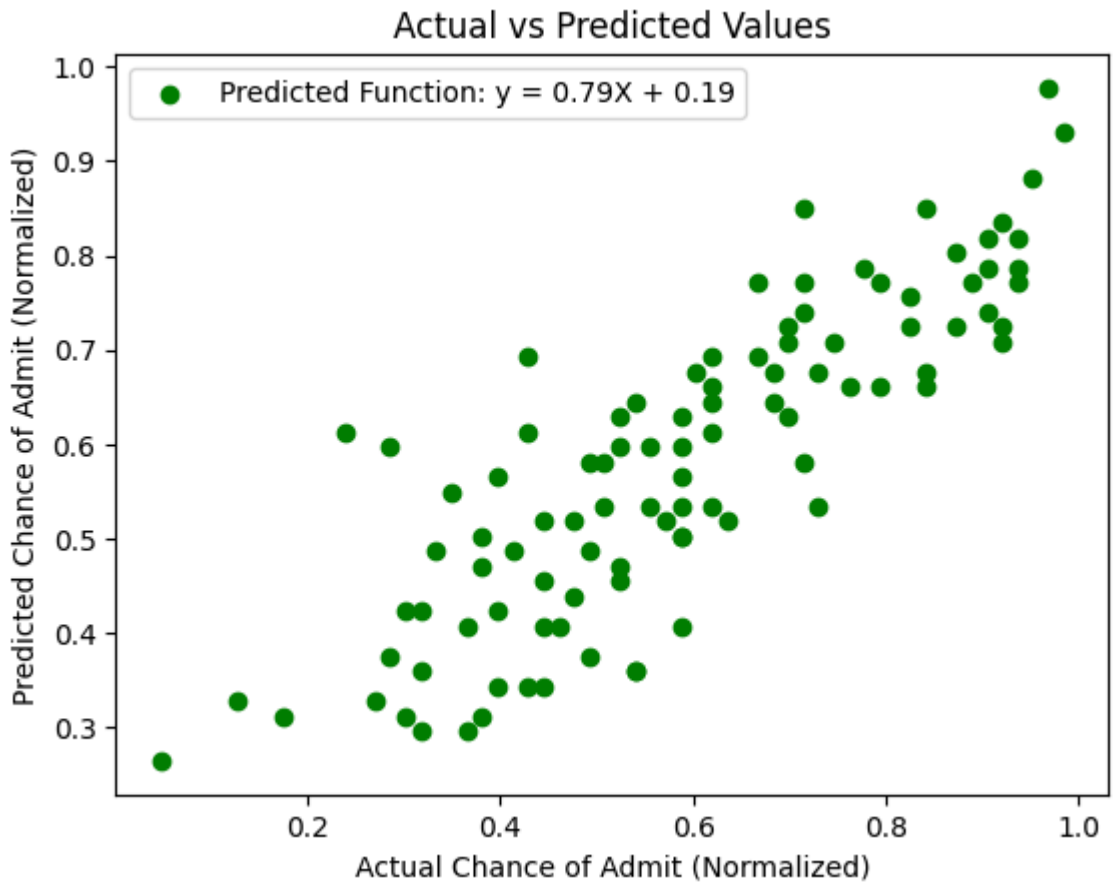
Part B:

For Part B, we applied the regression model to the dataset provided for the assignment. When manually inspecting the data, we can see that both GRE Score and TOEFL Score are relevant features for admission chances. The University Rating, Statement of Purpose(SOP), and Letter of Recommendation(LOR) scores indicate qualitative attributes that may influence admission changes and help capture differences between applicants. There is variation in this data but they are on the same scale, which makes comparisons straightforward.

To simplify the model, we are using only the GRE Score and Chance of Admit to drop redundant correlated features and examine a simple linear relationship.



The scatter plot of normalized GRE Score vs Chance of Admit with the fitted regression line (in red) indicates a positive linear relationship. The linear regression line seems to fit the general trend.



Mean Squared Error: 0.013
R-squared: 0.709

Standardizing or normalizing the data before training is crucial to ensure that each feature contributes equally to the model, especially when features are on different scales. Without normalization, features with larger ranges can dominate the learning process, leading to biased results.

Classification Report

In this project, we used a dataset of applicant profiles to predict admission chances into three categories: Low, Medium, and High. Starting with the original "Chance of Admit" variable, which was continuous, we discretized it into three classes to make classification more interpretable. Key features in the dataset included scores such as GRE, TOEFL, CGPA, and other factors like research experience and recommendation ratings.

```
Classification Report:
              precision    recall  f1-score   support

         Low       0.72      0.68      0.70        34
      Medium       0.68      0.69      0.68        39
        High       0.89      0.93      0.91        27

    accuracy                           0.75       100
   macro avg       0.76      0.76      0.76       100
weighted avg       0.75      0.75      0.75       100
```

Decision Tree Analysis for Admissions Prediction

Step 1: Import the libraries

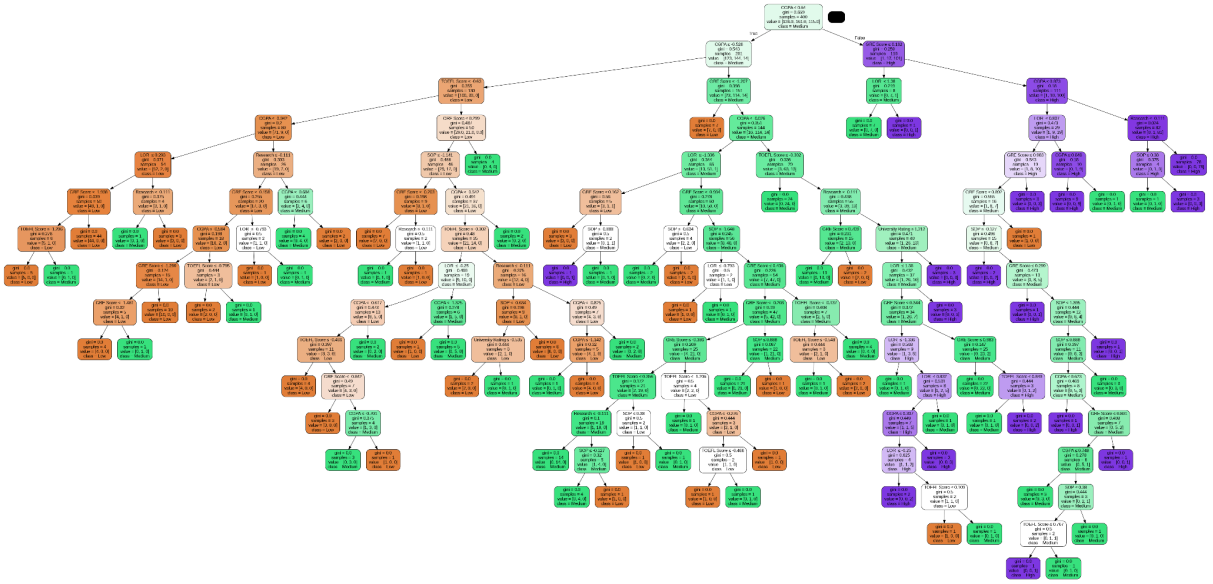
Step 2: Load and Preprocess Data
Load the data, discretize the "Chance of Admit" column, and split the data into training and test sets.

Step 3: Standardize the Features
Standardize the features for optimal model performance.

Step 4: Train the Decision Tree Classifier
Train the classifier on the training data.

Step 5: Evaluate the Model on Test Data
Evaluate the classifier and generate a classification report.

Step 6: Visualize the Decision Tree
Finally, create a visual of the decision tree using pydotplus and graphviz.



Valuable Rules:

1. If CGPA > 0.64 and GRE Score > 0.16, then the Chance of Admit is high (class 2).
2. If CGPA <= 0.64, TOEFL Score <= -0.63, and Research is 0, then the Chance of Admit is low (class 0).
3. If CGPA <= -0.53, GRE Score <= 0.30, and SOP is below average (-1.14), then the Chance of Admit is low (class 0).

Part C:

For part C, we are given instructions on how to work through a classification tree exercise found in the “Entropy_IDE_Exercise.pdf.” The objective is to develop a classification decision tree using the ID3 algorithm utilizing a provided dataset. The dataset features attributes regarding color, shape, size, and class. The goal is to build a decision tree that effectively classifies the item to its attributes.

Color	Shape	Size	Class
Red	Square	Big	+
Blue	Square	Big	+
Red	Round	Small	-
Green	Square	Small	-
Red	Round	Big	+
Green	Round	Big	-

Step 1: Initial Entropy Calculation

The initial entropy of the dataset is calculated to assess the impurity of the class distribution.

Entropy Calculation

Using the following formula for entropy E(t)E(t)E(t):

E(t)=−∑p(j∣t)log2p(j|t)E(t) = - \sum p(j|t) \log_2 p(j|t)E(t)=−∑p(j∣t)log2p(j∣t)

where:

- p(+)p(+)p(+) = probability of positive instances
- p(−)p(-)p(−) = probability of negative instances

Given:

- Number of positive instances = 3
- Number of negative instances = 3
- Total instances = 6

Step 2: Information Gain Calculation

Next, we calculate the entropy and information gain for each attribute, color, shape, and size. An average Entropy Calculation Function measures the uncertainty in our set of data. Data sets allow us to prepare and organize data for analysis or modeling. Information gain is a metric that gauges the usefulness of an attribute for classifying a dataset. It reflects the decrease in randomness about the target variable after observing the attribute's value. Information gain results show the extent to which an attribute aids in predicting the target variable within a dataset.

Step 3: Best Attribute Selection

The attribute with the highest information gain is Color, which will be chosen as the first splitting attribute for the decision tree.

Decision Tree Structure

The resulting decision tree based on the **Color** attribute is as follows:

```
      Color

    /   |   \

  Red  Blue  Green
```

/ \ | \

+ + + -

Step 4: Impact of Adding a New Attribute

Introducing a new attribute, such as "Pattern of Shirt" with possible values like "Checked," "Striped," and "Solid," could lead to significant changes in the decision tree.

Possible Changes:
The new attribute might result in the creation of additional branches in the decision tree, leading to new splits. Furthermore, existing nodes may need to be updated if the new attribute enhances classification. Depending on its relevance, the overall complexity of the model could increase, or its accuracy could improve.

Consequences of Ignoring the New Attribute:
If a data scientist overlooks this new attribute, several issues could occur. Firstly, the model's performance may decline due to the exclusion of relevant information, resulting in inaccurate predictions. This could have financial repercussions, as incorrect predictions might lead to poor production decisions and misallocated resources. Additionally, discovering the importance of the new attribute later on could yield surprising insights and prompt strategic changes in business decisions.

Task Division:

Dhruv Sharma: Worked on part a of the project by creating several linear regression models, for our dataset

Rick Clinger: Worked on part b-regression of the project which was to use the dataset provided and manually inspect the data to determine if preprocessing needed to be executed before training the model. After training the model I made a visualization of the data and calculated the line of regression and the mean squared error.

Prabhash Paila: Worked on Part b- Classification of the project. Used a decision tree model to classify applicants' admission chances into Low, Medium, and High categories based on features like GRE Score, TOEFL Score, CGPA, and research experience. After discretizing the "Chance of Admit" variable and standardizing features, we trained the model and evaluated its performance using a classification report. Visualized the decision tree, allowing us to interpret its decision paths and extract key rules that indicate the factors most strongly associated with admission probability.

Brian Hert: Worked on part c of the project which was the classification tree homework. The analysis demonstrates the effectiveness of the ID3 algorithm in creating a decision tree based on the provided dataset. Additionally, it highlights the need to take all relevant attributes into account to achieve accurate predictions and make well-informed business decisions.