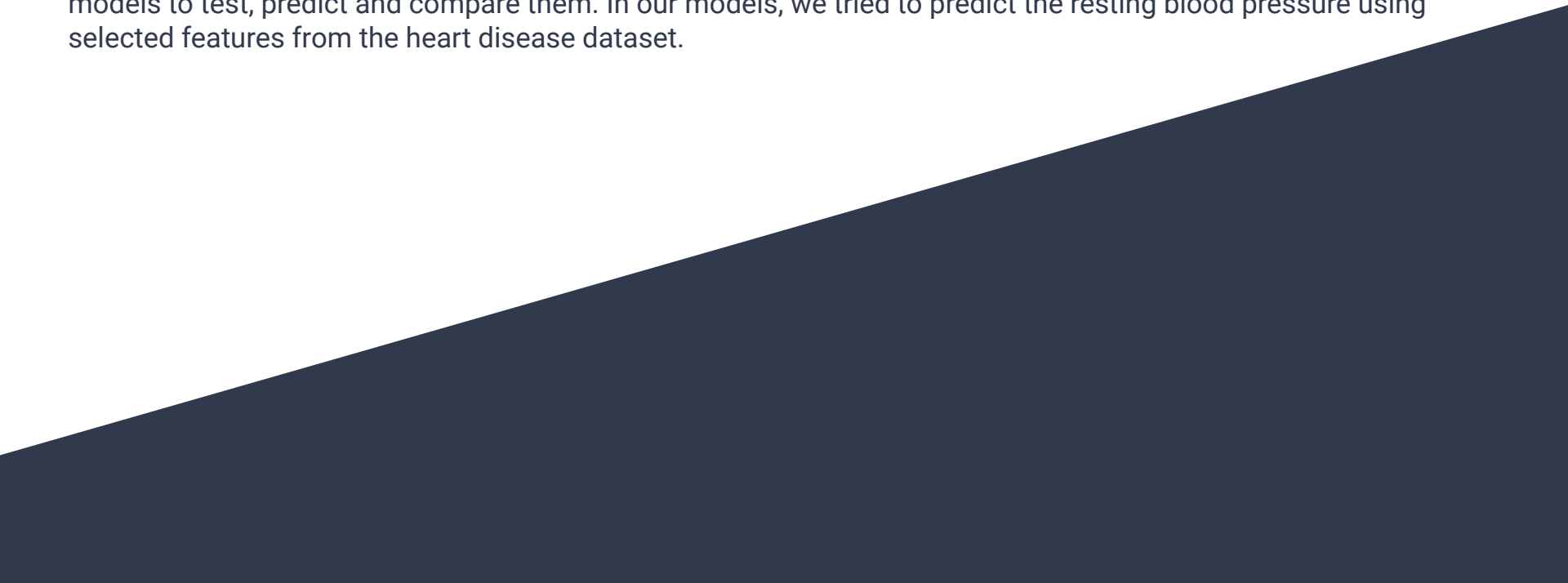


Linear Regression Project

By:
Dhruv Sharma,
Richard Clinger,
Prabhash Paila, &
Brian Hert

Part A

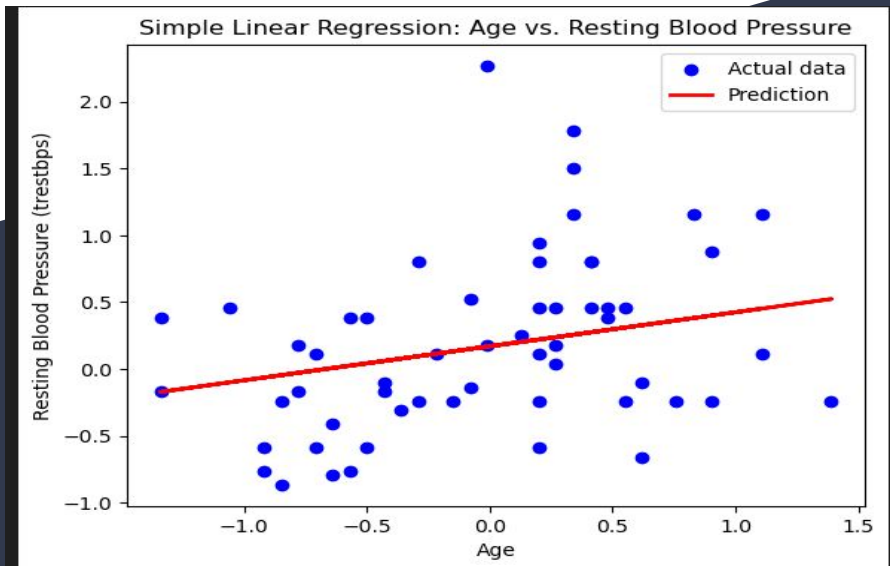
For part a, we used the dataset we preprocessed in the last project, and created several linear regression models to test, predict and compare them. In our models, we tried to predict the resting blood pressure using selected features from the heart disease dataset.

A large, dark blue, curved shape that starts from the bottom left and extends diagonally upwards towards the right, filling the lower half of the slide.

Simple Regression Model

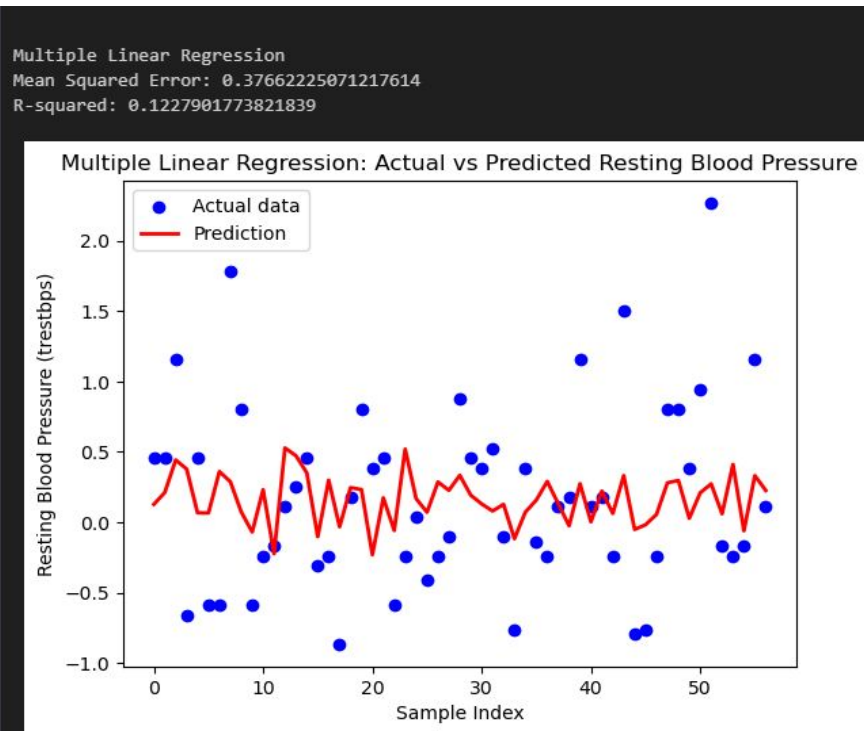
For the Simple Regression model, we used age as our predictor and the resting blood pressure (trestbps) as our target. We also used MSE and R-squared to evaluate our models

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



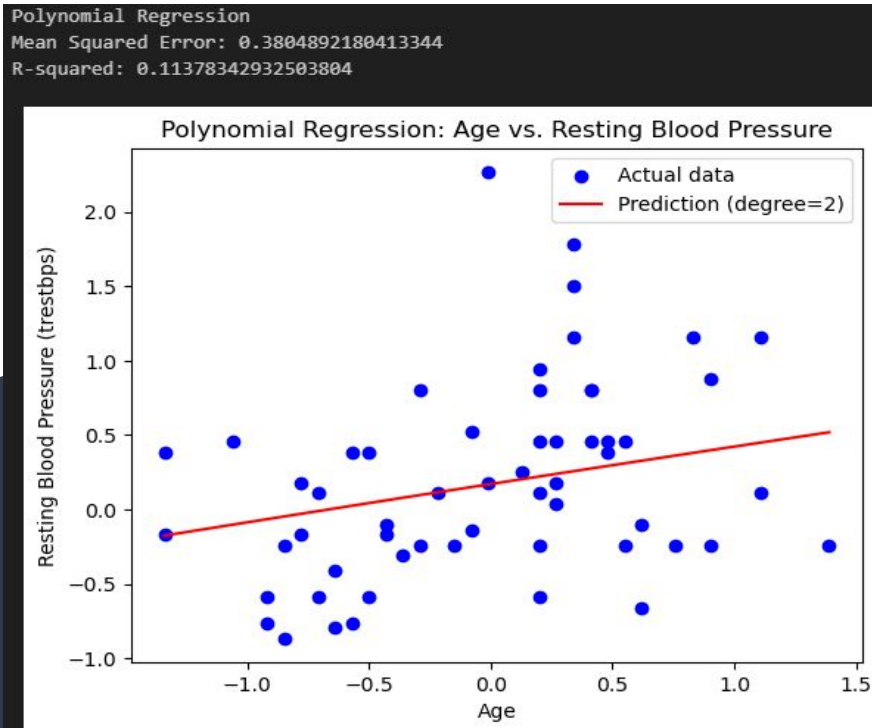
Multiple Regression Model

For Multiple Regression, we used age, chol and thalach as our predictors for the target, trestbps.



Polynomial Regression

In our polynomial model, we used a degree of 2, and MSE and R-squared to evaluate our models



Ridge, Lasso and Decision Trees

Ridge adds a penalty to reduce overfitting

Lasso adds a penalty and can eliminate less important features

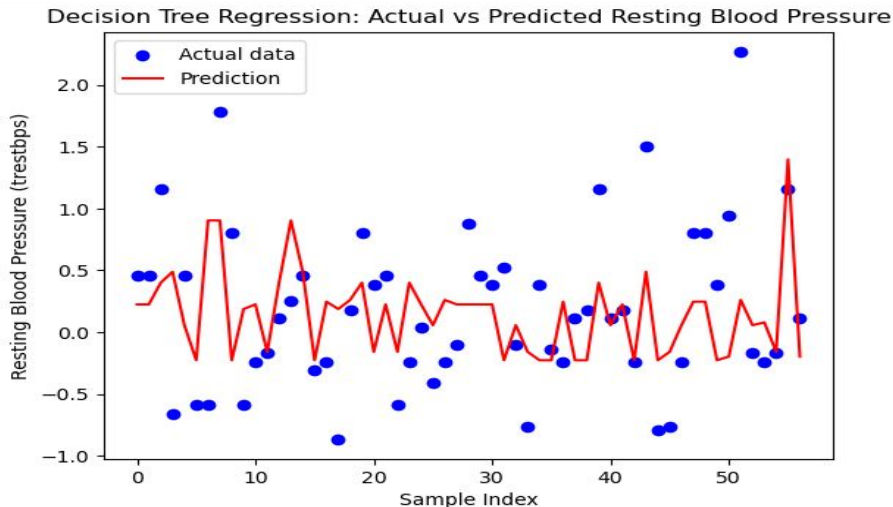
Decision Tree Regression was used to capture complex relationships that are not representable in linear splits.

For all the models, we used MSE and R-Squared to evaluate all the models

```
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
```



Classification Report

Precision, Recall, and F1-Score: The classification report provides precision (accuracy of positive predictions), recall (ability to identify all actual positives), and F1-score (balance of precision and recall) for each class (Low, Medium, High).

Model Performance: High scores for the High and Low classes indicate the model effectively distinguishes applicants with strong or weak profiles, while the Medium class shows slightly lower accuracy due to its intermediate nature.

Overall Accuracy: The report also includes overall accuracy, showing how well the model classifies applicants into the correct admission category across all classes.

Classification Report:

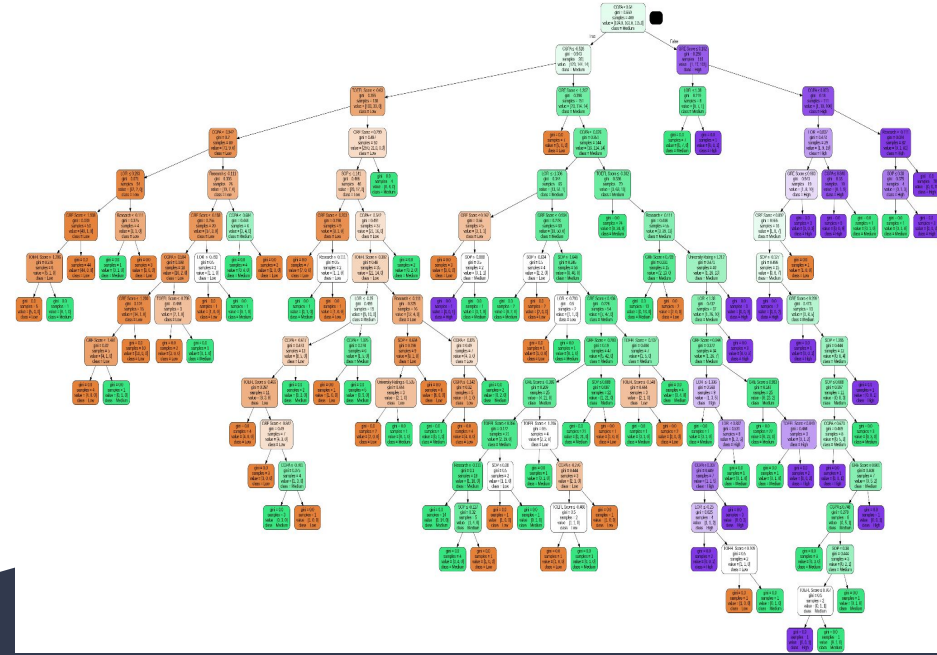
	precision	recall	f1-score	support
Low	0.72	0.68	0.70	34
Medium	0.68	0.69	0.68	39
High	0.89	0.93	0.91	27
accuracy			0.75	100
macro avg	0.76	0.76	0.76	100
weighted avg	0.75	0.75	0.75	100

Decision Tree

Interpretable Classification: The decision tree model uses key features like CGPA, GRE Score, and TOEFL Score to classify applicants into Low, Medium, and High admission probability categories, creating an interpretable, visual flow of decisions.

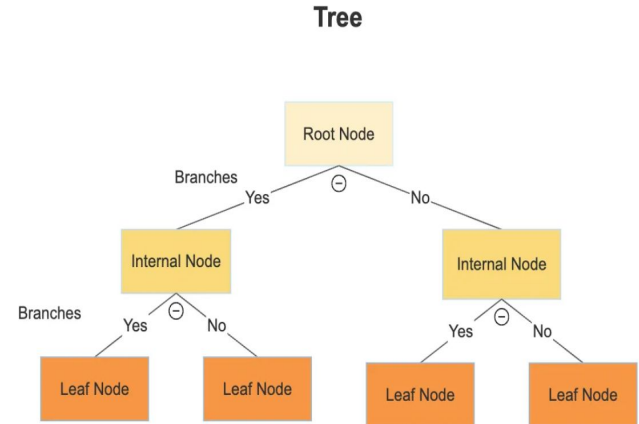
Rule Extraction: By analyzing the tree structure, we extracted valuable decision rules, such as "High CGPA and GRE Score" indicating a high chance of admission, which provides actionable insights for applicants to improve their profiles.

Visual and Practical Insights: The tree's branches and thresholds reveal important factors in admissions decisions, highlighting specific feature combinations that most influence admission likelihood, making it useful for understanding admission trends.



Decision Tree Analysis Classification and Insights

This slide explores the ID3 decision tree algorithm, highlighting its effectiveness in classifying data based on various attributes. We will analyze a dataset containing information about color, shape, and size, and examine how the introduction of a new attribute can impact the decision tree structure and predictive accuracy. Stay tuned as we uncover the significance of information gain in model selection and its implications for decision-making.



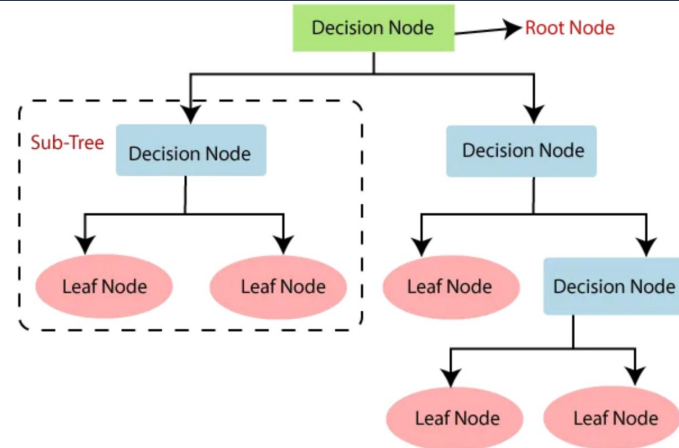
Dataset Overview

The dataset includes four important attributes: Color, Shape, Size, and Class, with Class serving as the target attribute for classification. Each attribute is critical for distinguishing among the different categories present in the dataset. For instance, Color and Shape highlight the visual features, while Size indicates the physical dimensions of the objects. Analyzing these attributes together allows us to build predictive models that accurately categorize the data.

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

Initial Entropy and Information Gain

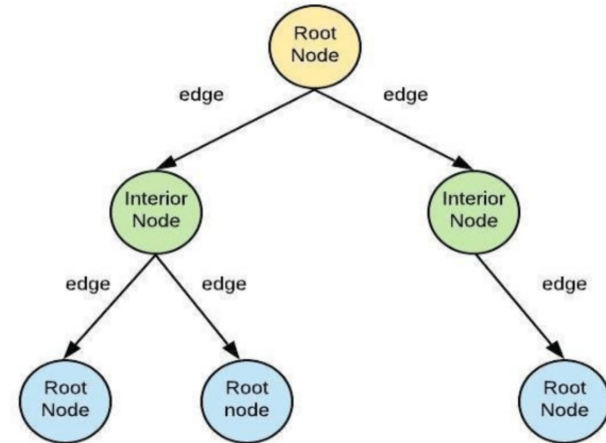
The initial entropy of the dataset is calculated to be 1.0, indicating a high level of disorder or uncertainty in the class distribution. The information gains for each attribute are as follows: Color has an information gain of 0.54, Shape has 0.08, and Size has 0.46. Among these, Color provides the highest information gain, making it the best attribute for splitting the dataset. This means that using Color to partition the data will most effectively reduce uncertainty and improve the accuracy of the classification.



Decision Tree Classification Flowchart

Decision Tree Structure

The decision tree is organized around the attribute Color, which acts as the initial decision node. It branches into three outcomes: if the Color is Red, the classification is positive (Class +); if the Color is Blue, the classification is again positive (Class +); and if the Color is Green, the classification is negative (Class -). This clear and concise structure demonstrates how Color alone can determine an item's class, effectively showcasing the decision-making process.



Impact of Adding a New Attribute

The impact of adding a new attribute, such as Pattern of Shirt, which includes values like "checked," "striped," and "solid." This addition could lead to significant changes in the decision tree, including new splits based on the attribute and potential revisions to existing nodes if it improves classification accuracy. While this may increase the model's complexity, it could also enhance its predictive performance. Conversely, overlooking this attribute might result in inaccurate predictions, financial repercussions from misguided production decisions, and unexpected insights if the attribute is discovered later, potentially affecting strategic planning.