

DATA WAREHOUSE **AND MINING**

MINI PROJECT REPORT

REGRESSION **ALGORITHM**

Made by:

- **Surya Pratap – 8367**
- **Rachel Jose – 8361**
- **Nolita Rego – 8363**

Under the guidance of: Mrs. Sujata Deshmukh

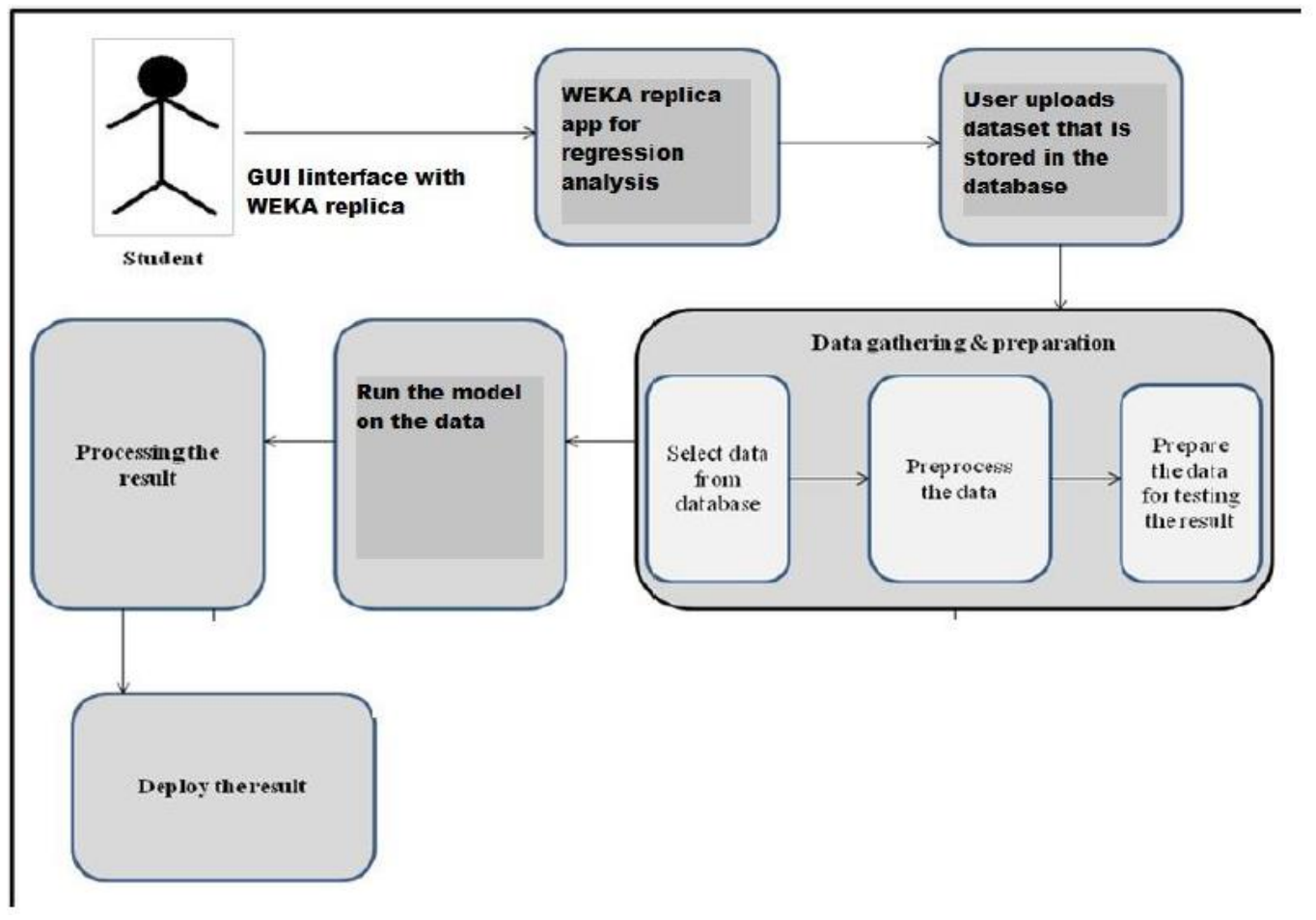
■ **Problem Definition:** -

The system takes in any dataset (for regression model) and evaluated the dataset of Linear Regression model after analyzing the dataset and displays the resulting graphs and other details. And finally gives the accuracy percentage of the model on that particular dataset.

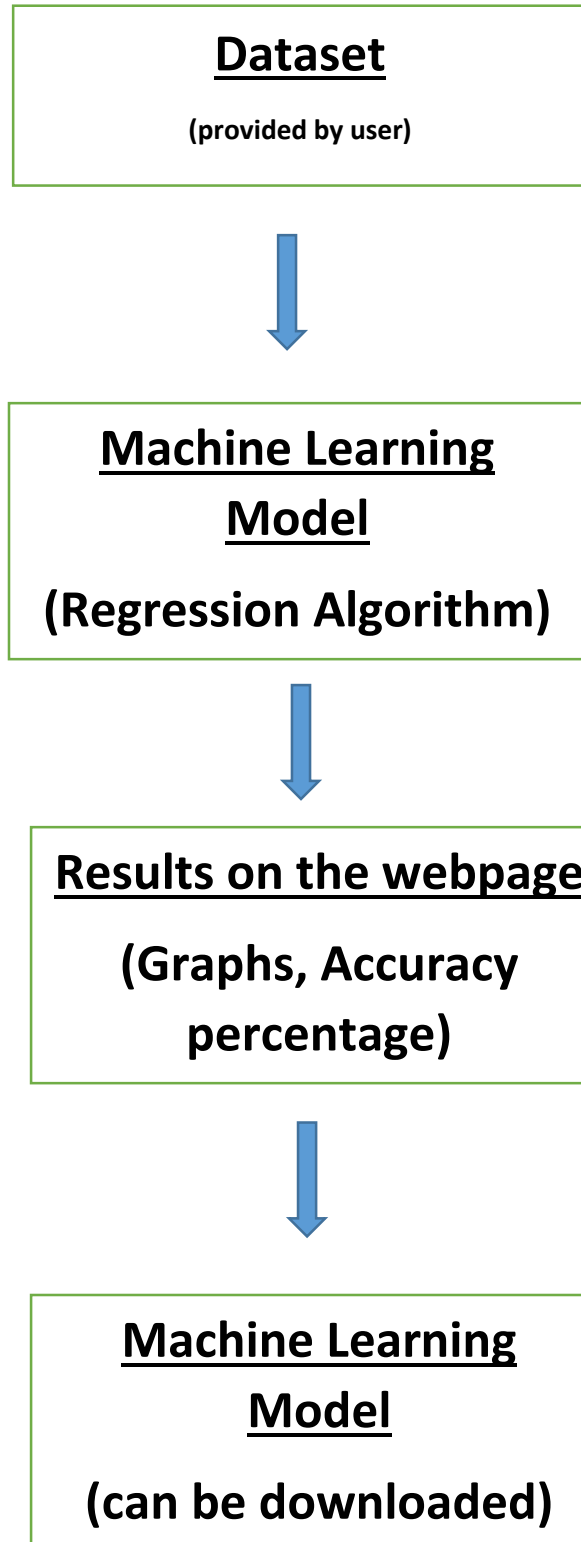
■ **Requirements:** -

- **Dataset** – The dataset must be .csv file and the output variable to be predicted must be at the last column of the dataset table.
- **Python libraries** – pandas, numpy, matplotlib, seaborn, scikit-learn, flask, werkzeug, os, pathlib, joblib and basic libraries

■ Architecture Diagram: -



■ Flow Diagram: -

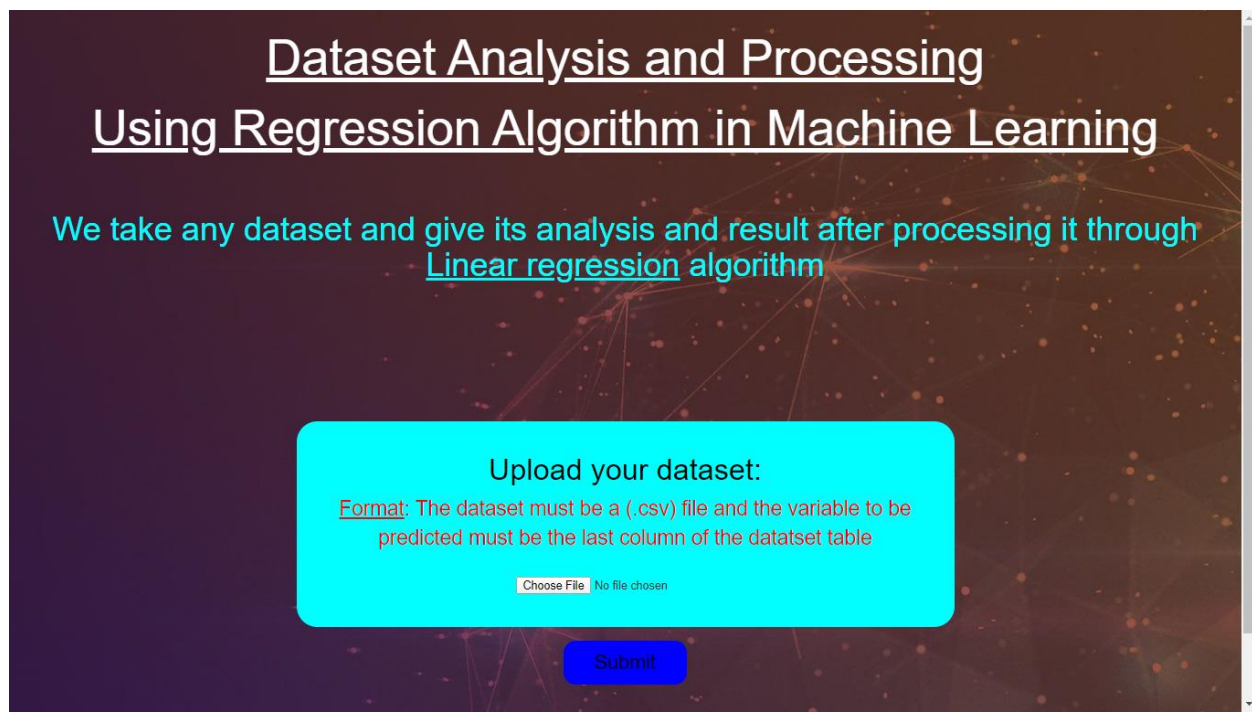


- **Code**: - Uploaded in the .zip file on Moodle. Open the main.py file and run using the server.

- **Screenshots**: -

- **Main page**

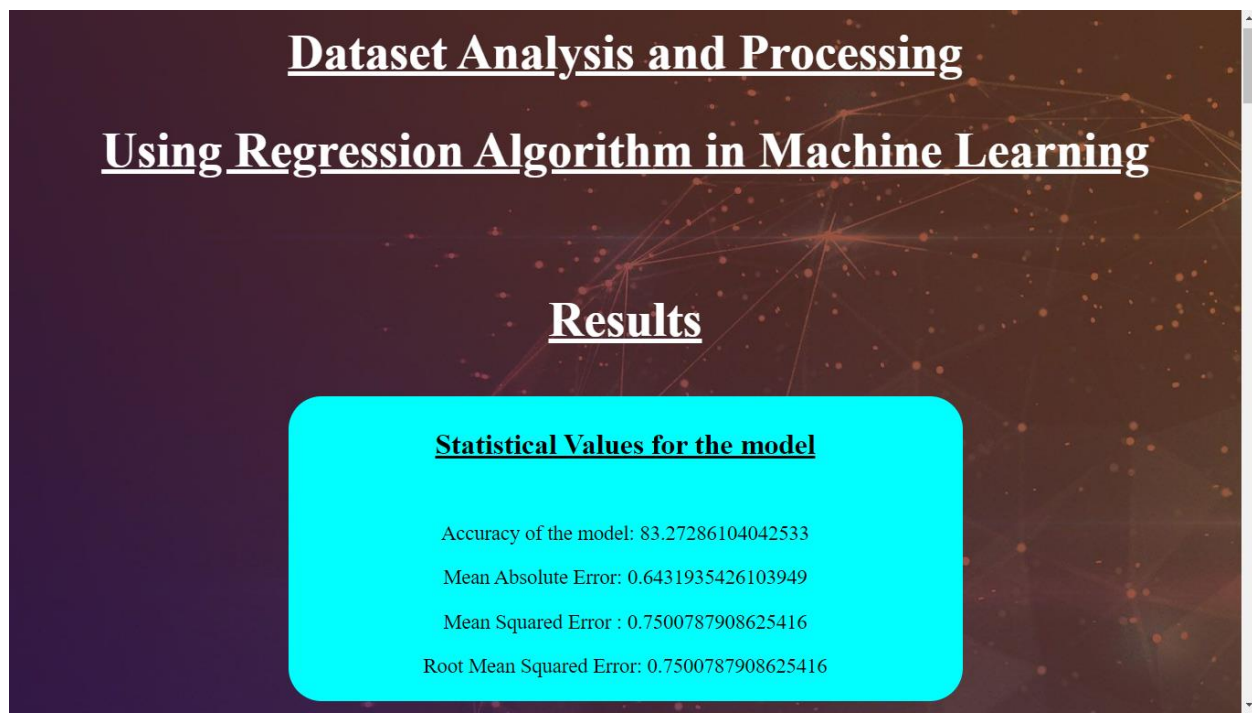
User uploads the dataset.



The screenshot shows a web application interface with a dark blue background featuring a network of orange dots and lines. The title 'Dataset Analysis and Processing' is displayed in white, followed by the subtitle 'Using Regression Algorithm in Machine Learning'. Below this, a cyan-colored text box contains the message: 'We take any dataset and give its analysis and result after processing it through Linear regression algorithm'. In the center, a large yellow rounded rectangle contains the text 'Upload your dataset:'. Below this, a red text box provides the format instructions: 'Format: The dataset must be a (.csv) file and the variable to be predicted must be the last column of the dataset table'. At the bottom of the yellow box is a file upload button labeled 'Choose File' with the text 'No file chosen' next to it. Below the yellow box is a red 'Submit' button.

○ Results page

The Model runs on the given dataset and gives the results in three forms- Statistical values, Dataset Analysis and Results of Linear Regression.

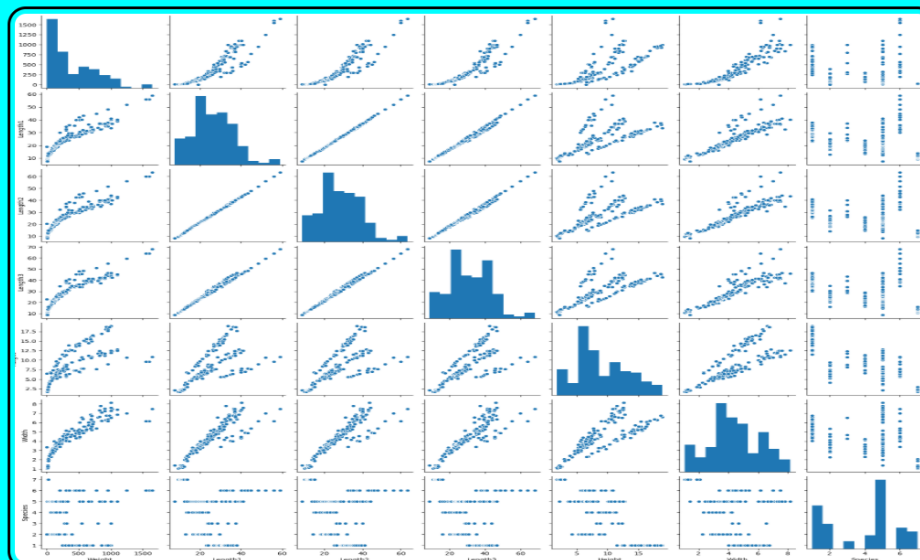


Dataset analysis

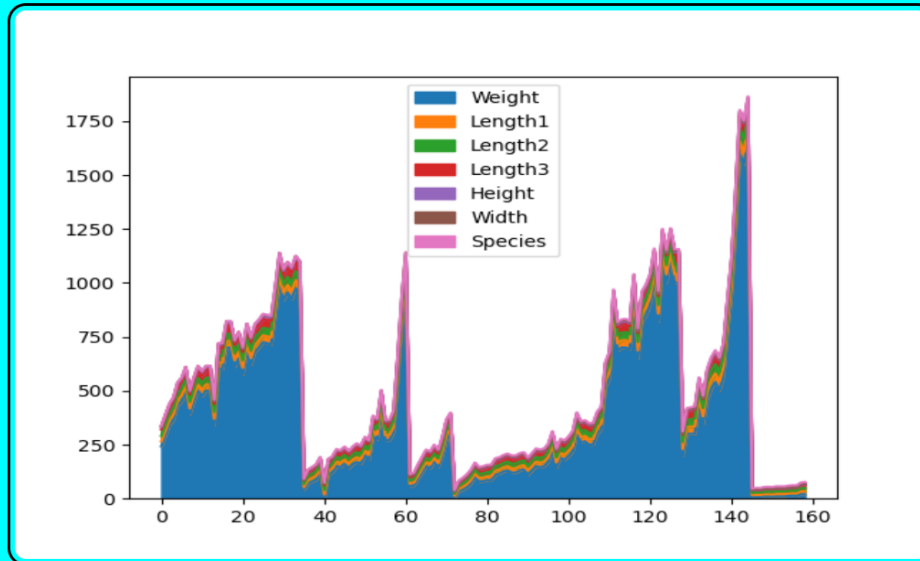
Weights Table

	Coefficient
Weight	0.002754156746377741
Length1	0.5350458769783998
Length2	0.9119974428142997
Length3	-1.2937220137104026
Height	-0.011263283804232742
Width	-0.8992161348433028

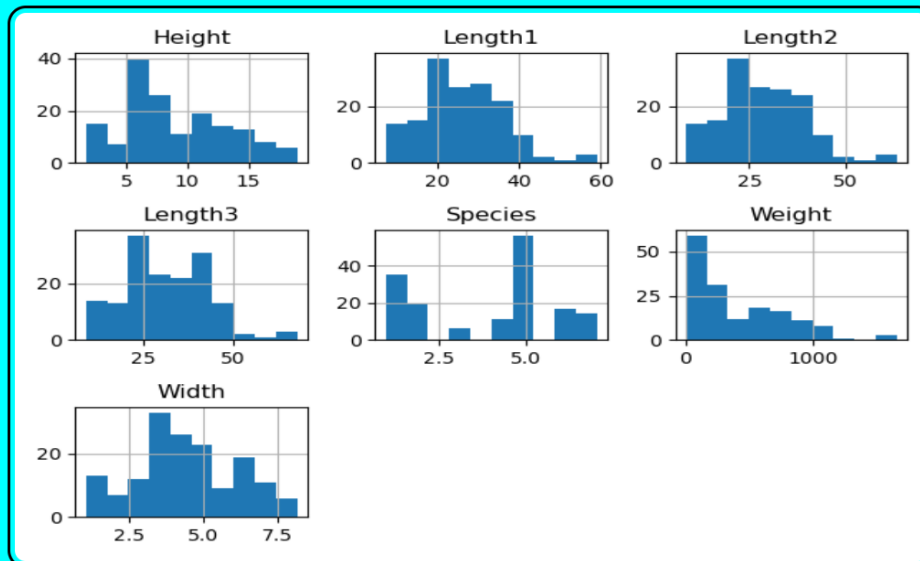
Pair Plots



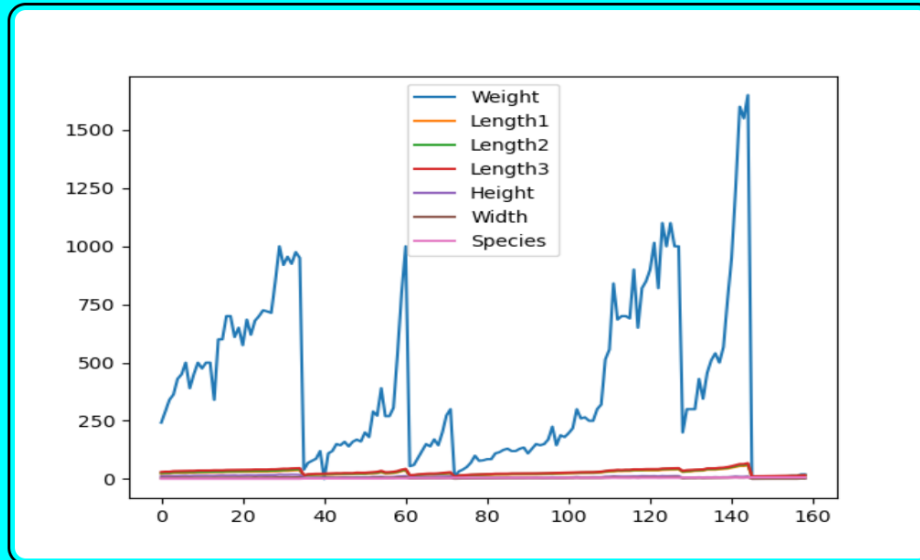
Area Plot



Histogram

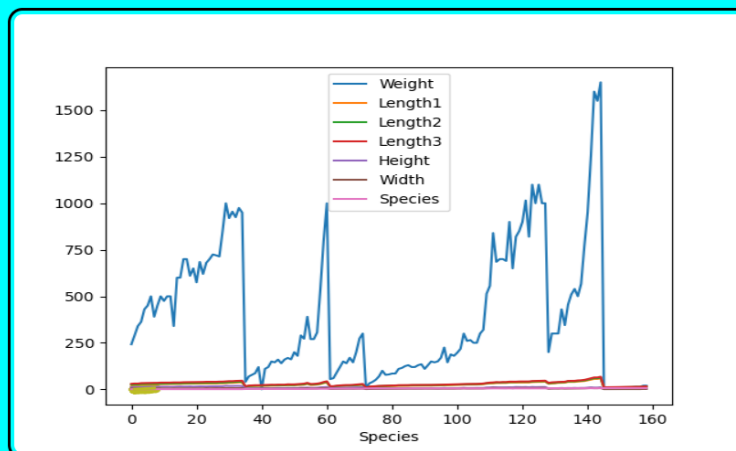


Line Plot

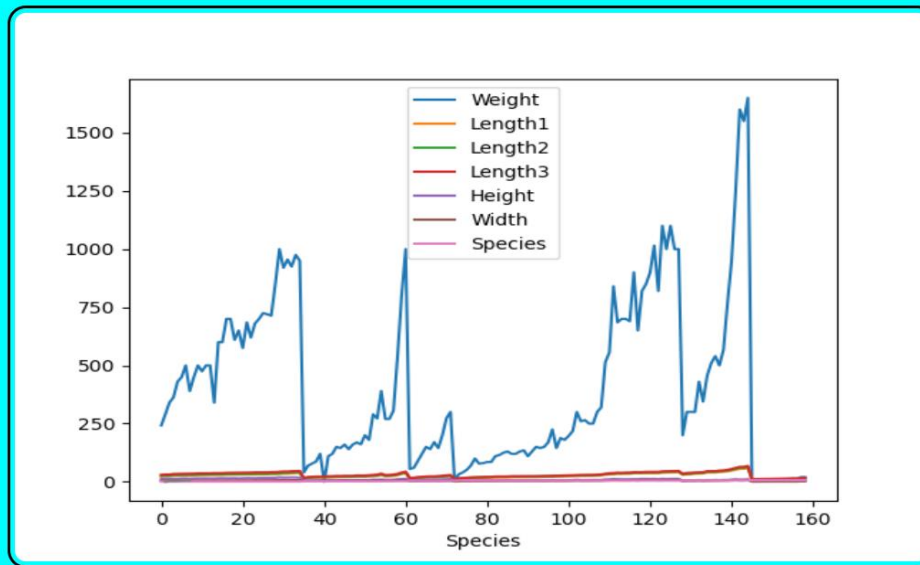


Results of Linear Regression

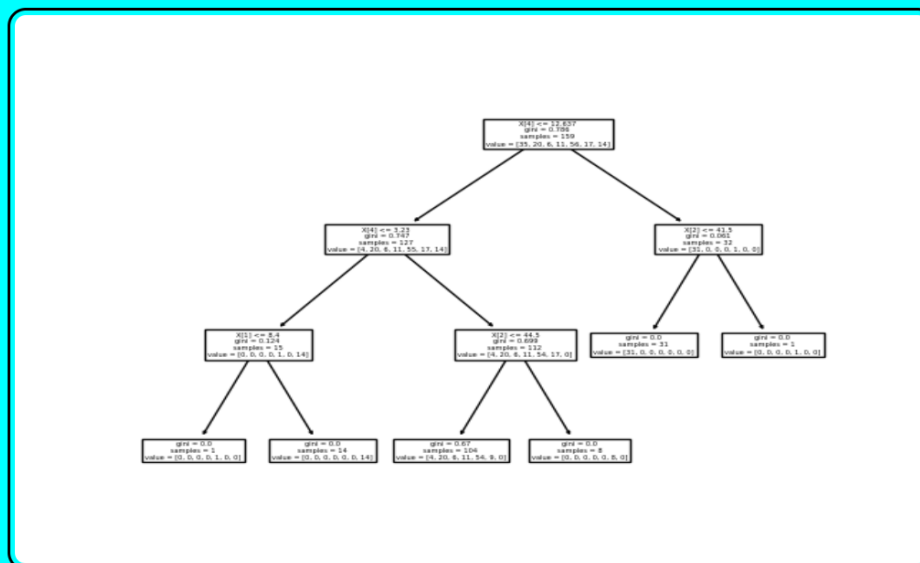
Regression Plot



Line Plot



Decision Tree



○ Download Option

There is an option for the user to download the model (. sav) and save it in his system for personal use.

The screenshot displays a web browser window at the URL 127.0.0.1:5000/upload. The main content area features a decision tree diagram with the following structure:

- Root Node: $X[4] \leq 3.23$, $gini = 0.767$, $samples = 127$, $value = [14, 20, 6, 11, 10, 17, 14]$
 - Left Child: $X[1] \leq 6.4$, $gini = 0.124$, $samples = 12$, $value = [0, 0, 0, 1, 0, 14]$
 - Left Leaf: $gini = 0.0$, $samples = 1$, $value = [0, 0, 0, 0, 1, 0, 0]$
 - Right Leaf: $gini = 0.0$, $samples = 11$, $value = [0, 0, 0, 0, 0, 14]$
 - Right Child: $X[2] \leq 44.3$, $gini = 0.899$, $samples = 112$, $value = [4, 20, 6, 11, 14, 17, 0]$
 - Left Leaf: $gini = 0.0$, $samples = 104$, $value = [14, 20, 6, 11, 14, 9, 0]$
 - Right Leaf: $gini = 0.0$, $samples = 8$, $value = [0, 0, 0, 0, 0, 0, 0]$
- Right Child: $X[2] \leq 41.3$, $gini = 0.001$, $samples = 12$, $value = [12, 0, 0, 0, 1, 0, 0]$
 - Left Leaf: $gini = 0.0$, $samples = 11$, $value = [12, 0, 0, 0, 0, 0, 0]$
 - Right Leaf: $gini = 0.0$, $samples = 1$, $value = [0, 0, 0, 0, 1, 0, 0]$

Below the decision tree, there is a blue button labeled "Download Model". At the bottom of the browser window, a file named "model (4).sav" is shown in the download bar. The Windows taskbar at the bottom indicates the time is 7:28 AM on 4/15/2020.

- **Conclusion:** - We have successfully made a Linear Regression model that takes in any dataset (in the mentioned format) and displays the results in three forms – Statistical Analysis, Dataset Analysis and Results of Linear Regression on that model. Graphs, and values are displayed for the dataset as results. We have also provided an option to save the model.

Thank You!!