**REPORT**

PART 1:

* Method used for Data pre-processing –
  + Understood the overview of the dataset.
  + Dropped columns which were irrelevant for prediction.
  + Check for columns with Nan (empty) values or mismatching data types.
  + Replace them with mean and mode and refactored invalid data
  + Checked for outliers and removed them.
  + Converted Objects-data types to categorical values using one-hot encoding, label encoding and frequency encoding.
* Data Set –
  + **Breeding Bird Atlas** 
    - Domain - Breeding bird observations based on geographical locations
    - Type of Data - The dataset consists of categorical and numerical data
    - Features – 15 features
    - Number of Samples – 361582
    - Mean for each numerical column:
      * Fed. Region – 5.85
      * Month – 49.77
      * Day – 49.5
      * Year – 1964.18
      * Temperature – 49.46
      * Average UB Student – 2.85
    - Standard Deviation for each numerical column:
      * Fed. Region – 5.83
      * Month – 28.65
      * Day – 28.79
      * Year – 190.06
      * Temperature – 17.32
      * Average UB Student – 0.49
    - Missing Values for each column:
      * Fed. Region – 5795
      * Block ID – 2718
      * Map Link – 4717
      * County – 10602
      * Common Name - 10530
      * Scientific Name – 7485
      * NYS Protection Status – 8470
      * Family Name – 2456
      * Family Description -4733
      * Breeding Behavior -5183
      * Month – 358156
      * Day – 352244
      * Year – 10480
    - Graph-
      * Box Plot are used for detecting outliers, for column ‘Year’, we could see there are few values which lies beyond the first and third quartiles

A graph with numbers and a number of objects

Description automatically generated with medium confidence

* + - * For Fed. Region we do not have any outliers

A blue rectangular object with black lines

Description automatically generated

* + - * Bar Plot for County, here the data is linearly reducing and the most frequent value being Parulidae

A graph of a bar plot

Description automatically generated

* + - * Comparing the feature ‘NYC protection status’ with ‘Breeding status’ column we see the most repeating Status is ‘Protected’

A graph of breeding status

Description automatically generated

* + - * When comparing ‘year’ with ‘Breeding Status’, we could see at year 1984 there were lot of breeding with ‘confirmed’ status occurred

A graph of breeding status

Description automatically generated

* + - * Correlation matrix here gives us the relation of each feature with the target column ‘Breeding Status’, the breeding behavior column had the most relation with the target column and the least being Temperature which provides just 0.1% relation

A chart with numbers and a number of numbers

Description automatically generated with medium confidence

* + Dataset 2 -
  + Dataset 3 -

PART 2: Penguin Dataset

Best Accuracy – 89.55 %

Loss Graph -

A graph with a blue line

Description automatically generated

Analysis – The loss graph for the model with a learning rate of 0.005 and 100,000 iterations shows that the model started with an initial loss of approximately 1.8. After a few iterations, the loss dropped significantly to 0.7, indicating rapid initial learning. The subsequent result suggests the model has reached a point where further training yields minimal improvements

Graphs for various Learning Rate and Iterations –

* Learning Rate – 0.001, Iterations – 100000, Accuracy – 79.1%

A graph with a blue line

Description automatically generated

* Learning Rate – 0.02 and Iterations – 200000, Accuracy – 89.55%

A graph with a blue line

Description automatically generated

* Learning Rate - 0.01 and Iterations – 250000, Accuracy – 88.06%

A graph with a line

Description automatically generated

* Learning Rate - 0.4 and Iterations – 300000, Accuracy – 88.06%

A graph with numbers and lines

Description automatically generated

* + A lower learning rate of 0.001 results in gradual convergence, which can prevent overshooting the minimum but may require more iterations
  + A medium learning rate of 0.005 often provides a good balance, leading to quicker convergence and improved accuracy
  + A high learning rate of 0.4 can lead to rapid learning but risks overshooting the optimal solution, causing fluctuations in loss and potentially lowering accuracy.
  + Increasing the number of iterations allows the model to refine its weights further, leading to lower loss. However, after a certain point, the benefits diminish, especially if the learning rate is not set appropriately.

Benefits / Drawbacks of Linear Regression –

* + Linear Regression performs requires less computational power and memory, making it suitable for smaller datasets or when computational resources are limited.
  + It performs well for linearly separable features making it an effective choice for binary classification tasks
  + Logistic regression is primarily used for binary classification problems
  + They are very sensitive to outliers which will affect the model’s accuracy