Report: Logistic Regression Model for Iris Virginica Classification

1. Objective

The primary objective of this project was to develop a machine learning model capable of accurately classifying an iris flower as belonging to the **Iris virginica** species or not, based on four physical features: sepal length, sepal width, petal length, and petal width. A **Logistic Regression** algorithm was selected for this binary classification task.

---

2. Exploratory Data Analysis (EDA) and Visualization

An initial analysis was performed on the Iris dataset, which consists of 150 samples with 50 samples for each of the three species (Setosa, Versicolor, and Virginica).

- **Key Findings:** The dataset was found to be clean with **no missing values**. The class distribution is perfectly balanced, which is ideal for training a classification model.
- **Visual Insights:**
  - A **pair plot** revealed strong visual separation between the species. Iris virginica consistently showed the largest petal length and petal width, making these features highly predictive.
  - A **correlation heatmap** indicated a very high positive correlation (+0.96) between petal length and petal width, suggesting that as one increases, the other does as well.
  - **Box plots** further confirmed that the distribution of petal measurements for Iris virginica is distinct from the other two species.

---

3. Data Preprocessing

To prepare the data for the logistic regression model, the following steps were taken:

1. **Binarization of Target:** The multi-class target variable ('species') was converted into a binary target. A value of **1** was assigned to `virginica` and **0** to the other two species (`setosa` and `versicolor`).
2. **Train-Test Split:** The dataset was split into an 80% training set (120 samples) and a 20% testing set (30 samples) to ensure the model could be evaluated on unseen data.
3. **Feature Scaling:** The features were scaled using **`StandardScaler`**. This process normalizes the data to have a mean of 0 and a standard deviation of 1, which is crucial for the optimal performance of a logistic regression model. The scaler was fitted only on the training data to prevent data leakage.

---

4. Model Performance and Results

The Logistic Regression model was trained on the scaled training data and evaluated on the scaled test data. The performance was exceptionally strong.

- **Accuracy:** The model achieved a **perfect accuracy of 1.00 (100%)** on the test set.
- **Classification Report:** The report below shows perfect precision, recall, and f1-score for both classes, indicating that the model made no errors in distinguishing 'Virginica' from 'Not Virginica' flowers.

Class	Precision	Recall	F1-Score	Support
Not Virginica	1.00	1.00	1.00	20
Virginica	1.00	1.00	1.00	10
Overall	**1.00**	**1.00**	**1.00**	**30**

Export to Sheets

- **Confusion Matrix:** The confusion matrix visually confirms the perfect performance. All 20 'Not Virginica' samples were correctly predicted, and all 10 'Virginica' samples were correctly predicted. There were zero false positives and zero false negatives.

5. Conclusion

The Logistic Regression model proved to be **highly effective** for this classification task. The distinct nature of the Iris virginica's physical features, combined with proper data preprocessing, allowed the model to achieve 100% accuracy on the unseen test data. This demonstrates that even a relatively simple linear model can be extremely powerful when the underlying data is well-structured and clearly separable. The model is reliable and robust for identifying Iris virginica flowers.