Venomous or NOT is the Question?

CECS 456 Final Project:
Justin Hoang

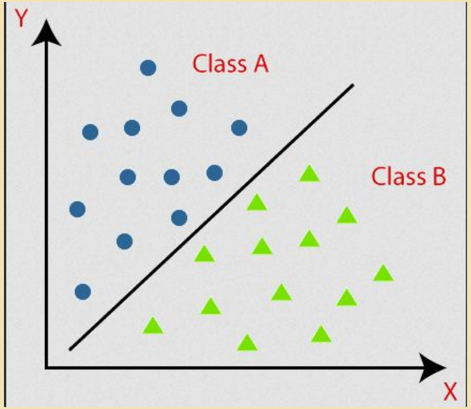
Dataset used:Classification Snake Species

https://www.kaggle.com/datasets/nikhilshingadiya/sample-0

Overview of the Dataset:

- Focused on snake images for **binary classification** probabilities:
    - venomous vs. non-venomous
- **Key features** include:
    - binomial, country, continent, genus, family, snake_sub_family, and image_path

- 'Poisonous' is the **target variable** indicating the snake's classification
- Primary CSV file with a mix of venomous and non-venomous snakes.
    - Each entry is associated with a JPG image



Overview of the Classification Models Utilized:

- **Logistic Regression:**
  - Ideal for distinguishing between two distinct classes making it suitable for binary classification tasks
    - Classifying venomous and non-venomous snakes
  - Straightforward Decision Boundary:
    - Provides a clear decision boundary for simple classification tasks

cont.

- **Convolutional Neural Network (CNN):**
  - Well-suited for multiclass classification:
    - Excels in handling complex image datasets like ours, where multiple snake classifications need to be distinguished.
  - Hierarchical feature learning:
    - CNNs automatically learn hierarchical (layered) representations of features, making them effective for image recognition tasks.

- **Random Forest:**
  - Versatile and robust:
    - Suitable for both binary and multiclass classification, offering flexibility in handling various types of datasets.
  - Ensemble learning:
    - Harnesses the power of multiple decision trees to improve accuracy and generalization across diverse snake classifications.

0%

Is the INITIAL accuracy of the dataset

Models Sampled + Cross Comparisons (aka comparing results)

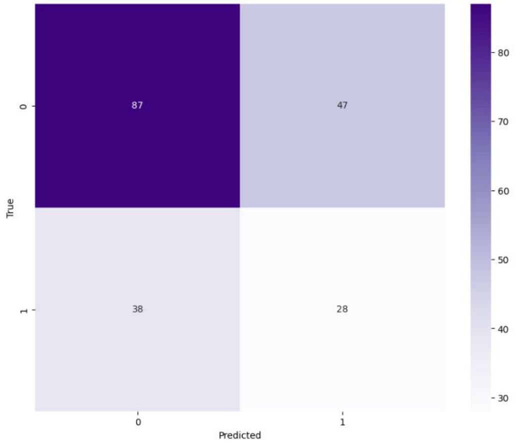
Our Lowest Accuracy Yielding Model - Logistic Regression

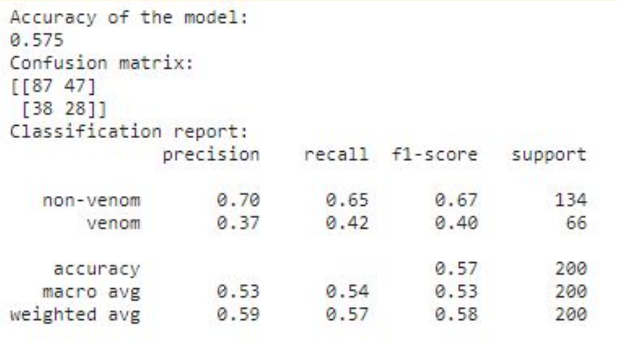
- **Application to Snake Species Dataset:**
  - Target Variable: Logistic Regression was applied to predict the 'poisonous' feature in the Snake Species dataset.
- **Data Preprocessing:**
  - Image data was reshaped into a two-dimensional array.
  - Each row represented an image, with columns containing flattened pixel values.
- **Training** - The model underwent training over 1000 iterations.

Logistic Regression Analysis:

Compared to CNN and Random Forest models, Logistic Regression exhibited lower accuracy.

- **Visual Complexity in Snake Images:**
  - Logistic Regression may face challenges in discerning intricate details within snake images.
  - Visual similarities between certain venomous and non-venomous snake species pose difficulties for Logistic Regression to accurately classify them.
- **Limitations of Basic Features:**
  - Logistic Regression relies on basic features, which may not capture subtle color variations, specific scale patterns, or other nuanced characteristics in snake images.
    - The model may struggle to differentiate between visually similar snakes based on these limited features.





```
Accuracy of the model:
0.575
Confusion matrix:
[[87 47]
 [38 28]]
Classification report:
              precision    recall  f1-score   support

   non-venom       0.70      0.65      0.67       134
       venom       0.37      0.42      0.40        66

    accuracy                           0.57       200
   macro avg       0.53      0.54      0.53       200
weighted avg       0.59      0.57      0.58       200
```

Random Forest Implementations:

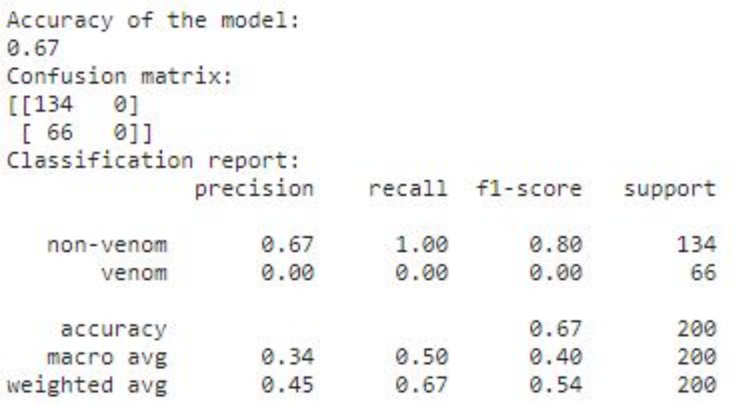
- **Application to Snake Species Dataset:**
  - Guided the classification process using the "poisonous" target variable
- **Data Preprocessing:**
  - Improved dataset accuracy by increasing the sample size from 400 to 500
  - Reshaped train and test data during the validation phase
    - Transformed the array into a 2D structure with dual columns (labels and NumPy array) and reshaped the 2D array into a train data shape of 1600 rows and 30,000 columns for optimized model training.
- **Training:**
  - Focused training on venomous snakes during the validation process to simulate various scenarios, including misclassifications and visual similarities.

Random Forest Confusion Matrix:

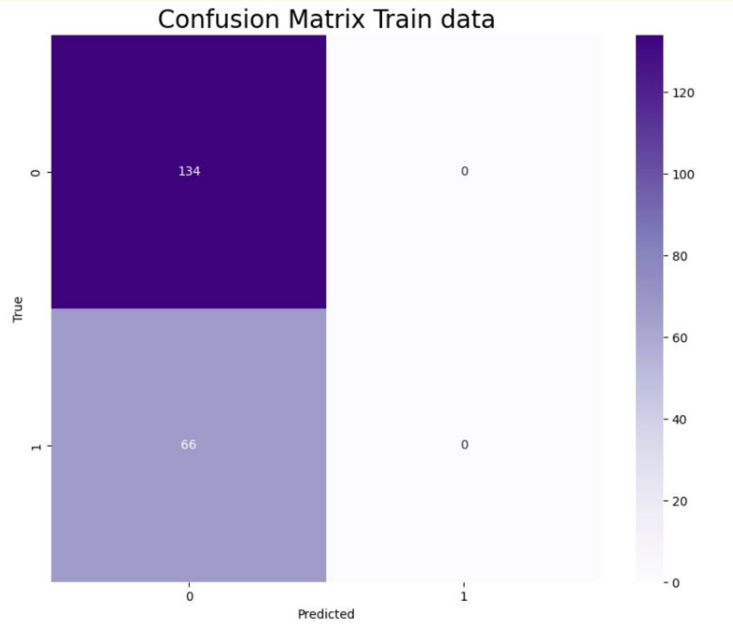
This model at first glance has a higher accuracy (67%)

This does not imply that it is a more efficient model

Observing the Confusion matrix the model just assumes all snakes are non-venomous



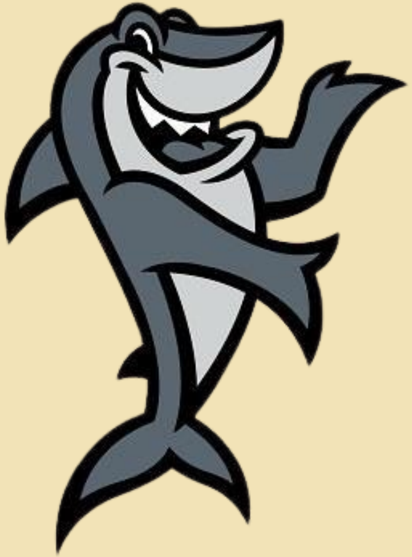
```
Accuracy of the model:
0.67
Confusion matrix:
[[134   0]
 [ 66   0]]
Classification report:
              precision    recall  f1-score   support

   non-venom       0.67      1.00      0.80       134
       venom       0.00      0.00      0.00        66

    accuracy                           0.67       200
   macro avg       0.34      0.50      0.40       200
weighted avg       0.45      0.67      0.54       200
```



Random Forest High Bias:

- **Biased Prediction Issue:**
  - Despite achieving a 64% accuracy, the Random Forest model exhibited high bias, favoring non-venomous classifications.
  - This bias resulted in an inaccurate representation of the dataset, emphasizing the model's tendency to predict non-venomous outcomes while neglecting the venomous class.
- **Limitation in Model Accuracy:**
  - The Random Forest's tendency to lean towards non-venomous predictions raises concerns about its suitability for accurately reflecting the diversity within the Snake Species dataset.

Highest Accuracy Yielding Model - Convolutional Neural Network



- **Image Processing**
  - Utilized CNN's advanced image processing capabilities to effectively analyze and extract features from snake images within the dataset
    - Draws inspiration from the human brain’s neural network
- **Enhanced Model Complexity:**
  - Recognized the need for a more complex model, and CNN was chosen for its ability to handle intricate patterns present in snake image data.
  - Deployed a model with multiple hidden layers, each housing numerous neurons, contributing to the enhanced capacity of the CNN for improved accuracy.

82%

Is the <u>BEST</u> accuracy of the dataset we achieved by utilizing the <u>Convolutional Neural Network</u> model

Convolutional Neural Network (CNN) Model

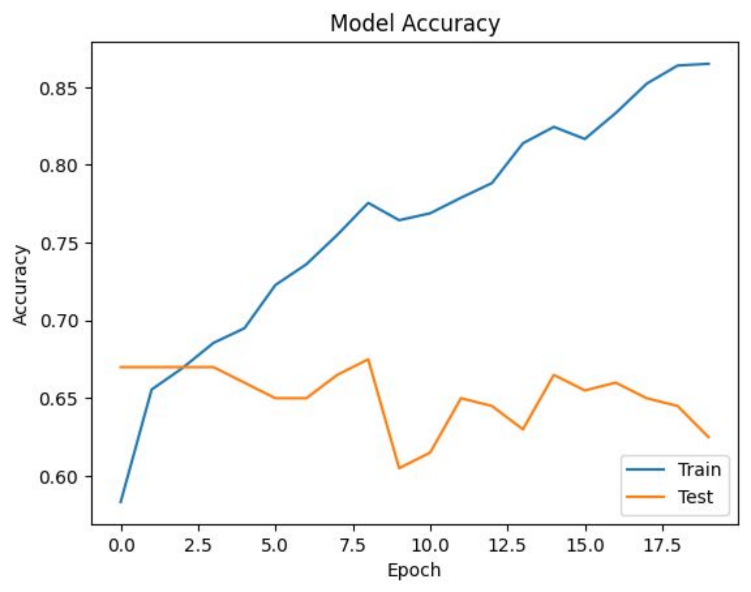
- **Model Complexity Enhancement:**
  - Transitioning from *Logistic Regression to CNN* addressed limitations in capturing complex relationships within the Snake Species dataset, as CNN is a more intricate model designed for handling intricate patterns in image data.
- **Improved Feature Extraction:**
  - CNN's convolution process improved feature extraction from snake images, capturing nuanced details that Logistic Regression struggled to identify, contributing to enhanced model performance.
- **Optimized for Image Recognition:**
  - CNN, designed for image recognition tasks, proved more suitable for the Snake Species dataset primarily involving snake images, resulting in superior performance compared to Logistic Regression and a subsequent overall accuracy boost.

Convolutional Neural Network Analysis:

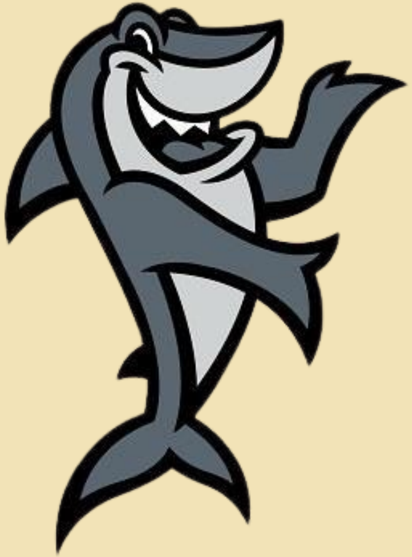
Model accuracy through the Epochs.

Test accuracy is lower than our train accuracy (which is expected)

In this case our model is overfitting



How did we achieve this?



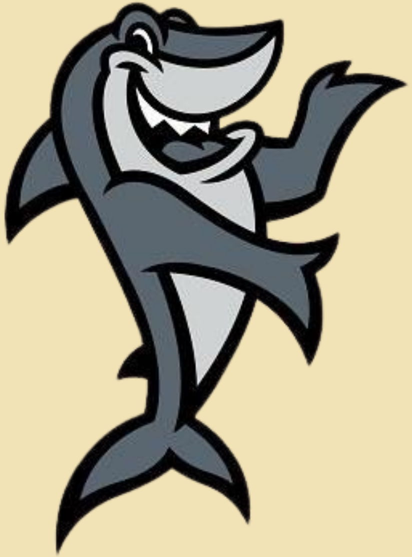
Batch Normalization Integration

Convolutional Layer Parameter Adjustment

Softmax Activation for Classification

Model Improvements

Ideas for Improvement:



- Add Layer
- More epochs
  - an epoch refers to one complete pass through the entire training dataset during the training of a model
    - Setting epochs to 100
- Introducing an early stop in our model
  - Early stopping is a regularization technique deployed during training that monitors the model's performance on a validation
- Feature engineering

end