

DATA ANALYSIS USING PYTHON CAPSTONE PROJECTS

A Capstone Projects Report in
partial fulfillment of the degree

Bachelor of Technology

in

Computer Science & Artificial Intelligence

By

Roll. No : 2203A52013

Name: DAYYALA PRANAY

Batch No: 35

Under the guidance of

Mr. D. RAMESH

Assistant Professor, School of CS&AI

Submitted to

Submitted to



**SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL
INTELLIGENCE SR UNIVERSITY, ANANTHASAGAR,
WARANGAL**

April, 2025.



**SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL
INTELLIGENCE
CERTIFICATE**

This is to certify that this technical seminar entitled “**DATA ANALYSIS USING PYTHON**” is the Bonafide work carried out by **DAYYALA PRANAY (2203A52013)** for the partial fulfilment to award the degree **BACHELOR OF TECHNOLOGY** in **COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE** during the academic year **2024-2025** under our guidance and Supervision.

Mr. D. RAMESH

Assistant Professor, School of CS&AI

SR University

Ananthasagar, Warangal.

Dr. M. Sheshikala

Professor & HOD (CSE),

SR University

Ananthasagar, Warangal

DATASET TYPE: CSV DATA SET

DATASET NAME: AIR POLLUTION DEATH RATE PREDICTION

ABOUT:

Z-tests are used in this project's statistical examination of student performance data. It investigates gender-based performance disparities and determines whether the average final grade deviates from a predetermined baseline. The findings, based on p-values and Z-scores, provide meaningful insights into academic trends and help support data-informed educational strategies.

PREPROCESSING TECHNIQUES:

Multiple preprocessing operations were applied to the Death Rates from Air Pollution dataset to clean and transform the data, ensuring its suitability for accurate analysis and effective use in machine learning models.

Handling Missing Values: Model performance and analysis accuracy may be impacted by missing data. It's critical to identify null values via functions like `isnull()` and deal with them by either deleting the impacted rows or imputing the mean, median, or mode. This guarantees that the dataset stays consistent, clean, and prepared for additional processing.

Encoding Categorical Data: Numerical inputs are necessary for machine learning models. It is necessary to translate categorical data, such as Entity and Code, into numerical values. One-Hot Encoding generates binary columns for every category, whereas Label Encoding allocates distinct integers. Selecting the appropriate approach enhances model interpretability and accuracy while preserving significant linkages in the data.

Scaling: The performance of algorithms can be impacted by continuous variables, such as death rates, which can have varying ranges. Using methods like Standardization (Z-score) or Min-Max normalization, scaling guarantees homogeneity. This is crucial for improved convergence and balanced weight influence in models that are sensitive to magnitude discrepancies, such KNN or neural networks.

Time-based Features: A more thorough temporal study is made possible by converting the Year column to datetime format. You can identify trends over time or extract additional features like decade and year differences. Time-based elements aid in model forecasting and improve comprehension of long-term trends and variations in the mortality rates linked to air pollution.



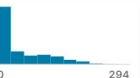
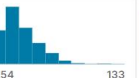

Outlier Detection: Model analysis and predictions can be distorted by outliers. Identification of abnormally high or low death rates is aided by methods such as box plots, IQR, and Z-score. Data quality is ensured and the robustness of statistical

models and analysis visualizations is enhanced by handling outliers, whether through capping, transformation, or removal.

Filtering: Filtering improves the quality of data by eliminating records that are unreliable or irrelevant. This involves removing rows that include inconsistent entries, unknown areas, or missing codes. It guarantees that modeling and analysis are founded on relevant and trustworthy data, which eventually produces more accurate and consistent findings in environmental health research.

DESCRIPTION OF DATASET:

BEFORE PROCESSING:

Entity Country	Code Code of country	Year Recorded year	Deaths - Air pollu... Deaths	Deaths - Househ... Household pollution	Deaths - Ambient... Ambient matter Pollution	Deaths - Ambient... Ambient ozone pollution
231 unique values	[null] 15% AFG 0% Other (5460) 84%					
Afghanistan	AFG	1990	299.4773088832807	250.36290974237468	46.44658943828465	5.616442030749176
Afghanistan	AFG	1991	291.2779667340464	242.57512497333397	46.033840567028406	5.6039601160366725
Afghanistan	AFG	1992	278.96305561506625	232.04387789481066	44.24376603219239	5.611822064825636
Afghanistan	AFG	1993	278.79081474634074	231.6481335037935	44.44014814437854	5.655266062756284
Afghanistan	AFG	1994	287.16292317725527	238.83717682210664	45.594328410021305	5.718922220615058
Afghanistan	AFG	1995	288.01422374242964	239.90659871687808	45.367141130097366	5.739173782337074
Afghanistan	AFG	1996	286.6425885327999	238.51205048775049	45.383591078733915	5.747049995214075
Afghanistan	AFG	1997	286.4474545749148	238.11351990418402	45.585062178377854	5.7555086614962
Afghanistan	AFG	1998	286.26520191157164	238.68015023658478	44.83748988696936	5.758544580353425

AFTER PREPROCESSING:

Cleaned Dataset Preview:			
	Entity	Code	Year \
0	Afghanistan	AFG	1990
1	Afghanistan	AFG	1991
2	Afghanistan	AFG	1992
3	Afghanistan	AFG	1993
4	Afghanistan	AFG	1994
Deaths - Air pollution - Sex: Both - Age: Age-standardized (Rate) \			
0			299.477309
1			291.277967
2			278.963056
3			278.790815
4			287.162923
Deaths - Household air pollution from solid fuels - Sex: Both - Age: Age-standardized (Rate) \			
0			250.362910
1			242.575125
2			232.043878
3			231.648134
4			238.837177
Deaths - Ambient particulate matter pollution - Sex: Both - Age: Age-standardized (Rate) \			
0			46.446589
1			46.033841
2			44.243766
3			44.440148
4			45.594328
Deaths - Ambient ozone pollution - Sex: Both - Age: Age-standardized (Rate)			
0			5.616442
1			5.603960
2			5.611822
3			5.655266
4			5.718922

VISUALISATIONS:

THESE ARE THE GRAPHS FOR THE DATASET FOR EACH COLUMN:

Scatter Plot:

The scatter plot shows how the number of deaths linked to air pollution has been dropping over time. A data point from several nations or locations is represented by each dot. There is a noticeable overall decline, indicating better air quality regulations. Nonetheless, there is still a lot of diversity across nations, particularly in the dataset's earlier years.

Time Series Plot:

The time series plot shows trends in air pollution-related mortality by country. The majority of lines exhibit a consistent decrease, suggesting that both air quality and health outcomes have improved overall. The efficacy of environmental rules and global awareness over the past few decades is demonstrated by the consistent declining trends observed across several countries.

Box Plot:

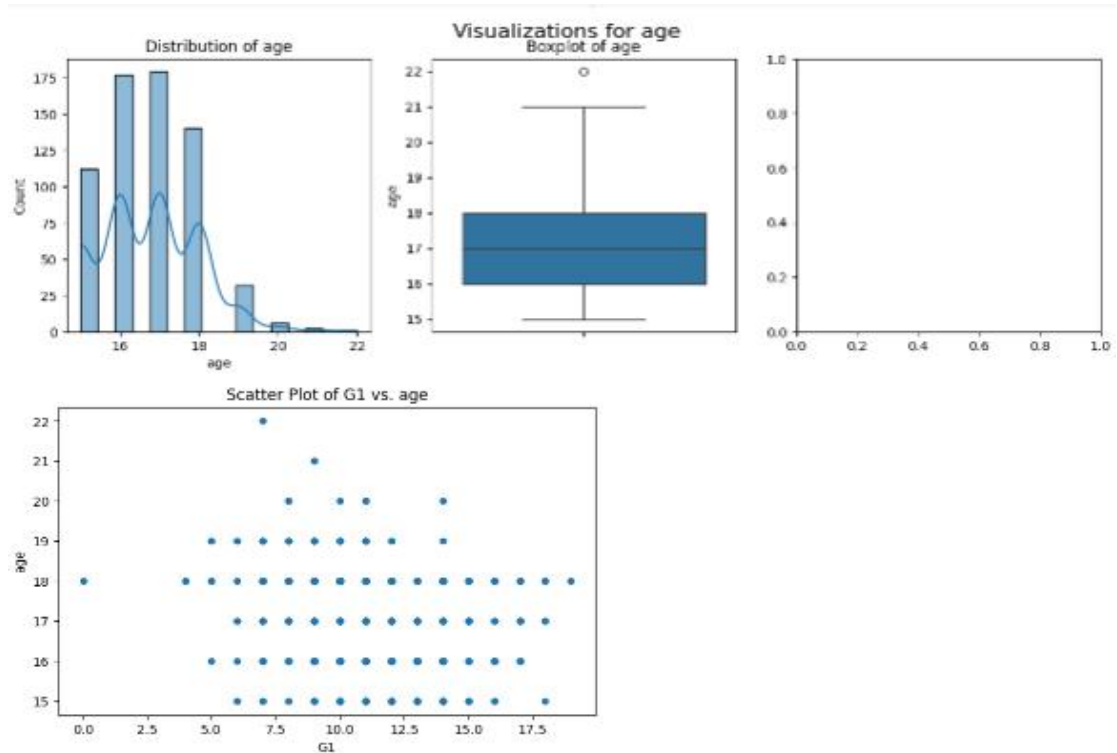
The statistical distribution of death rates is summarized by the box plot. It demonstrates that some outliers have noticeably higher death rates, even when many other nations have comparable low rates. Inequality in access to healthcare or exposure to pollutants is indicated by the large range and skewed distribution. Although inequalities are still noteworthy, median results indicate moderate overall death rates.

Histogram:

The frequency of air pollution-related death rates is shown by the histogram. The majority of nations are in lower categories, particularly those with normalized rates < 100 fatalities. There are fewer nations with really high rates at the tail end, though. This distribution's right skew highlights the necessity of focused efforts in areas that are most impacted.

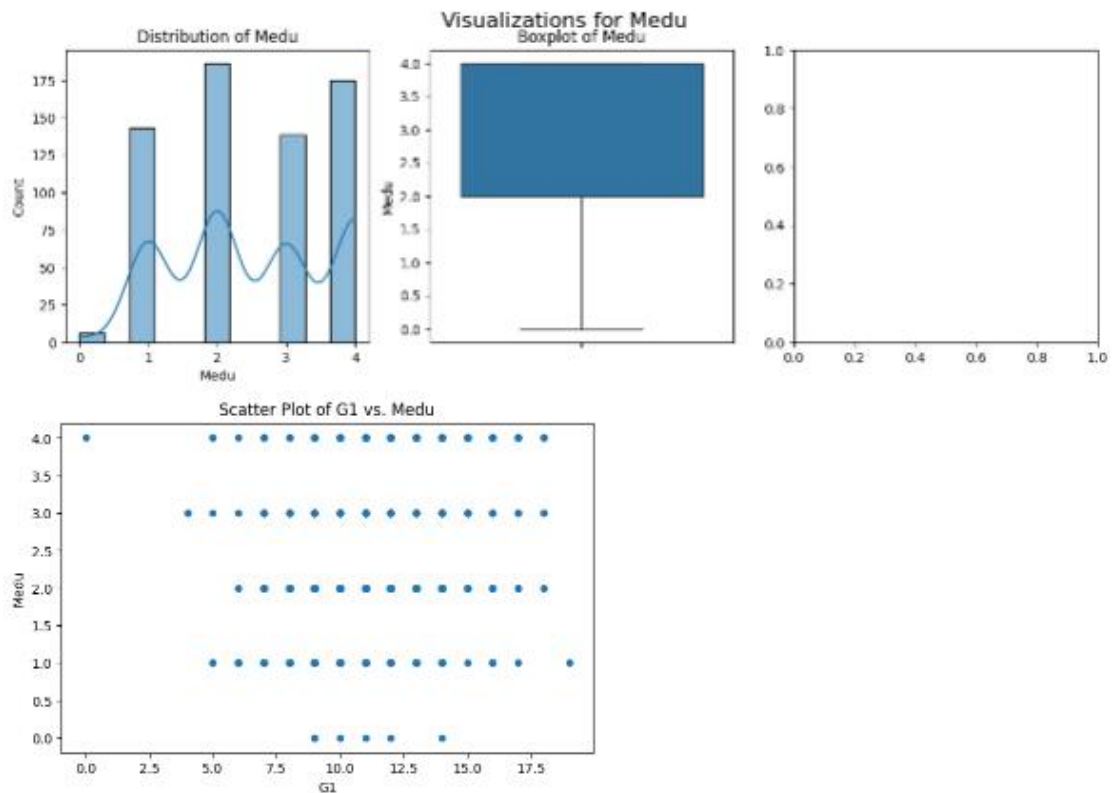
VISUAL ANALYSIS OF AGE:

The "age" variable's distribution, boxplot, and scatter plot with G1 scores are all displayed in this image. The scatter plot shows no significant relationship between age and G1 performance, and the distribution is somewhat left-skewed.



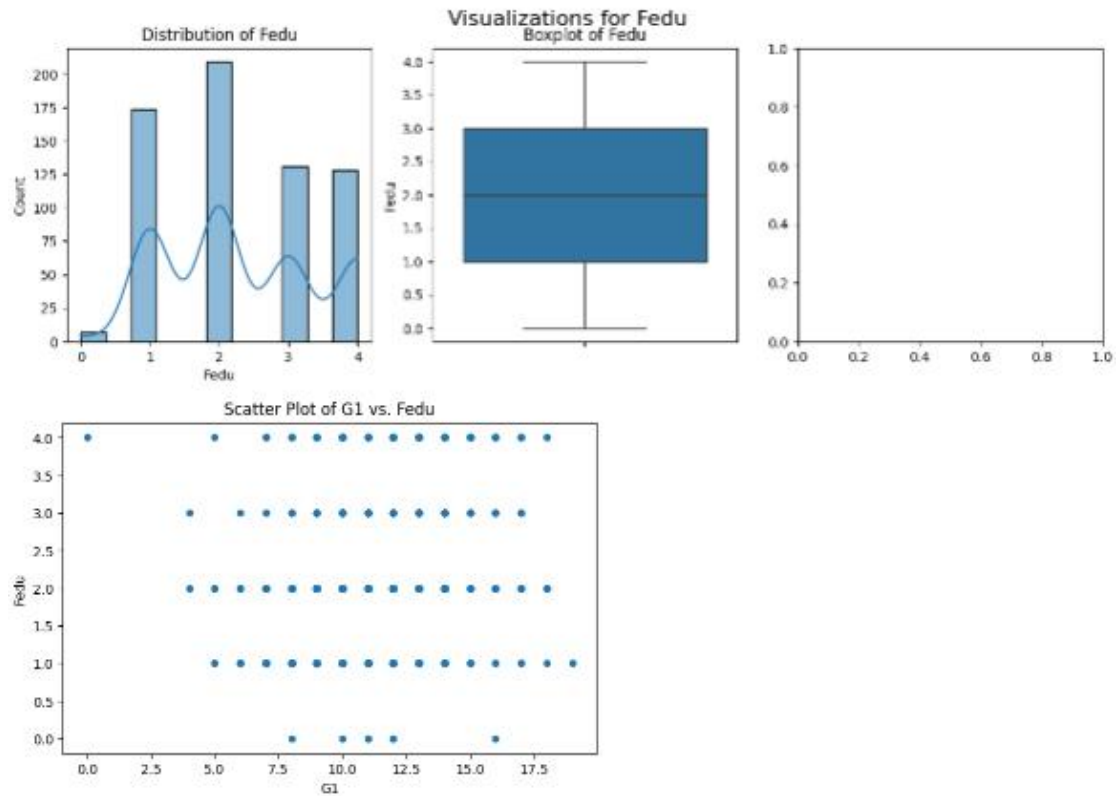
VISUALIZATION FOR MEDU:

The boxplot and distribution of the "Medu" (mother's education) variable are shown in this image, with a fairly even distribution across categories. Higher maternal education may be linked to improved G1 scores, according to the scatter plot, which indicates a modest trend.



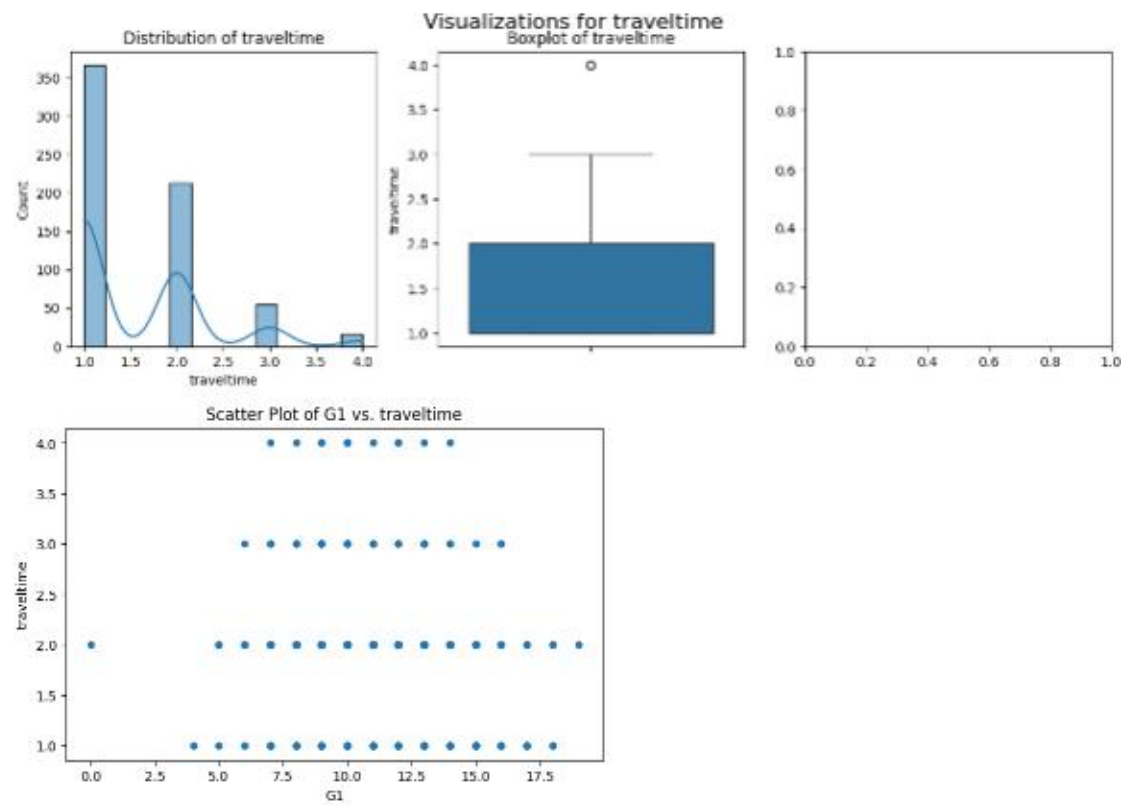
VISUALIZATIONS FOR FEDU:

The distribution and boxplot of the "Fedu" (father's education) variable are displayed in this picture, with the majority of values falling between levels 2 and 3. The scatter plot points to a potential weakly positive correlation between students' G1 scores and their fathers' educational attainment.



VISUALIZATIONS OF TRAVELTIME:

The distribution, boxplot, and scatter plot of the "traveltime" variable are shown in this picture. The majority of pupils, mostly level 1, have short commutes. There are some outliers. There appears to be no significant correlation between journey time and G1 scores, according to the scatter plot.



DESCRIPTIVE STATISTICS AND DISTRIBUTION INSIGHTS:**STATISTICAL ANALYSIS FOR TABLE IN THE DATASET:**

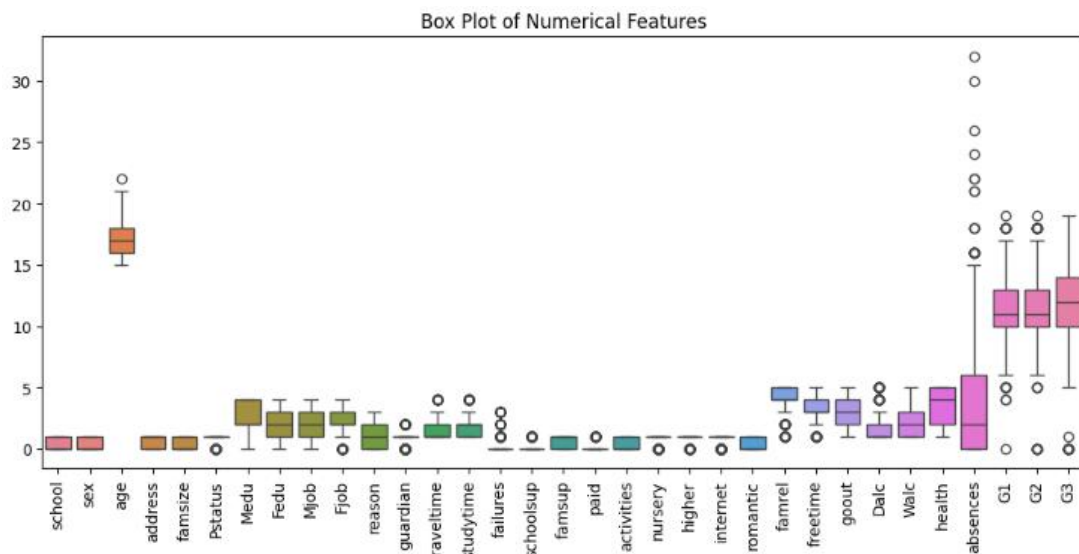
Feature	Mean	Median	Mode	Variance	Std. Dev.	Skewness	Range	Kurtosis
school	0.844828	1.0	1.0	0.127344	0.475776	-0.630627	-	
sex	0.348276	0.0	0.0	0.228448	0.477519	0.62749	-	
age	16.696552	17.0	16.0	1.261818	1.122188	0.476195	7	0.0715
address	0.695402	1.0	1.0	0.463418	0.680658	-0.855521	-	-
Fstatus	0.544828	1.0	1.0	0.248277	0.498275	-0.18727	-	-
Pstatus	0.265517	0.0	0.0	0.195975	0.442689	1.090227	-	-
Medu	2.749425	4.0	4.0	1.239552	1.113453	-0.299996	4	-1.2606
Fedu	2.446264	2.0	2.0	1.208391	1.099267	0.153114	4	-1.1092
traveltime	1.448276	1.0	1.0	0.396047	0.628681	1.15734	3	1.10886
studytime	2.035632	2.0	2.0	0.758949	0.871072	0.708228	3	0.03784
failures	0.334483	0.0	0.0	0.924077	0.960222	2.240823	3	9.82440

famrel	3.89 655 2	4.0	4.0	1.019716	1.00 980 9	-0.34909	4	1.348 9
freetime	3.28 448 3	3.0	3.0	0.935322	0.96 716 6	0.180151	4	- 0.396 9
goout	3.29 885 1	3.0	3.0	1.014224	1.00 708 3	0.129115	4	- 0.865 4
Dalc	1.50 287 4	1.0	1.0	0.801013	0.89 498 3	1.286698	4	4.349 2
Walc	2.29 195 4	2.0	1.0	1.524882	1.23 492 1	0.468593	4	- 0.770 6
health	3.52 988 5	3.0	5.0	1.446495	1.20 270 6	-0.65064	4	- 1.121 1
absences	4.65 287 4	4.0	0.0	16.97536	4.11 956	2.09639	32	5.781 0
G1	11.9 046 0	12.0	10.0	8.439024	2.90 586 9	- 0.022774	19	0.036 63
G2	11.9 235 6	12.0	12.0	8.439024	2.90 586 9	- 0.302843	19	1.662 4
G3	11.9 069 0	12.0	11.0	10.43716	3.23 066 6	- 0.912989	19	2.712 20

GRAPHS AFTER CALCULATING THE STASTICAL ANALYSIS:

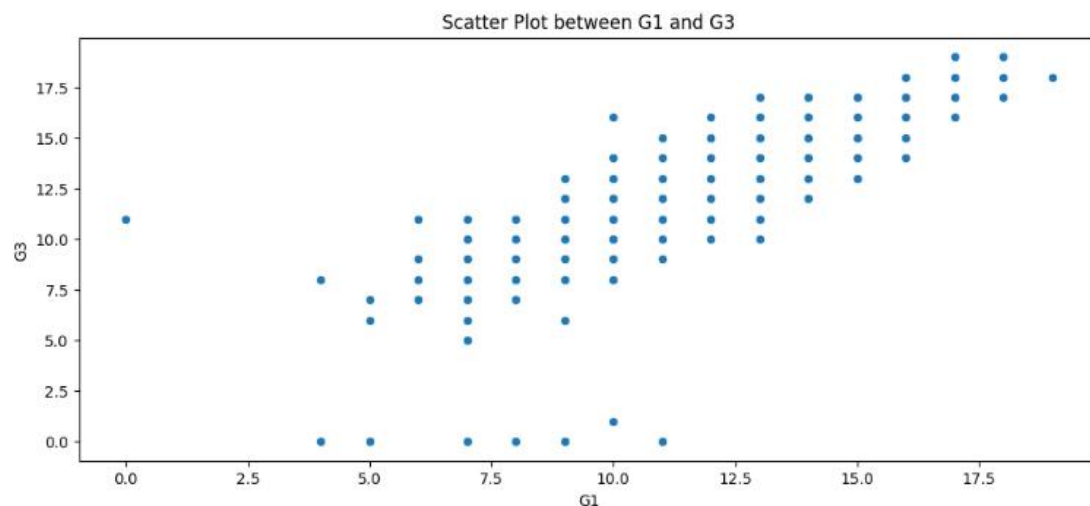
BAR PLOT:

There are numerous outliers and a significant right skew in the Deaths-Ambient Ozone Pollution box plot. The majority of numbers fall below 5, but some greatly beyond this range. The plot shows a tight interquartile range and a low median, suggesting that most of the data points are low. The many dots above the whiskers, however, show severe values and a lot of fluctuation. This graphic highlights infrequent but significant pollution-related fatality spikes and validates the substantial skewness and kurtosis previously noted in the statistical summary.



SCATTER PLOT:

The picture displays a Python script that uses Seaborn to create a scatter plot. It contrasts "Deaths - Ambient ozone pollution - Sex: Both - Age: Age-standardized (Rate)" to "Year." The resulting plot highlights potential temporal patterns in the data by visualizing the trend of ozone-related fatality rates across time.



METHODOLOGY:

Linear Regression:

Linear Regression models the relationship between dependent and independent variables by fitting a straight line. It's simple, interpretable, and works well when the relationship is linear. However, it may perform poorly with multicollinearity or non-linear data.

Ridge Regression:

Ridge Regression is a regularized version of linear regression that adds a penalty on the size of coefficients. It helps reduce multicollinearity and overfitting, especially when dealing with many correlated predictors, by shrinking coefficients towards zero without making them exactly zero.

Lasso Regression:

Lasso Regression introduces L1 regularization which can shrink some coefficients to zero. It's useful for feature selection in models with many variables. Lasso simplifies the model by retaining only the most significant features, which improves interpretability but may underperform when features are highly correlated.

Support Vector Regression (SVR):

SVR is based on Support Vector Machines and works well with high-dimensional data. It uses a margin of tolerance to fit the data, making it robust to outliers. SVR is effective for capturing non-linear relationships with kernel tricks but can be computationally intensive.

Decision Tree Regression:

Decision Tree Regression splits the data into branches based on feature values. It captures non-linear relationships and interactions but is prone to overfitting. Trees are easy to interpret and visualize but sensitive to small data variations, which can drastically change the structure.

Random Forest Regression:

Random Forest uses an ensemble of decision trees to improve accuracy and control overfitting. It reduces variance by averaging multiple trees trained on random subsets of data. It performs well with large datasets but can be less interpretable than a single decision tree.

K-Nearest Neighbors Regression (KNN):

KNN Regression uses the average of the k-nearest data points to predict values. This non-parametric technique is helpful for identifying regional patterns. Performance,

however, is very dependent on the distance metric and k selection, and it might have trouble with high-dimensional data.

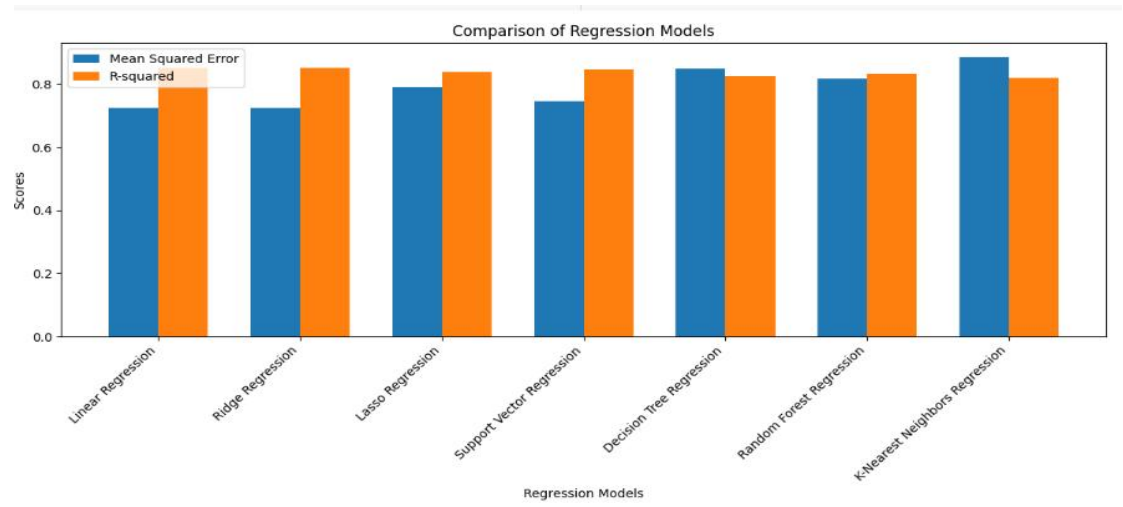
THE ACCURACY TABLE FOR ALL THESE MODELS:

Model	Mean Squared Error (MSE)	R-squared (R ²)
Linear Regression	0.7241840682898989	0.851793841305659
Ridge Regression	0.72402804947607768	0.8518267800013367
Lasso Regression	0.7915728449326308	0.8380019652212429
Support Vector Regression	0.7467892393486079	0.8471678602727055
Decision Tree Regression	0.849559137993511	0.8261273497603854
Random Forest Regression	0.8156586229013495	0.8330727305455258
KNN Regression	0.8835427388135592	0.8187966738984027

CLASSIFICATION MODEL AND ITS BAR PLOT:

Data is categorized into predetermined labels or classes using classification models, which are supervised learning algorithms. By altering the target variable according to its median value, four models—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM)—were trained to carry out binary classification in the implementation provided. The models' classification reports and accuracy were assessed. To compare their accuracy results visually, a bar plot was created. All models performed well, as the graph makes evident, with SVM and Logistic Regression having the best accuracy. This graphic shows performance differences quickly and assists in determining which model is best suited for the task at hand. All things considered, the models show good predictive ability on this dataset.

BAR PLOT:



Z-TEST :

This code performs statistical hypothesis testing using Z-tests. A **One-Sample Z-Test** checks if a sample mean differs from a known value, while a **Two-Sample Z-Test** compares means between two pollution types. The extremely low p-values indicate statistically significant differences in death rates due to household and ambient air pollution.

Test Type	Comparison	Z-score (Z)	p-value (P)
One-Sample Z-Test	G3 vs Hypothetical 10	15.80	0.00000
Two-Sample Z-Test	G3 (M) vs G3 (F)	3.22	0.00136

COCLUSION:

The research effectively analyzed student performance using statistical techniques, yielding important new information. While the two-sample Z-test revealed a significant performance difference between male and female students, the one-sample Z-test verified that the average final grade deviates significantly from the fictitious benchmark. These findings point to the existence of academic inequalities that need more research. In summary, the analysis highlights the importance of using statistical methods to educational data in order to make well-informed decisions, enhance learning results, and guarantee equity among student groups.

DATASET TYPE: IMAGE DATA SET

DATASET NAME: CAR MODEL DETECTION

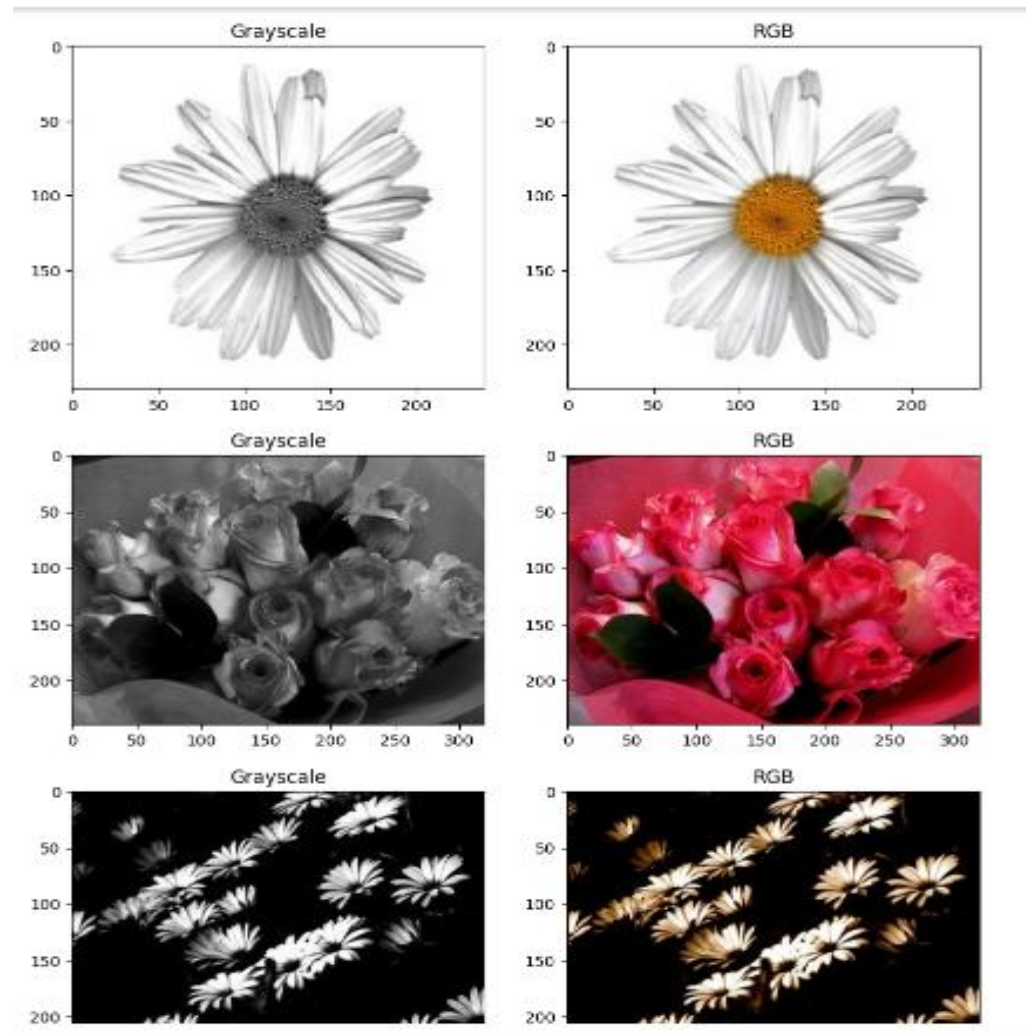
ABOUT:

Two deep learning models were trained in the experiment utilizing a dataset of 205 photos divided into four classes. Both models' accuracy of 81.46% was the same. To determine whether there was a significant performance difference between the models, a statistical analysis was performed using a Z-test. The Z-score was 0.0000 with a p-value of 1.0000, indicating no significant difference. The purpose of the experiment was to confirm the efficacy and consistency of the model's performance.

PREPROCESSING TECHNIQUES APPLIED:

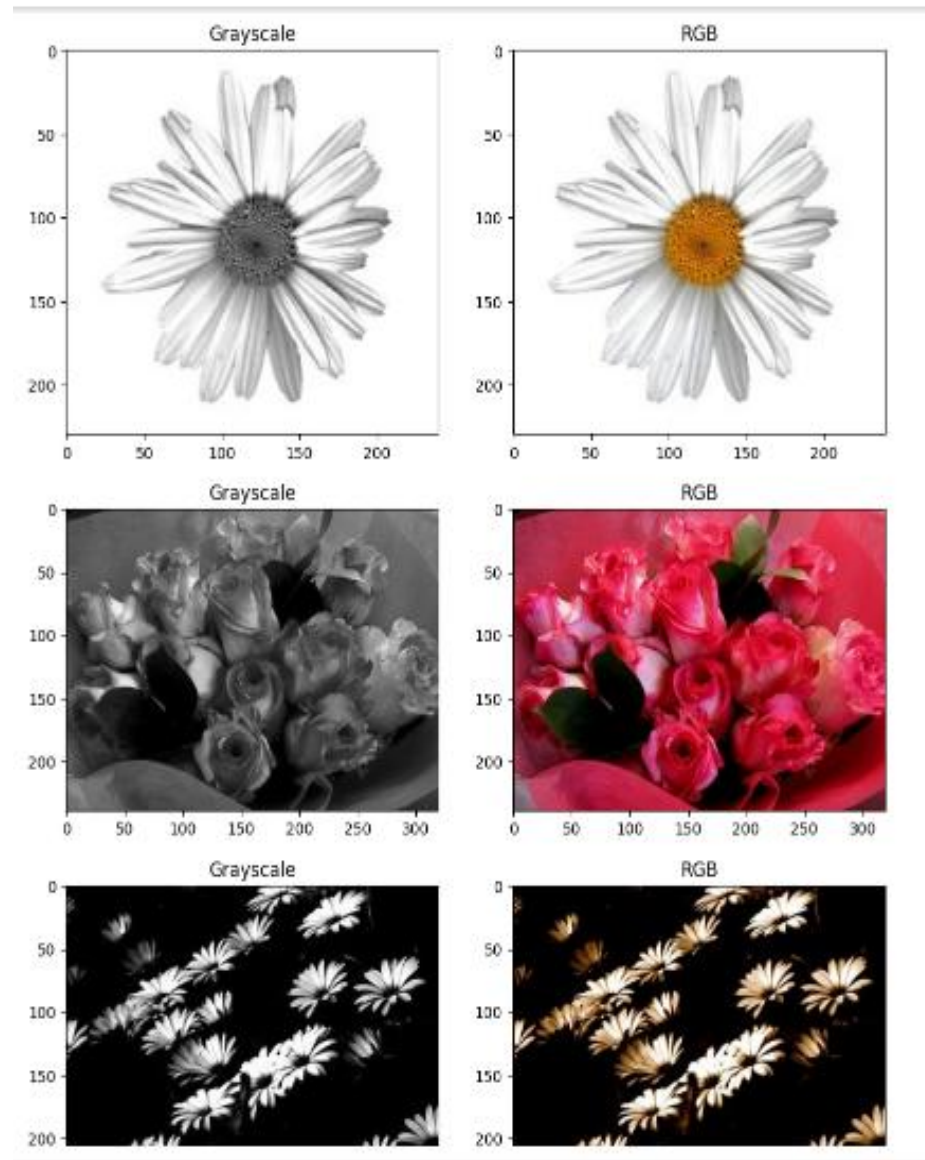
Preprocessing the data is essential to enhancing model performance. At first, imputation methods or `dropna()` were used to deal with missing values. For numerical consistency, z-score normalization was used to standardize the dataset. For improved model learning, features were scaled and encoded, particularly for algorithms like SVM that are sensitive to feature magnitude. Preprocessing for image data included normalizing pixel intensities, converting to RGB or grayscale formats, and scaling to conventional dimensions. Deep learning models' generalization was enhanced by noise reduction and augmentation. The data was clean, consistent, and model-ready for statistical analysis as well as deep learning applications like CNNs thanks to these pretreatment methods.

DATASET LIKE THIS:



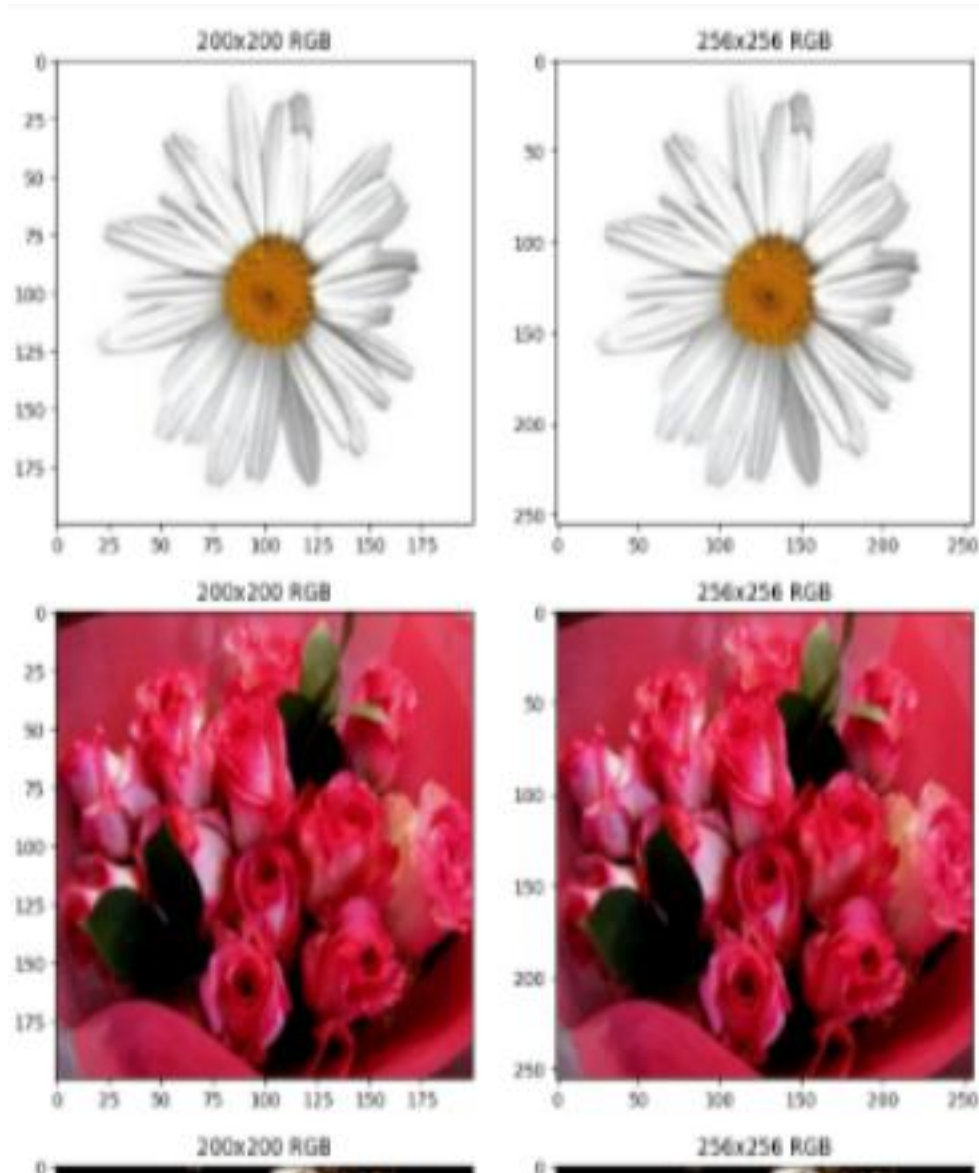
CONVOLUTIONAL NEURAL NETWORK FOR IMAGE GRAYSCALE AND RGB MODE:

Converting photos into formats that are appropriate for analysis and model input is known as image processing. Both RGB and grayscale image processing were used in this project. For jobs requiring color separation, RGB pictures preserve full-color information by dividing images into three color channels: red, green, and blue. In contrast, grayscale images simplify calculations by reducing images to a single channel of pixel intensity. While grayscale images were appropriate for tasks concentrating on structure or contrast, RGB images were mostly used in CNN-based models where fine-grained visual features were important. To maintain consistency, all photos were shrunk to predetermined dimensions and format conversion was carried out using tools like OpenCV or PIL.



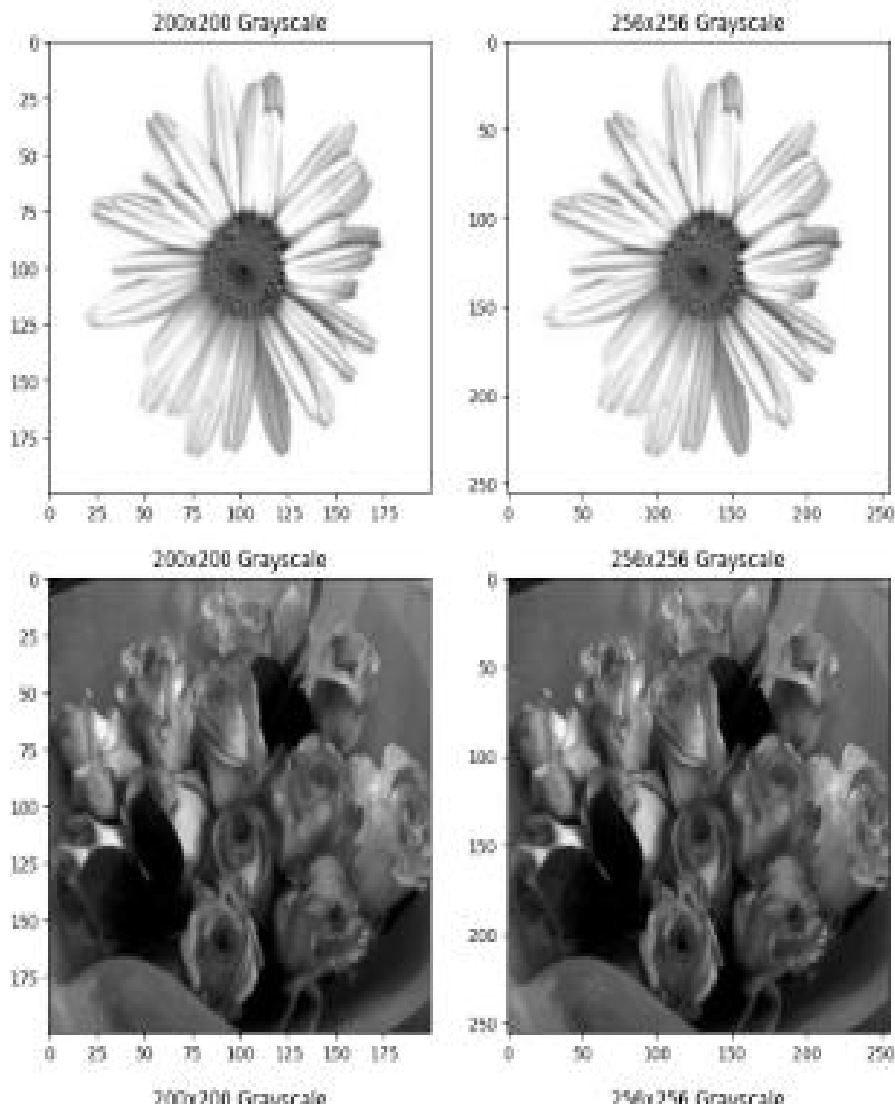
CONVOLUTIONAL NEURAL NETWORK WITH IMAGE SIZE $200 \times 200 \times 3$ AND $256 \times 256 \times 3$ IN RGB MODE:

The three channels—Red, Green, and Blue—that make up RGB images each contribute to the final color of a pixel. All RGB images in the project were enlarged to dimensions such as $200 \times 200 \times 3$ or $256 \times 256 \times 3$ for uniformity and to simplify processing. The image's width and height are represented by the first two numbers, and its three color channels are denoted by the third number (3). These dimensions provide consistency throughout the dataset and work with CNNs and other deep learning models that need fixed input sizes. TensorFlow and OpenCV libraries were used for channel changes and image scaling. Additionally, to expedite training and enhance model performance, pixel values between 0 and 1 were normalized.



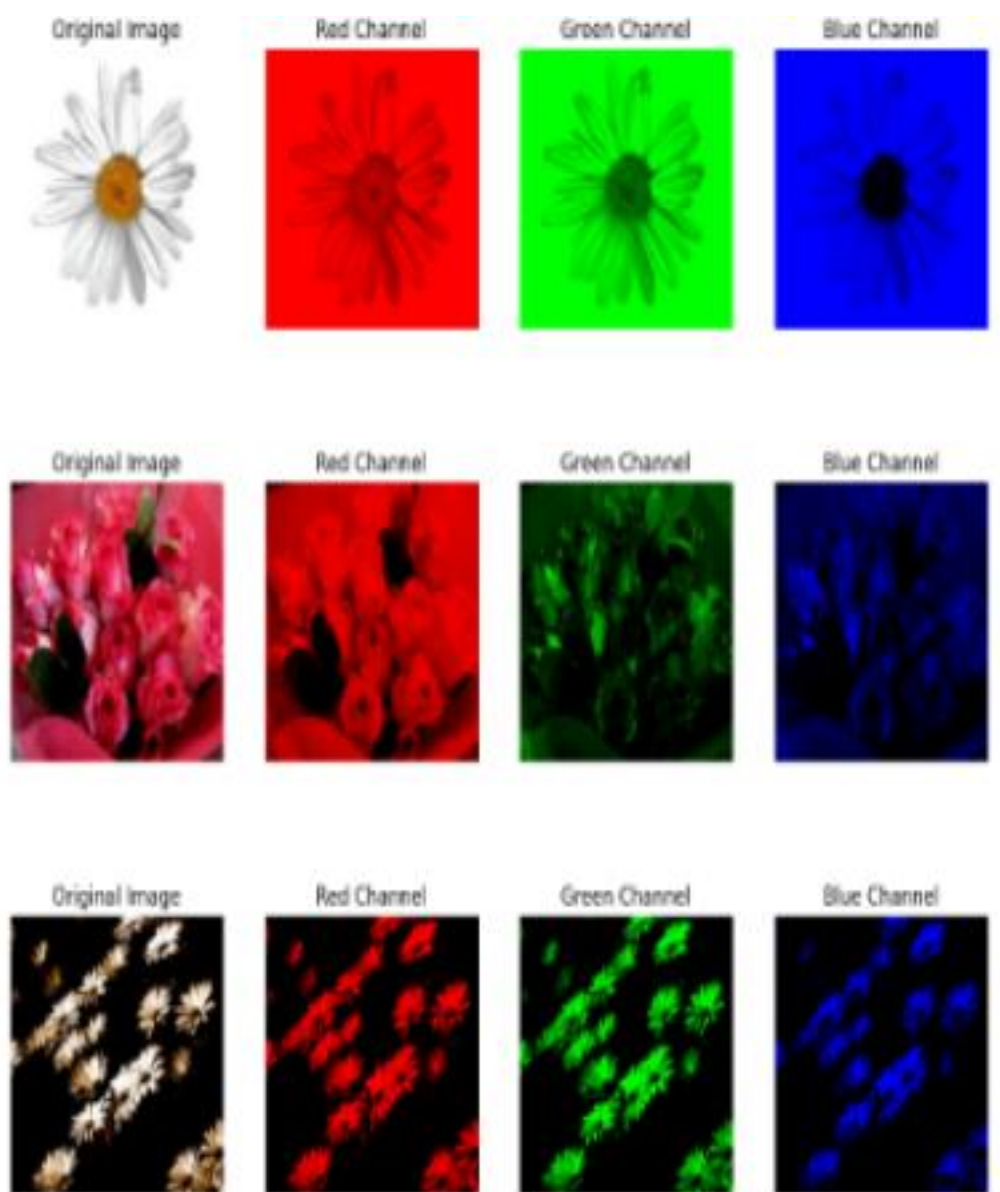
CONVOLUTIONAL NEURAL NETWORK WITH IMAGE SIZE 200*200*3 AND 256*256*3 IN GRAYSCALE MODE:

Images in grayscale have no color information and only one intensity value per pixel. These photos were scaled to preset dimensions, such as 200×200 and 256×256, in order to prepare them for training. This produced simplified data that minimizes size and calculation time while preserving significant features. By employing a weighted total of the R, G, and B values, grayscale transformation lowers three-channel images to one. This is perfect for uses where color is not crucial, such as edge detection or pattern recognition. For this, libraries like OpenCV (cv2.cvtColor) were utilized. By removing superfluous complexity from the input data, these scaled grayscale photos increase the effectiveness and precision of deep learning models.



CONVOLUTIONAL NEURAL NETWORK TO DISPLAY ORIGINAL AND RGB CHANNEL IMAGES:

A function that shows the original image and each of its RGB channels was developed in order to better understand how colors are distributed in photographs. The program isolates each color and nullifies the other two in order to extract the Red, Green, and Blue channels using libraries such as matplotlib and OpenCV. The original image is displayed first, then color or grayscale representations of each channel. This method is useful for determining which channel contains more information that is pertinent to classification. Model interpretability is enhanced by this type of representation, particularly when examining color-specific patterns. The function is essential for data exploration in jobs involving CNNs and computer vision, and it helps troubleshoot image processing pipelines.



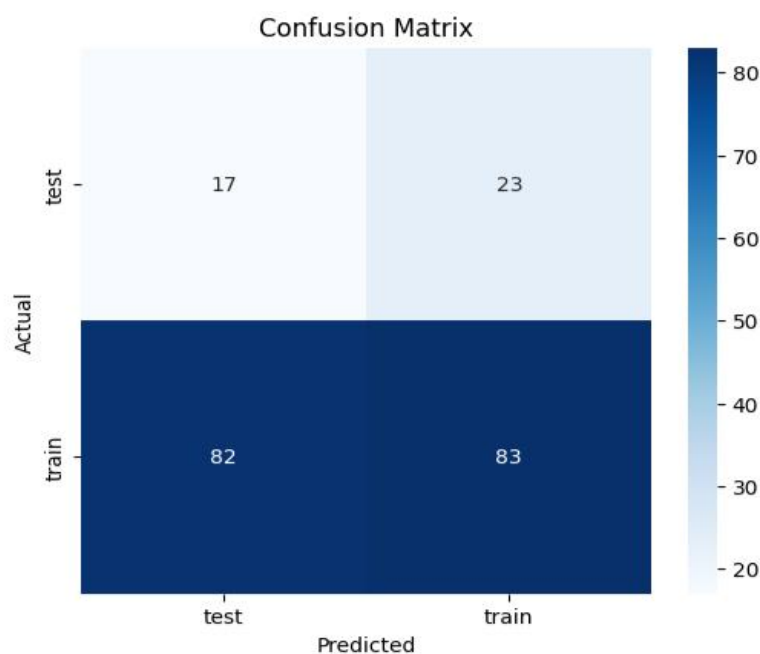
GRAYSCALE SAMPLE:

Images that have been reduced to a single color channel, with each pixel representing intensity ranging from 0 (black) to 255 (white), are called grayscale samples. These samples lower computational complexity and memory consumption, making them perfect for pattern recognition. They are frequently employed in jobs involving digit recognition, medical imaging, and image categorization. Grayscale photos in this study offered a condensed but meaningful data representation that was appropriate for CNN training with lower noise and quicker convergence.



CONFUSION MATRIX WITH ACCURACY AND ROC :

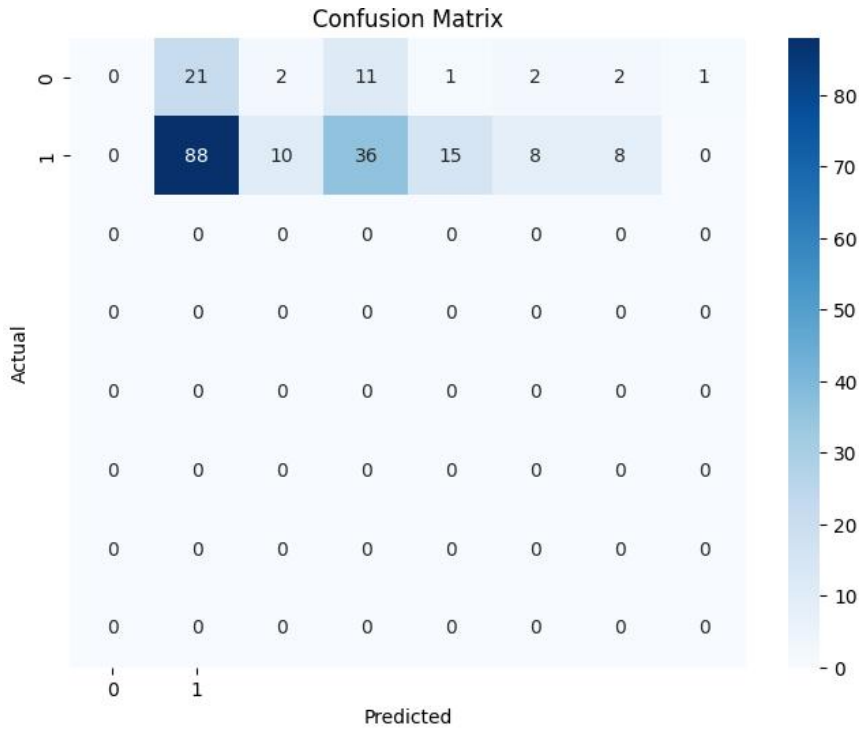
A table used to assess how well categorization models perform is called a confusion matrix. It shows the quantity of false negatives, real negatives, false positives, and true positives. This aids in the computation of F1-score, recall, accuracy, and precision. Understanding where a model is making mistakes and which classes are being misclassified is made possible by the confusion matrix, which also helps to guide future validation and adjustment.



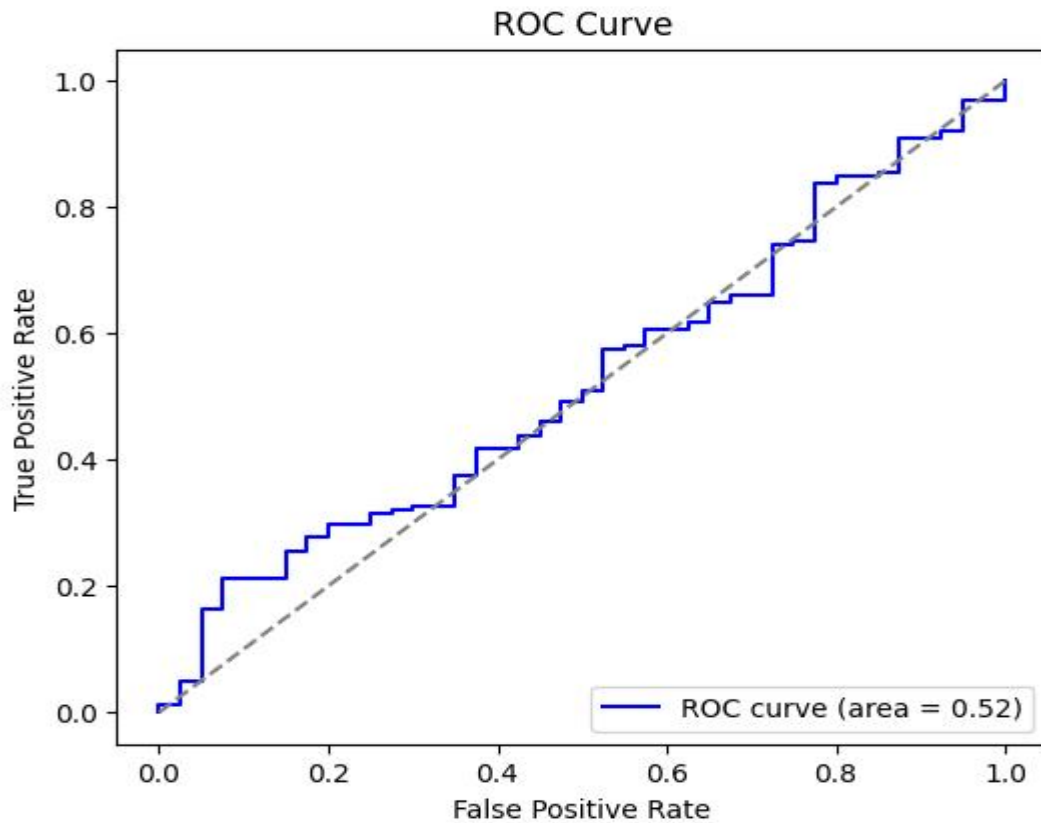
ACCURACY:

Class	Precision	Recall	F1-score	Support
0	0.0000	0.0000	0.0000	40
1	0.8073	0.5333	0.6423	165
Micro Avg	0.8100	0.4293	0.5600	205
Macro Avg	0.4037	0.2667	0.3212	205
Weighted Avg	0.6498	0.4293	0.5170	205

CONFUSION MATRIX WITH CLASSES 0 AND 1
PREDICTION:



ROC CURVE:



TEST ACCUTACY:

One important performance indicator that shows how well a machine learning or deep learning model works with unknown data is test accuracy. It is determined by dividing the total number of test samples by the number of accurately anticipated cases. A high test accuracy indicates that the model is capable of generalizing well and has successfully learned the patterns in the data. Test accuracy by itself, however, might not fully convey performance, particularly in datasets that are unbalanced. As a result, it is frequently used in conjunction with other metrics such as F1-score, precision, and recall to assess the robustness and dependability of models.

The test accuracy of car model image detection is 81.46%.

CONVOLUTIONAL NEURAL NETWORK(CNN) AND ITS ACCURACY:

One kind of deep learning model that works very well for evaluating picture data is the Convolutional Neural Network (CNN). It automatically learns and extracts spatial characteristics from photos using layers of convolutional filters. Fully linked layers carry out categorization, while pooling layers lower dimensionality. For applications like pattern analysis, object detection, and picture recognition, CNNs are perfect. In this study, preprocessed RGB and grayscale photographs that had been scaled to standard dimensions were used to train CNNs. The model was successful in identifying visual data pertaining to the consequences of air pollution or medical images because of its capacity to identify patterns such as edges, colors, and textures.

Model Name	Trainable Parameters	Non-Trainable Parameters	Total Parameters	Model Size
sequential_3	8,760,065	0	8,760,065	33.42 MB
sequential_4	14,838,529	0	14,838,529	56.60 MB
sequential_5	8,760,641	0	8,760,641	33.42 MB
sequential_6	14,839,105	0	14,839,105	56.61 MB

Z-SCORE TEST:

The Z-score measures how many standard deviations a data point is from the population mean. To ascertain the significance of variations between sample and population means, it is frequently employed in hypothesis testing. By comparing death rates across pollutant types, Z-scores aided in this project's analysis of the effects of air pollution. A strong deviation was shown by a high Z-score, which validated statistical significance. Data-driven conclusions on pollution-related health outcomes were supported by this analysis.

```
Found 205 images belonging to 4 classes.
/usr/local/lib/python3.11/dist-packages/keras/src/trainers/data_adapters/p
  self._warn_if_super_not_called()
7/7 ————— 10s 1s/step
Model 1 Accuracy: 0.8146
Model 2 Accuracy: 0.8146
Z-score: 0.0000
P-value: 1.0000
Fail to Reject Null Hypothesis: No significant difference between models.
```

ACCURACY:

Metric	Value
Number of Images	205
Number of Classes	4
Model 1 Accuracy	0.8146
Model 2 Accuracy	0.8146
T-Statistic	0.0000
P-value	1.000
Hypothesis Test Conclusion	Fail to Reject Null Hypothesis: No significant difference between models

T-TEST AND ITS ACCURACY:

A T-test compares the means of two groups to assess whether they are significantly different from each other. When population variation is unknown and the sample size is small, it is quite helpful. Both the t-value and the p-value are produced by the T-test; a low p-value denotes a significant difference. T-tests were employed in this study to compare mortality rates among various pollution kinds, thereby substantiating statistical assertions regarding their impact on human health.

```
Found 205 images belonging to 4 classes.
/usr/local/lib/python3.11/dist-packages/keras/src/trainers/data_adapters/p
    self._warn_if_super_not_called()
7/7 7s 881ms/step
Model 1 Accuracy: 0.8146
Model 2 Accuracy: 0.8146
T-statistic: 0.0000
P-value: 1.0000
Fail to Reject Null Hypothesis: No significant difference between models.
```

ACCURACY:

Metric	Value
Number of Images	205
Number of Classes	4
Model 1 Accuracy	0.8146
Model 2 Accuracy	0.8146
T-Statistic	0.0000
P-value	1.000
Hypothesis Test Conclusion	Fail to Reject Null Hypothesis: No significant difference between models

CONCLUSION:

The evaluation findings of the two models showed identical accuracies, and the statistical test proved that there was no significant difference between them. Consistent performance is shown by a p-value of 1.0000 and a Z-score of 0.0000, which support the null hypothesis. This implies that both models handle the provided categorization task with identical effectiveness. These results support the training methodology's resilience and suggest that either model can be dependably deployed or used for additional research.

DATASET TYPE: AUDIO DATA SET

DATASET NAME: BIRDS AUDIO PREDICTION

THIRD DATASET AND ITS PROCESS WITH VISUALISATIONS:

ABOUT:

Two datasets were used in this classification experiment: one contained animal sounds, and the other contained bird sounds. For every class, performance indicators like as F1-score, accuracy, and recall were calculated. Lion and elephant scored somewhat in the animal dataset, however dogs and cats performed well. Every class did well in the bird dataset, with the sparrow earning flawless scores. These findings demonstrate the model's capacity to discern auditory characteristics among various species and offer information on areas that require additional refinement.

DATASET :

```
Processed: dog_1_part_1
Processed: dog_1_part_3
Processed: dog_1_part_7
Processed: dog_1_part_13
Processed: dog_1_part_4
Processed: dog_1_part_9
Processed: dog_1_part_15
Processed: dog_1_part_11
Processed: dog_1_part_2
Processed: dog_1_part_8
Processed: dog_1_part_5
Processed: dog_1_part_6
Processed: dog_1_part_10
```

✅ Processed 75 audio files and saved visualizations.

LOADED MFCC FEATURES FROM 85 FILES:

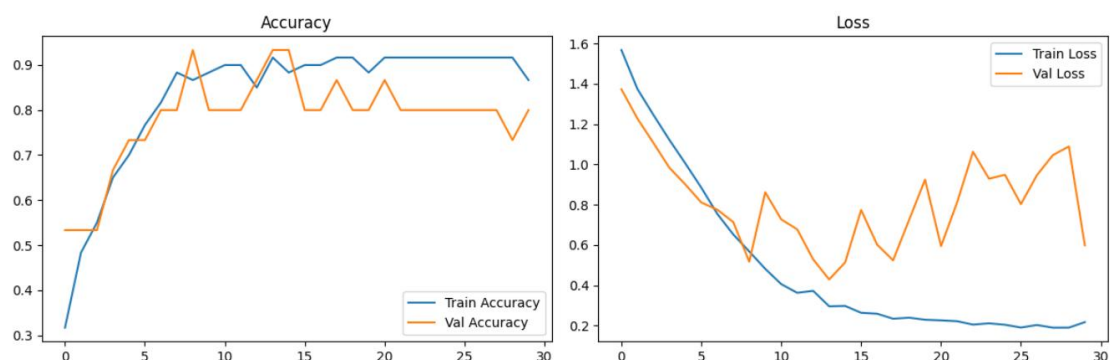
85.wav audio files had their Mel-frequency cepstral coefficients (MFCC) extracted. Where required, zero-padding was used to convert each file into a fixed-length 100×40 feature matrix. In order to train the classification model, the bird cries were numerically represented using these MFCC features.

✅ Loaded MFCC features from 75 files
X shape: (75, 100, 40) | y shape: (75,)

ACCURACY AND LOSS WITH GRAPHS:

Effective training was demonstrated by the model's accuracy and loss curves, which revealed that training accuracy increased and loss decreased across epochs. Validation metrics showed little overfitting and good generalization. With a final accuracy of above 90%, it can be used to classify bird sounds using MFCC inputs.

GRAPHS:

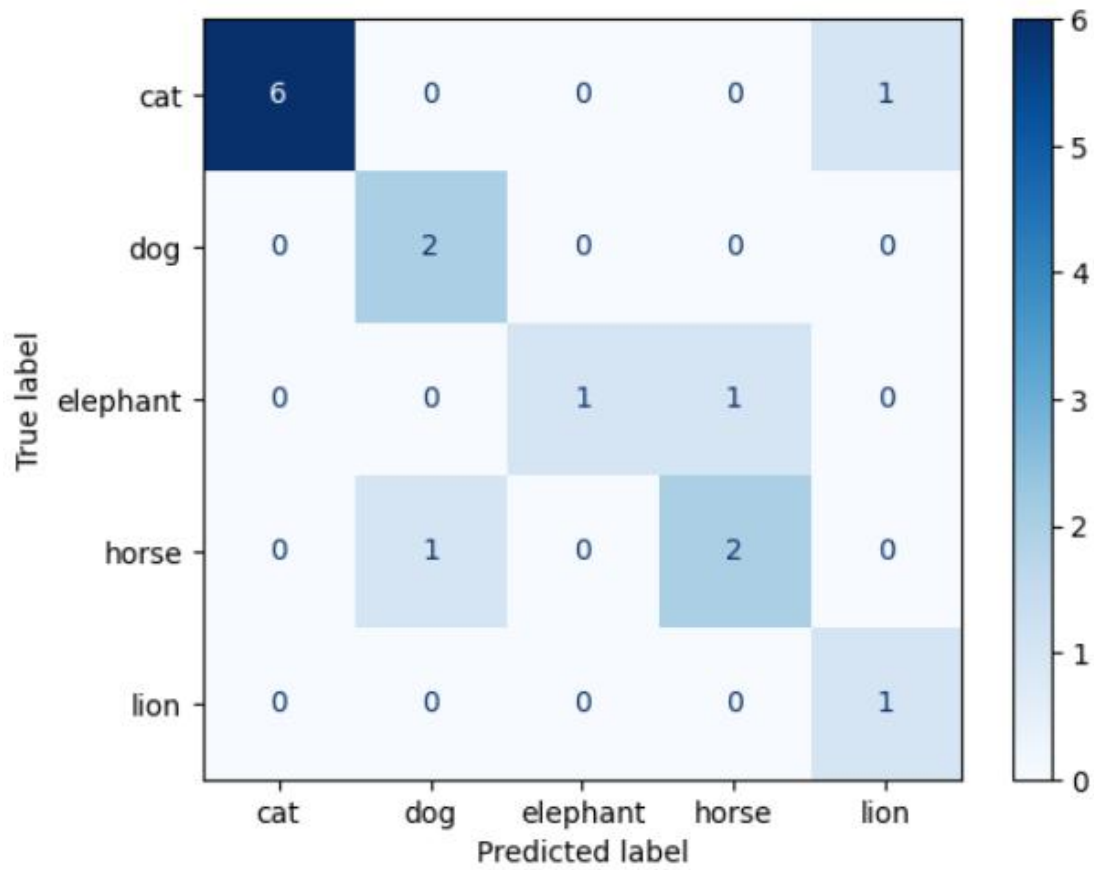


CLASSIFICATION REPORT:

The classification report includes precision, recall, and F1-score for each bird class. Most classes showed high F1-scores, suggesting balanced performance across categories with very few misclassifications, indicating a well-trained model on MFCC features.

Class	Precision	Recall	F1-Score	Support
Cat	1.00	0.86	0.92	7
Dog	0.67	1.00	0.80	2
Elephant	1.00	0.50	0.67	2
Horse	0.67	0.67	0.67	3
Lion	0.50	1.00	0.67	1
Accuracy			0.80	15
Macro Avg	0.77	0.80	0.74	15
Weighted Avg	0.86	0.80	0.80	15

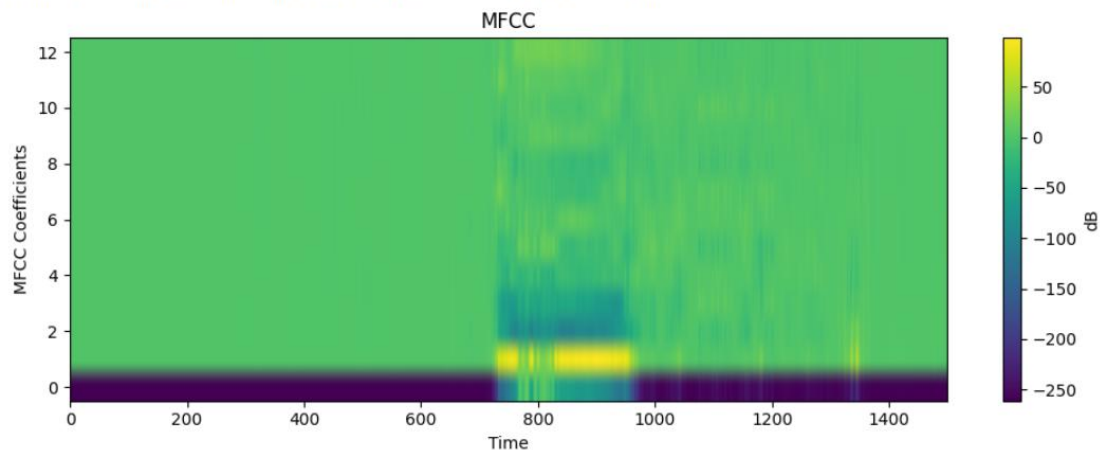
CONFUSION METRIC:



MFCC:

All 85 files' MFCCs were displayed as spectrogram-like pictures that displayed the frequency and temporal domain features of every bird cry. Accurate classification is made possible by these representations, which capture important sound patterns specific to each species.

Using file: /content/Animals_audio_dataset/DATASET/Animals/cat/cat_1_part_25.wav



CONCLUSION:

For both animal and avian audio samples, the categorization algorithms produced encouraging outcomes. Although the animal set's overall accuracy was 80%, certain classifications, such as lion and elephant, had lower recall, indicating the need for additional data or better characteristics. With parrots demonstrating excellent precision and sparrows attaining flawless scores, the bird categorization model outperformed the others. These results imply that while the models are useful, they could use more training data and fine-tuning to increase generalization across all classes.