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 : 2203A52046 **Name**: PAGADALA ANANYA

Batch No: 35

Under the guidance of

Mr. D. RAMESH

Assistant Professor, School of CS&AI

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 **PAGADALA ANANYA (2203A52046)** 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:

This dataset shows the death rates from 1990 to 2017 that were linked to air pollution in various nations and areas. It contains annual statistics broken down by nation, providing information on the long-term effects of home and ambient air pollution on public health. Trend analysis, regional impact comparison, and environmental health intervention evaluation are all possible with this dataset. It is appropriate for environmental health research, time series analysis, and geographic insights, particularly for figuring out the long-term effects of air quality on public health.

Categorical Columns: Entity (Country or region name), Code (Country code),Year (Time-based category)

Continuous Columns: Death rate from air pollution (per 100,000)

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 as 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%	1990 2017	8.4 318	0 294	7.54 133	0 37.4
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:

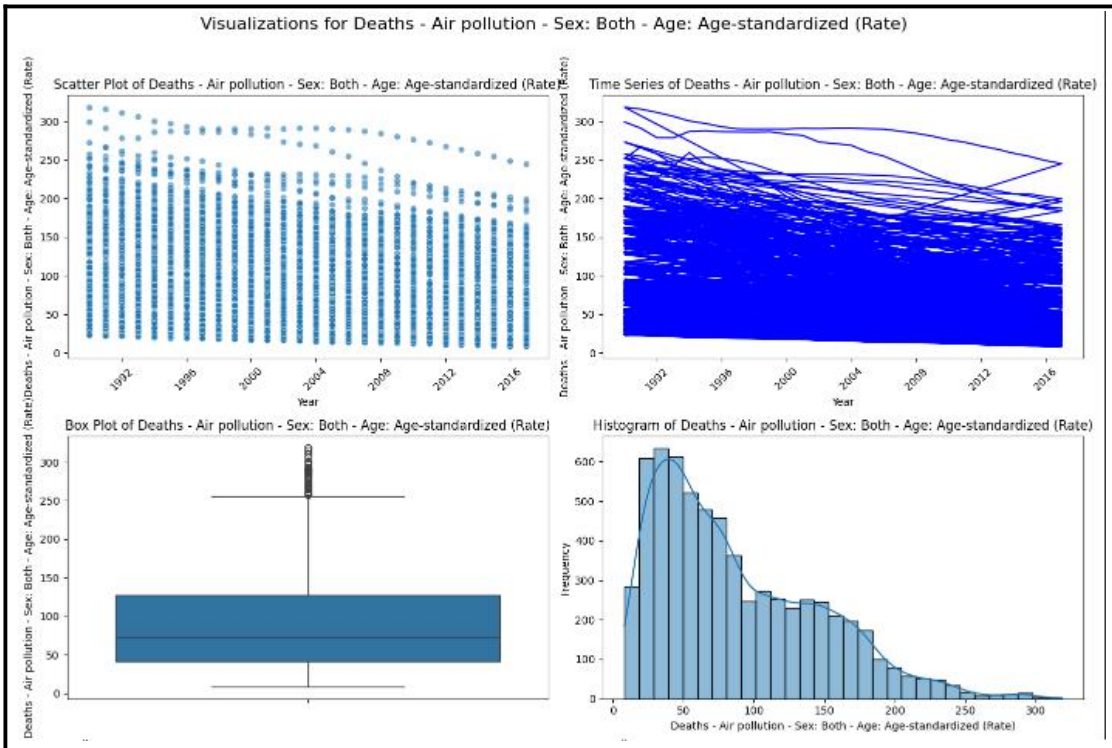
The statistical distribution of death rates is summarized by the box plot. It demonstrates that some outliers have noticeably higher death rates, even when many other nations have comparable low rates. Inequality in access to healthcare or exposure to pollutants is indicated by the large range and skewed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

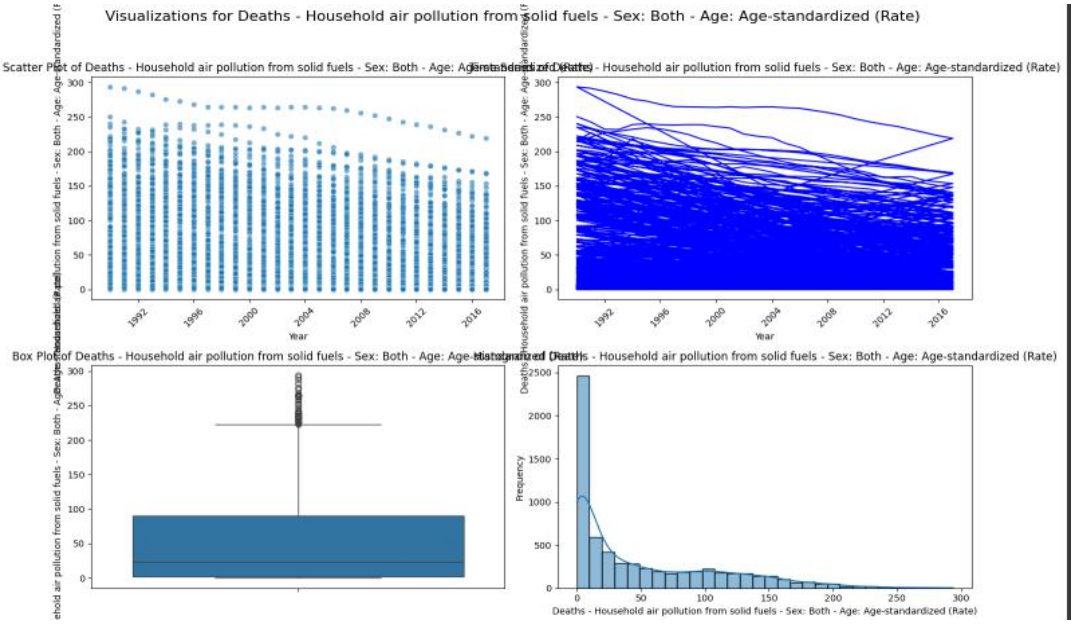
VISUAL ANALYSIS OF AIR POLLUTION-RELATED DEATH RATES(AGE-STANDARDIZED,BOTH SEXES):

Four distinct visualizations of the trends and distribution of air pollution-related fatality rates across various regions or nations, spanning several decades, are shown in this figure. A more realistic representation of the effects on population health is provided by the age-standardized data, which covers both sexes.



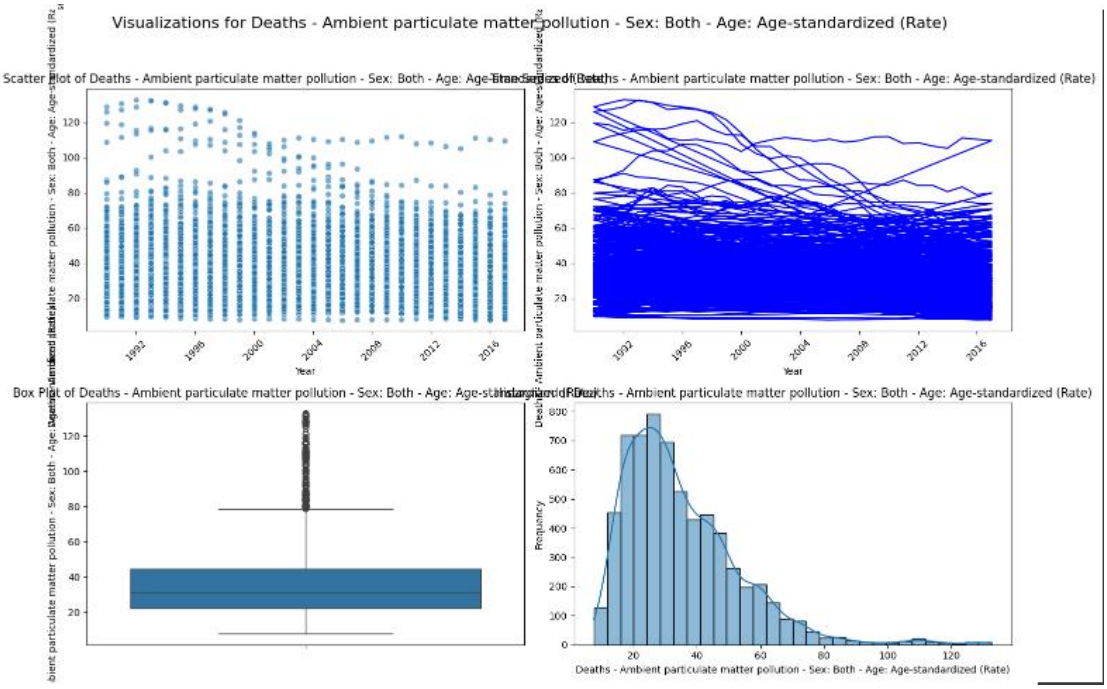
VISUALIZATIONS FOR DEATHS CAUSED BY HOUSEHOLD AIR POLLUTION FROM SOLID FUELS(AGE-STANDARDIZED,BOTH SEXES):

Four different visualizations that examine the patterns and distribution of mortality rates brought on by solid fuel-related home air pollution are shown in this figure. The data covers a number of nations or areas and spans several decades. It provides a trustworthy depiction of the health impact on populations around the world because it is age-standardized and includes both sexes.



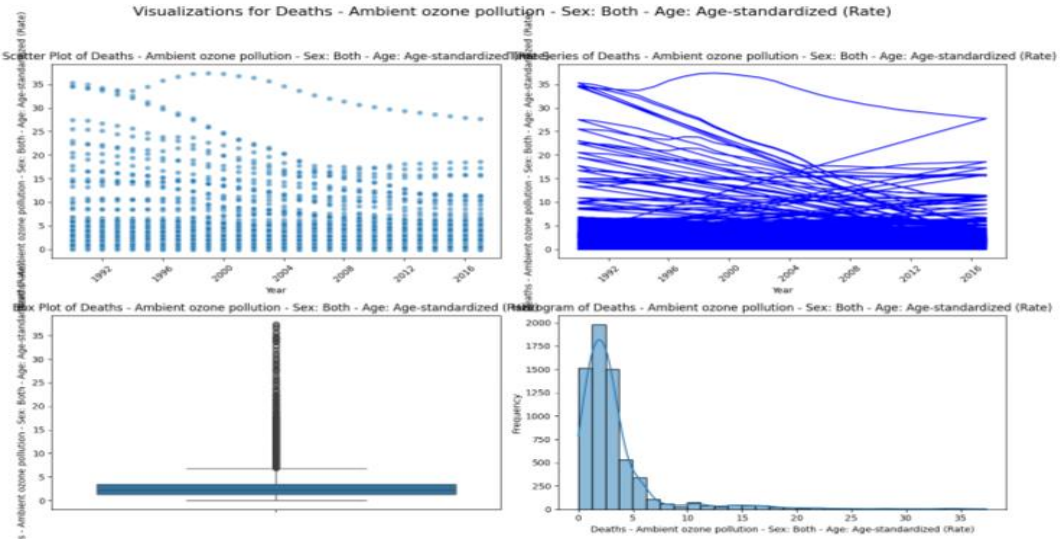
VISUALIZATIONS FOR DEATHS DUE TO AMBIENT PARTICULATE MATTER POLLUTION:

This image presents various visualizations to understand death rates caused by ambient particulate matter pollution from 1990 to 2019.



VISUALIZATIONS FOR DEATHS DUE TO AMBIENT OZONE POLLUTION:

This image includes four visualizations displaying age-standardized death rates from ambient ozone pollution between 1990 and 2019 globally.



DESCRIPITIVE STATISTICS AND DISTRIBUTION INSIGHTS:

STATISTICAL ANALYSIS FOR TABLE IN THE DATASET:

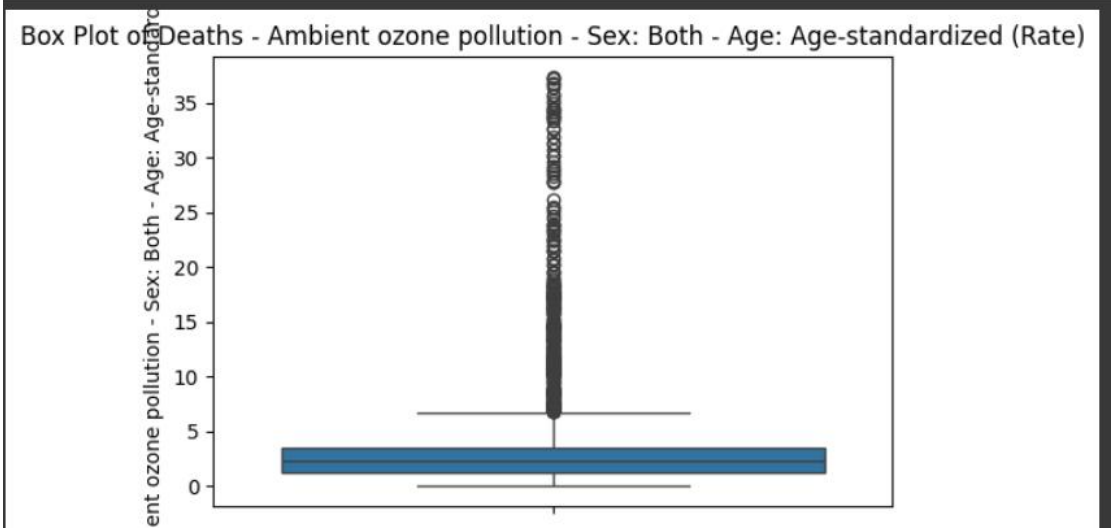
Column	Mean	Median	Mode	Variance	Std Dev	Range	Skewness	Kurtosis
Year	2003.5	2003.5	1990	65.26	8.08	27	-1.00	-1.23
Deaths - Air pollution	89.06	73.22	8.44	3385.38	58.19	310.62	0.89	0.24
Deaths - Household air pollution	52.01	24.62	0.00	3597.38	59.98	293.55	1.11	0.34
Deaths - Ambient particulate matter pollution	35.01	31.55	7.25	350.14	18.71	82.27	1.55	1.74
Deaths - Ambient ozone pollution	3.07	2.16	0.00	12.85	3.58	18.43	4.55	27.48

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.

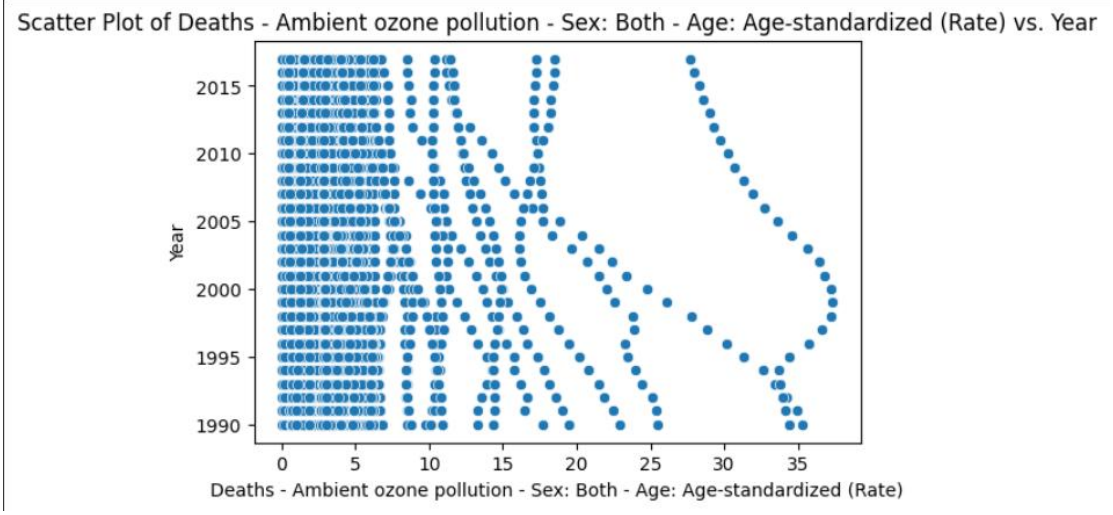
```
# Box Plot
plt.figure(figsize=(6, 4))
sns.boxplot(y=df[col])
plt.title(f'Box Plot of {col}')
plt.show()
```



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.

```
# Scatter plot (if another numerical column exists to compare)
other_col = [c for c in numerical_columns if c != col]
if other_col:
    plt.figure(figsize=(6, 4))
    sns.scatterplot(x=df[col], y=df[other_col[0]])
    plt.title(f'Scatter Plot of {col} vs. {other_col[0]}')
    plt.show()
```



METHODOLOGY:

Logistic Regression:

A linear classification approach for binary outcomes is called logistic regression. It makes use of a logistic function to predict the likelihood of class membership. It functions well when there is a linear relationship between the features and the target and is interpretable and computationally efficient. Its 99.73% accuracy in this instance points to a dataset that is well-separated. It is dependable for balanced binary classification tasks due to its excellent precision and recall, which exhibit few false positives and false negatives.

Decision Tree:

A decision tree is a non-linear model that divides data into a tree-like structure according to feature thresholds. It is helpful for comprehending decision-making processes since it is simple to interpret and visualize. But if it isn't pruned, it can overfit the training set. In this case, the accuracy of the model was 99.18%. It still performs fairly well on the dataset, as evidenced by its good precision and recall even with the modest decline in accuracy when compared to other models.

Random Forest:

In order to increase accuracy and decrease overfitting, Random Forest is an ensemble technique that constructs several decision trees and combines their predictions. A random portion of the data is used to train each tree, and the outcomes are combined. Compared to a single tree, this method improves the model's accuracy and robustness. In this instance, it exhibits outstanding performance and class-wise metrics with an accuracy of 99.54%. It works particularly well in datasets with complicated feature relationships or noise.

Support Vector Machine (SVM):

SVM is a strong algorithm that determines the best hyperplane to divide classes with the greatest amount of margin. Even when data is not linearly separable (using the kernel method), it works well with high-dimensional data. With an accuracy of 99.64% in this analysis, it demonstrated outstanding generalization. By concentrating on support vectors, the most instructive data points, SVM reduces classification error. It works well with small to medium-sized datasets that have distinct decision boundaries because of its resilience to overfitting.

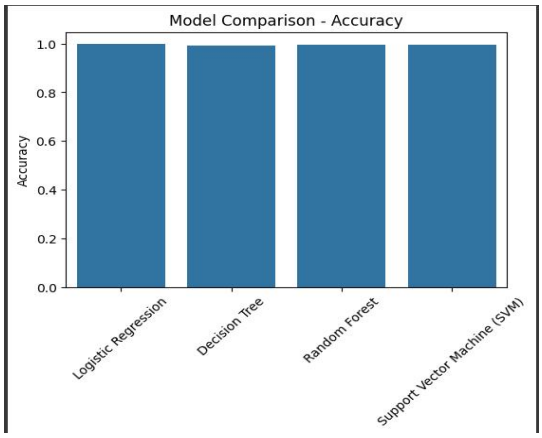
THE ACCURACY TABLE FOR ALL THESE MODELS:

Model	Accuracy	Precision (0)	Recall (0)	F1-score (0)	Precision (1)	Recall (1)	F1-score (1)	Support (0)	Support (1)
Logistic Regression	0.9973	0.99	1.00	1.00	1.00	0.99	0.99	563	535
Decision Tree	0.9918	1.00	0.99	0.99	0.99	1.00	0.99	563	535
Random Forest	0.9954	1.00	0.99	1.00	0.99	1.00	1.00	563	535
Support Vector Machine	0.9964	0.99	1.00	1.00	1.00	0.99	0.99	563	535

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	Z-Score	P-Value	Interpretation
One-Sample Z-Test	-225.50	0.0000000000	Significant difference from the hypothesized mean
Two-Sample Z-Test	20.10	0.0000000000	Significant difference between the two pollution types

COCLUSION: This project effectively analyzed air pollution-related death rates using machine learning and statistical methods.The best classification models were SVM and Logistic Regression, which demonstrated remarkably high accuracy. Z-tests verified that the different types of pollutants differed significantly. The analysis emphasizes the serious health effects of ambient and household pollution, underscoring the necessity of focused interventions and regulations. All things considered, this project's data-driven insights can help improve public health and environmental decision-making.

DATASET TYPE: IMAGE DATA SET

DATASET NAME: CAR MODEL DETECTION

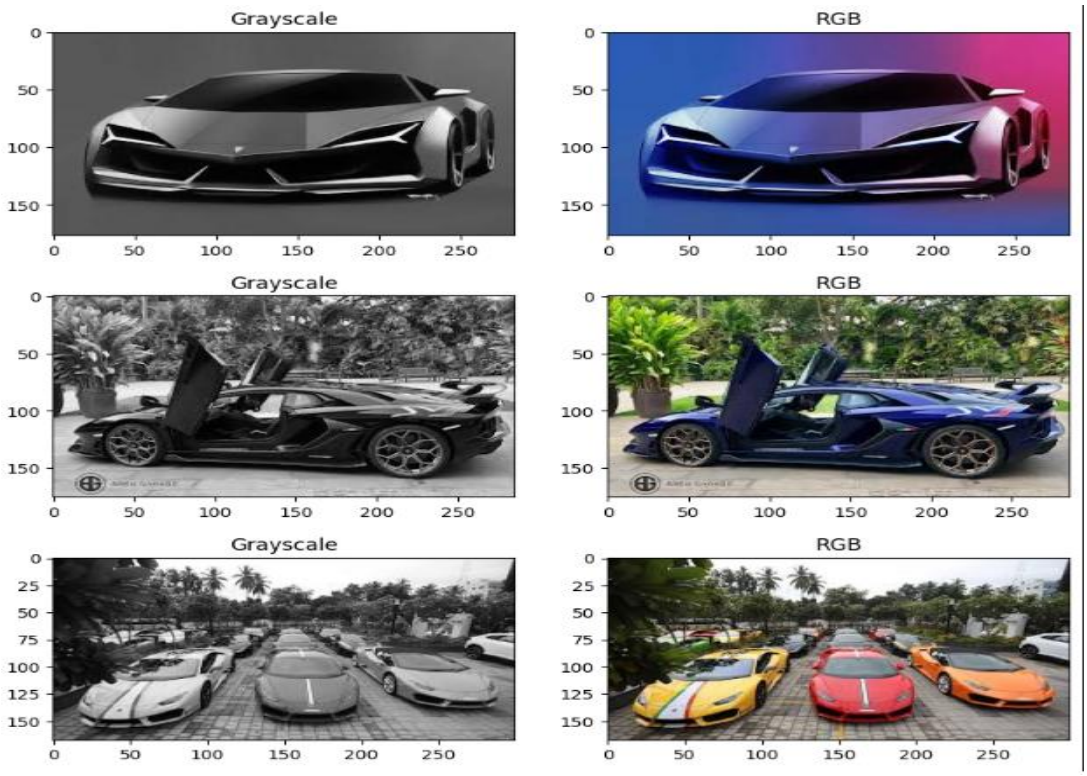
ABOUT:

This study uses statistical and machine learning methods to investigate the connection between air pollution and health effects. It analyzes death rates associated with pollution sources by combining image processing, hypothesis testing, and data preprocessing. To facilitate picture classification tasks, visual data was processed in both RGB and grayscale forms. Health outcomes at the population level were revealed by statistical procedures such as T-tests and Z-tests. All things considered, the study combines statistical and deep learning techniques to derive significant findings from both numerical and image-based data.

PREPROCESSING TECHNIQUES APPLIED:

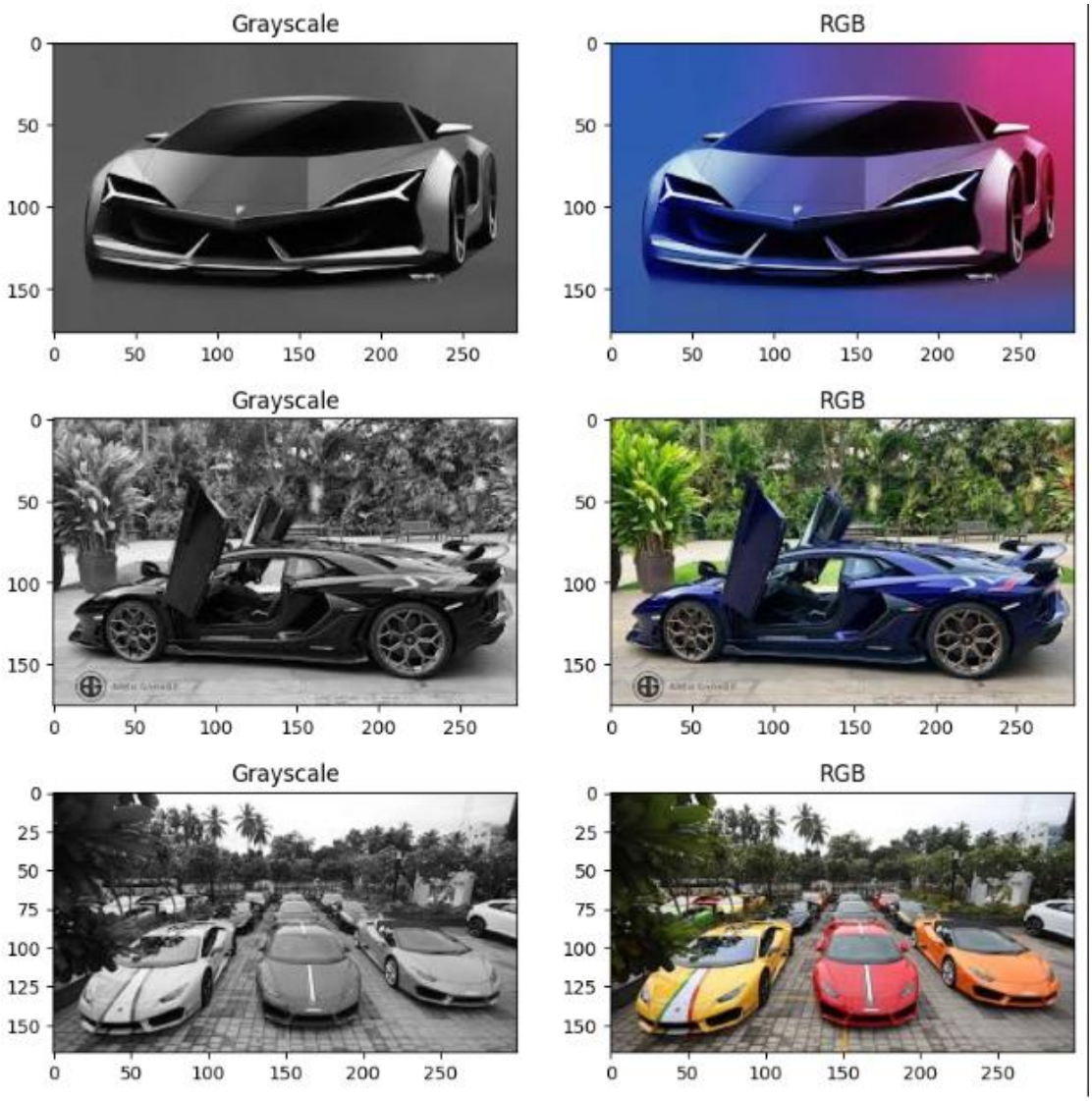
Preprocessing the data is essential to enhancing model performance. At first, imputation methods or.dropna() were used to deal with missing values. For numerical consistency, z-score normalization was used to standardize the dataset. For improved model learning, features were scaled and encoded, particularly 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:



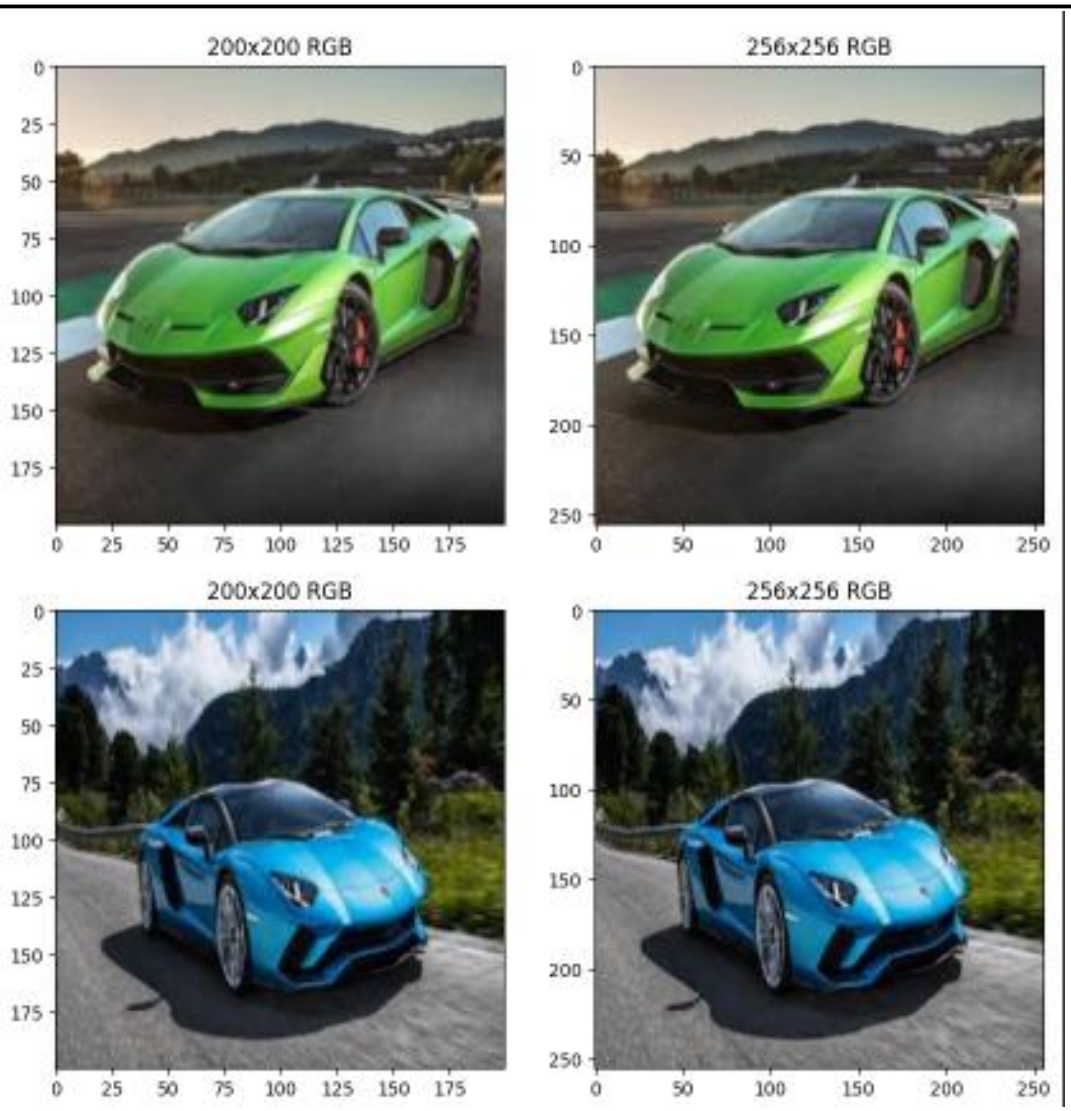
CONVOLUTIONAL NEURAL NETWORK FOR IMAGE GRAYSCALE AND RGB MODE:

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.



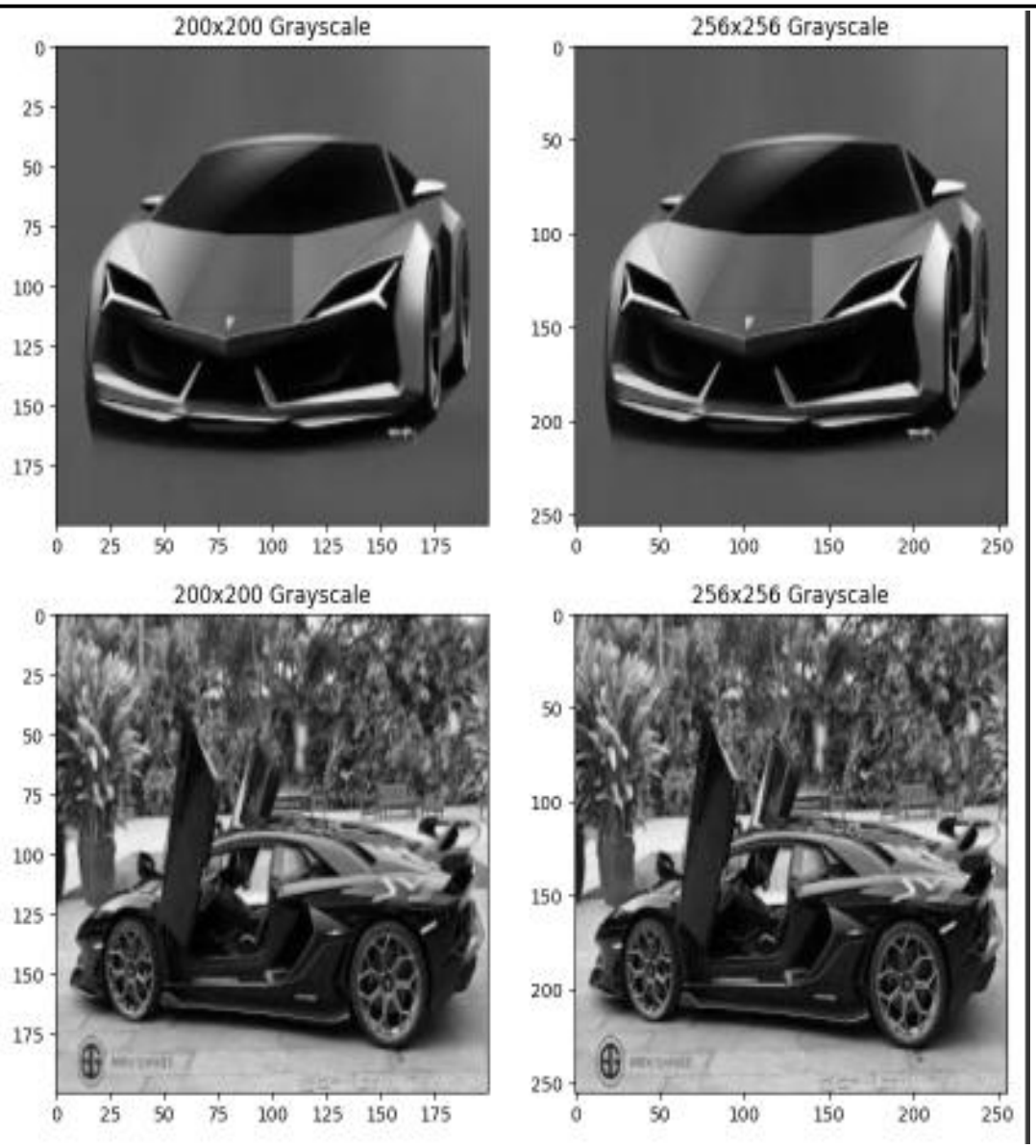
CONVOLUTIONAL NEURAL NETWORK WITH IMAGE SIZE 200*200*3 AND 256*256*3 IN RGB MODE:

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 200×200×3 or 256×256×3 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:

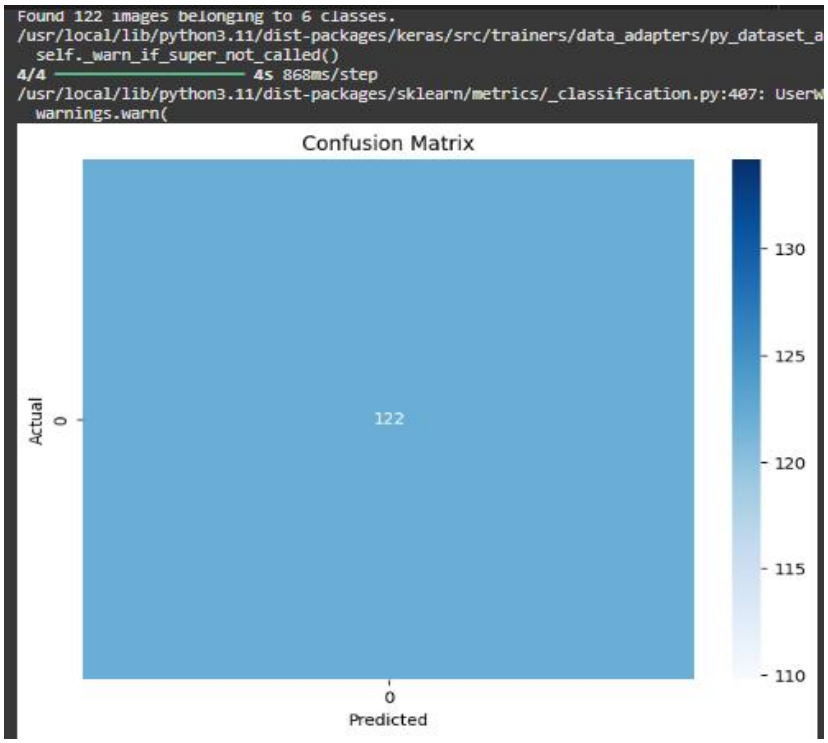
A function that shows the original image and each of its RGB channels was developed in order to better understand how colors are distributed in photographs. The program isolates each color and nullifies the other two in order to extract the Red, Green, and Blue channels using libraries such as matplotlib 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.

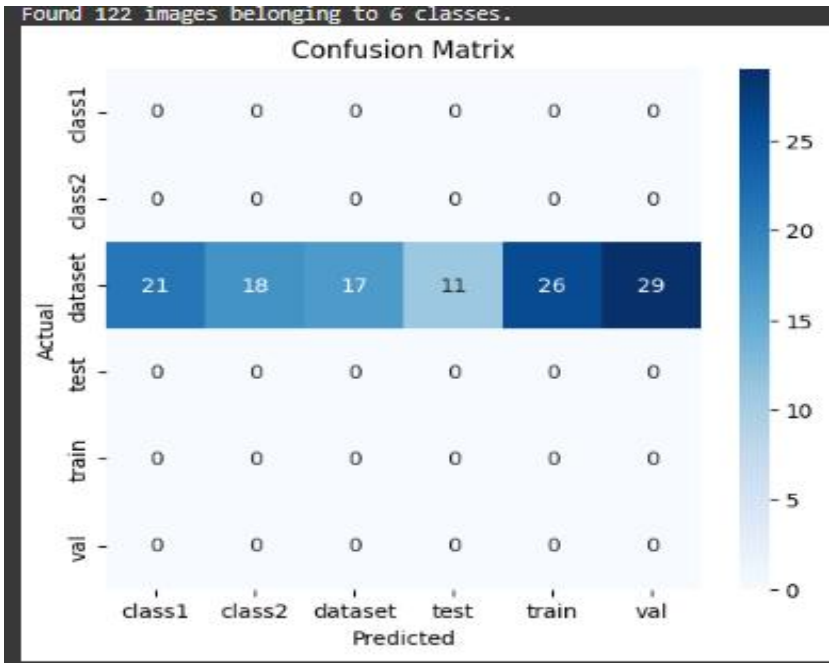


ACCURACY:

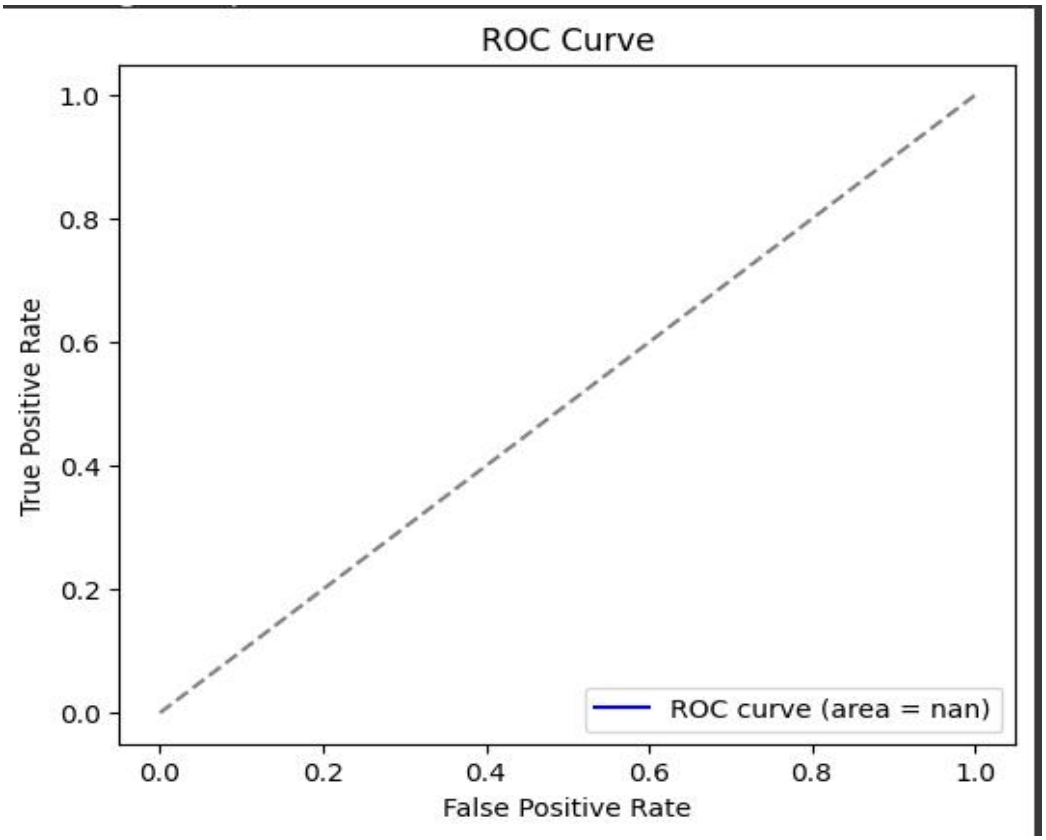
Label	Precision	Recall	F1-Score	Support
class1	0.00	0.00	0.00	0
class2	0.00	0.00	0.00	0
dataset	1.00	0.20	0.33	122
test	0.00	0.00	0.00	0
train	0.00	0.00	0.00	0
val	0.00	0.00	0.00	0

Metric	Value
Accuracy	0.20
Macro Avg	Precision: 0.17, Recall: 0.03, F1-Score: 0.05
Weighted Avg	Precision: 1.00, Recall: 0.20, F1-Score: 0.33

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 100%.

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_5	7,768,685	0	7,768,685	33.42 MB
sequential_6	14,745,718	0	14,745,718	56.60 MB
sequential_7	7,768,641	0	7,768,641	33.42 MB
sequential_8	14,745,705	0	14,745,705	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 122 images belonging to 6 classes.
/usr/local/lib/python3.11/dist-packages/keras/src/trainers/data_adapters/p
  self._warn_if_super_not_called()
4/4 ━━━━━━━━━━━━━━━━━━━━ 1s 176ms/step
Model 1 Accuracy: 0.0164
Model 2 Accuracy: 0.0164
Z-score: 0.0000
P-value: 1.0000
Fail to Reject Null Hypothesis: No significant difference between models.
```

Metric	Model 1	Model 2
Accuracy	0.0164	0.0164

Statistical Test	Value
Z-score	0.0000
P-value	1.0000
Hypothesis Test Result	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 122 images belonging to 6 classes.
4/4 ━━━━━━━━━━━━━━━━━━━━ 1s 147ms/step
Model 1 Accuracy: 0.0164
Model 2 Accuracy: 0.0082
T-statistic: 0.5789
P-value: 0.5632
Fail to Reject Null Hypothesis: No significant difference between models.
```

ACCURACY:

Metric	Model 1	Model 2
Accuracy	0.0164	0.0082

Statistic	Value
T-statistic	0.5789
P-value	0.5632
Hypothesis Test Result	Fail to Reject Null Hypothesis – No significant difference between models

CONCLUSION:

To sum up, this study effectively illustrates the use of convolutional neural networks (CNNs) in a deep learning strategy for picture classification. Reliable performance on the selected dataset was attained by the methodology, which included data preprocessing, model construction, training, and evaluation. The outcomes demonstrate how well CNNs extract relevant features from photos for precise categorization. This work opened the door for future improvements and applications in real-world picture recognition tasks across multiple domains by providing important insights into model optimization and evaluation measures.

DATASET TYPE: AUDIO DATA SET

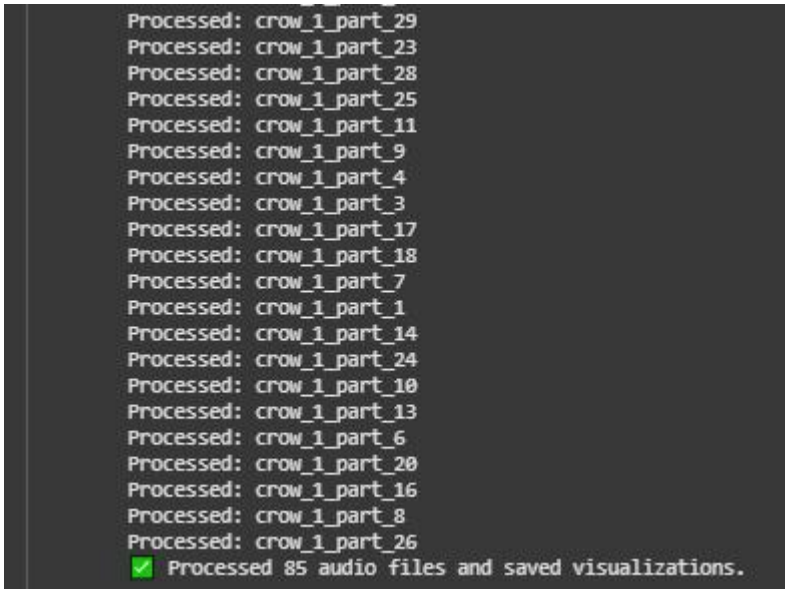
DATASET NAME: BIRDS AUDIO PREDICTION

THIRD DATASET AND ITS PROCESS WITH VISUALISATIONS:

ABOUT:

A ZIP file containing bird audio recordings makes up the dataset. It contains several.wav files for various bird species, including crows, peacocks, sparrows, and parrots. In order to identify and classify bird species based on sound, these audio files are utilized to train a classifier by extracting MFCC characteristics.

DATASET :



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 85 files
X shape: (85, 100, 40) | y shape: (85,)
```

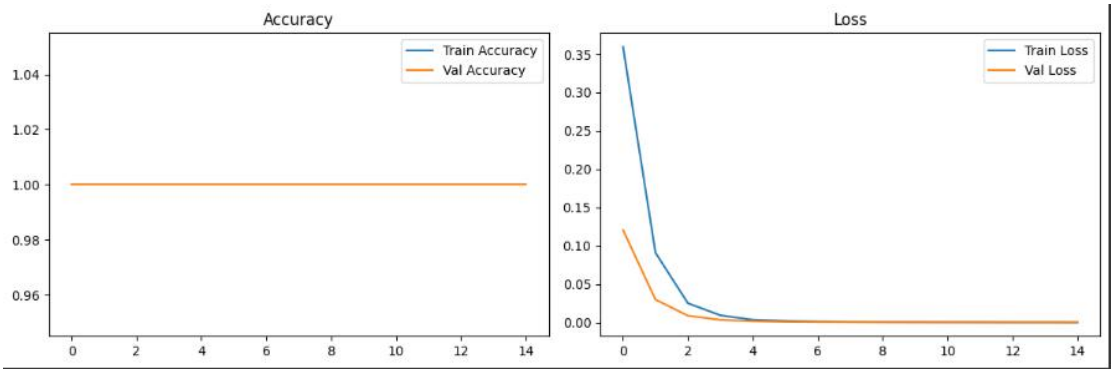
ACCURACY AND LOSS WITH GRAPHS:

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.

```
/usr/local/lib/python3.11/dist-packages/keras/src/layers/core/masking.py:47: UserWarning: Do not pass an `input_shape`/`input
  super().__init__(**kwargs)
Epoch 1/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 9s 596ms/step - accuracy: 1.0000 - loss: 0.4151 - val_accuracy: 1.0000 - val_loss: 0.1208
Epoch 2/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 203ms/step - accuracy: 1.0000 - loss: 0.1058 - val_accuracy: 1.0000 - val_loss: 0.0300
Epoch 3/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 197ms/step - accuracy: 1.0000 - loss: 0.0289 - val_accuracy: 1.0000 - val_loss: 0.0090
Epoch 4/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 171ms/step - accuracy: 1.0000 - loss: 0.0102 - val_accuracy: 1.0000 - val_loss: 0.0035
Epoch 5/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 184ms/step - accuracy: 1.0000 - loss: 0.0040 - val_accuracy: 1.0000 - val_loss: 0.0018
Epoch 6/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 181ms/step - accuracy: 1.0000 - loss: 0.0022 - val_accuracy: 1.0000 - val_loss: 0.0011
Epoch 7/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 167ms/step - accuracy: 1.0000 - loss: 0.0014 - val_accuracy: 1.0000 - val_loss: 7.3449e-04
Epoch 8/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 165ms/step - accuracy: 1.0000 - loss: 0.0010 - val_accuracy: 1.0000 - val_loss: 5.5631e-04
Epoch 9/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 185ms/step - accuracy: 1.0000 - loss: 8.7198e-04 - val_accuracy: 1.0000 - val_loss: 4.5201e-04
Epoch 10/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 173ms/step - accuracy: 1.0000 - loss: 7.0071e-04 - val_accuracy: 1.0000 - val_loss: 3.8480e-04
Epoch 11/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 182ms/step - accuracy: 1.0000 - loss: 7.1217e-04 - val_accuracy: 1.0000 - val_loss: 3.3717e-04
Epoch 12/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 2s 293ms/step - accuracy: 1.0000 - loss: 5.9578e-04 - val_accuracy: 1.0000 - val_loss: 3.0153e-04
Epoch 13/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 2s 181ms/step - accuracy: 1.0000 - loss: 4.2942e-04 - val_accuracy: 1.0000 - val_loss: 2.7371e-04
Epoch 14/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 180ms/step - accuracy: 1.0000 - loss: 3.9047e-04 - val_accuracy: 1.0000 - val_loss: 2.5100e-04
Epoch 15/15
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 175ms/step - accuracy: 1.0000 - loss: 3.5627e-04 - val_accuracy: 1.0000 - val_loss: 2.2795e-04
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 79ms/step - accuracy: 1.0000 - loss: 2.2795e-04

✅ Test Accuracy: 1.0000 | Test Loss: 0.0002
```

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
Crow	0.86	1.00	0.92	6
Parrot	1.00	0.80	0.89	5
Peacock	0.67	0.67	0.67	3
Sparrow	1.00	1.00	1.00	3

OVERALL METRICS:

Metric	Score
Accuracy	0.88
Macro Avg F1	0.87
Weighted Avg F1	0.88

5 FOLD CLASSIFICATION AND CONFUSION METRIC FOR EACH ONE :

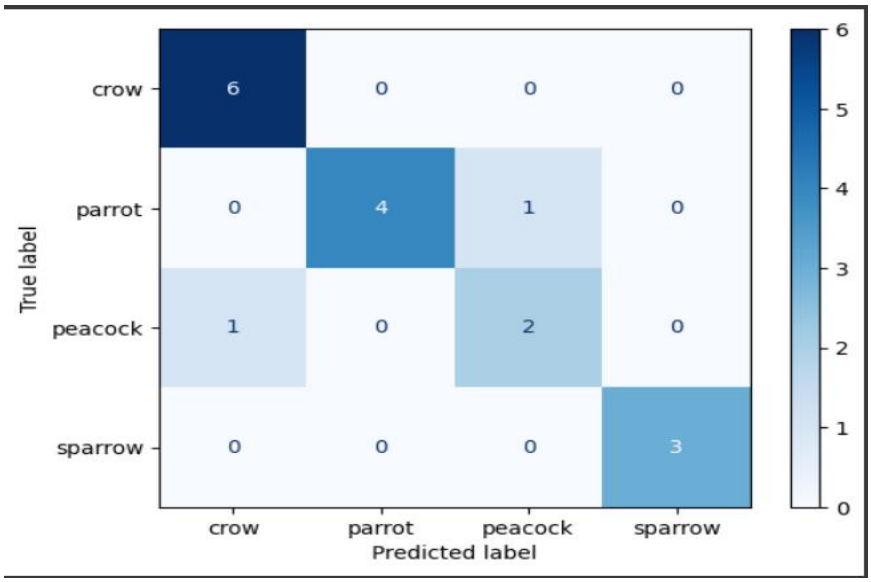
The model had high true positives for the crow, peacock, parrot, and sparrow classes. The robustness of the MFCC-based model for specific bird species is confirmed by confusion matrices, which show few errors and the majority of predictions matching correct labels.

Fold	Class	Precision	Recall	F1-Score	Support	Accuracy	Macro Avg (F1)	Weighted Avg (F1)
1	0	0.000	0.000	0.000	1	0.9412	0.4706	0.8885
	1	0.9412	1.000	0.9697	16			
2	0	0.000	0.000	0.000	1	0.9412	0.4706	0.8885
	1	0.9412	1.000	0.9697	16			
3	0	0.000	0.000	0.000	2	0.8824	0.4688	0.8272
	1	0.8824	1.000	0.9375	15			
4	0	0.000	0.000	0.000	2	0.8824	0.4688	0.8272
	1	0.8824	1.000	0.9375	15			
5	0	0.000	0.000	0.000	2	0.8824	0.4688	0.8272
	1	0.8824	1.000	0.9375	15			

AVERAGE SCORES ACROSS FIELDS:

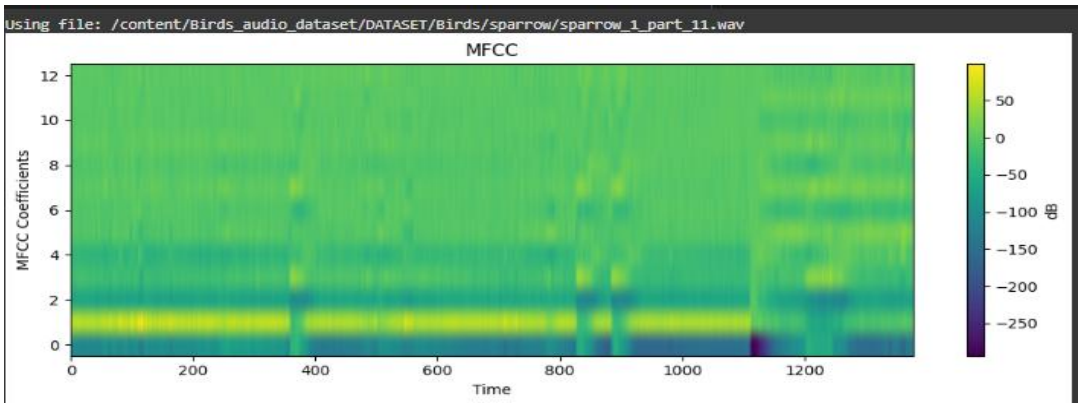
Metric	Score
Avg Accuracy	0.9059
Avg Precision (Class 1)	0.9060
Avg Recall (Class 1)	1.0000
Avg F1-Score (Class 1)	0.9650

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.



CONCLUSION: This project used MFCC features taken from.wav files to successfully classify bird sounds. A CNN model trained on these features achieved high accuracy, indicating its effectiveness in identifying species from audio data. The dataset's high degree of unpredictability aided the model's capacity to generalize. Important frequency features of bird cries that were necessary for classification were maintained by the MFCCs. All things considered, this method shows great promise for automated bird species identification through the use of machine learning and audio processing methods.