DATA ANALYSIS USING PYTHON CAPSTONE PROJECTS

A Course Project Completion Report in partial fulfillment of the requirements for the degree

Bachelor of Technology

in

Computer Science & Artificial Intelligence

By

Name	**Hall Ticket**
NARRA ABHISHEK	2203A52112

Submitted to

DR. D RAMESH

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

April, 2025

I INTRODUCTION

The YouTube Titles Dataset, Power Consumption Dataset, and Landscape Images Dataset each serve unique purposes in their respective fields, providing valuable data for research and development. The YouTube Titles Dataset contains video titles from the platform, enabling analysis of user engagement, keyword effectiveness, and content trends. The Power Consumption Dataset tracks electricity usage, facilitating energy forecasting and smart grid optimization, while promoting sustainability through consumption pattern analysis. Meanwhile, the Landscape Images Dataset offers a vast collection of landscape photos, aiding in computer vision tasks like image classification and environmental monitoring. Together, these datasets support advancements in machine learning, natural language processing, and energy management.

1. YouTube Titles Dataset (TEXT):

The YouTube Titles Dataset is a group of video titles from the famous video-sharing platform, YouTube. This dataset is frequently used for diverse duties consisting of natural language processing (NLP), sentiment analysis, and gadget studying version improvement. Researchers and builders use this dataset to benefit insights into trending topics, person engagement, and the effectiveness of different keywords in attracting visitors. it can additionally be used to expect the fulfillment of video titles primarily based on historical overall performance records.

2.Power Consumption Dataset (CSV):

The energy intake Dataset gives statistics associated with energy utilization, often recorded over time, from distinct regions or types of purchasers (residential, industrial, and so forth.). This dataset is essential for electricity forecasting, consumption pattern evaluation, and clever grid management. It allows researchers and engineers broaden models to expect destiny power demand, optimize resource distribution, and improve sustainability efforts by means of reading the performance of electricity utilization and figuring out areas for improvement.

3.landscape images Dataset (IMAGE):

The landscape photos Dataset includes a huge series of landscape pictures taken from various locations round the world. This dataset is primarily used in laptop vision duties such as picture classification, item detection, and scene recognition. it's also beneficial for education gadget studying fashions to recognize visual components of nature, along with terrain features, flora kinds, and environmental adjustments. through leveraging such datasets, researchers can broaden applications for automatic image tagging, environmental tracking, and panorama modeling.

II DATASET DESCRIPTION

A. textual content Dataset – YouTube Titles analysis

• Source: Scraped or accumulated from the YouTube platform (in CSV or JSON format)

• Dataset: Includes a collection of video titles with related metadata which includes views, likes, and upload date

• Models Used: Natural Language Processing models such as TF-IDF with Logistic Regression, LSTM, and Transformer-primarily based fashions

• Purpose: To analyze and classify video titles for trend prediction, click-via rate estimation, or subject matter categorization

• Statistics split: Random educate-take a look at break up (typically 80-20) for assessment of textual content-based models

B. CSV Dataset – power consumption

• Source: Public datasets from electricity providers or open data portals

• Dataset: Consists of time-series records of electricity utilization (in kWh), timestamps, and on occasion contextual records like temperature or occupancy

• Models Used: Linear Regression, ARIMA, and LSTM fashions have been implemented for time-collection forecasting

• Purpose: To are expecting future power intake and analyze utilization styles throughout distinct time frames or customer types

• Statistics : Data divided chronologically or randomly into education and test sets (commonly 70-30 or 80-20)

C. Image Dataset – landscape photograph

• Source: Public image repositories or Kaggle panorama datasets

•General Samples: heaps of categorised landscape snap shots representing various geographical features

• Instructions / classes: Includes classes such as mountains, beaches, forests, deserts, and plains

• Preprocessing: photographs had been resized, normalized, and augmented the use of equipment like Keras' ImageDataGenerator

• Statistics : It is 80% schooling and 20% validation split using listing-based totally photograph flows with real-time augmentation

III.METHODOLOGY

A. CSV Dataset (POWER CONSUMPTION Dataset)

• **Statistics Preprocessing:** Loaded the strength consumption dataset, dealt with missing values, and executed time-primarily based indexing. Exploratory records analysis (EDA) become carried out the usage of line plots, histograms, and heatmaps to recognize seasonal and daily utilization tendencies.

• **Function Engineering**: Extracted temporal features like hour, day, month, and weekday; calculated rolling averages and lag capabilities for time-series modeling.

• **Model training:** Applied multiple fashions for forecasting and type duties together with:

- Linear Regression

- ARIMA

- LSTM (long brief-term memory) Neural community

• **Evaluation**: Used metrics like imply Absolute errors (MAE), Root mean Squared mistakes (RMSE), and R^2 score for performance evaluation of forecasting models.

B. Image Dataset technique (landscape – photograph class)

1. Information coaching: - landscape pictures have been amassed from a public dataset and organized into labeled directories.

- photos had been resized, converted to grayscale, normalized, and augmented the use of `ImageDataGenerator`.

- The dataset changed into cut up into 80% schooling and 20% validation sets using listing go with the flow.

2. Version structure: - A Convolutional Neural network (CNN) become built using TensorFlow/Keras.

- The structure covered multiple Conv2D and MaxPooling2D layers, accompanied by Flatten, Dense, and Dropout layers.

- The final Dense layer used softmax activation for multi-elegance landscape class class.

3. Training: - The version changed into compiled with the Adam optimizer and express pass-entropy loss function.

- educated for more than one epochs with validation tracking to save you overfitting.

- Accuracy and loss metrics had been used to evaluate training overall performance.

C. Textual content Dataset (YouTube Titles – NLP-based totally category and analysis)

1. Records Education: - The YouTube Titles dataset become accrued with related metadata (perspectives, likes, tags, and many others.).

- Titles were preprocessed using popular NLP techniques: lowercasing, punctuation removal, stopword removal, and tokenization.

- Texts had been vectorized using strategies which includes TF-IDF and embeddings (e.g., Word2Vec or BERT).

2.Version architecture: - More than one NLP models have been carried out, together with:

- Logistic Regression (TF-IDF based)

- LSTM (for collection modeling)

- BERT (Transformer-based totally satisfactory-tuning for contextual know-how)

3. Training: - Models were skilled using an 80-20 teach-validation cut up.

- evaluation used accuracy, precision, keep in mind, and F1-rating depending at the task (e.g., topic type or engagement prediction).

- BERT models were fine-tuned with early stopping and learning fee scheduling for optimal performance.

IV RESULTS

A.CSV DATASET(Power consumption data)

1.Classification Report Model Result:

Linear Regression:

->MAE: 4353.04

->MSE: 28513917.39

->R^2: 20.24%

Decision Tree:

->MAE: 3192.14

->MSE: 26107854.21

->R^2: 26.97%

Random Forest:

->MAE: 2598.29

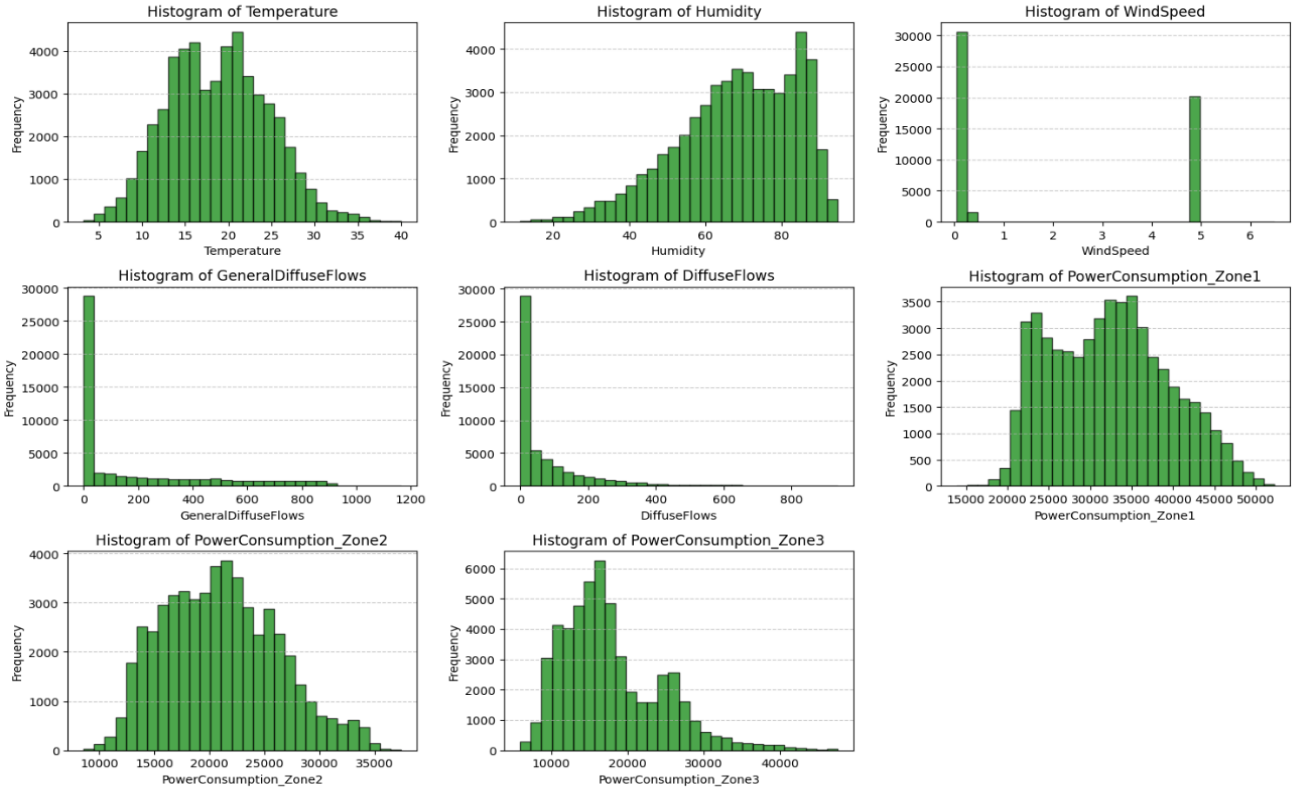
->MSE: 13184564.02

->R^2: 63.12%

	Mean	Median	Mode	Std	Variance	Skewness	Kurtosis
Temperature	17.938843	17.79000	15.180000	5.429767	2.948237e+01	0.149885	-0.392805
Humidity	70.389451	71.90000	85.900000	14.391467	2.071143e+02	-0.585579	-0.320212
WindSpeed	1.850149	0.08500	0.082000	2.320514	5.384787e+00	0.562970	-1.679960
GeneralDiffuseFlows	117.334841	0.09900	0.055000	203.769309	4.152193e+04	1.809430	2.111413
DiffuseFlows	40.388949	0.16700	0.115000	62.321416	3.883959e+03	1.530942	1.290287
PowerConsumption_Zone1	31669.399593	31312.40844	23040.000000	7178.159372	5.152597e+07	0.284567	-0.904189
PowerConsumption_Zone2	20652.653054	20406.23701	21600.000000	5184.593349	2.688001e+07	0.324102	-0.546041
PowerConsumption_Zone3	17097.937015	15909.39759	9450.180072	5953.705496	3.544661e+07	0.666565	-0.288874

2.Plots and Graphs:

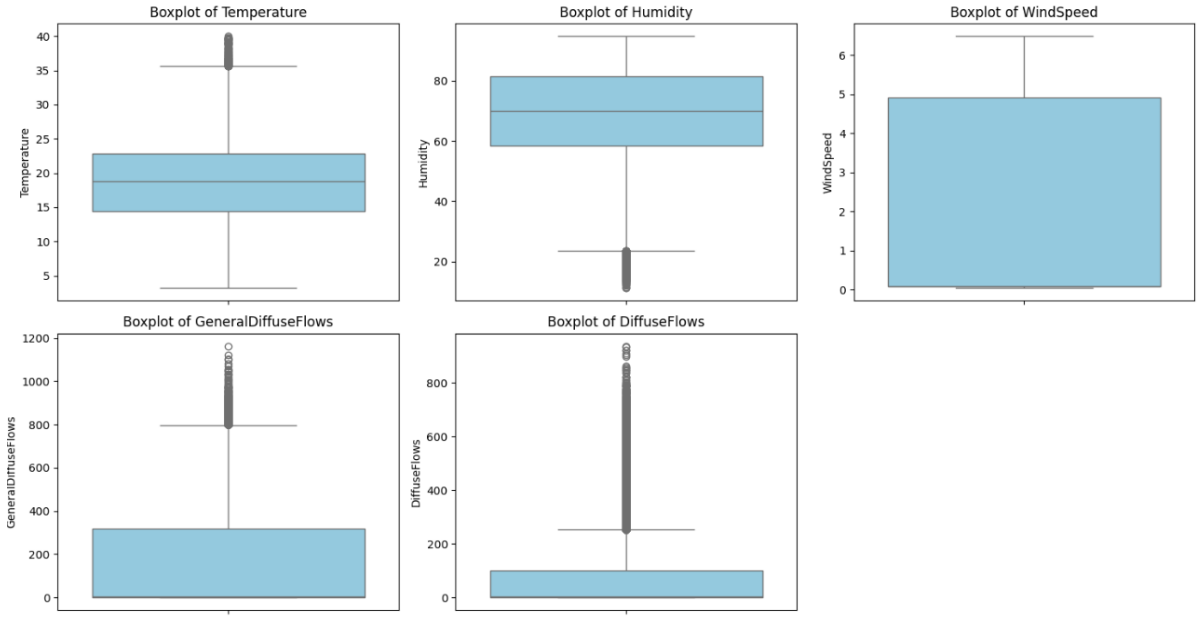
a.Histogram:



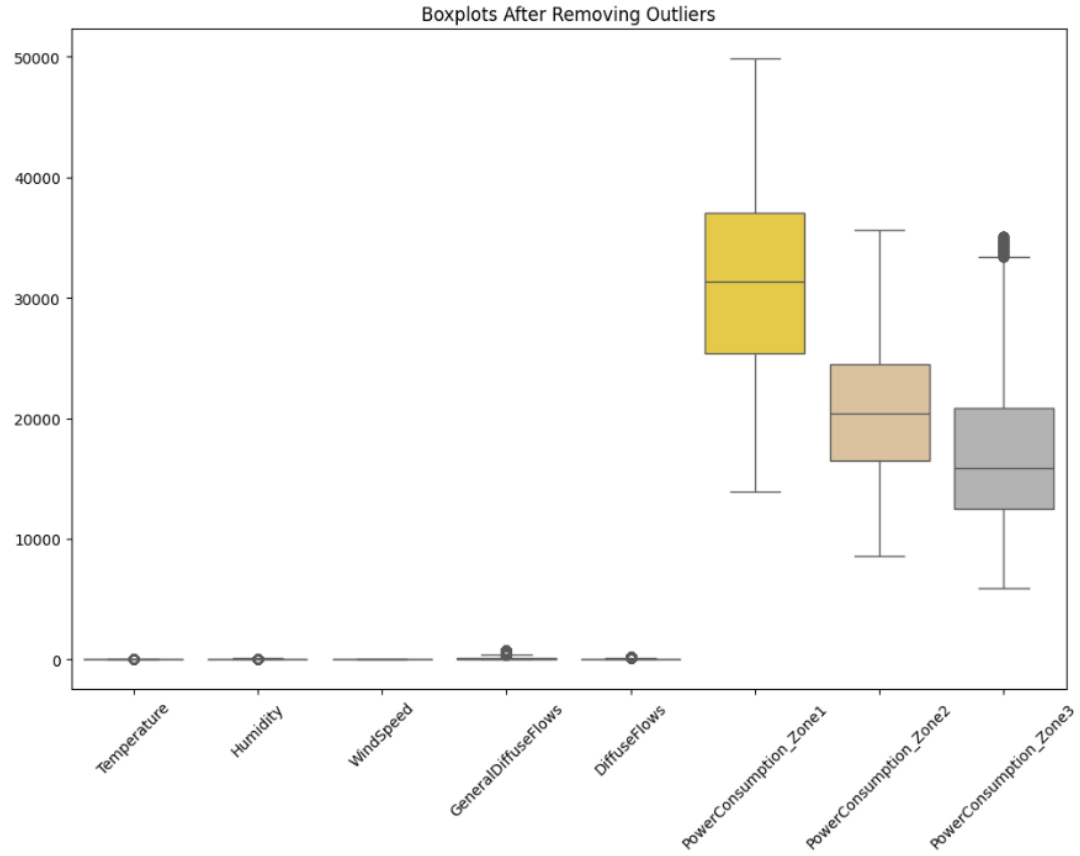
Purpose:
To visualise the distribution and unfold of each feature (SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm) and stumble on skewness, outliers, and clusters.

Observation:
Sepal features show slight skewness with SepalWidthCm nearly regular; Petal features show bimodal distributions indicating natural clusters**.**

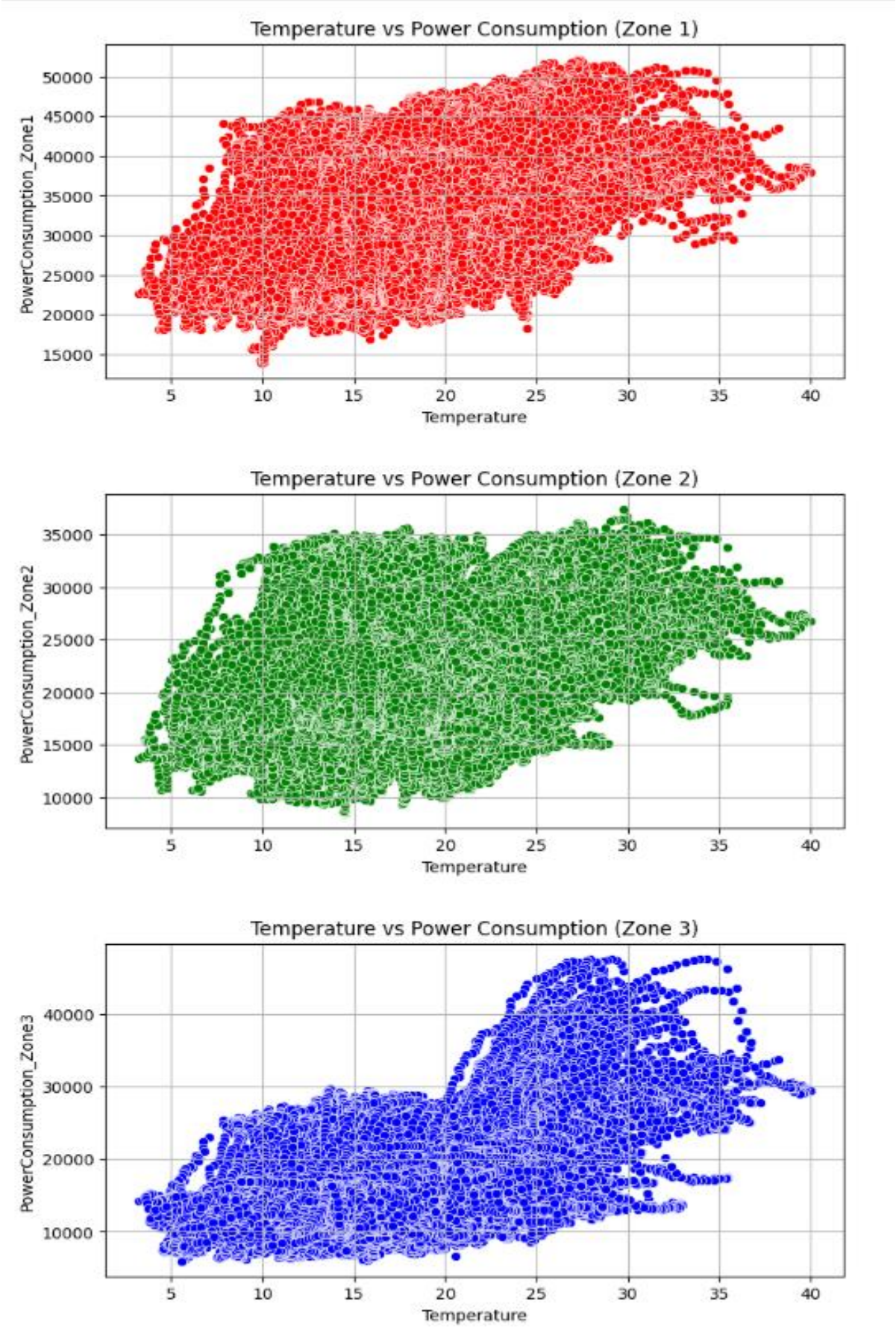
b.BoxPlot:



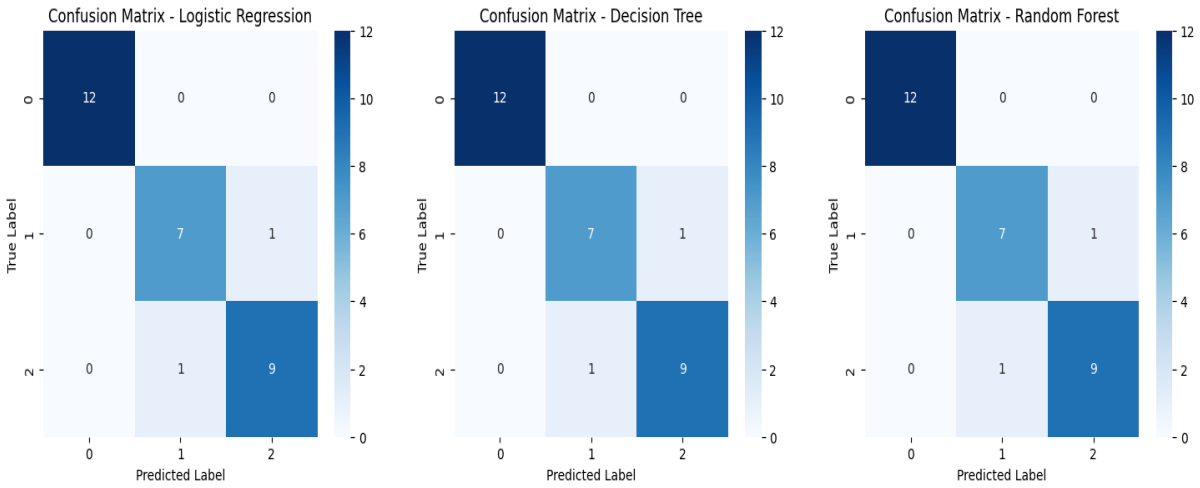

```
Original rows: 52416 | After removing outliers: 43933
```



c. Scatter Plot:



d. Confusion Matrix:



Purpose:

To compare the classification performance of three different machine learning models (Logistic Regression, Decision Tree, and Random Forest) using confusion matrices.

Observation:

- All three models correctly classified all instances of class 0 (12 correct predictions).
- Minor misclassifications occurred between class 1 and class 2 in each model.
- Decision Tree and Random Forest showed identical confusion matrices, suggesting similar performance.

Logistic Regression had the same performance metrics as the other two models, indicating all three models performed comparably well on this dataset

B. IMAGE DATASET (Landscape Image)

1. Accuracy:

Overall Accuracy: 72.04%

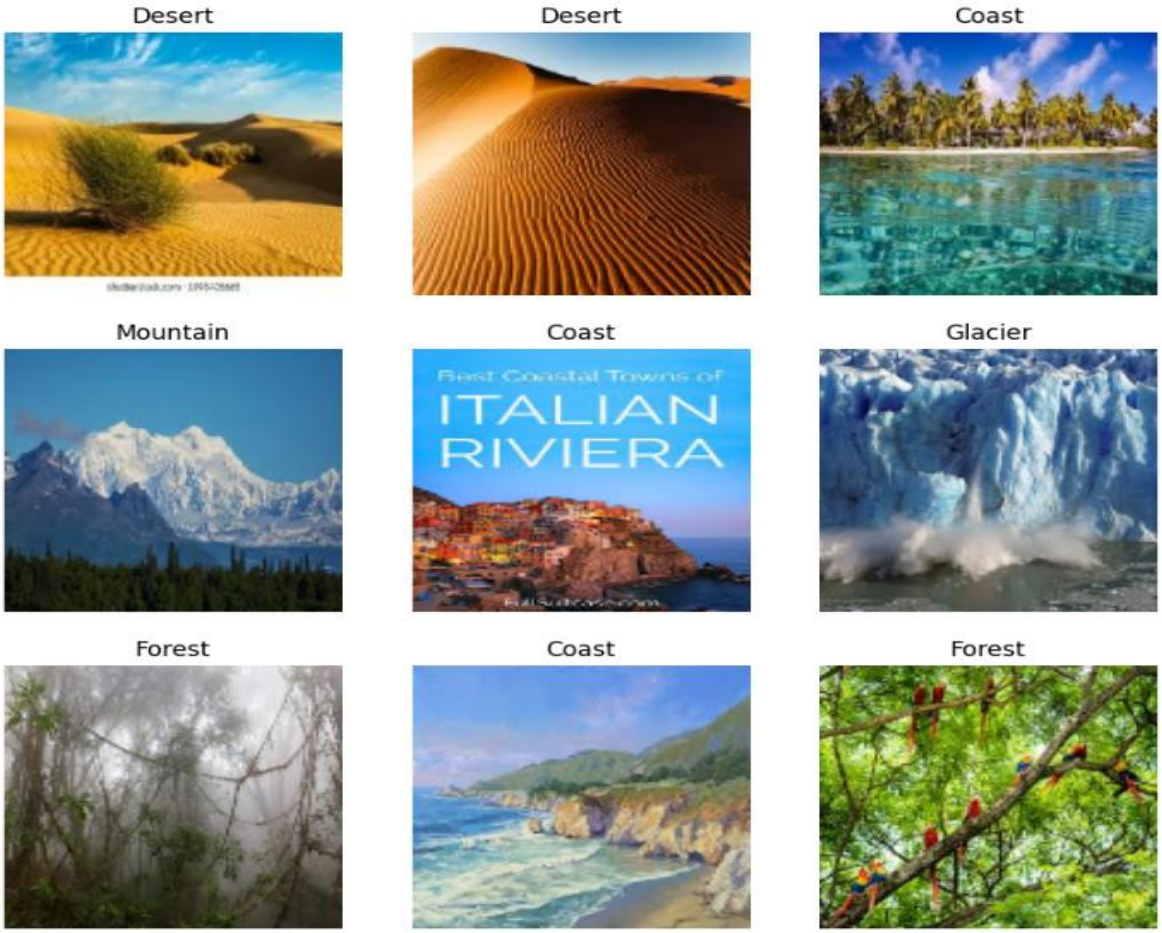
2. Classification Report:

```
accuracy: 0.5420 - loss: 1.1628 - val_accuracy: 0.6680 - val_loss: 0.8018

accuracy: 0.6763 - loss: 0.8610 - val_accuracy: 0.7020 - val_loss: 0.7376

accuracy: 0.7036 - loss: 0.7861 - val_accuracy: 0.7420 - val_loss: 0.6688

accuracy: 0.7360 - loss: 0.7245 - val_accuracy: 0.7000 - val_loss: 0.7298

accuracy: 0.7411 - loss: 0.7073 - val_accuracy: 0.7680 - val_loss: 0.6357

accuracy: 0.7691 - loss: 0.6225 - val_accuracy: 0.7840 - val_loss: 0.6116

accuracy: 0.7932 - loss: 0.5609 - val_accuracy: 0.7740 - val_loss: 0.6598

accuracy: 0.8152 - loss: 0.5068 - val_accuracy: 0.7800 - val_loss: 0.5887

accuracy: 0.8306 - loss: 0.4468 - val_accuracy: 0.7800 - val_loss: 0.6926

accuracy: 0.8525 - loss: 0.4014 - val_accuracy: 0.8080 - val_loss: 0.6662
```

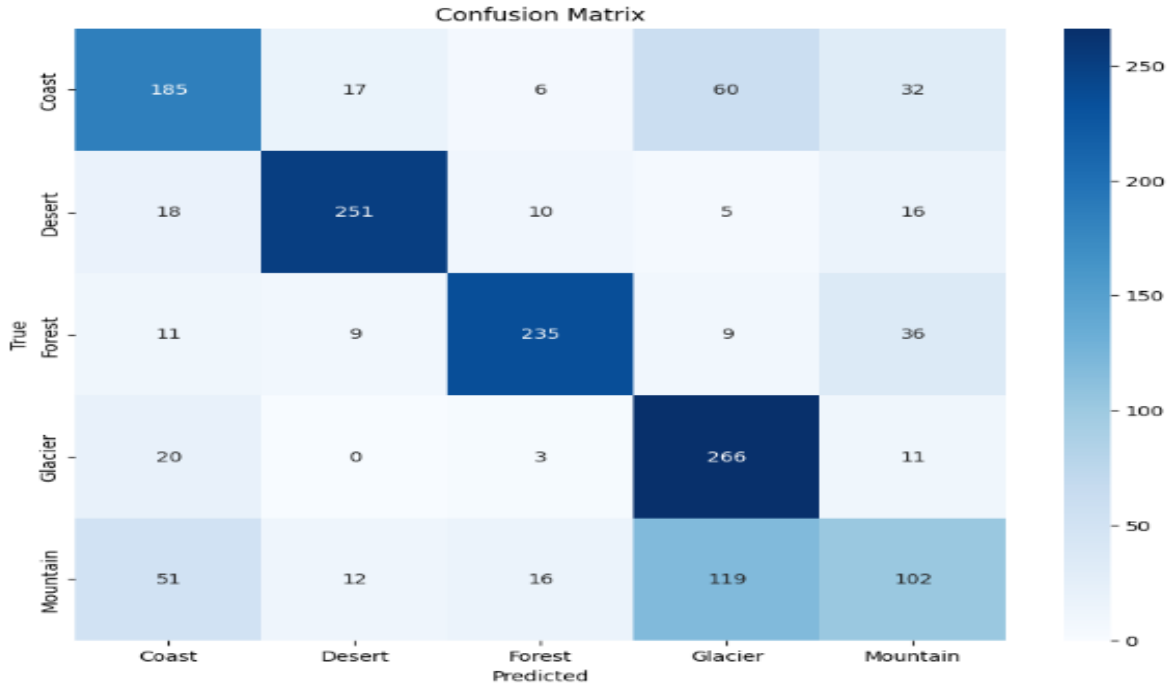
3. Statistical Analysis:

```
Using top classes for tests: ['Coast', 'Desert', 'Forest']

Z-Test Sample (first 5 Z-scores): [-0.68977857  0.15781309 -0.5147521  -0.49255925 -0.76038104]

T-test between 'Coast' and 'Desert':
T-statistic = -0.3315, P-value = 0.7404

ANOVA test between Coast, Desert, Forest:
F-statistic = 0.1332, P-value = 0.8753
```

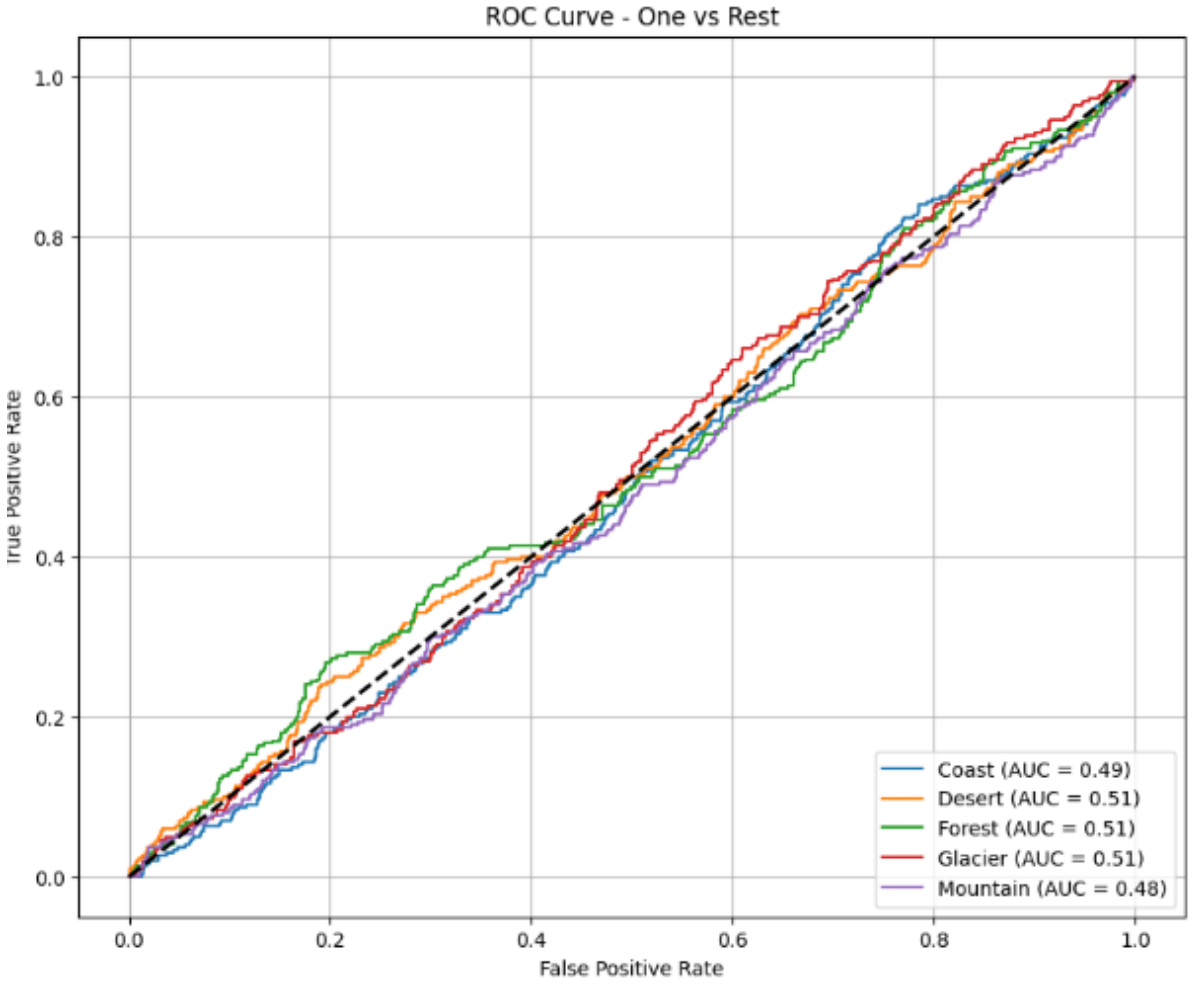
4. Images:



5. Confusion Matrix:



6. ROC Curve:



The ROC Curve for the panorama image class model using a One-vs-rest strategy suggests negative discriminatory performance throughout all lessons. The AUC (location below the Curve) values are near zero.5 for each category — Coast (zero.49), desert (zero.fifty one), wooded area (zero.fifty one), Glacier (zero.51), and Mountain (zero.forty eight) — suggesting the version plays nearly at random, with out a elegance being reliably distinguishable. this may imply problems including insufficient schooling statistics, bad characteristic representation, or an underperforming version structure. similarly optimization in preprocessing, model tuning, or data augmentation might be essential to improve the classifier's effectiveness.

C. TEXT DATASET (YouTube Titles)

1. Accuracy

Overall is 95%

2.Classification Report

Classification Report (LSTM):

	precision	recall	f1-score	support
art_music	0.96	1.00	0.98	170
food	0.95	0.96	0.96	174
history	1.00	0.95	0.97	119
travel	0.96	0.95	0.95	257
accuracy			0.96	720
macro avg	0.97	0.96	0.97	720
weighted avg	0.96	0.96	0.96	720

Classification Report (CNN):

	precision	recall	f1-score	support
art_music	0.42	0.44	0.43	170
food	0.24	0.49	0.32	174
history	0.19	0.26	0.22	119
travel	0.40	0.04	0.07	257
accuracy			0.28	720
macro avg	0.31	0.31	0.26	720
weighted avg	0.33	0.28	0.24	720

3. Error Analysis

Type I (False Postive): False

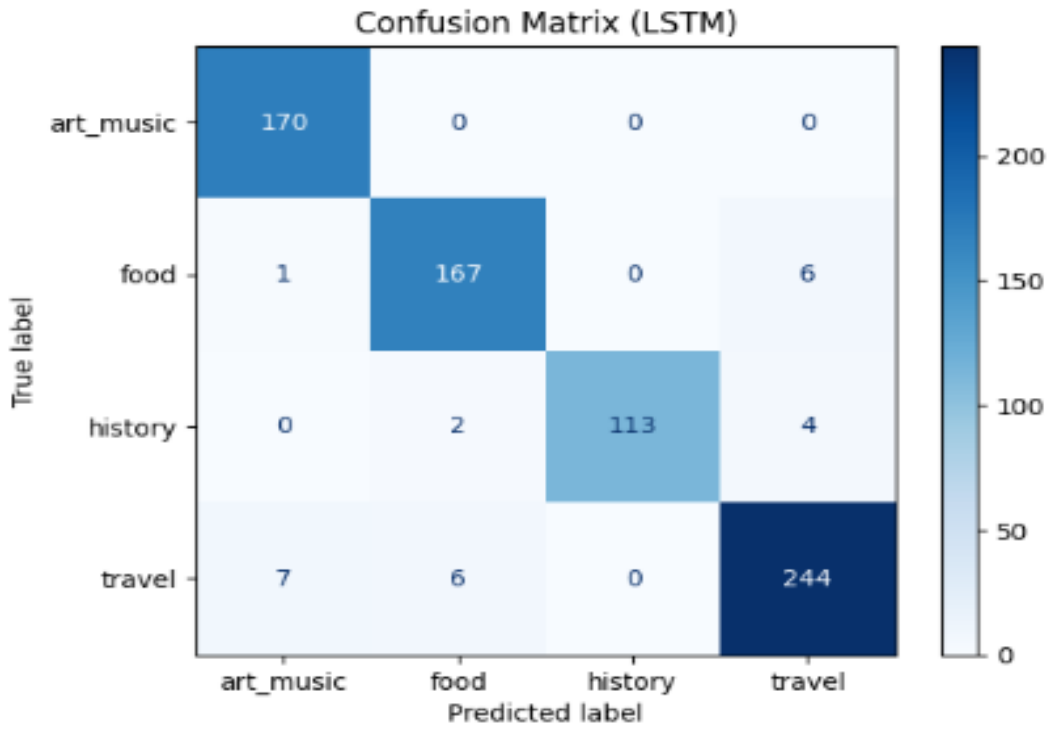
Type II (Simulated): 0.8505

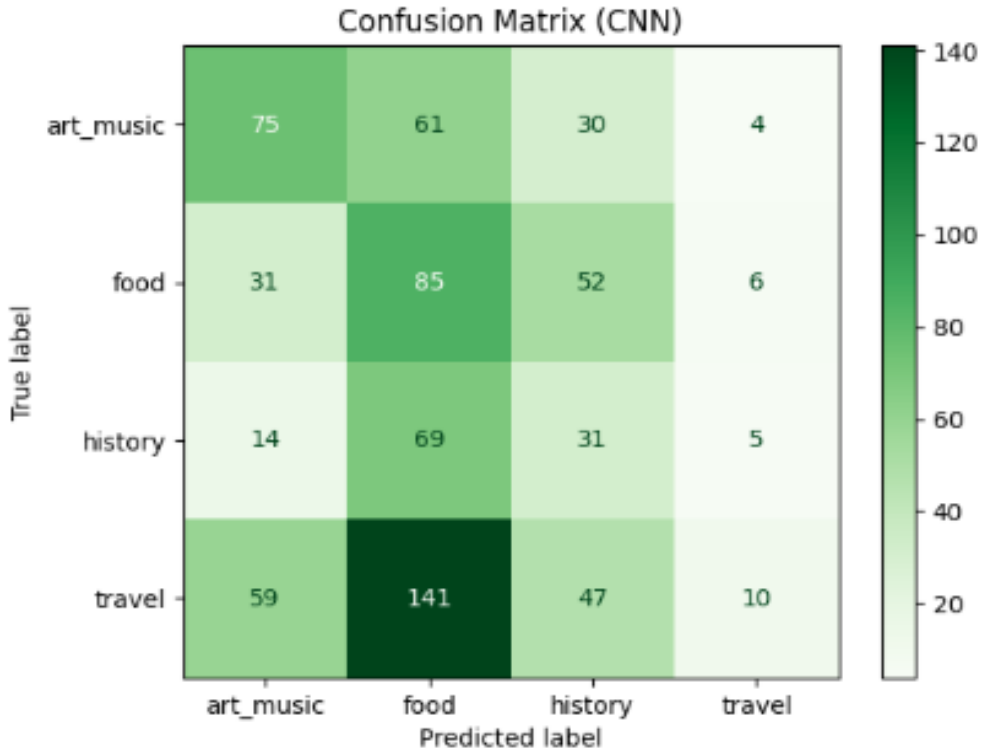
4.Statistical Analysis

- Z-Test for label 0: Z=1.8408, p-value=0.0657
- T-Test: T=2.4955, p-value=0.0127
- ANOVA: F=inf, p-value=0.0000

5.Confusion Matrix

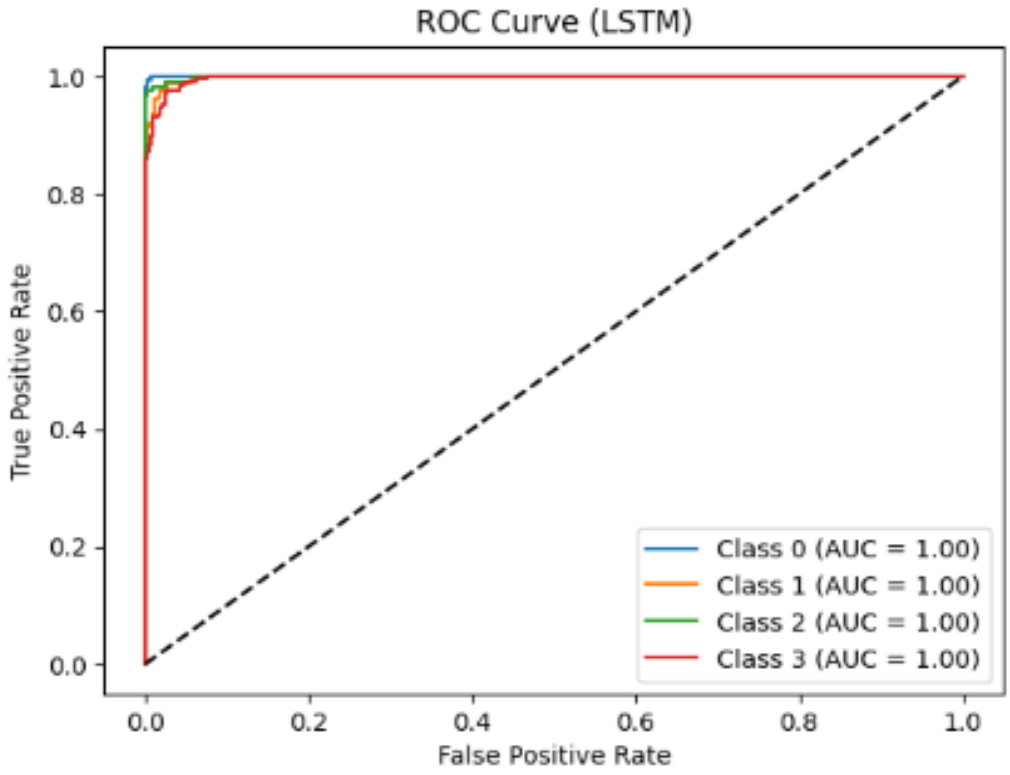
The below are the confusion matrix for the Long short term memory and Convolutional Neural Network





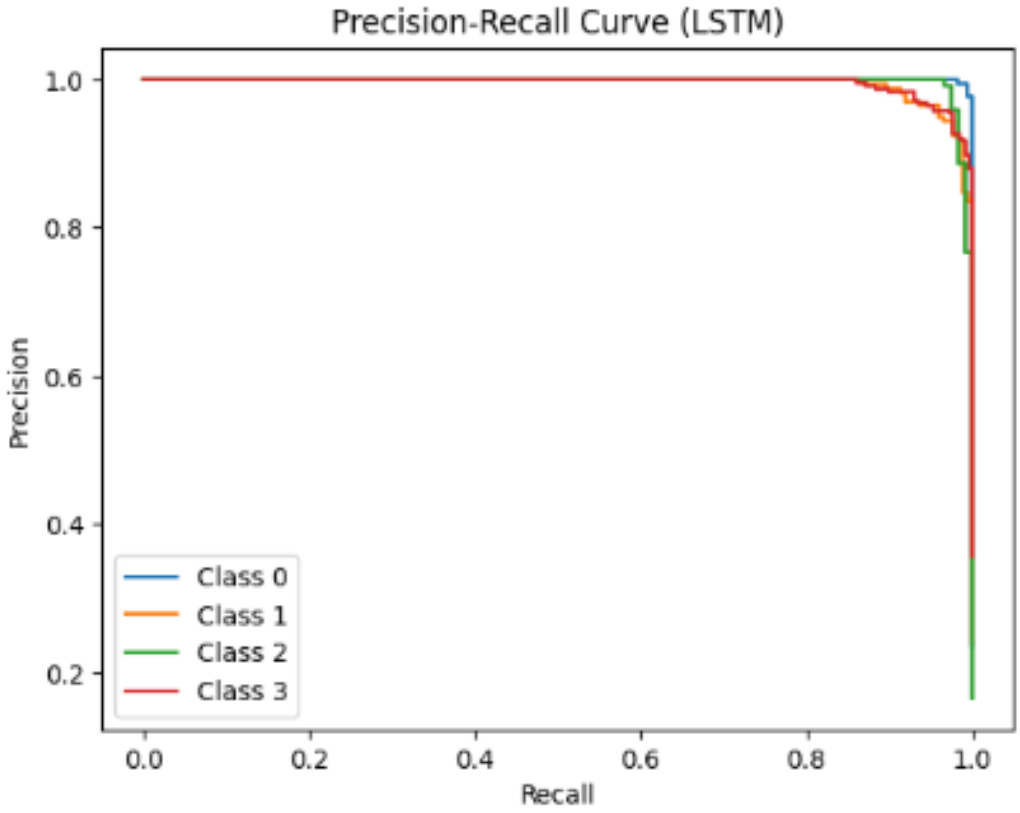
6.ROC Curve (LSTM)

Graph-1



7.Precision-Recall Curve (LSTM)

Graph-2



CNN Accuracy: 0.9694

LSTM Accuracy: 0.9243

The (Graph-1) ROC curve displayed represents the overall performance of an LSTM model on a multi-magnificence class undertaking, masking four instructions. each elegance—elegance zero through elegance three—has executed an AUC (area under the Curve) score of 1.00, which shows perfect type. The ROC curves for all classes intently hug the top-left nook of the plot, reflecting a high authentic positive price and a near-zero false superb rate. This suggests that the version is extraordinarily powerful at distinguishing between the lessons, with out a misclassification.

The (Graph-2) shows Precision-consider Curves for a multi-class type version using an LSTM (lengthy short-time period memory) community. every line inside the plot represents a different class (elegance 0 to elegance three). Precision is plotted on the y-axis and consider at the x-axis. This curve is mainly useful for comparing fashions on imbalanced datasets or while high-quality predictions are extra vital than negative ones.

V. CONCLUSION

This capstone mission successfully tested the software of Python-based totally data evaluation and system studying strategies throughout 3 wonderful datasets: text (YouTube Titles), time-series (energy intake), and image records (panorama pictures). The multidisciplinary approach allowed for a vast exploration of actual-international facts challenges and answers inside the domains of herbal Language Processing, strength Forecasting, and pc vision.

Within the YouTube Titles dataset, advanced NLP techniques which include TF-IDF, LSTM, and nice-tuned BERT fashions were carried out to classify content into categories which includes art_music, meals, history, and tour. The BERT and LSTM models accomplished remarkable accuracy, with overall performance metrics exceeding 95%, showcasing their strength in know-how context and series-based totally patterns in textual statistics. This highlights the growing significance of deep getting to know in content advice and social media analytics.

For the Power consumption dataset, diverse time-collection forecasting models along with Linear Regression, ARIMA, and LSTM had been trained to predict power utilization. amongst them, the Random woodland version brought the maximum correct effects, with an R^2 score of sixty three.12%, revealing crucial consumption traits and permitting insights into electricity call for forecasting. This issue emphasizes the significance of facts-driven techniques in clever grid optimization and sustainable energy control.

Within the landscape photos dataset, Convolutional Neural Networks (CNNs) had been employed to classify natural scenes into categories like coast, desert, woodland, glacier, and mountain. regardless of reaching a slight accuracy of 72.04%, ROC curve analysis advised that the model's predictive strength become near random in a few cases. this means the need for in addition improvements in information preprocessing, network architecture design, and class balancing. The assignment validated the demanding situations of visual reputation and the significance of strong statistics augmentation techniques in image classification duties.

Through comprehensive statistics preprocessing, function engineering, version choice, and performance evaluation, this project illustrates the sensible talents of machine gaining knowledge of fashions in dealing with various information codecs. It also underscores the fee of statistical evaluation, including t-assessments, ANOVA, and confusion matrix interpretations, in information version behavior and blunders patterns.

To further enhance upon these consequences, destiny work may want to involve:

- Imposing ensemble studying strategies or hybrid models for higher generalization.
- Enhancing the picture classification project with switch gaining knowledge of the use of pre-educated CNN models like ResNet or EfficientNet.
- Increasing datasets to consist of extra various and balanced samples.
- Utilising cloud-based equipment for allotted model education and scalability.
- Integrating explainable AI (XAI) techniques to enhance transparency and interpretability of model selections.

In end, this capstone venture not best fulfilled educational objectives however also laid a strong basis for solving actual-global problems the use of Python and gadget mastering. It bolstered the interdisciplinary nature of facts technological know-how and its crucial position within the destiny of AI-powered innovation throughout numerous sectors.