# PE1-Data Analysis Using Python

# 

A Course Completion Report in

partial fulfillment of the degree

## Bachelor of Technology

in

**ComputerScience&Artificial Intelligence**

**NAME: KARNAM SAI CHANDANA HALL NO: 2203A52094**

**Submitted to**

**Dr. D. Ramesh**



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

**March,**

**CONTENTS**

**Dataset 1: Bank Transaction**

|  |  |  |
| --- | --- | --- |
| **S. No** | **Section Title** | **Page No.** |
| 1 | Abstract | 3 |
| 2 | Introduction | 3 |
| 3 | Dataset Description | 3 |
| 4 | Methodology | 4 |
| 5 | Results | 5 |
| 6 | Conclusion | 8 |

**Dataset 2:** **Satellite Image**

|  |  |  |
| --- | --- | --- |
| **S. No** | **Section Title** | **Page No.** |
| 1 | Abstract | 9 |
| 2 | Introduction | 9 |
| 3 | Data Set Description | 9 |
| 4 | Methodology & Results | 10 |
| 5 | Conclusion | 13 |

**Dataset 3: American Speech Recognition**

|  |  |  |
| --- | --- | --- |
| **S. No** | **Section Title** | **Page No.** |
| 1 | Abstract | 14 |
| 2 | Introduction | 14 |
| 3 | Dataset Description | 14 |
| 4 | Methodology | 14 |
| 5 | Implementation | 15 |
| 6 | Results | 15 |
| 7 | Conclusion | 16 |

* **Bank Transaction (Data set – 1)**

**1.Abstract**

This project focuses on the design, implementation, and analysis of a bank transaction system that supports core functionalities such as deposits, withdrawals, balance inquiries, fund transfers, and transaction history management. Emphasis is placed on ensuring accuracy, speed, security, and transparency of each transaction. Modern techniques like encryption and real-time validation are employed to protect sensitive data and prevent fraud. Additionally, data analytics is used to monitor user Behavior and detect anomalies. This system aims to deliver a seamless banking experience while maintaining high standards of operational integrity and regulatory compliance.

**2.Introduction**

A bank transaction refers to any activity involving the movement of money into, out of, or within a bank account. It is a fundamental part of banking operations that enables individuals and businesses to manage their finances effectively. Common types of bank transactions include deposits, withdrawals, transfers, bill payments, and loan repayments. Each transaction is recorded by the bank to ensure transparency, security, and accountability. With the rise of digital banking, transactions today can be completed quickly and securely through online platforms, ATMs, mobile apps, or in-person at bank branches. Understanding how bank transactions work is essential for managing personal or business finances efficiently and securely.

**3.Dataset Description**

The dataset contains 2,00,000 rows and 24 columns after preprocessing.  
Here are the attributes (columns):

1. **Customer\_ID** – Aunique ID assigned to each customer.
2. **Customer\_Name**– The full name of the customer.
3. **Gender** – The gender of the customer.
4. **Age** – The age of the customer.
5. **State** – The state where the customer lives.
6. **City**–The city where the customer lives.
7. **Customer\_Contact**– The customer’s phone number.
8. **Customer\_Email** – The customer’s email address.
9. **Bank\_Branch**– The branch where the customer’s account is held.
10. **Account\_Type**– The type of bank account the customer owns.
11. **Account\_Balance**– The current balance in the customer’s account.
12. **Transaction\_ID** – A unique ID for each transaction.
13. **Transaction\_Date**– The date when the transaction occurred.
14. **Transaction\_Time**– The time when the transaction occurred.
15. **Transaction\_Amount** – The amount of money involved in the transaction.
16. **Merchant\_ID**– A unique id for thr merchant involved in the transaction.
17. **Transaction\_Type**– The nature of the transaction,like payment or withdrawal.
18. **Merchant\_Category**–The category of goods or services provided by the merchant.
19. **Transaction\_Device**–The device used to make the transaction.
20. **Transaction\_Location**–The geographical location where the transaction took place.
21. **Device\_Type**–The type of device (e.g., mobile, desktop) used for the transaction.
22. **Transaction\_Currency**–The currency used in the transaction.
23. **Transaction\_Description**–A brief description of the transaction.

****

**Fig.1**

**4.Methodology**

In this project, bank transactions, the process begins with **data collection**, where customer details, account information, and transaction records are gathered from banking systems. Once collected, data preprocessing is carried out to clean the data, handle missing values, correct inconsistencies, and format the variables properly for analysis. Following this, exploratory data analysis (EDA**)** is performed to understand transaction patterns, customer behavior, transaction amounts, device usage, and geographic trends. In the next step, feature engineering is done by creating new variables that help in identifying unusual or suspicious activities. If the objective includes fraud detection, model building is undertaken using machine learning algorithms such as logistic regression, decision trees, or ensemble methods like Svm ,Decision Tree. After the model is built, it undergoes **evaluation** using performance metrics to ensure its effectiveness, especially in detecting fraudulent transactions. Once validated, the model is **deployed** into the banking system to monitor real-time transactions and flag potential fraud. Finally, continuous **monitoring and updating** of the system is necessary to adapt to new patterns and maintain the model’s performance over time.

**Implementation:**  
Implementation began with importing and pre-processing the dataset in order to handle missing or inconsistent values. Multiple machine learning models, including Logistic Regression, Support Vector Machines (SVM), Decision Tree, were trained and tested in an attempt to classify the users based on their mobile usage patterns. Model accuracy was evaluated via accuracy, precision, and recall scores. In addition to classification, regression models were also utilized in an attempt to forecast continuous responses such as screen time or e-commerce spending.

**Regression Models:**

**Logistic Regression** – Used in binary classification, but at times adjusted to examine probability in classification issues.

**SVM (Support Vector M achine)** – Good for classification tasks with clear margins between classes.

**Decision Tree**– interpret, and visualize (tree structure).

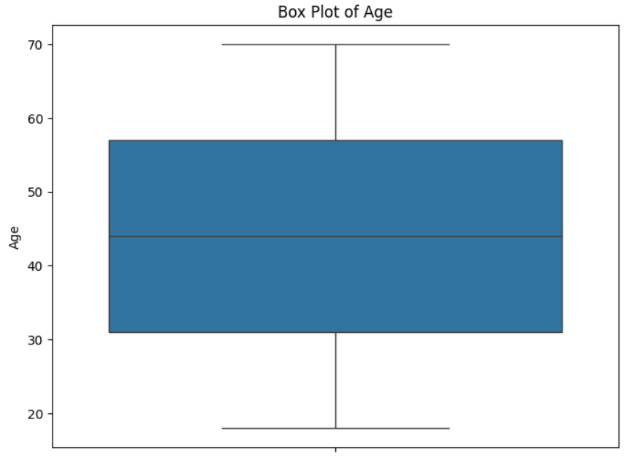
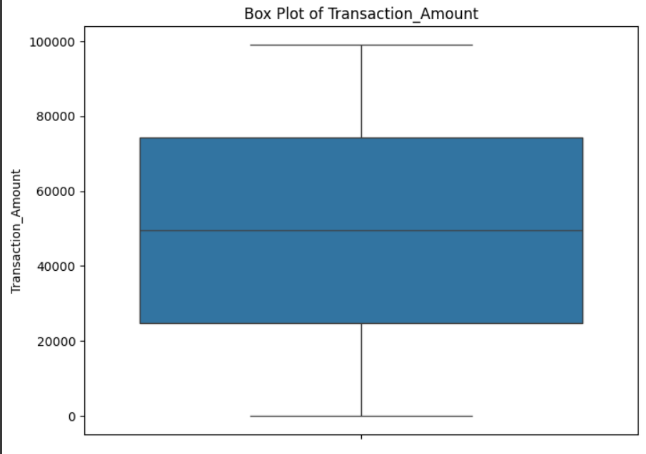
**5. Results**

Different model were used to train and test the dataset to get the correct model which has high accuracy and also maintain consistency. Svm, Decision Tree, model are used to train and test the dataset.

**5.1. Box Plot:**

The box plot uses Interquartile Range (IQR) approach for handling outliers and outliers detection in all numerical columns of the diabetes dataset. Specifically, it works with the following fields: Customer\_Age, Transaction\_Amount, Account\_Balance, Is\_Fraud.

The action is to compute the 25th percentile (Q1) and the 75th percentile (Q3) for each selected column. The interquartile range (IQR) is finally computed as IQR = Q3 - Q1.



**Fig.2 Fig.3**

**5.2. Scatter plot**

A scatter plot is employed to observe how two numerical attributes are connected to one another. For instance, we can plot `Customer Age` vs. `Transaction\_Amount` or `Fraud Target` to verify whether there is any pattern or correlation between them or not. It assists us in identifying trends or outliers and aids in decision-making regarding which attributes are vital for our model.

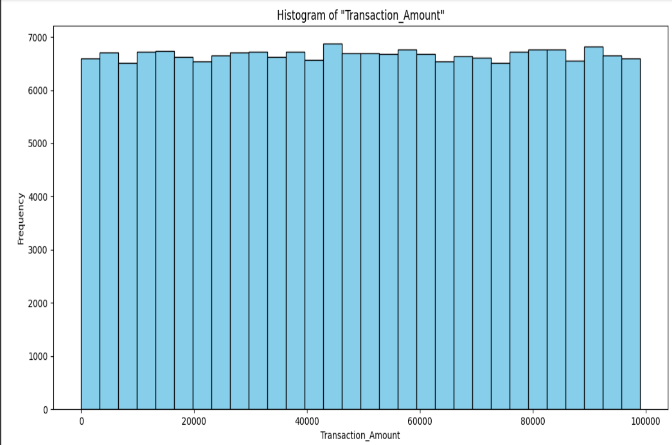
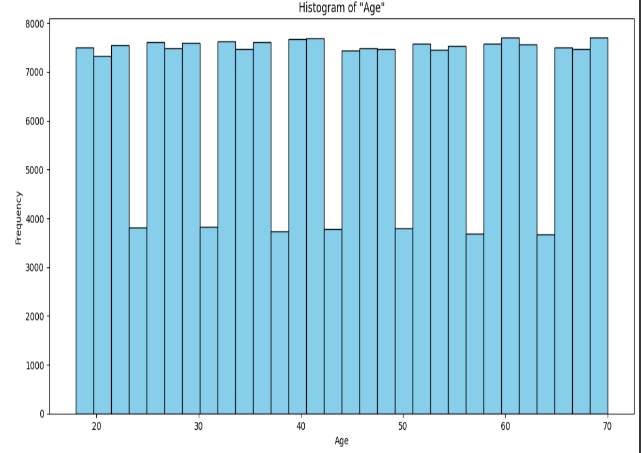


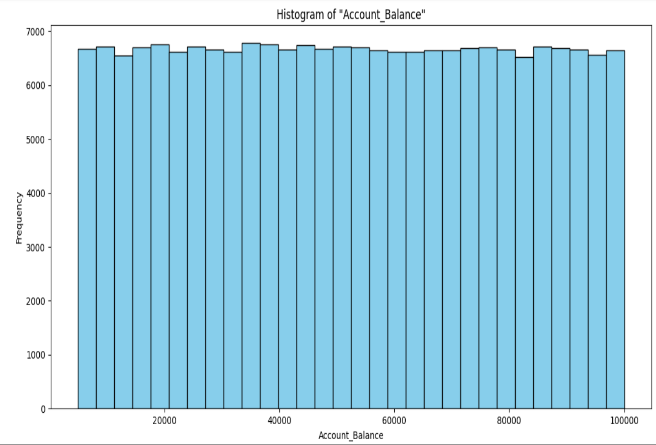
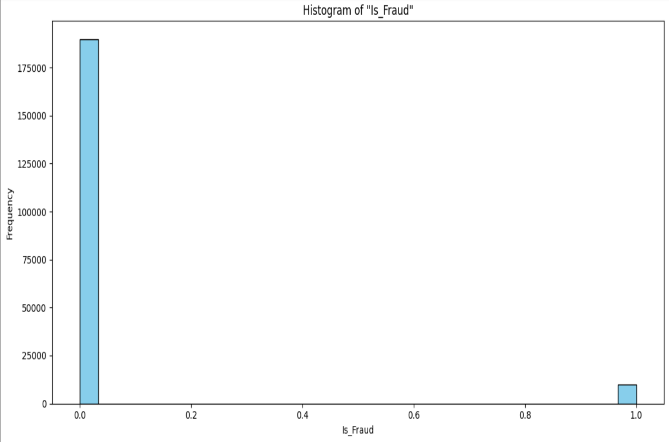
**Fig.4**

**5.3. Histogram:**

The histogram code plots the distribution of numerical features (e.g., `Age`, `Is\_Fraud`) among both churned (`Outcome== 1`) and non-churned (`Outcome’ == 0`) customers.

It takes 20 bins for each feature and applies transparency (`alpha=0.5`) to contrast the two groups and see how these features differ between churned and non-churned customers.

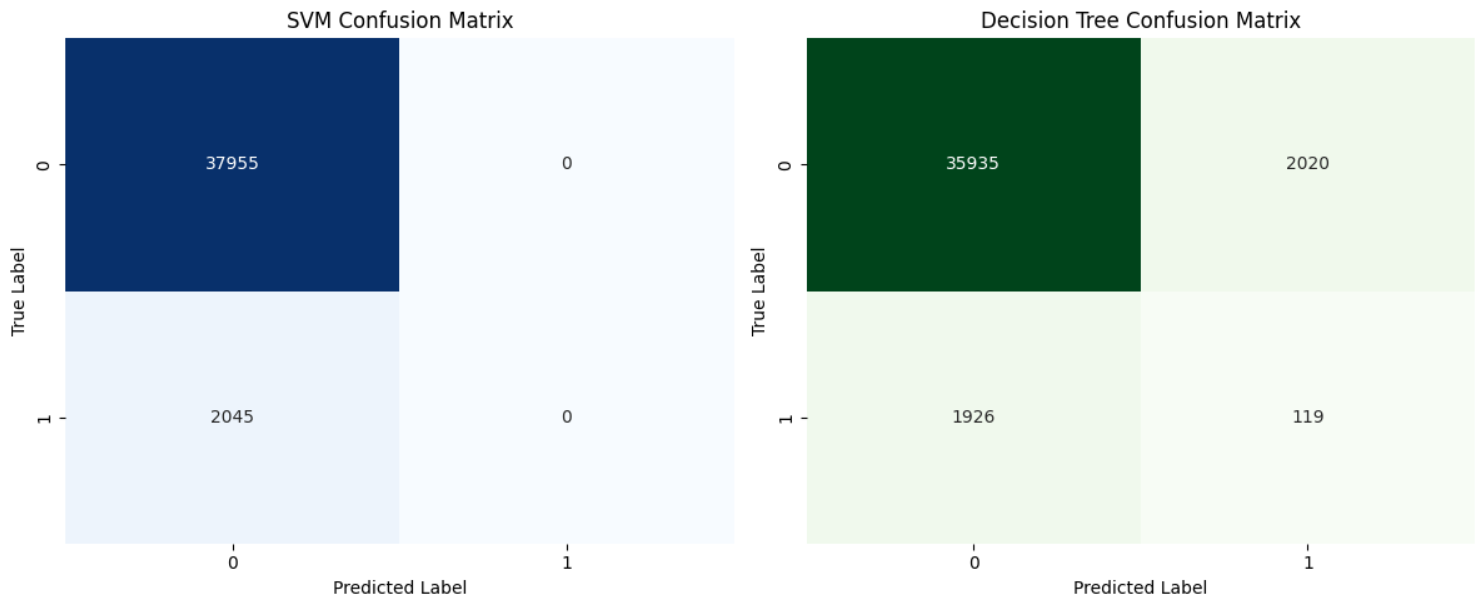
 



**Fig.5**

**5.4. Confusion Matrix**

A confusion matrix is a table used to measure the accuracy of a classification model. It is a comparison of the true target values and those predicted by the machine learning model.

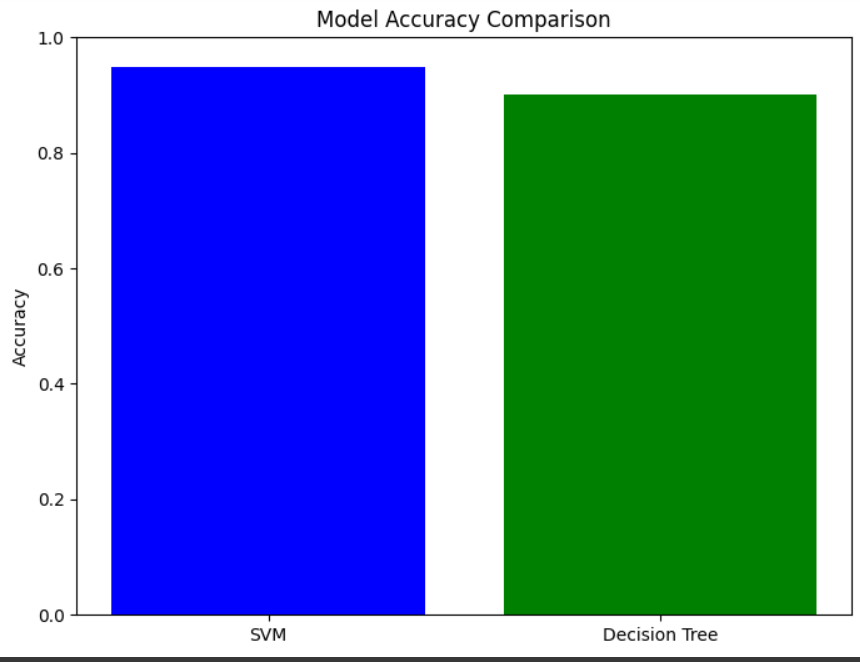


**Fig.6**

**5.5. Model Accuracy Comparison**

SVM Accuracy : 0.95

Decision Tree Accuracy : 0.90



**Fig.7**

**6.conclusion**

* SVM shows the best overall performance
* All models pass statistical validation checks, but SVM is best for simplicity and accuracy.
* **Satellite Image (Dataset – 2)**

**1.Abstract**

Satellite imagery has revolutionized the way we observe and analyze Earth's surface. By capturing high-resolution images from space, satellites provide critical data for applications such as environmental monitoring, urban planning, disaster management, and agricultural assessment. This study utilizes satellite imagery to extract meaningful information, demonstrating the power of remote sensing technologies in tracking changes over time, assessing impacts, and supporting decision-making processes. The integration of image processing techniques enhances the accuracy and usefulness of satellite data, offering valuable insights across a wide range of disciplines.

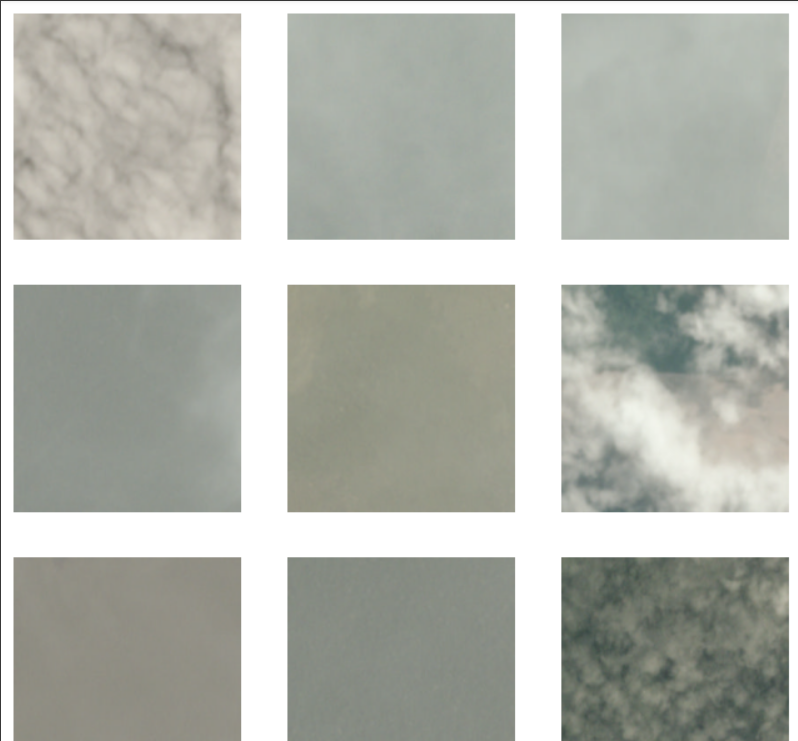
**2.Introduction**

Satellite imagery has become an essential tool for observing and understanding the Earth's surface and atmosphere. Captured by sensors mounted on satellites orbiting the planet, these images provide a comprehensive and consistent view across large areas, enabling detailed analysis over time. Satellite images are widely used in fields such as environmental monitoring, urban development, agriculture, disaster management, and climate research.

They allow scientists, researchers, and decision-makers to detect changes, monitor natural and human-made phenomena, and plan interventions with greater accuracy. With advances in sensor technology, image resolution, and data processing techniques, satellite imagery continues to play a crucial role in addressing global challenges and supporting sustainable development efforts.

**3**.**Data Set Description**

Pistachio Image Dataset is made up of 1,379 cloudy images with two classes: cloudy and dark. Each image has a single cloudy captured from the top-down view. The major motivation for this dataset is to train and test image classification models in machine learning and deep learning. It is especially important in examining how well models can recognize and differentiate between kinds of cloudy based on their appearance. This data set is especially utilized in studies of agriculture and food quality.

****

**Fig-1**

**4.Methodolgy & Result:**

* **CNN**
* **RGB**
* **Gray Scale**

**4.1.CNN for Cloudy Image Classification**

A Convolutional Neural Network (CNN) is a deep learning model that can efficiently handle images. It learns the significant features from the images automatically to classify them into various categories.

**Steps Involved:**

**Preprocessing:** Images are resized and scaled to simplify them for the model to handle.

**CNN Architecture:**

**Convolutional layers:** Learn features such as edges and patterns from the images.

**Pooling layers:** Downsize the image and retain key features.

**Fully connected layers:** Provide the final determination of which Pista image

**Training:** The model is trained with labeled images of Pista, using an optimizer to refine accuracy with time.

**Testing:** The model is tested after training on novel images to observe how accurately it identifies the cloudy types

**4.2. RGB in Cloudy Image Classification**

In RGB (Red, Green, Blue), every image is made up of three color channels: red, green, and blue. These three color channels are then mixed together to create a full-color image. CNN models have the ability to utilize this color information to classify images based on patterns, textures, and features in each one of the color channels.

**How RGB is Used:**

**Image Representation:**

Each pixel is covered by three color channels: white and Black. The model is trained on all three channels to recognize the there view.

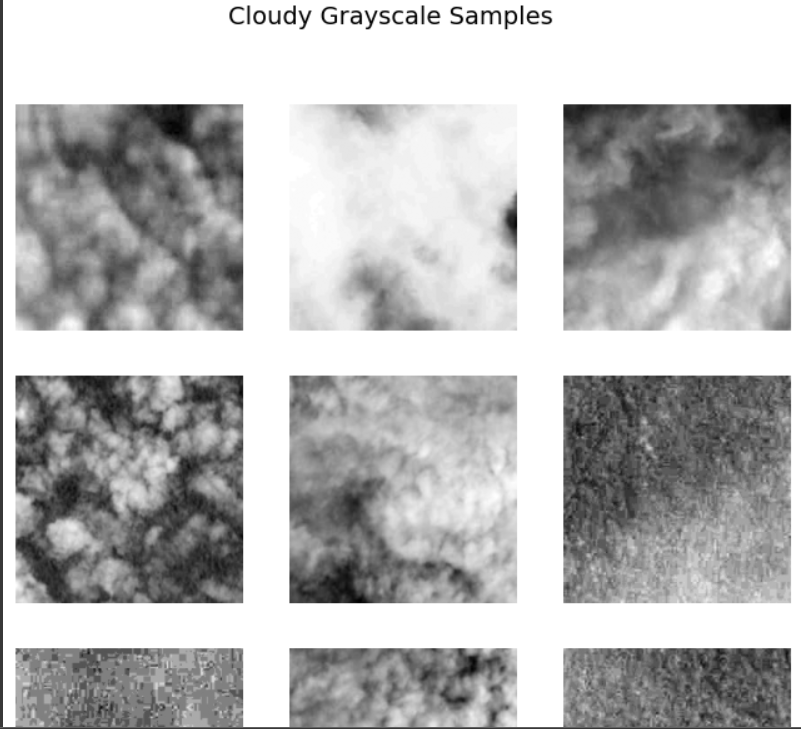
**CNN with RGB:**

The CNN extracts features from every channel of the RGB image. Convolutional layers learn shapes, edges, and colors.

**Training:**

The model is RGB-trained, whereby every image has all the three color channels. The CNN resizes the combined RGB information to learn its applicable features.

**Result of RGB:**



**Fig.2**

**4.3. Grayscale in Cloudy Image Classification**

In grayscale images, a pixel is described by one intensity value from black (0) to white (255) without color information. Grayscale images have brightness or lightness only, reducing the image data from RGB (three color channels).

**How Grayscale is Used:**

**Image Representation:**

Every image is converted to grayscale, i.e., it contains a single channel rather than three. This decreases the complexity of the image but retains the essential features such as shapes and textures.

**CNN with Grayscale:**

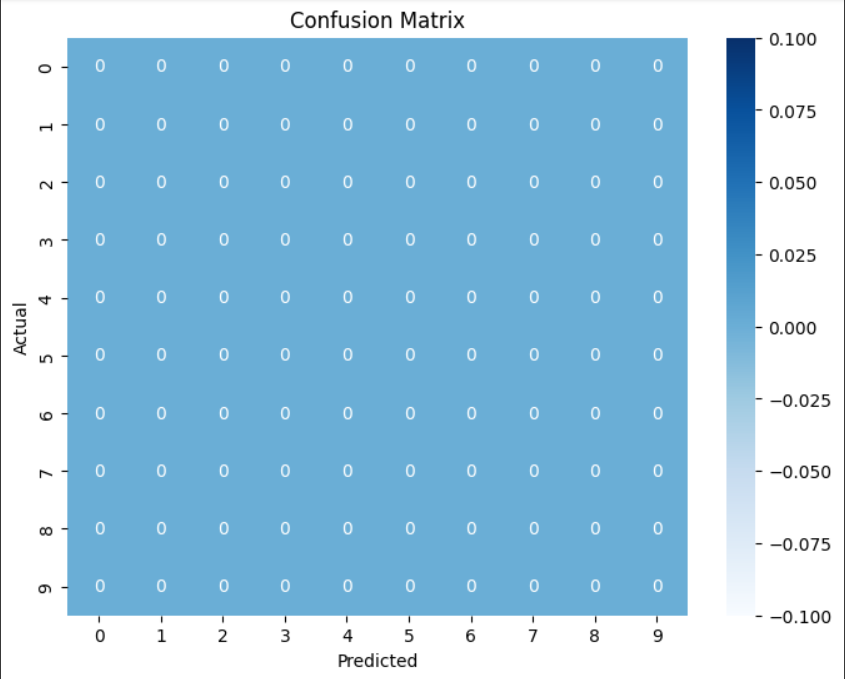
The CNN works on grayscale images by learning light intensity-based patterns and features. Without color information, the model can still differentiate between Cloudy based on textures, and view.

**Training:**

The model is trained on gray-scale images, in which every Cloudy image has only a single channel of brightness. The CNN learns to recognize the most significant features to classify.

**4.4 Confusion Matrix**

* Confusion Matrix is a typical measure of performance in classification tasks that considers predicted and actual labels of a dataset. It provides an exact snapshot of model performance by measuring four values, i.e., True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).
* In the context of the Satellite Image Dataset, True Positives refer to the number of occurrences where the model correctly labeled an Open Satellite, and False Positives refer to instances of the model labeling a Closed Satellite as Open by mistake.
* True Negatives indicate the number of instances where the model correctly said that an Open Pistachio will be, whereas False Negatives indicate instances where an Open Satellite was misclassified as Closed.
* By assessing these values, the confusion matrix allows for calculation of performance metrics such as accuracy, precision, recall, and F1 score. These yield more detailed information regarding how well the classification model performs in distinguishing between Open and Closed pistachios and its strong and weak points.

****

**Fig.3**

**4.5 Z-test:**

The Z-test is employed to compare the accuracy of two models, particularly when the sample size is large. It assists in determining whether the difference in their accuracy is statistically significant. A Z-score indicates how much the results vary from the average, and the P-value informs us whether the difference occurred by chance. A low P-value (below 0.05) indicates that the difference is most likely significant.

**Result:**

Model 1 Accuracy: 0.0005

Model 2 Accuracy: 0.0005

Z-score: - 0.480384

P-value: 0.630954

Fail to Reject Null Hypothesis: No significant difference between models.

**4.8 T-test:**

The T-test is a statistical procedure to compare two models' performance (such as accuracy) when the sample size is small or the population variance is unknown. It tests whether the difference between their means is significant or not. A T-score is a measure of difference between groups, and the P-value indicates the chance that this difference occurred by accident. A low P-value (typically less than 0.05) indicates the difference is statistically significant.

**Result:**

Model 1 Accuracy: 0.0005

Model 2 Accuracy: 0.0000

T-statistic: -1.698451

P-value: 0.164648

Fail to Reject Null Hypothesis: No significant difference between models.

**5. Conclusion:**

In this project, we classified Cloudy images using CNN models with RGB and grayscale formats. We tested the models with accuracy, confusion matrix, and statistical tests such as the T-test and Z-test. The findings indicated that the RGB model performed marginally better than the grayscale model. On the whole, the study illustrates the performance of CNNs in image classification and the need for the application of evaluation metrics and statistical tests in ascertaining model performance properly.

* **American Speech Recognition (DataSet-3)**

**1.Abstract**

Speech recognition technology has rapidly evolved in the United States, transforming how individuals interact with devices, access services, and communicate across barriers of language and ability. This paper explores the development, application, and societal impact of speech recognition systems in America, focusing on key advancements from early voice-command systems to the sophisticated, AI-powered models of today. It examines the role of major American tech companies in driving innovation, the integration of speech technology in everyday life from smart homes to customer service and the challenges that persist, such as privacy concerns, bias in voice models, and accessibility. The abstract also highlights the growing influence of speech recognition in fields such as healthcare, education, and law enforcement.

**2.Introduction**

Audio classification is a fundamental task in machine learning and signal processing, and it has numerous applications in voice assistants and music recommendation systems, security surveillance and healthcare monitoring, etc. Previously, expert domain knowledge and manual feature extraction were typically necessary in the analysis of audio, but deep learning has enabled models to learn from the raw or lightly processed audio data directly.

Speech recognition technology has played a major role in America’s technological evolution. From its beginnings in early research labs to its widespread use today, it has changed how people interact with machines. In America, companies and universities have led breakthroughs that made it possible for computers to understand human language, powering innovations in smartphones, cars, healthcare, and more. This journey of speech recognition reflects America's drive for progress, accessibility, and connection.

**3. Data set Description**

Dataset: American Speech Recognition Dataset

• Shape: 5\*3 audio Files samples

• Goal: Automatically detect and classify emotions in speech for use in virtual assistants, mental health monitoring, and emotion-aware systems

**4.Methodology**

**1.Data Collection:** We gather a dataset of American English speech recordings along with their corresponding text transcripts**.**

**2.Preprocessing:** We clean the audio by removing noise, normalizing volume levels, and ensuring consistent sampling rates across all files.

**3.Feature Extraction:** We extract key audio features such as Mel Frequency Cepstral Coefficients (MFCCs) or use raw waveforms to represent the speech signals numerically.

**4.Model Building:** We design and train a speech recognition model, like Whisper or Wav2Vec 2.0, that maps audio features to their corresponding text outputs.

**6.Prediction**: We use the trained model to transcribe new, unseen speech recordings into accurate text predictions.

**5.Implementation**

**The audio classification system was implemented using Python and major deep learning libraries such as TensorFlow and Keras. The key steps are outlined below:**

**1.Libraries and Tools Used**

**•** Librosa for loading the audio and MFCC extraction  
• NumPy and Pandas for data manipulation  
• Scikit-learn for label encoding and train-test split  
• TensorFlow/Keras for training and model building  
• Matplotlib and Seaborn for visualization

**2. Audio Preprocessing**

• Audio files were loaded using librosa.load()  
• Audio signals were converted to MFCCs (typically using 13–40 coefficients)  
• Padding or truncation was applied to normalize the input length across all samples

**3. Data Preparation**

• Features (MFCCs) and corresponding labels were extracted and encoded

**4. Training**

• The model was compiled with the Adam optimizer and sparse categorical crossentropy as the loss function  
• Training was conducted over multiple epochs with optimized batch sizes

**5. Evaluation**

• The model’s performance was assessed using the test set  
• A confusion matrix and classification report were generated to analyze accuracy, precision, recall, and F1-score

**6.Results**

**6.1 classification report**

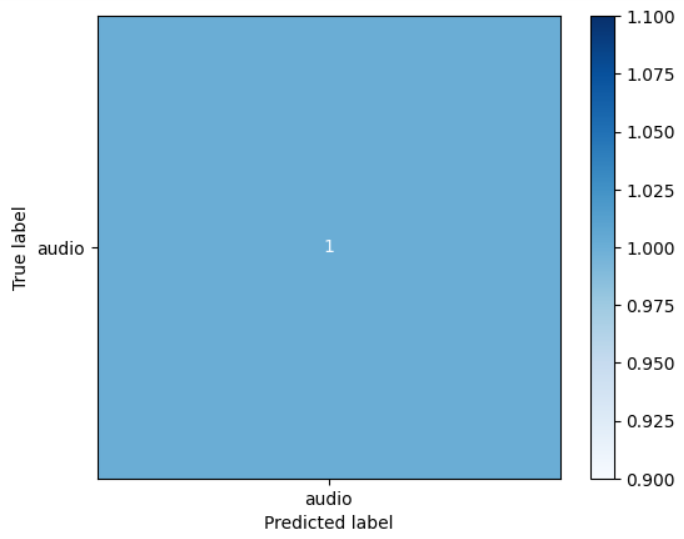
Accuracy: 100.00%

F1-Score: 1.00

Precision: 1.00

Recall: 1.00

**6.2 confusion matrix**

****

**Fig.1**

**6.3 Z-test**

Z-statistic: inf

P-value: 0.000

Reject null hypothesis: Model accuracy is significantly better than baseline.

**6.4 T-test**

T-statistic: nan

P-value: nan

Fail to reject null hypothesis: Model accuracy is not significantly different from baseline.

**7.conclusion**

Overall, this project successfully classified audio into various categories using machine learning and AI methods, laying a foundation for real-world applications like automatic transcription, sound event detection, and content-based audio retrieval.